# Advanced Probability and Applications

Olivier Lévêque, IC–LTHI, EPFL
(with special thanks to Simon Guilloud for the figures)

August 9, 2023

## Contents

## Basic terminology and conventions

- A *discrete* set means a set in bijection with a subset of $\mathbb{N}$ (so either finite or countable).

- Capital letters $X, Y, Z$ refer to random variables, while small letters $x, y, z$ refer to numbers.

- A number $x \in \mathbb{R}$ is said to be *non-negative* if $x \geq 0$, and *positive* if $x > 0$.

- Likewise, a function $f : \mathbb{R} \to \mathbb{R}$ is said to be *non-decreasing* if $f(x_1) \leq f(x_2)$ as soon as $x_1 < x_2$, and *increasing* if $f(x_1) < f(x_2)$ as soon as $x_1 < x_2$.

- Closed intervals are denoted as $[a, b]$; open intervals are denoted as $]a, b[$.

- For two sets $A, B$, "$A \subset B$" means "either $A$ is a (strict) subset of $B$ or $A = B$".

- If $x$ is an element of a set $A$, then $\{x\}$ denotes the subset of $A$ containing the only element $x$ (=singleton).

- And an important remark: a necessary preliminary to Probability Theory is Measure Theory; likewise, a necessary preliminary to Measure Theory is Topology, and it is probably fair to say also that a necessary preliminary to Topology is Set Theory. As we cannot cover everything in these notes, some facts will be stated without proof in order to avoid opening too many Pandora's boxes... Readers interested in gaining a deeper understanding of the field are of course encouraged to search for other more detailed references on the subject.

# 1 $\sigma$-fields and random variables

## 1.1 $\sigma$-fields

In probability, the *fundamental set* $\Omega$ describes the set of all possible *outcomes* (or *realizations*) of a given experiment. It might be *any* set, without any particular structure, such as for example $\Omega = \{1, \ldots, 6\}$ representing the outcomes of a die roll, or $\Omega = [0, 1]$ representing the outcomes of a concentration measurement of some chemical product. Note moreover that the set $\Omega$ need not be composed of numbers exclusively; it would be for example perfectly valid to consider the set $\Omega = \{\text{banana, apple, orange}\}$.

Given a fundamental set $\Omega$, it is important to describe what *information* does one have on the system, namely on the outcomes of the experiment. This notion of information is well captured by the mathematical notion of *$\sigma$-field*, which is defined below. Note that in elementary probability courses, it is generally assumed that the information one has about a system is *complete*, so that it becomes useless to introduce the concept below.

**Definition 1.1.** Let $\Omega$ be a set. A *$\sigma$-field* (or *$\sigma$-algebra*) on $\Omega$ is a collection $\mathcal{F}$ of subsets of $\Omega$ (or *events*) satisfying the following properties or *axioms*:

(i) $\emptyset, \Omega \in \mathcal{F}$.

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

(iii) If $A_1, \ldots, A_n \in \mathcal{F}$, then $\bigcup_{j=1}^n A_j \in \mathcal{F}$.

(iii') If $(A_n,\, n \geq 1)$ is a sequence of subsets of $\Omega$ and $A_n \in \mathcal{F}$ for every $n \geq 1$, then $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$.

Using De Morgan's law: $\bigcap_{j=1}^n A_j = \left(\bigcup_{j=1}^n A_j^c\right)^c$, the above properties imply that

(iv) If $A_1, \ldots, A_n \in \mathcal{F}$, then $\bigcap_{j=1}^n A_j \in \mathcal{F}$.

(iv') If $(A_n,\, n \geq 1)$ is a sequence of subsets of $\Omega$ and $A_n \in \mathcal{F}$ for every $n \geq 1$, then $\bigcap_{n=1}^\infty A_n \in \mathcal{F}$.

(v) Also, if $A, B \in \mathcal{F}$, then $B \backslash A = B \cap A^c \in \mathcal{F}$.

**Terminology.** The pair $(\Omega, \mathcal{F})$ is called a *measurable space* and the events belonging to $\mathcal{F}$ are said to be *$\mathcal{F}$-measurable*, that is, they are the events that one can decide on whether they happened or not, given the information $\mathcal{F}$. In other words, if one knows the information $\mathcal{F}$, then one is able to tell to which events of $\mathcal{F}$ (= subsets of $\Omega$) does the realization of the experiment $\omega$ belong.

**Example 1.2.** For a generic set $\Omega$, the following are always $\sigma$-fields:

$\mathcal{F}_0 = \{\emptyset, \Omega\}$ (= *trivial* $\sigma$-field).
$\mathcal{P}(\Omega) = \{\text{all subsets of } \Omega\}$ (= *complete* $\sigma$-field).

**Example 1.3.** Let $\Omega = \{1, \ldots, 6\}$. The following are $\sigma$-fields on $\Omega$:

$\mathcal{F}_1 = \{\emptyset, \{1\}, \{2, \ldots, 6\}, \Omega\}$.
$\mathcal{F}_2 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$.
$\mathcal{F}_3 = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \Omega\}$.

**Example 1.4.** Let $\Omega = [0, 1]$ and $I_1, \ldots, I_n$ be a family of disjoint intervals in $\Omega$ such that $I_1 \cup \ldots \cup I_n = \Omega$ ($\{I_1, \ldots, I_n\}$ is also called a *partition* of $\Omega$). The following is a $\sigma$-field on $\Omega$:

$$\mathcal{F}_4 = \{\emptyset, I_1, \ldots, I_n, I_1 \cup I_2, \ldots, I_1 \cup I_2 \cup I_3, \ldots, \Omega\} \quad \text{(NB: there are } 2^n \text{ events in total in } \mathcal{F}_4)$$

In the discrete setting (that is, in a $\sigma$-field with a finite or countable number of elements), the smallest elements contained in a $\sigma$-field are called the *atoms* of the $\sigma$-field. Formally, $F \in \mathcal{F}$ is an atom of $\mathcal{F}$ if for any $G \in \mathcal{F}$ such that $G \subset F$, it holds that either $G = \emptyset$ or $G = F$. In the above example with $\Omega = [0, 1]$, the atoms of $\mathcal{F}_4$ are therefore $I_1, \ldots, I_n$. Note moreover that a $\sigma$-field with $n$ atoms has $2^n$ elements, so that the number of elements of a finite $\sigma$-field is always a power of 2.

## 1.2 $\sigma$-field generated by a collection of events

An event carries in general more information than itself. As an example, if one knows whether the result of a die roll is odd (corresponding to the event $\{1,3,5\}$), then one also knows of course whether the result is even (corresponding to the event $\{2,4,6\}$). It is therefore convenient to have a mathematical description of the information generated by a single event, or more generally by a family of events.

**Definition 1.5.** Let $\mathcal{A} = \{A_i,\, i \in I\}$ be a collection of subsets of $\Omega$ (where $I$ need not be finite nor countable). The *$\sigma$-field generated by $\mathcal{A}$* is the smallest $\sigma$-field on $\Omega$ containing all the events $A_i$. It is denoted as $\sigma(\mathcal{A})$. Note that the $\sigma$-field $\sigma(A)$ can equivalently be defined as the *intersection* of all $\sigma$-fields that contain the collection $\mathcal{A}$, as yes, any intersection of $\sigma$-fields is still a $\sigma$-field (exercise!).

**Remark.** A natural question is whether such a vague definition makes sense. Observe first that there is always at least one $\sigma$-field containing $\mathcal{A}$: it is $\mathcal{P}(\Omega)$. Then, one can show that an arbitrary intersection of $\sigma$-fields is still a $\sigma$-field. One can therefore provide the following alternative definition of $\sigma(\mathcal{A})$: it is the intersection of *all* $\sigma$-fields containing the collection $\mathcal{A}$, which is certainly a well-defined object.

**Example.** Let $\Omega = \{1,\dots,6\}$ (cf. Example 1.3).

Let $\mathcal{A}_1 = \{\{1\}\}$. Then $\sigma(\mathcal{A}_1) = \mathcal{F}_1$.
Let $\mathcal{A}_2 = \{\{1,3,5\}\}$. Then $\sigma(\mathcal{A}_2) = \mathcal{F}_2$.
Let $\mathcal{A}_2 = \{\{1,2,3\}\}$. Then $\sigma(\mathcal{A}_3) = \mathcal{F}_3$.
Let $\mathcal{A} = \{\{1\},\dots,\{6\}\}$. Then $\sigma(\mathcal{A}) = \mathcal{P}(\Omega)$.

**Exercise.** Let $\mathcal{A} = \{\{1,2,3\},\{1,3,5\}\}$. Compute $\sigma(\mathcal{A})$.

**Example.** Let $\Omega = [0,1]$ and let $\mathcal{A}_4 = \{I_1,\dots,I_n\}$ (cf. Example 1.4). Then $\sigma(\mathcal{A}_4) = \mathcal{F}_4$. This is a particular instance of the fact that in the discrete case, a $\sigma$-field is always generated by the collection of its atoms.

**Borel $\sigma$-field on $[0,1]$.** A very important example of generated $\sigma$-field on $\Omega = [0,1]$ is the following:

$$\mathcal{B}([0,1]) = \sigma(\{\{0\},\{1\},\,]a,b[\,:\ a,b \in [0,1],\ a < b\})$$

is the *Borel $\sigma$-field* on $[0,1]$ and elements of $\mathcal{B}([0,1])$ are called the *Borel subsets* of $[0,1]$. As surprising as it may be, it turns out that $\mathcal{B}([0,1]) \neq \mathcal{P}([0,1])$ [without proof], which generates some difficulties from the theoretical point of view. Nevertheless, it is quite difficult to construct explicit examples of subsets of $[0,1]$ which are *not* in $\mathcal{B}([0,1])$. Note indeed that

a) All singletons belong to $\mathcal{B}([0,1])$. Indeed, for any $0 < x < 1$, $\{x\} = \bigcap_{n\geq 1} ]x - \frac{1}{n}, x + \frac{1}{n}[$ belongs to $\mathcal{B}([0,1])$, by the property seen above and the fact that the Borel $\sigma$-field is by definition the smallest $\sigma$-field containing all open intervals.

b) Therefore, all closed intervals, being unions of open intervals and singletons, also belong to $\mathcal{B}([0,1])$.

c) Likewise, all countable intersections of open intervals $\mathcal{B}([0,1])$, as well as all countable unions of closed intervals belong to $\mathcal{B}([0,1])$.

d) The story goes on with countable unions of countable intersections of open intervals, etc. Even though the list is quite long, not all the subsets of $[0,1]$ are part of $\mathcal{B}([0,1])$, as mentioned above.

**Remark.** In general, the $\sigma$-field generated by a collection of events contains many more elements than the collection itself! The Borel $\sigma$-field is a good example. In the finite case, you will observe the same phenomenon while computing $\sigma(\{\{1,2,3\},\{1,3,5\}\})$ on $\Omega = \{1,\dots,6\}$.

**Remark.** It can be easily checked that the *atoms* of $\mathcal{B}([0,1])$ are the singletons $\{x\}$, $x \in [0,1]$. Nevertheless, one can check that $\mathcal{B}([0,1])$ is *not* generated by its atoms (as it is not a discrete $\sigma$-field). As a proof of this (exercise), compute what $\sigma(\{\{x\},\, x \in [0,1]\})$ is.

**Borel $\sigma$-field on $\mathbb{R}$ and $\mathbb{R}^2$.**

**Definition 1.6.** On the set $\mathbb{R}$, one defines

$$\mathcal{B}(\mathbb{R}) = \sigma(\{]a,b[ :\ a,b \in \mathbb{R},\ a < b\})$$

The elements of $\mathcal{B}(\mathbb{R})$ are called *Borel sets* on $\mathbb{R}$. Again, note that $\mathcal{B}(\mathbb{R})$ is strictly included in $\mathcal{P}(\mathbb{R})$.

**Definition 1.7.** On the set $\mathbb{R}^2$, one defines

$$\mathcal{B}(\mathbb{R}^2) = \sigma(\{]a,b[\times]c,d[ :\ a,b,c,d \in \mathbb{R},\ a < b,\ c < d\})$$

Note that even though $\mathcal{B}(\mathbb{R}^2)$ is generated by rectangles only, it contains all kinds of shapes in $\mathbb{R}^2$, including in particular discs and triangles (because every disc and triangle can be seen as a countable union of rectangles). Here again, one sees that the $\sigma$-field generated by a collection of events is much larger than the collection of events itself.

Finally, note that a straightforward generalization of the above definition allows to define $\mathcal{B}(\mathbb{R}^n)$ for arbitrary $n$. Even more generally, $\mathcal{B}(\Omega)$ can be defined for $\Omega$ a Hilbert / metric / topological space.

## 1.3 Sub-$\sigma$-field

One may have more or less information about a system. In mathematical terms, this translates into the fact that a $\sigma$-field contains more or less elements. It is therefore convenient to introduce a (partial) ordering on the ensemble of existing $\sigma$-fields, in order to establish a *hierarchy* of information. This notion of hierarchy is important and will come back when we will be studying stochastic processes that evolve in time.

**Definition 1.8.** Let $\Omega$ be a set and $\mathcal{F}$ be a $\sigma$-field on $\Omega$. A *sub-$\sigma$-field of* $\mathcal{F}$ is a collection $\mathcal{G}$ of events such that:

(i) If $A \in \mathcal{G}$, then $A \in \mathcal{F}$.

(ii) $\mathcal{G}$ is itself a $\sigma$-field.

**Notation.** $\mathcal{G} \subset \mathcal{F}$.

**Remark.** Let $\Omega$ be a generic set. The trivial $\sigma$-field $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is a sub-$\sigma$-field of any other $\sigma$-field on $\Omega$. Likewise, any $\sigma$-field on $\Omega$ is a sub-$\sigma$-field of the complete $\sigma$-field $\mathcal{P}(\Omega)$.

**Example.** Let $\Omega = \{1, \ldots, 6\}$ (cf. Example 1.3). Note that $\mathcal{F}_1$ is *not* a sub-$\sigma$-field of $\mathcal{F}_2$ (even though $\{1\} \subset \{1,3,5\}$), nor is $\mathcal{F}_2$ a sub-$\sigma$-field of $\mathcal{F}_1$. In general, note that

1) If $A \in \mathcal{G}$ and $\mathcal{G} \subset \mathcal{F}$, then it is true that $A \in \mathcal{F}$.

but

2) $A \subset B$ and $B \in \mathcal{G}$ together do *not* imply that $A \in \mathcal{G}$.

**Example.** Let $\Omega = [0,1]$ (cf. Example 1.4). Then $\mathcal{F}_4$ is a sub-$\sigma$-field of $\mathcal{B}([0,1])$. Also, if $\mathcal{F}_5 = \sigma(J_1, \ldots, J_m)$, where $\{J_1, \ldots, J_m\}$ represents a finer partition of the interval $[0,1]$ (i.e., each interval $I$ of $\mathcal{F}_4$ is a disjoint union of intervals $J$), then $\mathcal{F}_4 \subset \mathcal{F}_5$.

## 1.4 Random variables

The notion of random variable is usually introduced in elementary probability courses as a vague concept, essentially characterized by its distribution. In mathematical terms however, random variables do exist prior to their distribution: they are functions from the fundamental set $\Omega$ to $\mathbb{R}$ satisfying a *measurability property*.

**Definition 1.9.** Let $(\Omega, \mathcal{F})$ be a measurable space. A *random variable* on $(\Omega, \mathcal{F})$ is a map $X : \Omega \to \mathbb{R}$ satisfying

$$\{\omega \in \Omega \,:\, X(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}) \tag{1}$$

**Notation.** One often simply denotes the set $\{\omega \in \Omega \,:\, X(\omega) \in B\} = \{X \in B\} = X^{-1}(B)$: it is called the inverse image of the set $B$ through the map $X$ (watch out that $X$ need *not* be a bijective function in order for this set to be well defined).

**Terminology.** The above random variable $X$ is sometimes called *$\mathcal{F}$-measurable*, in order to emphasize that if one knows the information $\mathcal{F}$, then one knows the value of $X$.

**Example.** If $\mathcal{F} = \mathcal{P}(\Omega)$, then condition (1) is always satisfied, so every map $X : \Omega \to \mathbb{R}$ is an $\mathcal{F}$-measurable random variable. On the contrary, if $\mathcal{F} = \{\emptyset, \Omega\}$, then the only random variables which are $\mathcal{F}$-measurable are the maps $X : \Omega \to \mathbb{R}$ which are constant.

**Remark.** Condition (1) can be shown to be equivalent to the following condition: [without proof]

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

which is significantly easier to check.

**Definition 1.10.** Let $(\Omega, \mathcal{F})$ be a measurable space and $A \in \mathcal{F}$ be an event. Then the map $\Omega \to \mathbb{R}$ defined as

$$\omega \mapsto 1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

is a random variable on $(\Omega, \mathcal{F})$. It is called the *indicator function* of the event $A$.

**Example 1.11.** Let $\Omega = \{1, \ldots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (cf. Example 1.3). Then $X_1(\omega) = \omega$ and $X_2(\omega) = 1_{\{1,3,5\}}(\omega)$ are both random variables on $(\Omega, \mathcal{F})$. Moreover, $X_2$ is $\mathcal{F}_2$-measurable, but note that $X_1$ is neither $\mathcal{F}_1$- nor $\mathcal{F}_2$-measurable.

**Example 1.12.** Let $\Omega = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$ (cf. Example 1.4). Then $X_3(\omega) = \omega$ and $X_4(\omega) = \sum_{j=1}^{n} x_j 1_{I_j}(\omega)$ are both random variables on $(\Omega, \mathcal{F})$. Note however that only $X_4$ is $\mathcal{F}_4$-measurable.

We will need to consider not only random variables, but also functions of random variables. This is why we introduce the following definition.

**Definition 1.13.** A map $g : \mathbb{R} \to \mathbb{R}$ such that

$$\{x \in \mathbb{R} \,:\, g(x) \in B\} \in \mathcal{B}(\mathbb{R}), \quad \forall B \in \mathcal{B}(\mathbb{R})$$

is called a *Borel-measurable function* on $\mathbb{R}$.

**Remark.** A Borel-measurable function on $\mathbb{R}$ is therefore nothing but a random variable on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

**Notation.** Again, one often uses the shorthand notations $\{x \in \mathbb{R} \,:\, g(x) \in B\} = \{g \in B\} = g^{-1}(B)$, but this does not mean that $g$ is invertible!

As it is difficult to construct explicitly sets which are not Borel sets, it is equally difficult to construct functions which are not Borel-measurable. Nevertheless, one often needs to check that a given function is Borel-measurable. A useful criterion for this is the following [without proof].

**Proposition 1.14.** If $g : \mathbb{R} \to \mathbb{R}$ is continuous, then it is Borel-measurable.

Finally, let us mention this useful property of functions of random variables.

**Proposition 1.15.** If $X$ is an $\mathcal{F}$-measurable random variable and $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, then $Y = g(X)$ is also an $\mathcal{F}$-measurable random variable.

*Proof.* Let $B \in \mathcal{B}(\mathbb{R})$. Then

$$\{Y \in B\} = \{g(X) \in B\} = \{X \in g^{-1}(B)\} \in \mathcal{F}$$

since $X$ is an $\mathcal{F}$-measurable random variable and $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$ by assumption. □

The above proposition is saying no more than the following: assume that knowing the information $\mathcal{F}$ allows you to determine the value of $X$. Then knowing this same information $\mathcal{F}$ also gives you the value of $Y = g(X)$.

## 1.5 $\sigma$-field generated by a collection of random variables

The amount of information contained in a random variable, or more generally in a collection of random variables, is given by the definition below.

**Definition 1.16.** Let $(\Omega, \mathcal{F})$ be a measurable space and $\{X_i,\, i \in I\}$ be a collection of random variables on $(\Omega, \mathcal{F})$. The *$\sigma$-field generated by $X_i$, $i \in I$*, denoted as $\sigma(X_i,\, i \in I)$, is the smallest $\sigma$-field $\mathcal{G}$ on $\Omega$ such that all the random variables $X_i$ are $\mathcal{G}$-measurable.

**Remark.** Note that

$$\sigma(X_i,\, i \in I) = \sigma(\{\{X_i \in B\},\, i \in I,\, B \in \mathcal{B}(\mathbb{R})\})$$

where the right-hand side expression refers to Definition 1.5. It turns out that one also has [without proof]

$$\sigma(X_i,\, i \in I) = \sigma(\{\{X_i \le t\},\, i \in I,\, t \in \mathbb{R}\})$$

**Example.** Let $(\Omega, \mathcal{F})$ be a measurable space. If $X_0$ is a constant random variable (i.e. $X_0(\omega) = c \in \mathbb{R},\ \forall \omega \in \Omega$), then $\sigma(X_0) = \{\emptyset, \Omega\}$.

**Example.** Let $\Omega = \{1, \ldots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (cf. Examples 1.3 and 1.11). Then $\sigma(X_1) = \mathcal{P}(\Omega)$ and $\sigma(X_2) = \mathcal{F}_2$.

**Example.** Let $\Omega = [0,1]$ and $\mathcal{F} = \mathcal{B}([0,1])$ (cf. Examples 1.4 and 1.12). Then $\sigma(X_3) = \mathcal{B}([0,1])$ and $\sigma(X_4) = \mathcal{F}_4$ (if all values $x_j$ are distinct).

The $\sigma$-field $\sigma(X)$ can be seen as the *information carried by the random variable $X$*. By definition, a random variable $X$ is always $\sigma(X)$-measurable. Following the proof of Proposition 1.15, one can also show the proposition below.

**Proposition 1.17.** If $X$ is a random variable on a measurable space $(\Omega, \mathcal{F})$ and $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, then $Y = g(X)$ is a $\sigma(X)$-measurable random variable, which is equivalent to saying that $\sigma(Y) \subset \sigma(X)$: the information carried by $Y$ is in general less than that carried by $X$.

Note that it can be strictly less: if you think e.g. about the case $Y = X^2$, then the information about the sign of $X$ is lost in $Y$; on the other hand, if the function $g$ is invertible (meaning that one can write $X = g^{-1}(Y)$), then $\sigma(Y) = \sigma(X)$.

A further generalization of Proposition 1.17 is the following: if $g : \mathbb{R}^2 \to \mathbb{R}$ is Borel-measurable and $Y = g(X_1, X_2)$, where $X_1, X_2$ are two random variables, then $Y$ is a $\sigma(X_1, X_2)$-measurable random variable, or put differently, $\sigma(Y) \subset \sigma(X_1, X_2)$. The other inclusion $\sigma(X_1, X_2) \subset \sigma(Y)$ is of course not true in general, as the two random variables $(X_1, X_2)$ carry potentially more information than the single random variable $Y$.

**Final remark.** It turns out that the reciprocal statement of Proposition 1.17 is also true: if $Y$ is a $\sigma(X)$-measurable random variable, then there exists a Borel-measurable function $g : \mathbb{R} \to \mathbb{R}$ such that $Y = g(X)$ [without proof].

# 2 Probability measures and distributions

## 2.1 Probability measures

**Definition 2.1.** Let $(\Omega, \mathcal{F})$ be a measurable space. A *probability measure* on $(\Omega, \mathcal{F})$ is a map $\mathbb{P} : \mathcal{F} \to [0,1]$ satisfying the following axioms:

(i) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(ii) If $A_1, \ldots, A_n \in \mathcal{F}$ are disjoint (i.e. $A_j \cap A_k = \emptyset$ for all $1 \leq j \neq k \leq n$), then $\mathbb{P}(\cup_{j=1}^n A_j) = \sum_{j=1}^n \mathbb{P}(A_j)$.

(ii') If $(A_n, \, n \geq 1)$ is a collection of disjoint events in $\mathcal{F}$, then $\mathbb{P}(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mathbb{P}(A_n)$.

The following properties can be further deduced from the above axioms (proofs are left as exercise):

(iii) If $A_1, \ldots, A_n \in \mathcal{F}$, then $\mathbb{P}(\cup_{j=1}^n A_j) \leq \sum_{j=1}^n \mathbb{P}(A_j)$.

(iii') If $(A_n, \, n \geq 1)$ is a collection of events in $\mathcal{F}$, then $\mathbb{P}(\cup_{n=1}^\infty A_n) \leq \sum_{n=1}^\infty \mathbb{P}(A_n)$.

(iv) If $A, B \in \mathcal{F}$ and $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$ and $\mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$. Also, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(v) If $A, B \in \mathcal{F}$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. This formula generalizes to the countable union of an arbitrary number of sets: it is called the inclusion-exclusion formula.

(vi) If $(A_n, \, n \geq 1)$ is a collection of events in $\mathcal{F}$ such that $A_n \subset A_{n+1}$, $\forall n \geq 1$, then $\mathbb{P}(\cup_{n=1}^\infty A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

(vi') If $(A_n, \, n \geq 1)$ is a collection of events in $\mathcal{F}$ such that $A_n \supset A_{n+1}$, $\forall n \geq 1$, then $\mathbb{P}(\cap_{n=1}^\infty A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

**Terminology.** The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*. Properties (ii), resp. (ii'), are referred to as the *additivity*, resp. *$\sigma$-additivity*, of probability measures. Properties (iii), resp. (iii'), are referred to as the *subadditivity*, resp. *sub-$\sigma$-addivity*, of probabilty measures (or more prosaically as the *union bound* sometimes).

**Example.** Let $\Omega = \{1, \ldots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ be the measurable space associated to a die roll. The probability measure associated to a balanced die is defined as

$$\mathbb{P}_1(\{i\}) = \frac{1}{6}, \ \forall i \in \{1, \ldots, 6\}$$

and is extended by additivity to all subsets of $\Omega$. E.g.,

$$\mathbb{P}_1(\{1,3,5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

The probability measure associated to a loaded die is defined as

$$\mathbb{P}_2(\{6\}) = 1 \quad \text{and} \quad \mathbb{P}_2(\{i\}) = 0, \ \forall i \in \{1, \ldots, 5\}$$

and is extended by additivity to all subsets of $\Omega$, so that for $A \subset \Omega$, $\mathbb{P}_2(A) = 1$ if $6 \in A$ and $\mathbb{P}_2(A) = 0$ otherwise.

In a discrete $\sigma$-field, once a probability measure is defined on the atoms of the $\sigma$-field, it is always possible to extend it by ($\sigma$-)additivity to the whole $\sigma$-field. In the general case, a similar statement holds true, but the extension procedure is much more complicated.

**Example.** Let $\Omega = [0,1]$ and $\mathcal{F} = \mathcal{B}([0,1])$. One defines the following probability measure on the subintervals of $[0,1]$:

$$\mathbb{P}(\,]a,b[\,) = b - a$$

**Fact.** [without proof] Carathéodory's extension theorem (see Appendix A.1 at the end of this document) states that $\mathbb{P}$ can be extended uniquely by $\sigma$-additivity to all Borel subsets of $[0,1]$. It is called the

*Lebesgue measure* on $[0,1]$ and is sometimes denoted as $\mathbb{P}(B)=|B|$. Note that it corresponds also to the *uniform distribution* on $[0,1]$.

**Examples.** - Let $\Omega=\mathbb{R}$ and $\mathcal{F}=\mathcal{B}(\mathbb{R})$. One can define the following probability measure on open intervals:

$$\mathbb{P}(\,]a,b[\,)=\int_a^b dx\,\frac{1}{\sqrt{2\pi}}\,\exp(-x^2/2)$$

Such a measure can be uniquely extended to all Borel subsets of $\mathbb{R}$: it is called the (normalized) Gaussian measure on $\mathbb{R}$.

- Let $\Omega=\mathbb{R}^2$ and $\mathcal{F}=\mathcal{B}(\mathbb{R}^2)$. One can define the following probability measure on open rectangles:

$$\mathbb{P}(\,]a,b[^2)=\int_{]a,b[^2} dxdy\,\frac{1}{2\pi}\,\exp(-(x^2+y^2)/2)$$

which can again be extended to all Borel subsets of $\mathbb{R}^2$.

**Remarks.** - One can also define the following measure on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$, by setting on open intervals:

$$\mathbb{P}(\,]a,b[\,)=b-a$$

This measure can be again uniquely extended to all Borel subsets of $\mathbb{R}$. It is however *not* a probability measure, as with this definition, one sees (using the above properties) that

$$\mathbb{P}(\mathbb{R})=\lim_{n\to\infty}\mathbb{P}(\,]-n,+n[\,)=\lim_{n\to\infty}2n=+\infty$$

This measure is called the *Lebesgue measure* on $\mathbb{R}$ and is again denoted as $\mathbb{P}(B)=|B|$ for $B\in\mathcal{B}(\mathbb{R})$.

- We see here that defining first $\mathbb{P}$ on the singletons $\{x\}$ (which are the atoms of $\mathcal{B}(\mathbb{R})$) instead of the open intervals $]a,b[$ would not be a good idea, as we would have $\mathbb{P}(\{x\})=0$, $\forall x\in\mathbb{R}$ for *both* the Gaussian measure and the Lebesgue measure on $\mathbb{R}$, although these are clearly different.

**Definition 2.2.** Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. An event $A\in\mathcal{F}$ is said to be *negligible* if $\mathbb{P}(A)=0$, resp. *almost sure* (abbreviated a.s.) if $\mathbb{P}(A)=1$.

**Remark.** The wording “almost sure” is far from ideal, but has been commonly agreed upon (Jacob Bernoulli would call such sets “morally certain” instead).

It should be emphasized that a negligible event need not be empty, nor need an almost sure event be equal to the whole space $\Omega$. Here are examples:

- In the probability space of a loaded die (see above), the set $\{1,2,3,4,5\}$ is a negligible event, while the singleton $\{6\}$ is an almost sure event.

- In the probability space $([0,1],\,\mathcal{B}([0,1]),\mathbb{P}=\text{Lebesgue measure})$, any singleton $\{x\}$ is negligible.

Here is moreover a general statement that can be made about negligible and almost sure sets.

**Proposition 2.3.**
- Let $(A_n,\,n\geq 1)$ be a collection of negligible events in $\mathcal{F}$. Then $\bigcup_{n\geq 1}A_n$ is also negligible.

- Let $(B_n,\,n\geq 1)$ be a collection of almost sure events in $\mathcal{F}$. Then $\bigcap_{n\geq 1}B_n$ is also almost sure.

*Proof.* By the sub-$\sigma$-additivity property (property (iii’) above),

$$\mathbb{P}\Big(\bigcup_{n\geq 1}A_n\Big)\leq\sum_{n\geq 1}\mathbb{P}(A_n)=\sum_{n\geq 1}0=0$$

which proves the first claim. The second claim is a consequence of the first one: consider $A_n = B_n^c$; then $\mathbb{P}(A_n) = \mathbb{P}(B_n^c) = 0$ by assumption and

$$\mathbb{P}\Big(\bigcap_{n\geq 1} B_n\Big) = 1 - \mathbb{P}\Big(\bigcup_{n\geq 1} A_n\Big) \geq 1 - \sum_{n\geq 1} \mathbb{P}(A_n) = 1 - 0 = 1$$

□

As a consequence, *any countable set* in $[0,1]$ is negligible with respect to the Lebesgue measure. In particular, $\mathbb{Q} \cap [0,1]$ is negligible! Perhaps more surprisingly, there exist also *uncountable* sets in $[0,1]$ which are negligible with respect to the Lebesgue measure (see below).

## 2.2 Distribution of a random variable

**Definition 2.4.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X$ be a random variable defined on this probability space. The *distribution* of $X$ is the map $\mu_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ defined as

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

**Remark.** The fact that $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$ implies that $\mu_X$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The triple $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$ forms therefore a new probability space.

**Notation.** If a random variable $X$ has distribution $\mu$, this is denoted as $X \sim \mu$. Likewise, if two random variables $X$ and $Y$ share the same distribution $\mu$, then they are said to be *identically distributed* and this is denoted as $X \sim Y \sim \mu$.

**Example 2.5.** The probability space describing two independent (and balanced) die rolls is $\Omega = \{1,\ldots,6\} \times \{1,\ldots,6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and

$$\mathbb{P}(\{(i,j)\}) = \frac{1}{36}, \quad \forall (i,j) \in \Omega$$

Let $X_1(i,j) = i$ be the result of the first die, and $Y(i,j) = i+j$ be the sum of the two dice. Then

$$\mu_{X_1}(\{i\}) = \mathbb{P}(\{X_1 = i\}) = \mathbb{P}(\{(i,1),\ldots,(i,6)\}) = \frac{6}{36} = \frac{1}{6}, \quad \forall i \in \{1,\ldots,6\}$$

and

$$\mu_Y(\{2\}) = \mathbb{P}(\{Y=2\}) = \mathbb{P}(\{(1,1)\}) = \frac{1}{36}, \quad \mu_Y(\{3\}) = \mathbb{P}(\{Y=3\}) = \mathbb{P}(\{(1,2),(2,1)\}) = \frac{1}{18}$$

More generally:

$$\mu_Y(\{i\}) = \frac{6 - |7-i|}{36}, \quad i \in \{2,\ldots,12\}$$

## 2.3 Cumulative distribution function

**Definition 2.6.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X$ be a random variable defined on this probability space. The *cumulative distribution function* (cdf) of $X$ is the map $F_X : \mathbb{R} \to [0,1]$ defined as

$$F_X(t) = \mu_X(\,]-\infty, t]) = \mathbb{P}(\{X \leq t\}), \quad t \in \mathbb{R}$$

**Fact.**[without proof] The knowledge of $F_X$ is equivalent to the knowledge of $\mu_X$. This fact is of course related to the fact that $\mathcal{B}(\mathbb{R}) = \sigma(\{\,]-\infty, t[,\ t \in \mathbb{R}\})$, but this observation alone does not allow to conclude directly.

From the properties of probability measures, one deduces that the cdf of a random variable satisfies the following properties:

(i) $\lim_{t\to-\infty} F_X(t) = 0$, $\lim_{t\to+\infty} F_X(t) = 1$.

(ii) $F_X$ is *non-decreasing*, i.e. $F_X(s) \le F_X(t)$ for all $s < t$.

(iii) $F_X$ is *right-continuous* on $\mathbb{R}$, i.e. $\lim_{\varepsilon\downarrow 0} F_X(t+\varepsilon) = F_X(t)$, for all $t \in \mathbb{R}$.

Indeed:

(i) $\lim_{t\to\infty} F_X(t) = \lim_{n\to\infty} F_X(n) = \lim_{n\to\infty} \mathbb{P}(\{X \le n\}) = 1$, as the sequence of events $\{X \le n\}$ is an increasing sequence with $\cup_{n\ge1}\{X \le n\} = \Omega$. The result then follows from the use of property (vi) listed above for probability measures. A similar reasoning shows that $\lim_{t\to-\infty} F_X(t) = \lim_{n\to\infty} F_X(-n) = \lim_{n\to\infty} \mathbb{P}(\{X \le -n\}) = 0$, by the fact that $\{X \le -n\}$ is this time a decreasing sequence of events with $\cap_{n\ge1}\{X \le -n\} = \emptyset$ and the use of property (vi') of probability measures.

(ii) If $s \le t$, then $\{X \le s\} \subset \{X \le t\}$, so $F_X(s) = \mathbb{P}(\{X \le s\}) \le \mathbb{P}(\{X \le t\}) = F_X(t)$.

(iii) For any $t \in \mathbb{R}$, we have $\lim_{\varepsilon\downarrow 0} F_X(t+\varepsilon) = \lim_{n\to\infty} F_X(t+\frac{1}{n}) = \lim_{n\to\infty} \mathbb{P}(\{X \le t+\frac{1}{n}\}) = \mathbb{P}(\{X \le t\}) = F_X(t)$, as again the sequence $\{X \le t+\frac{1}{n}\}$ is a decreasing sequence of events with $\cap_{n\ge1}\{X \le t+\frac{1}{n}\} = \{X \le t\}$.

**Important remarks.**
- Any function $F : \mathbb{R} \to \mathbb{R}$ satisfying the above properties (i), (ii) and (iii) is the cdf of a random variable [without proof].

- One cannot show that the cdf of a random variable is left-continuous (and therefore continuous) in general. Indeed, repeating the above argument, we obtain: for any $t \in \mathbb{R}$, $\lim_{\varepsilon\downarrow 0} F_X(t-\varepsilon) = \lim_{n\to\infty} F_X(t-\frac{1}{n}) = \lim_{n\to\infty} \mathbb{P}(\{X \le t-\frac{1}{n}\}) = \mathbb{P}(\{X < t\})$, as $\{X \le t-\frac{1}{n}\}$ is a increasing sequence of events with $\cup_{n\ge1}\{X \le t-\frac{1}{n}\} = \{X < t\}$. But $\mathbb{P}(\{X < t\}) \ne F_X(t)$ in general. It is wrong in particular for discrete random variables (see below).

- Any cdf $F_X$ has at most a countable number of jumps on the real line [without proof]. If $F_X$ has a jump of size $p \in [0,1]$ at $t \in \mathbb{R}$, this actually means that $\mathbb{P}(\{X = t\}) = F_X(t) - \lim_{\varepsilon\downarrow 0} F_X(t-\varepsilon) = p$. This implies in particular that a discrete random variable cannot take more than a countable number of values.

## 2.4 Two important classes of random variables

**Discrete random variables.**

**Definition 2.7.** $X$ is a *discrete random variable* if it takes values in a discrete (i.e., finite or countable) subset $D$ of $\mathbb{R}$, that is, $X(\omega) \in D$ for every $\omega \in \Omega$.

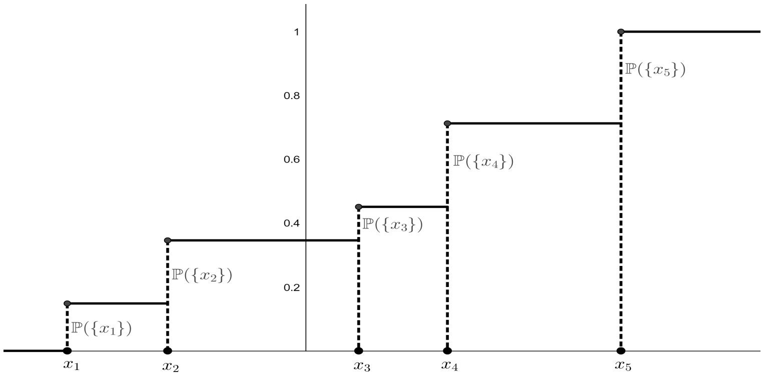
The distribution of a discrete random variable is entirely characterized by the numbers $p_x = \mathbb{P}(\{X = x\})$, where $x \in D$. Note that $0 \le p_x \le 1$ for all $x \in D$ and that $\sum_{x\in D} p_x = \mathbb{P}(\{X \in D\}) = 1$. The sequence of numbers $(p_x,\, x \in D)$ is sometimes called the *probability mass function* (pmf) of the random variable $X$. It should not be confused with the *probability density function* (pdf) defined below for continuous random variables only. One further has:

$$\mu_X(B) = \mathbb{P}(\{X \in B\}) = \sum_{x\in D\cap B} p_x, \quad \forall B \in \mathcal{B}(\mathbb{R})$$

and

$$F_X(t) = \mathbb{P}(\{X \le t\}) = \sum_{x\in D,\, x\le t} p_x, \quad \forall t \in \mathbb{R}$$

is a step function, as illustrated below:



**Example.** A binomial random variable $X$ with parameters $n \geq 1$ and $p \in [0,1]$ (denoted as $X \sim \text{Bi}(n,p)$) takes values in $\{0,\ldots,n\}$ and is characterized by the numbers

$$p_k = \mathbb{P}(\{X = k\}) = \binom{n}{k}\, p^k\,(1-p)^{n-k}, \quad k \in \{0,\ldots,n\}$$

where $\displaystyle\binom{n}{k} = \frac{n!}{k!(n-k)!}$ are the binomial coefficients.

**Continuous random variables.**

**Definition 2.8.** $X$ is a *continuous random variable* if $\mathbb{P}(\{X \in B\}) = 0$ whenever $B \in \mathcal{B}(\mathbb{R})$ is such that $|B| = 0$ (remember that $|B|$ is the Lebesgue measure of $B$).

In particular, this implies that if $X$ is a continuous random variable, then $\mathbb{P}(\{X = x\}) = 0\ \forall x \in \mathbb{R}$ (as $|\{x\}| = 0\ \forall x \in \mathbb{R}$). But please note that this last condition is *not* sufficient to guarantee that $X$ is a continuous random variable.

**Fact.** [without proof][1] If $X$ is a continuous random variable according to the above definition, then there exists a Borel-measurable function $p_X : \mathbb{R} \to \mathbb{R}$, called the *probability density function* (pdf) of $X$, such that $p_X(x) \geq 0\ \forall x \in \mathbb{R}$, $\int_{\mathbb{R}} p_X(x)\,dx = 1$ and

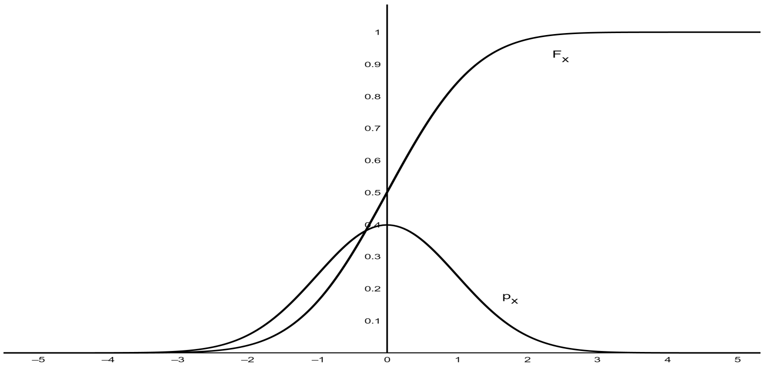
$$\mu_X(B) = \mathbb{P}(\{X \in B\}) = \int_B p_X(x)\,dx, \quad \forall B \in \mathcal{B}(\mathbb{R})$$

Moreover,

$$F_X(t) = \mathbb{P}(\{X \leq t\}) = \int_{-\infty}^{t} p_X(x)\,dx, \quad \forall t \in \mathbb{R}$$

is a continuous and "differentiable" function (whose "derivative" is $F'_X(t) = p_X(t)$, but watch out that $p_X$ need not be continuous in general, but only Borel-measurable, so $F_X$ is not differentiable in the classical sense). This is illustrated below:

[1] This fact is known as the Radon-Nikodym theorem. It has many different formulations and important applications in probability theory.



**Important remarks.**
- $p_X(x) \neq \mathbb{P}(\{X = x\})$, simply because $\mathbb{P}(\{X = x\}) = 0$ for all $x \in \mathbb{R}$.

- $p_X(x) \geq 0$, but as this quantity is not a probability, it is perfectly possible that $p_X(x) > 1$ for some values of $x$. The only requirement is that the *integral* of $p_X(x)$ over $\mathbb{R}$ is equal to 1.

**Example.** A Gaussian random variable $X$ with mean $\mu$ and variance $\sigma^2$ (denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$) takes values in $\mathbb{R}$ and has pdf

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

so in particular, $p_X(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} > 1$ if $\sigma < \frac{1}{\sqrt{2\pi}}$.

**Remark.** One could think that the only existing distributions are either discrete or continuous, or a combination of these. It turns out that life is more complicated than that! Some distributions are neither discrete, nor continuous. A famous example is the distribution whose cdf is the *devil's staircase*, as we shall see below.

**Change of variables.** Let $X$ be a generic random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $f : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function and $Y = f(X)$. Assume we know $F_X$ the cdf of $X$; what can we say on $F_Y$ the cdf of $Y$? Assume for example that $f$ is increasing, i.e., that $f(x_1) < f(x_2)$ as soon as $x_1 < x_2$. Let also $D = \{y \in \mathbb{R} : \exists x \in \mathbb{R} \text{ such that } f(x) = y\}$ be the *range* of the function $f$. Then $f : \mathbb{R} \to D$ is (Borel-measurable and) invertible, so for each point $t \in D$, we have

$$F_Y(t) = \mathbb{P}(\{Y \leq t\}) = \mathbb{P}(\{f(X) \leq t\}) = \mathbb{P}(\{X \leq f^{-1}(t)\}) = F_X(f^{-1}(t))$$

If in addition $X$ is a continuous random variable with pdf $p_X$ and $f$ is differentiable (with $f'(x) > 0$ for all $x \in \mathbb{R}$, as $f$ is increasing), then $Y$ is a continuous random variable also and

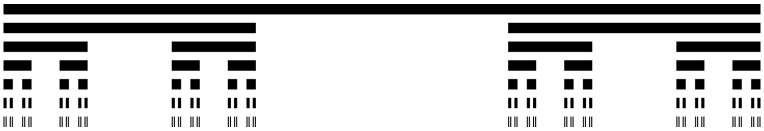
$$p_Y(t) = \frac{dF_Y(t)}{dt} = \frac{dF_X(f^{-1}(t))}{dt}\,\frac{df^{-1}(t)}{dt} = \frac{p_X(f^{-1}(t))}{f'(f^{-1}(t))}$$

Setting $x = f^{-1}(t)$ in the above relation, we get the more natural formula $p_X(x) = p_Y(f(x))\, f'(x)$. Similar reasonings allow to deal with $f$ decreasing or more general cases.

**Remark.** In the case where $X$ is a continuous random variable and $f$ is assumed to be non-decreasing only (i.e., $f(x_1) \leq f(x_2)$ if $x_1 < x_2$), note that $f$ is not necessarily invertible in this case, so that the above formulas do not hold. Also, $Y = f(X)$ need not be a continuous random variable in this case (consider for example the case where $f(x) = 1_{\{x \geq 0\}}$; $Y$ is a then discrete random variable taking values in $\{0, 1\}$ only).

## 2.5 The Cantor set and the devil's staircase

The Cantor set is a subset $C$ of $[0,1]$ obtained by removing recursively "middle intervals" as follows:



Mathematically, for $n \geq 1$, let

$$A_n = \bigcup_{a_1,\ldots,a_{n-1}\in\{0,2\}} \left]\sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{1}{3^n}, \sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{2}{3^n}\right[$$

be the set of (open) intervals removed at stage $n$. In particular, $A_1 = ]\frac{1}{3}, \frac{2}{3}[$, $A_2 = ]\frac{1}{9}, \frac{2}{9}[\bigcup]\frac{7}{9}, \frac{8}{9}[$, etc. Note also that $A_n$ and $A_m$ are *disjoint* for every $n \neq m$. The Cantor set $C$ is then defined as

$$C = [0,1]\backslash \bigcup_{n\geq 1} A_n$$

This set has strange properties: first observe that the Lebesgue measure of each $A_n$ is given by

$$|A_n| = \frac{2^{n-1}}{3^n} \quad \text{(each set } A_n \text{ is indeed made of } 2^{n-1} \text{ disjoint intervals, each of length } \frac{1}{3^n}\text{)}$$

so that, using the fact that the $A_n$ are disjoint as well as the formula for geometric series, we obtain:

$$|C| = 1 - \sum_{n\geq 1} \frac{2^{n-1}}{3^n} = 1 - \frac{1}{2}\left(\frac{1}{1-\frac{2}{3}} - 1\right) = 1 - \frac{1}{2}(3-1) = 0$$

The Cantor set $C$ has therefore Lebesgue measure 0. Surprisingly, $C$ is also uncountable. This can be seen as follows: any number $x \in [0,1]$ can be written using its binary decomposition:

$$x = \sum_{n\geq 1} \frac{b_n}{2^n}, \quad \text{where} \quad b_n \in \{0,1\} \tag{2}$$

Likewise, any number $x \in [0,1]$ can also be written using its ternary decomposition:

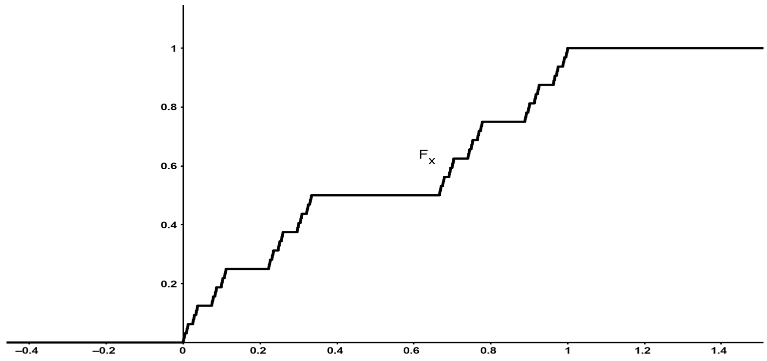
$$x = \sum_{n\geq 1} \frac{a_n}{3^n}, \quad \text{where} \quad a_n \in \{0,1,2\}$$

It is a fact [without proof] that any $x \in C$ can be written as

$$x = \sum_{n\geq 1} \frac{a_n}{3^n}, \quad \text{where} \quad a_n \in \{0,2\} \tag{3}$$

i.e., $x \in C$ if and only if $a_n \neq 1$ for every $n \geq 1$. Comparing formulas (2) and (3), we see that the sets $[0,1]$ and $C$ are in bijection with each other (the bijection being $b_n = 0 \longleftrightarrow a_n = 0$ and $b_n = 1 \longleftrightarrow a_n = 2$), proving that $C$ is uncountable, because $[0,1]$ is (the proof that the set $[0,1]$ is uncountable is by the way also due to Cantor and is called the *diagonalization argument*).

Let us now turn to the devil's staircase. This strange cdf has the following shape:



It can be defined recursively as follows: $F(t) = \frac{1}{2}$ for $t \in ]\frac{1}{3}, \frac{2}{3}[$, $F(t) = \frac{1}{4}$ for $t \in ]\frac{1}{9}, \frac{2}{9}[$, $F(t) = \frac{3}{4}$ for $t \in ]\frac{7}{9}, \frac{8}{9}[$, etc. Formally, one can define $F$ on the sets $A_n$ as follows:

$$F(t) = \sum_{k=1}^{n-1} \frac{a_k/2}{2^k} + \frac{1}{2^n} \quad \text{for } t \in \left] \sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{1}{3^n}, \sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{2}{3^n} \right[$$

It is then a fact that $F$ can be *extended by continuity* to all $t \in [0,1]$ [without proof, but the picture above should convince you, as well as the following argument: for $t, s \in [0,1]$, if $|t-s| \leq 1/3^n$, then $|F(t) - F(s)| \leq 1/2^n$; this implies that $F$ is not only continuous on $[0,1]$, but also *uniformly* continuous [2]].

Note now its strange properties: $F(0) = 0$, $F(1) = 1$, $F$ is non-decreasing on $[0,1]$, and on any set $A_n$, $F$ is *flat*, so that $F'(t) = 0$ for all $t \in \cup_{n \geq 1} A_n$, which is the complement of $C$ on $[0,1]$. This is saying more precisely that the set where $F$ is flat has full Lebesgue measure on the interval $[0,1]$, so that the function $F$ is *almost* flat. Moreover, we just said above that $F$ is *continuous* on the interval $[0,1]$.

If you think for a while, all these properties seem to contradict each other, but actually, they don't! At the beginning of the 20th century, the work of Cantor led to a revolution in mathematics...

The next question is: where to classify the devil's staircase, i.e., is it the cdf of a continuous or of a discrete random variable?

Let us first try to see it as the cdf of a continuous variable. We have seen that $F'(t) = 0$ for any $t \notin C$. Therefore, if $F$ were to admit a pdf, this pdf would be equal to 0 almost everywhere on $[0,1]$. But then, such a function cannot integrate to 1 on the interval $[0,1]$. $F$ is therefore not the cdf of a continuous random variable (even though it is itself a continuous function).

Let us now try to view $F$ as the cdf of a discrete random variable and look for the corresponding pmf. From the definition of $F$, it is clear that the pmf assigns no weight to elements $t \notin C$. Using the symmetry of the function $F$, one could then perhaps argue that the pmf should be the uniform distribution on $C$. But as we have seen above, $C$ is uncountable, so in particular infinite. Such a uniform discrete distribution on $C$ does therefore not exist!

One may still argue that $F$ is the cdf of the uniform distribution on $C$, as well as $F(t) = t$ is the cdf of the uniform distribution on $[0,1]$. This raises the question: what does that mean to pick a point uniformly in $C$? The same question is equally valid with $C$ replaced by $[0,1]$, actually...

[2] But please note that the function $F$ is not *absolutely* continuous: this would actually mean that $F$ admits a pdf.

# 3 Independence

The notion of independence is a central notion in probability. It is usually defined for events and random variables in elementary probability courses. Nevertheless, as it will become clear below, the independence between *σ-fields* turns out to be the most natural concept (remembering that a $\sigma$-field is related to the amount of information one has on a system).

In the following subsections, all events, random variables and sub-$\sigma$-fields are defined in a *common* probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

## 3.1 Independence of two events

**Definition 3.1.** Two events $A_1, A_2 \in \mathcal{F}$ are *independent* if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)$.

**Remark.** Although this definition is quite standard, a more explanatory definition of independence of two events is given by using conditional probabilities: we say that $A_1$ and $A_2$ are independent if $\mathbb{P}(A_1|A_2) = \mathbb{P}(A_1)$, which is saying that the realization of event $A_2$ has no influence on the probability that $A_1$ happens. Using the formula for the conditional probability $\mathbb{P}(A_1|A_2) = \mathbb{P}(A_1 \cap A_2)/\mathbb{P}(A_2)$, we then recover the above definition. We will come back later to conditional probability, which is a central concept in probability theory.

**Notation.** $A_1 \perp\!\!\!\perp A_2$.

**Proposition 3.2.** If two events $A_1, A_2 \in \mathcal{F}$ are independent, then it also holds that

$$\mathbb{P}(A_1 \cap A_2^c) = \mathbb{P}(A_1)\,\mathbb{P}(A_2^c), \quad \mathbb{P}(A_1^c \cap A_2) = \mathbb{P}(A_1^c)\,\mathbb{P}(A_2) \quad \text{and} \quad \mathbb{P}(A_1^c \cap A_2^c) = \mathbb{P}(A_1^c)\,\mathbb{P}(A_2^c)$$

*Proof.* We show here the first equality (noticing that the other two can be proved in a similar way):

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2^c) &= \mathbb{P}(A_1 \backslash (A_1 \cap A_2)) = \mathbb{P}(A_1) - \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) - \mathbb{P}(A_1)\,\mathbb{P}(A_2)\\ &= \mathbb{P}(A_1)\,(1 - \mathbb{P}(A_2)) = \mathbb{P}(A_1)\,\mathbb{P}(A_2^c)\end{aligned}$$

□

Note that the above proposition says actually something very natural. Let us assume for example that one rolls a balanced die with four faces. Then the events {the outcome is 1 or 2} and {the outcome is even} are independent; more precisely, the different *informations* associated with these events are. So the events {the outcome is 1 or 2} and {the outcome is odd} are also independent. This will motivate the extension of the definition of independence to $\sigma$-fields below.

## 3.2 Independence of two random variables

**Definition 3.3.** Two $\mathcal{F}$-measurable random variables $X_1, X_2$ are independent if

$$\mathbb{P}(\{X_1 \le t_1, X_2 \le t_2\}) = \mathbb{P}(\{X_1 \le t_1\})\,\mathbb{P}(\{X_2 \le t_2\}), \quad \forall t_1, t_2 \in \mathbb{R}$$

**Notation.** $X_1 \perp\!\!\!\perp X_2$.

**Example.** Let $X_0(\omega) = c \in \mathbb{R}, \forall \omega \in \Omega$ be a constant random variable. According to the above definition, $X_0$ is independent of any other random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

An immediate question following this definition is the following: let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be two Borel-measurable functions and $Y_1 = f_1(X_1)$, $Y_2 = f_2(X_2)$. Are $Y_1$ and $Y_2$ also independent? A priori, answering this question requires computing the joint distribution of $Y_1$ and $Y_2$ (which might be difficult depending on the functions $f_1$ and $f_2$), but the following proposition [given here without proof, but connected to the fact that the knowledge of a distribution $\mu_X$ is equivalent to that of its cdf $F_X$] allows for a much cleaner answer.

**Proposition 3.4.** $X_1, X_2$ are independent if and only if

$$\mathbb{P}(\{X_1 \in B_1, X_2 \in B_2\}) = \mathbb{P}(\{X_1 \in B_1\})\,\mathbb{P}(\{X_2 \in B_2\}), \quad \forall B_1, B_2 \in \mathcal{B}(\mathbb{R})$$

From this, we deduce the following:

**Proposition 3.5.** Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be two Borel-measurable functions. If $X_1, X_2$ are independent random variables, then $Y_1 = f_1(X_1)$ and $Y_2 = f_2(X_2)$ are also independent random variables.

*Proof.* From the assumption made, we have for every $B_1, B_2 \in \mathcal{B}(\mathbb{R})$:

$$\begin{aligned}\mathbb{P}(\{Y_1 \in B_1, Y_2 \in B_2\}) &= \mathbb{P}(\{f_1(X_1) \in B_1, f_2(X_2) \in B_2\}) = \mathbb{P}(\{X_1 \in f_1^{-1}(B_1), X_2 \in f_2^{-1}(B_2)\}) \\ &= \mathbb{P}(\{X_1 \in f_1^{-1}(B_1)\})\,\mathbb{P}(\{X_2 \in f_2^{-1}(B_2)\}) = \mathbb{P}(\{f_1(X_1) \in B_1\})\,\mathbb{P}(\{f_2(X_2) \in B_2\}) \\ &= \mathbb{P}(\{Y_1 \in B_1\})\,\mathbb{P}(\{Y_2 \in B_2\})\end{aligned}$$

Note that $f_1, f_2$ need not be invertible for the above equalities to hold: $f_i^{-1}(B_i)$ is just a notation for the pre-image of $B_i$ via the function $f_i$. □

Further simplifications of Definition 3.3 occur in the two following situations [again, without proof]:

- Assume $X_1, X_2$ are two discrete random variables, taking values in a common discrete set $D$ [3]. Then $X_1, X_2$ are independent if and only if

$$\mathbb{P}(\{X_1 = x_1, X_2 = x_2\}) = \mathbb{P}(\{X_1 = x_1\})\,\mathbb{P}(\{X_2 = x_2\}), \quad \forall x_1, x_2 \in D$$

**Example.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space describing two independent die rolls in Example 2.5 and let $X_1(i,j) = i$ and $X_2(i,j) = j$. One verifies below that these two random variables are indeed independent. It was already shown that $\mathbb{P}(\{X_1 = i\}) = \frac{1}{6}$, $\forall i \in \{1, \ldots, 6\}$. Likewise, $\mathbb{P}(\{X_2 = j\}) = \frac{1}{6}$, $\forall j \in \{1, \ldots, 6\}$ and

$$\mathbb{P}(\{X_1 = i, X_2 = j\}) = \mathbb{P}(\{(i,j)\}) = \frac{1}{36} = \mathbb{P}(\{X_1 = i\})\,\mathbb{P}(\{X_2 = j\}), \quad \forall (i,j) \in \Omega$$

so $X_1$ and $X_2$ are independent.

- Assume now $X_1, X_2$ are *jointly continuous* random variables, that is, there exists a Borel-measurable function $p_{X_1,X_2} : \mathbb{R}^2 \to \mathbb{R}_+$ (=joint pdf) such that

$$\mathbb{P}(\{(X_1, X_2) \in B\}) = \int_B p_{X_1,X_2}(x_1, x_2)\,dx_1 dx_2, \quad \forall B \in \mathcal{B}(\mathbb{R}^2)$$

Then $X_1, X_2$ are independent if and only if the function $p_{X_1,X_2}$ can be factorized as follows:

$$p_{X_1,X_2}(x_1, x_2) = p_{X_1}(x_1)\,p_{X_2}(x_2), \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

## 3.3 Independence of two sub-$\sigma$-fields

The above two definitions of independence for events and random variables can actually be seen as particular instances of a more general definition, concerning the independence of sub-$\sigma$-fields, that is to say, the independence of two different types of *information* one may have on a system.

**Definition 3.6.** Two sub-$\sigma$-fields $\mathcal{G}_1, \mathcal{G}_2$ of $\mathcal{F}$ are independent if

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2), \quad \forall A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2$$

[3] This can always be assumed for discrete random variables: indeed, if $X_1 \in D_1$ and $X_2 \in D_2$, then simply consider $D = D_1 \cup D_2$, which is also discrete.

**Notation.** $\mathcal{G}_1 \perp\!\!\!\perp \mathcal{G}_2$.

One can readily check (using in particular Proposition 3.2 for the first line) that

- $A_1, A_2$ are independent according to Definition 3.1 if and only if $\sigma(A_1), \sigma(A_2)$ are independent according to Definition 3.6.

- $X_1, X_2$ are independent according to Definition 3.3 if and only if $\sigma(X_1), \sigma(X_2)$ are independent according to Definition 3.6.

## 3.4 Independence of more sub-$\sigma$-fields

The notion of independence of more than two $\sigma$-fields generalizes easily as follows.

**Definition 3.7.** Let $\{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$ be a finite collection of sub-$\sigma$-fields of $\mathcal{F}$. This collection is independent if

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n), \quad \forall A_1 \in \mathcal{G}_1, \ldots, A_n \in \mathcal{G}_n$$

A finite collection of events $\{A_1, \ldots, A_n\}$ is declared to be independent if $\{\sigma(A_1), \ldots, \sigma(A_n)\}$ is independent, which is equivalent to saying that

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

where $A_i^* =$ either $A_i$ or $A_i^c$, $i \in \{1, \ldots, n\}$. Note that for $n > 2$, verifying only that

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

does not suffice to guarantee independence of the whole collection of events (i.e., Proposition 3.2 does not generalize to the case $n > 2$). One can actually show that independence of $\{A_1, \ldots, A_n\}$ holds if and only if

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j), \quad \forall J \subset \{1, \ldots, n\} \text{ such that } J \neq \emptyset$$

Note also that pairwise independence (i.e., $\mathbb{P}(A_j \cap A_k) = \mathbb{P}(A_j)\,\mathbb{P}(A_k)$ for every $j \neq k$) is not enough to ensure the independence of the whole collection.

A finite collection of random variables $\{X_1, \ldots, X_n\}$ is similarly declared to be independent if $\{\sigma(X_1), \ldots, \sigma(X_n)\}$ is independent, which is equivalent to saying that

$$\mathbb{P}(\{X_1 \in B_1, \ldots, X_n \in B_n\}) = \mathbb{P}(\{X_1 \in B_1\}) \cdots \mathbb{P}(\{X_n \in B_n\}), \quad \forall B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$$

and all the simplifications seen above apply similarly.

Finally, one can further generalize independence to an *arbitrary* (i.e., not necessarily countable) collection of sub-$\sigma$-fields of $\mathcal{F}$.

**Definition 3.8.** Let $\{\mathcal{G}_i,\, i \in I\}$ be an arbitrary collection of sub-$\sigma$-fields of $\mathcal{F}$. This collection is independent if *any finite sub-collection* $\{G_{i_1}, \ldots, \mathcal{G}_{i_n}\}$ is independent.

Infinite collections of sub-$\sigma$-fields or random variables occur in various contexts, but most prominently when dealing with *stochastic processes*, as we shall see during this course.

**Remark.** For a countable collection of events $(A_n,\, n \in \mathbb{N})$, it would not be a good idea to define independence directly as

$$\mathbb{P}(\cap_{n \in \mathbb{N}} A_n^*) = \prod_{n \in \mathbb{N}} \mathbb{P}(A_n^*)$$

as in many cases, this equality would simply read “$0 = 0$”.

## 3.5 Do independent random variables really exist ???

An innocent sentence such as "Let $X_1, X_2, X_3, \ldots$ be an infinite collection of independent random variables. . . " immediately raises a question: does there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which these random variables could be defined altogether? The answer is (fortunately for the remainder of this course. . . ): yes, but as we will see, the set $\Omega$ needed to fit all these independent random variables is quite large (to say the least!).

- Let us start by exploring what $\Omega$ is needed for only two independent random variables $X_1 \sim \mu_1$ and $X_2 \sim \mu_2$, with $\mu_1, \mu_2$ two distributions on $\mathbb{R}$. In this case, the set $\Omega = \mathbb{R}^2$ suffices. Indeed, for $\omega = (\omega_1, \omega_2) \in \mathbb{R}^2$, let us first define $X_1(\omega) = \omega_1$ and $X_2(\omega) = \omega_2$. Now, what about $\mathcal{F}$ and $\mathbb{P}$? For $\mathcal{F}$, we will consider the $\sigma$-field generated by the "rectangles" of the form

$$B_1 \times B_2, \quad \text{with } B_1, B_2 \in \mathcal{B}(\mathbb{R})$$

which is nothing but the already encountered Borel $\sigma$-field $\mathcal{B}(\mathbb{R}^2)$. As for $\mathbb{P}$, we set it to be given by

$$\mathbb{P}(B_1 \times B_2\,) = \mu_1(B_1)\,\mu_2(B_2)$$

on the "rectangles". Carathéodory's extension theorem then ensures (see Appendix A.1) that $\mathbb{P}$ can be uniquely extended to $\mathcal{B}(\mathbb{R}^2)$. With these definitions in hand, one can check that for every $B_1, B_2 \in \mathcal{B}(\mathbb{R})$, one has

$$\mathbb{P}(\{X_1 \in B_1, X_2 \in B_2\}) = \mathbb{P}(\{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_1 \in B_1, \omega_2 \in B_2\}) = \mathbb{P}(B_1 \times B_2) = \mu_1(B_1) \cdot \mu_2(B_2)$$

while

$$\mathbb{P}(\{X_1 \in B_1\}) = \mathbb{P}(\{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_1 \in B_1\}) = \mathbb{P}(B_1 \times \mathbb{R}) = \mu_1(B_1) \cdot \mu_2(\mathbb{R}) = \mu_1(B_1)$$

and similarly, $\mathbb{P}(\{X_2 \in B_2\}) = \mu_2(B_2)$, proving the claim that $X_1$ and $X_2$ are independent.

- For an infinite (yet countable, but this can be further generalized) collection of random variables, things are slightly more complicated, but the basic principle remains the same. First, the set $\Omega$ needed in this case becomes

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots) : \omega_n \in \mathbb{R},\, \forall n \geq 1\} = \mathbb{R}^{\mathbb{N}^*}$$

which can be viewed either as the set of infinite sequences of real numbers, or equivalently as the set of functions from $\mathbb{N}^*$ to $\mathbb{R}$. We then define

$$X_n(\omega) = \omega_n, \quad \forall n \geq 1$$

Now comes the trouble: what about $\mathcal{F}$ and $\mathbb{P}$? For $\mathcal{F}$, we take as before the $\sigma$-field generated by the "rectangles" in $\mathbb{R}^{\mathbb{N}^*}$, which are of the form:

$$B_1 \times B_2 \times \ldots \times B_n \times \mathbb{R} \times \mathbb{R} \times \ldots$$

where $n$ is now an arbitrary positive integer. Note that the $\sigma$-field $\mathcal{F}$ is quite large! Nevertheless, defining $\mathbb{P}$ on the above rectangles remains simple:

$$\mathbb{P}(B_1 \times B_2 \times \ldots \times B_n \times \mathbb{R} \times \mathbb{R} \times \ldots) = \mu_1(B_1) \cdot \mu_2(B_2) \cdots \mu_n(B_n)$$

and Carathéodory's extension theorem ensures again that $\mathbb{P}$ can be uniquely extended to $\mathcal{F}$. It is then quite easy to see that with all these definitions,

$$\mathbb{P}(\{X_1 \in B_1,\, X_2 \in B_2, \ldots, X_n \in B_n\}) = \mathbb{P}(\{X_1 \in B_1\}) \cdot \mathbb{P}(\{X_2 \in B_2\}) \cdots \mathbb{P}(\{X_n \in B_n\})$$

for any fixed $n \geq 1$, which was our aim.

### 3.6 Convolution

In general, we can say that the distribution of the sum of two *independent* random variables $X_1, X_2$ is the *convolution* of the distributions of $X_1$ and $X_2$. Let us first see what this means in two special cases.

**Discrete case.** Let $X_1, X_2$ be two independent discrete random variables, both with values in $\mathbb{Z}$. Let us compute for $k \in \mathbb{Z}$:

$$\begin{aligned}\mathbb{P}(\{X_1 + X_2 = k\}) &= \sum_{j\in\mathbb{Z}} \mathbb{P}(\{X_1 = j, X_1 + X_2 = k\}) = \sum_{j\in\mathbb{Z}} \mathbb{P}(\{X_1 = j, X_2 = k - j\}) \\ &= \sum_{j\in\mathbb{Z}} \mathbb{P}(\{X_1 = j\})\, \mathbb{P}(\{X_2 = k - j\})\end{aligned}$$

which is nothing but the discrete convolution of the probability mass functions of $X_1$ and $X_2$. Note that following the same reasoning, we also obtain for $t \in \mathbb{R}$:

$$\mathbb{P}(\{X_1 + X_2 \le t\}) = \sum_{j\in\mathbb{Z}} \mathbb{P}(\{X_1 = j\})\, \mathbb{P}(\{X_2 \le t - j\})$$

which may in turn be rewritten as

$$F_{X_1+X_2}(t) = \sum_{j\in\mathbb{Z}} \mathbb{P}(\{X_1 = j\}) F_{X_2}(t - j) \tag{4}$$

**Continuous case.** Let $X_1, X_2$ be two continuous and independent random variables, with joint pdf $p_{X_1,X_2}$. Then

$$F_{X_1+X_2}(t) = \mathbb{P}(\{X_1 + X_2 \le t\}) = \int_{\mathcal{D}(t)} dx_1 dx_2\, p_{X_1,X_2}(x_1, x_2)$$

where $\mathcal{D}(t) = \{(x_1, x_2) \in \mathbb{R}^2 \,:\, x_1 + x_2 \le t\}$. As $p_{X_1,X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2}(x_2)$, we further obtain:

$$F_{X_1+X_2}(t) = \int_{\mathbb{R}} dx_1\, p_{X_1}(x_1) \int_{-\infty}^{t-x_1} dx_2\, p_{X_2}(x_2)$$

which may be rewritten as

$$F_{X_1+X_2}(t) = \int_{\mathbb{R}} dx_1\, p_{X_1}(x_1)\, F_{X_2}(t - x_1) \tag{5}$$

from which we deduce[4] that the random variable $X_1 + X_2$ is also continuous with pdf

$$p_{X_1+X_2}(t) = \frac{d}{dt} F_{X_1+X_2}(t) = \int_{\mathbb{R}} dx_1\, p_{X_1}(x_1)\, \frac{d}{dt} F_{X_2}(t - x_1) = \int_{\mathbb{R}} dx_1\, p_{X_1}(x_1)\, p_{X_2}(t - x_1)$$

which is the classical convolution of $p_{X_1}$ and $p_{X_2}$.

**General case.** The expressions (4) and (5) share some similarity. In the general case, the cdf of the sum of two independent random variables $X_1, X_2$ can be expressed as

$$F_{X_1+X_2}(t) = \int_{\mathbb{R}} dF_{X_1}(x_1)\, F_{X_2}(t - x_1)$$

which coincides with expressions (4) and (5) in the discrete and continuous cases [the above integral $\int_{\mathbb{R}} dF_{X_1}(x) \ldots$, known as the *Lebesgue-Stieltjes integral*, is a generalization of the classical Riemann integral]. One uses the following short-hand notations for the above relation:

$$F_{X_1+X_2} = F_{X_1} * F_{X_2} \quad \text{or} \quad \mu_{X_1+X_2} = \mu_{X_1} * \mu_{X_2}$$

or also $p_{X_1+X_2} = p_{X_1} * p_{X_2}$ in the continuous case.

[4] modulo a permutation of derivative and integral, which we will simply admit here

**Example in the continuous case.** Let $X_1, X_2$ be two i.i.d.$\sim \mathcal{E}(1)$ random variables, with common cdf

$$F_{X_1}(t) = F_{X_2}(t) = \begin{cases} 1 - \exp(-t) & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

The cdf of $X_1 + X_2$ is then given by formula (5):

$$F_{X_1+X_2}(t) = \int_0^t dx_1 \, \exp(-x_1) \, (1 - \exp(-(t - x_1)) = 1 - \exp(-t) - t \, \exp(-t) \quad \text{for } t \geq 0$$

(and $F_{X_1+X_2}(t) = 0$ for $t < 0$). From this, we deduce that

$$p_{X_1+X_2}(t) = \frac{d}{dt} F_{X_1+X_2}(t) = t \, \exp(-t) \quad \text{for } t \geq 0$$

(and $p_{X_1+X_2}(t) = 0$ for $t < 0$). Of course, we could also have deduced this directly from the expression given for the convolution of the pdfs $p_{X_1}$ and $p_{X_2}$.

# 4 Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable random variable. We are interested in defining the *expectation* $\mathbb{E}(X)$ of this random variable, also known as its theoretical average. In the simple case where $X$ is a Bernoulli random variable with parameter $0 < p < 1$ (i.e., $\mathbb{P}(\{X = 1\}) = p = 1 - \mathbb{P}(\{X = 0\})$, the expectation of $X$ is given by

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(\{X = 1\}) + 0 \cdot \mathbb{P}(\{X = 0\}) = 1 \cdot p + 0 \cdot (1 - p) = p$$

*Remark.* What does it mean that $p$ is the average value of $X$, when $X$ actually *never* takes this precise value $p$? We will need the law of large numbers to answer this question: this is coming later.

In order to define $\mathbb{E}(X)$, we will aim here for a more ambitious goal, which is to define directly $\mathbb{E}(g(X))$, where $g : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function (recall from the first lecture that we know that $Y = g(X)$ is also an $\mathcal{F}$-measurable random variable in this case).

**Step 1.** Assume first that $g$ is a *simple non-negative function*, i.e., that

$$g(x) = \sum_{i \geq 1} y_i \, 1_{B_i}(x)$$

where $y_i$ are non-negative numbers and $B_i$ are disjoint Borel subsets of $\mathbb{R}$ (implying that $g$ is a Borel-measurable function). In this case, the expectation is defined as

$$\mathbb{E}(g(X)) = \sum_{i \geq 1} y_i \, \mathbb{P}(\{X \in B_i\}) = \sum_{i \geq 1} y_i \, \mathbb{P}(\{g(X) = y_i\})$$

Observe the following:

- As $X$ is $\mathcal{F}$-measurable and every $B_i$ is a Borel set by assumption, the sets $\{X \in B_i\} \in \mathcal{F}$.

- The above sum, being a sum of non-negative numbers, is non-negative. Being also a potentially infinite sum, it might take the value $+\infty$, which is OK.

- A simple example is the following: for a given $t \in \mathbb{R}$, consider $g(x) = 1_{\{x \leq t\}}$. Then

$$\mathbb{E}(g(X)) = \mathbb{E}(1_{\{X \leq t\}}) = \mathbb{P}(\{X \leq t\}) = F_X(t)$$

**Step 2.** Assume now that $g$ is a generic *non-negative Borel-measurable function.* For $n \geq 1$, we define

$$g_n(x) = \sum_{i \geq 1} \frac{i-1}{2^n} \, 1_{]\frac{i-1}{2^n}, \frac{i}{2^n}]}(g(x)), \quad x \in \mathbb{R}$$

Observe that these functions $g_n$ are simple non-negative functions. Also, for every value of $x \in \mathbb{R}$ and $n \geq 1$, it holds that

$$g_n(x) \leq g_{n+1}(x) \leq g(x)$$

For each $n \geq 1$, the definition of $\mathbb{E}(g_n(X))$ is given in Step 1:

$$\mathbb{E}(g_n(X)) = \sum_{i \geq 1} \frac{i-1}{2^n} \, \mathbb{P}\left(\left\{\frac{i-1}{2^n} < g(X) \leq \frac{i}{2^n}\right\}\right)$$

Because $g_n(x) \leq g_{n+1}(x)$ for every $x \in \mathbb{R}$ and $n \geq 1$, one can also check that $\mathbb{E}(g_n(X)) \leq \mathbb{E}(g_{n+1}(X))$ for every $n \geq 1$. This allows us to define

$$\mathbb{E}(g(X)) = \lim_{n \to \infty} \mathbb{E}(g_n(X))$$

Indeed, the above limit always exists (taking possibly the value $+\infty$), as $(\mathbb{E}(g_n(X)), \, n \geq 1)$ is a non-decreasing sequence of non-negative numbers.

**Step 3.** Consider finally the case where $g$ is a *generic Borel-measurable function*, taking possibly positive and negative values. Define in this case the *positive and negative parts* of $g$;

$$g^+(x) = \max(g(x), 0) \quad \text{and} \quad g^-(x) = \max(-g(x), 0)$$

Observe that for every $x \in \mathbb{R}$:

$$g^+(x) - g^-(x) = g(x) \quad \text{and} \quad g^+(x) + g^-(x) = |g(x)|$$

Also, $g^+$ and $g^-$ are themselves non-negative Borel-measurable functions (as well as $|g|$), so according to Step 2, both $\mathbb{E}(g^+(X))$ and $\mathbb{E}(g^-(X))$ are well defined, but take possibly the value $+\infty$. In order to avoid trouble when defining $\mathbb{E}(g(X))$, we will assume that both $\mathbb{E}(g^+(X)) < +\infty$ and $\mathbb{E}(g^-(X)) < +\infty$ (which is equivalent to assuming $\mathbb{E}(|g(X)|) < +\infty$) and define in this case:

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

This completes the definition of $\mathbb{E}(g(X))$ in the general case.

**Remark.** Please observe that $\mathbb{E}(g(X))$ is also well defined when *either* $\mathbb{E}(g^+(X)) = +\infty$ *or* $\mathbb{E}(g^-(X)) = +\infty$, but not both at the same time. In this case, we can write $\mathbb{E}(g(X)) = +\infty$ (resp., $\mathbb{E}(g(X)) = -\infty$). The only case which is to be avoided is the case $\infty - \infty$, which leads to an indetermination.

In the case of discrete and continuous random variables, we obtain the following:

**Proposition 4.1.** If $X$ is a discrete random variable with values in a set $D$ and pmf $\{p_x, \, x \in D\}$, then $\mathbb{E}(|g(X)|) = \sum_{x \in D} |g(x)| \, p_x$ and when this quantity is finite, we have

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

**Proposition 4.2.** If $X$ is a continuous random variable with pdf $p_X$, then $\mathbb{E}(|g(X)|) = \int_{\mathbb{R}} |g(x)| \, p_X(x) \, dx$ and when this quantity is finite, we have

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

**Notation in the general case.** What we have defined above is nothing but *Lebesgue's integral* (in the particular case where the underlying measure is a probability measure $\mathbb{P}$). Because it occurs in many different context, this integral has also *many* different (and of course, equivalent) notations that we list below for completeness:

$$\begin{aligned}\mathbb{E}(g(X)) &= \int_\Omega g(X(\omega)) \, d\mathbb{P}(\omega) = \int_\Omega g(X(\omega)) \, \mathbb{P}(d\omega) \\ &= \int_{\mathbb{R}} g(x) \, d\mu_X(x) = \int_{\mathbb{R}} g(x) \, \mu_X(dx) = \int_{\mathbb{R}} g(x) \, F_X(dx) = \int_{\mathbb{R}} g(x) \, dF_X(x)\end{aligned}$$

In particular, the last notation is the one that we used for convolution before (note that the positioning of $g(x)$ and $dF_X(x)$ after the integral is also irrelevant).

**Terminology.** - If $\mathbb{E}(|X|) < +\infty$, then $X$ is said to be an *integrable* random variable.
- If $\mathbb{E}(X^2) < +\infty$, then $X$ is said to be a *square-integrable* random variable.
- If there exists $0 \leq C < +\infty$ such that $|X(\omega)| \leq C$, $\forall \omega \in \Omega$, then $X$ is said to be a *bounded* random variable[5].
- If $\mathbb{E}(X) = 0$, then $X$ is said to be a *centered* random variable.
- If $X \sim -X$, the $X$ is said to be *symmetrically distributed.*

One has the following series of implications:

$$\begin{gathered} X \text{ is bounded} \quad \Rightarrow \quad X \text{ is square-integrable} \quad \Rightarrow \quad X \text{ is integrable} \\ X \text{ is integrable and } Y \text{ is bounded} \quad \Rightarrow \quad XY \text{ is integrable} \\ X, Y \text{ are both square-integrable} \quad \Rightarrow \quad XY \text{ is integrable} \\ X \text{ is symmetrically distributed (and integrable)} \quad \Rightarrow \quad X \text{ is centered} \end{gathered}$$

*Proof.* The fact that any bounded random variable is integrable follows from the simple fact that $\mathbb{E}(|X|) \leq C$ if $|X(\omega)| \leq C$ for all $\omega \in \Omega$. Any bounded random random variable $X$ is therefore also square-integrable (as $X^2$ is bounded if $X$ is bounded). Likewise, if $X$ is integrable and $|Y(\omega)| \leq C$ for all $\omega \in \Omega$, then $|X(\omega)\, Y(\omega)| \leq C\, |X(\omega)|$, so $XY$ is also integrable.

Besides, any square-integrable random variable is also integrable, because $|X| \leq \frac{X^2+1}{2}$. Likewise, the product $XY$ is integrable if both $X$ and $Y$ are square-integrable, because $|XY| \leq \frac{X^2+Y^2}{2}$.

Finally, if $X \sim -X$, then $\mathbb{E}(X) = -\mathbb{E}(X)$, implying that $\mathbb{E}(X) = 0$ and concluding the proof. □

**Basic properties.** [whose proofs can de deduced more or less straightforwardly from the above definition]

**Linearity.** If $c \in \mathbb{R}$ is a constant and $X$, $Y$ are integrable, then

$$\mathbb{E}(cX) = c\,\mathbb{E}(X) \quad \text{and} \quad \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

**Positivity.** If $X$ is integrable and non-negative, then $\mathbb{E}(X) \geq 0$.

**"Strict" positivity.** If $X$ is integrable, non-negative and $\mathbb{E}(X) = 0$, then $\mathbb{P}(\{X = 0\}) = 1$. One cannot indeed guarantee in this case that $X(\omega) = 0$ for all $\omega \in \Omega$, but just that the event $\{\omega \in \Omega : X(\omega) = 0\}$ is almost sure. Another way to say this is: "$X = 0$ almost surely", often abbreviated as "$X = 0$ a.s."

**Monotonicity.** If $X$, $Y$ are integrable and $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$.

**Jensen's inequality (baby version).** If $X$ is integrable, then $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.

**Variance, covariance and independence.**

**Definition 4.3.** Let $X, Y$ be two *square-integrable* random variables. The *variance* of $X$ is defined as

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

and the *covariance* of $X$ and $Y$ is defined as

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

so that $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$.

**Remark.** Observe that $\operatorname{Var}(X)$ is also well defined when $X$ is integrable but not square-integrable, in which case $\operatorname{Var}(X) = +\infty$. If $X$ is not integrable, then $\operatorname{Var}(X)$ is ill defined ($\infty - \infty$ indetermination).

[5] This condition is not to be mixed with the weaker condition sometimes encountered in the literature: $X$ is *finite* if $\mathbb{P}(\{|X| < +\infty\}) = 1$; this last condition is actually satisfied by *any* real-valued random variable $X$.

**Terminology.** If $\mathrm{Cov}(X,Y)=0$, then $X$ and $Y$ are said to be *uncorrelated.*

**Facts.** [without proofs] Let $c\in\mathbb{R}$ be a constant and $X,Y$ be square-integrable random variables.

a) $\mathrm{Var}(cX)=c^2\,\mathrm{Var}(X)$.

b) $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$.

In addition, if $X,Y$ are independent, then

c) $\mathrm{Cov}(X,Y)=0$, i.e. $\mathbb{E}(XY)=\mathbb{E}(X)\,\mathbb{E}(Y)$ (but the reciprocal statement is wrong).

d) $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$ (also true if $X$ and $Y$ are only uncorrelated).

# 5 Characteristic function

With the definition of expectation in hand, we can define an important object: the characteristic function.

As a preliminary, we need define $\mathbb{E}(g(X))$ for a complex-valued function $g$, but this is simple:
if $\mathbb{E}(|g(X)|)<+\infty$ (where $|\cdot|$ stands for the modulus), then $\mathbb{E}(g(X))=\mathbb{E}(\mathrm{Re}(g(X)))+i\,\mathbb{E}(\mathrm{Im}(g(X)))$.

**Definition 5.1.** The *characteristic function (or Fourier transform)* of a random variable $X$ is the mapping $\phi_X:\mathbb{R}\to\mathbb{C}$ defined as

$$\phi_X(t)=\mathbb{E}\left(e^{itX}\right),\quad t\in\mathbb{R}$$

Note that $|e^{itx}|=1$ for all $x\in\mathbb{R}$, so $\mathbb{E}(|e^{itX}|)=1<+\infty$ and $\phi_X(t)$ is well defined for all $t\in\mathbb{R}$.

For a discrete random variable $X$ with pmf $(p_x,\,x\in D)$, the above formula reads:

$$\phi_X(t).=\sum_{x\in D}e^{itx}\,p_x$$

and for a continuous random variable $X$ with pdf $P_X$:

$$\phi_X(t)=\int_{\mathbb{R}}e^{itx}\,p_X(x)\,dx$$

**Examples.** - Let $X\sim\mathrm{Bern}(p)$. Then

$$\phi_X(t)=p\,e^{it}+1-p$$

- Let $X$ be such that $\mathbb{P}(X=+1)=\mathbb{P}(\{X=-1\})=\frac{1}{2}$. Then

$$\phi_X(t)=\frac{e^{it}+e{-}it}{2}=\cos(t)$$

- Let $X\sim\mathcal{U}([a,b])$. Then

$$\phi_X(t)=\frac{1}{b-a}\int_a^b e^{itx}\,dx=\frac{e^{itb}-e^{ita}}{it(b-a)}\qquad\left(=\frac{\sin(t)}{t}\quad\text{in case } a=-1 \text{ and } b=+1\right)$$

- Let $X\sim\mathcal{N}(\mu,\sigma^2)$. Then noticing that $X\sim\mu+\sigma Z$, where $Z\sim\mathcal{N}(0,1)$, we obtain

$$\begin{aligned}\phi_X(t)&=\mathbb{E}\left(e^{it(\mu+\sigma Z)}\right)=\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}\,e^{it(\mu+\sigma z)-z^2/2}\,dz=e^{it\mu-t^2\sigma^2/2}\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}\,e^{-(z-it\sigma)^2/2}\,dz\\&=e^{it\mu-t^2\sigma^2/2}\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}\,e^{-\frac{w^2}{2}}\,dw=\exp(it\mu-t^2\sigma^2/2)\end{aligned}$$

**Remark.** The above "innocent" change of variable $z - it\sigma \mapsto w$ requires quite a bit of complex analysis to be fully justified!

**Properties.** The characteristic function of a random variable satisfies the following properties:

(i) $\phi_X(0) = 1$.

(ii) $\phi_X$ is *continuous* on $\mathbb{R}$.

(iii) $\phi_X$ is *positive semi-definite* on $\mathbb{R}$, i.e.

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

(iv) $\phi_X(-t) = \overline{\phi_X(t)}$, for all $t \in \mathbb{R}$.

(v) If $X \sim -X$, then $\phi_X(t) \in \mathbb{R}$, for all $t \in \mathbb{R}$.

(vi) $|\phi_X(t)| \leq \phi_X(0) = 1$, for all $t \in \mathbb{R}$.

(vii) $\phi_{cX}(t) = \phi_X(ct)$ for all $c, t \in \mathbb{R}$.

*Proof of (iii)*[6]. For every $n \geq 1$, $c_1, \ldots, c_n \in \mathbb{C}$, $t_1, \ldots, t_n \in \mathbb{R}$, we have

$$\sum_{j,k=1}^{n} c_j \, \overline{c_k} \, \phi_X(t_j - t_k) = \sum_{j,k=1}^{n} c_j \, \overline{c_k} \, \mathbb{E}\left(e^{i(t_j - t_k)X}\right)$$

$$= \mathbb{E}\left(\sum_{j=1}^{n} c_j \, e^{it_jX} \, \overline{\sum_{k=1}^{n} c_k \, e^{it_kX}}\right) = \mathbb{E}\left(\left|\sum_{j=1}^{n} c_j \, e^{it_jX}\right|^2\right) \geq 0$$

The following theorem, due to Bochner and given here without proof, says that the above properties fully characterize what characteristic functions are.

**Theorem.** Any function $\phi : \mathbb{R} \to \mathbb{C}$ satisfying properties (i), (ii), (iii) above is the characteristic function of a random variable $X$.

On top of that, the knowledge of a characteristic function $\phi_X$ fully characterizes the distribution of the random variable $X$, as the following inversion formula shows (given here without proof):

**Inversion formula.** To a given characteristic function $\phi_X$ corresponds a unique cdf $F_X$ given by

$$F_X(b) - F_X(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \, \phi_X(t) \, dt \quad \text{for all } a < b \in \mathbb{R} \text{ continuity points of } F_X$$

If in addition $\int_{\mathbb{R}} |\phi_X(t)| \, dt < +\infty$, then the cdf $F_X$ admits a pdf $p_X$ given by

$$F_X'(x) = p_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \, \phi_X(t) \, dt \quad \forall x \in \mathbb{R}$$

This last property shows that there is a relation between the integrability of the function $\phi_X$ and the regularity of the function $F_X$. The reciprocal statement does not hold, but a weaker statement holds, known as *the Riemann-Lebesgue theorem*: if $F_X$ admits a pdf $p_X$, then $\lim_{t \to \pm\infty} \phi_X(t) = 0$.

The properties below show that there are also interesting relations between the integrability of $F_X$ and the regularity of $\phi_X$:

[6] The other properties are quite straightforward, except for the continuity property, which requires the dominated convergence theorem (not seen in this course).

If $\mathbb{E}(|X|) = \int_{\mathbb{R}} |x| \, dF_X(x) < +\infty$, then $\phi_X$ is continuously differentiable on $\mathbb{R}$ and

$$\phi_X'(0) = i\,\mathbb{E}(X)$$

This last relation can be obtained via the following informal computation:

$$\frac{d}{dt}\phi_X(t)\bigg|_{t=0} = \mathbb{E}\left(\frac{d}{dt}e^{itX}\bigg|_{t=0}\right) = \mathbb{E}\left(\left(iX\,e^{itX}\right)\bigg|_{t=0}\right) = \mathbb{E}(iX)$$

More generally, it holds for $k \geq 1$ that if $\mathbb{E}(|X|^k) < +\infty$, then $\phi_X$ is $k$-times differentiable on $\mathbb{R}$ and

$$\frac{d^k\phi}{dt^k}(t)\big|_{t=0} = i^k\mathbb{E}(X^k)$$

Many other interesting relations can be found between $F_X$ and $\phi_X$, which we shall not list here.

**Factorization property.** Finally, here is a very useful property of characteristic functions:

$$X_1, X_2 \text{ are independent if and only if } \mathbb{E}(e^{it_1X_1+it_2X_2}) = \phi_{X_1}(t_1)\,\phi_{X_2}(t_2), \text{ for all } t_1, t_2 \in \mathbb{R}$$

*Proof.* We only prove the easy part here, namely that independence implies the property on the right-hand side: independence of $X_1$ and $X_2$ implies independence of $e^{it_1X_1}$ and $e^{it_2X_2}$, so

$$\mathbb{E}\left(e^{it_1X_1+it_2X_2)}\right) = \mathbb{E}\left(e^{it_1X_1}\,e^{it_1X_2}\right) = \mathbb{E}\left(e^{it_1X_1}\right)\,\mathbb{E}\left(e^{it_2X_2}\right)$$

which completes the proof. □

**Remark.** Considering the particular case $t_1 = t_2 = t$, we obtain that independence of $X_1$ and $X_2$ implies that $\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\,\phi_{X_2}(t)$ for all $t \in \mathbb{R}$ (but please note that this condition alone does *not* guarantee independence of $X_1$ and $X_2$). As the distribution of the sum of two independent random variables $X_1, X_2$ is given by the *convolution* $F_{X_1} * F_{X_2}$, we recover a classical result, expressed here in the context of probability distributions: the Fourier transform of the convolution of two distributions is the product of their Fourier transforms.

# 6 Random vectors and Gaussian random vectors

## 6.1 Random vectors

A $d$-dimensional random vector $X = (X_1, \ldots, X_d)$ is nothing but a random variable with values in $\mathbb{R}^d$ instead of $\mathbb{R}$. Here is the formal definition:

**Definition 6.1.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $d > 1$. An $\mathcal{F}$-measurable random vector with values in $\mathbb{R}^d$ is a mapping $X : \Omega \mapsto \mathbb{R}^d$ such that

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

Similarly to the scalar case, this condition can be shown to be equivalent to

$$\{\omega \in \Omega \,:\, X_1(\omega) \leq t_1, \ldots, X_d(\omega) \leq t_d\} \in \mathcal{F}, \quad \forall t_1, \ldots, t_d \in \mathbb{R}$$

which can in turn be rephrased as

$$X_j \text{ is an } \mathcal{F}\text{-measurable random variable}, \quad \forall j \in \{1, \ldots, d\}$$

**Joint distribution, multidimensional cdf and marginals.** The *joint distribution* of $X = (X_1, \ldots, X_d)$ is the probability measure $\mu_X$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ defined as

$$\mu_X(B) = \mathbb{P}(\{X \in B\}), \quad \text{for } B \in \mathcal{B}(\mathbb{R}^d)$$

As in the case of random variables, the knowledge of $\mu_X$ can be shown to be equivalent to the knowledge of its corresponding *multidimensional cdf*:

$$F_X(t_1, \ldots, t_d) = \mathbb{P}(\{X_1 \le t_1, \ldots, X_d \le t_d\}), \quad \text{for } t_1, \ldots, t_d \in \mathbb{R}$$

One also defines the *marginals* of the distribution $\mu_X$ as

$$\mu_{X_j}(B) = \mathbb{P}(\{X_j \in B\}) = \mathbb{P}(\{X_j \in B; X_k \in \mathbb{R}, \forall k \neq j\})$$

Watch out that the knowledge of all the marginals $\mu_{X_1}, \ldots, \mu_{X_d}$ is *not* equivalent to the knowledge of $\mu_X$. Put differently, the knowledge of all the marginal cdfs $F_{X_1}, \ldots, F_{X_d}$ is not equivalent to the knowledge of the multidimensional cdf $F_X$.

**Discrete and continuous random vectors.** - A *discrete random vector* is a random vector $X$ that takes values in a discrete set $D \subset \mathbb{R}^d$. This is equivalent to saying that each component $X_j$ is a discrete random variable.

- A *continuous random vector* is a random vector $X$ such that $\mathbb{P}(\{X \in B\}) = 0$ for every $B \in \mathcal{B}(\mathbb{R}^d)$ with Lebesgue measure zero, i.e., $|B| = 0$. In this case (by the same theorem as in the scalar case), there exists a (Borel-measurable) *joint pdf* $p_X = p_{X_1,\ldots,X_d} : \mathbb{R}^d \to \mathbb{R}_+$ such that

$$\mathbb{P}(\{X \in B\}) = \int_B p_X(x_1, \ldots, x_d)\, dx_1 \cdots dx_d, \quad \forall B \in \mathcal{B}(\mathbb{R}^d)$$

(with the integral over $B = \mathbb{R}^d$ being equal to 1). In the case where $p_X$ is continuous, it holds that

$$p_X(x_1, \ldots, x_d) = \frac{\partial^d F_X}{\partial x_1 \cdots \partial x_d}(x_1, \ldots, x_d), \quad \forall (x_1, \ldots, x_d) \in \mathbb{R}^d$$

Also, the following is true: if $X_1, \ldots, X_d$ are independent and continuous random variables, then $(X_1, \ldots, X_d)$ is a continuous random vector. But without the independence assumption, the statement is wrong, as we shall see.

In the present section, we will encounter a certain number of counter-intuitive facts. Here is the first one: it is not because $X_1, X_2$ are both continuous random variables that $X = (X_1, X_2)$ is necessarily a continuous random vector. Here is a counter-example, which will serve other purposes.

**Example 6.2.** Let $X_1 \sim \mathcal{N}(0,1)$, $Z$ be independent of $X_1$ and such that $\mathbb{P}(\{Z = +1\}) = \mathbb{P}(\{Z = -1\}) = \frac{1}{2}$, and finally $X_2 = Z\, X_1$. We see that the distribution of $X_2$ is the same as that of $X_1$:

$$\begin{aligned}
\mathbb{P}(\{X_2 \le t\}) &= \mathbb{P}(\{ZX_1 \le t\} \,|\, \{Z = +1\})\, \mathbb{P}(\{Z = +1\}) + \mathbb{P}(\{ZX_1 \le t\} \,|\, \{Z = -1\})\, \mathbb{P}(\{Z = -1\}) \\
&= \frac{1}{2}\, \mathbb{P}(\{X_1 \le t\} \,|\, \{Z = +1\}) + \frac{1}{2}\, \mathbb{P}(\{X_1 \ge -t\} \,|\, \{Z = +1\}) \\
&= \frac{1}{2}\, \mathbb{P}(\{X_1 \le t\}) + \frac{1}{2}\, \mathbb{P}(\{X_1 \ge -t\}) = \mathbb{P}(\{X_1 \le t\})
\end{aligned}$$

by the symmetry of the distribution of $X_1$. So $X_1, X_2$ are both continuous random variables. But $X = (X_1, X_2)$ is not a continuous random vector. Indeed, let $\Delta = \{(x_1, x_2) \in \mathbb{R}^2 \,:\, x_1 + x_2 = 0\}$. This diagonal line in $\mathbb{R}^2$ has Lebesgue measure 0 (i.e., $|\Delta| = 0$), but

$$\mathbb{P}(\{X \in \Delta\}) = \mathbb{P}(\{X_1 + X_2 = 0\}) = \mathbb{P}(\{X_1 + ZX_1 = 0\}) = \mathbb{P}(\{Z = -1\}) = \frac{1}{2} > 0$$

so $X$ is not a continuous random vector, according to the definition given above.

**Expectation and covariance.** - If $\mathbb{E}(|X_j|) < +\infty$ for every $j \in \{1, \ldots, d\}$, then we define the *expectation* of the random vector $X$ as

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d))$$

which is a $d$-dimensional vector. Similarly, if $\mathbb{E}(X_j^2) < +\infty$ for every $j \in \{1, \ldots, d\}$, then we define the *covariance* of the random vector $X$ as the $d \times d$ matrix:

$$\text{Cov}(X) = \Big\{ \text{Cov}(X_j, X_k) \Big\}_{j,k=1}^{d}$$

where we recall that $\text{Cov}(X_j, X_j) = \text{Var}(X_j)$.

**Proposition 6.3.** The $d \times d$ matrix $\text{Cov}(X)$ is symmetric and positive semi-definite, i.e.,

$$\text{Cov}(X_k, X_j) = \text{Cov}(X_j, X_k), \quad \forall j \neq k, \quad \text{and} \quad \sum_{j,k=1}^{d} c_j\, c_k\, \text{Cov}(X_j, X_k) \geq 0$$

for all $c_1, \ldots, c_d \in \mathbb{R}$.

*Proof.* The symmetry property follows from the very definition of $\text{Cov}(X_j, X_k)$. The second property follows from the bilinearity property of $\text{Cov}(X_j, X_k)$, i.e.,

$$\sum_{j,k=1}^{d} c_j\, c_k\, \text{Cov}(X_j, X_k) = \text{Cov}\left(\sum_{j=1}^{d} c_j X_j, \sum_{k=1}^{d} c_k X_k\right) = \text{Var}\left(\sum_{j=1}^{d} c_j X_j\right) \geq 0.$$

□

**Consequence (spectral decomposition).** Because the matrix $\text{Cov}(X)$ is symmetric and positive semi-definite, the spectral theorem states that it admits $d$ non-negative eigenvalues $\lambda_1, \ldots, \lambda_d$, as well as $d$ eigenvectors $v^{(1)}, \ldots, v^{(d)}$ forming an orthonormal basis of $\mathbb{R}^d$, i.e., $\text{Cov}(X)\, v^{(k)} = \lambda_k\, v^{(k)}$ for every $1 \leq k \leq d$. This can be rewritten as

$$\text{Cov}(X) = V \Lambda V^T \quad \text{(or componentwise: } \text{Cov}(X_j, X_\ell) = \sum_{k=1}^{d} \lambda_k\, v_j^{(k)}\, v_\ell^{(k)})$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ (i.e., the diagonal matrix whose diagonal entries are given by the $\lambda$'s) and $V$ is the matrix whose columns are the vectors $v^{(1)}, \ldots, v^{(d)}$. Because $(v^{(1)}, \ldots, v^{(d)})$ is an orthonormal basis of $\mathbb{R}^d$, the matrix $V$ is orthogonal, i.e. $VV^T = V^T V = I$, the identity matrix.

**Proposition 6.4.** Let $X = (X_1, \ldots, X_d)$ be a square-integrable random vector (i.e., all its components $X_1, \ldots, X_d$ are square-integrable). If $X_1, \ldots, X_d$ are independent random variables, then $\text{Cov}(X)$ is a diagonal matrix (i.e., a matrix whose off-diagonal entries $\text{Cov}(X_j, X_k) = 0$ for all $j \neq k$).

The above proposition follows from the fact mentioned earlier that if $X \perp\!\!\!\perp Y$, then $\text{Cov}(X, Y) = 0$. The reciprocal statement does not hold in general.

**Characteristic function.** Let us finally mention that one defines the characteristic function of a random vector $X$ in the same way as that of a random variable:

$$\phi_X(t_1, \ldots, t_d) = \mathbb{E}(\exp(i(t_1 X_1 + \ldots + t_d X_d))) \quad \text{for } (t_1, \ldots, t_d) \in \mathbb{R}^d$$

which is always a well defined expectation, as $t_1 X_1 + \ldots + t_d X_d \in \mathbb{R}$.

## 6.2 Gaussian random vectors

**Reminder.** A Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, admits as pdf

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

By extension, we say that $X$ is Gaussian, and write $X \sim \mathcal{N}(\mu, 0)$, if $X(\omega) = \mu$ for every $\omega \in \Omega$.

**Definition 6.5.** A random vector $X = (X_1, \ldots, X_d)$ is a $d$-dimensional *Gaussian random vector* if

$$c_1 X_1 + \ldots + c_d X_d \quad \text{is a Gaussian random variable} \quad \forall c_1, \ldots, c_d \in \mathbb{R}$$

**Notation.** If $\mathbb{E}(X) = \mu$ and $\mathrm{Cov}(X) = A$, we write $X \sim \mathcal{N}(\mu, A)$.

**Proposition 6.6.** If $X_1, \ldots, X_d$ are independent Gaussian random variables, then $X = (X_1, \ldots, X_d)$ is a Gaussian random vector.

**Remark.** From Proposition 6.4 above (and the fact that Gaussian random variables are square-integrable), we also know that in this case, the $d \times d$ matrix $\mathrm{Cov}(X)$ is diagonal.

*Proof in the case $d = 2$.* Assume $X_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$ and $X_1 \perp\!\!\!\perp X_2$. Then

$$\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\,\phi_{X_2}(t) = e^{i\mu_1 t - \sigma_1^2 t^2/2}\, e^{i\mu_1 t - \sigma_1^2 t^2/2} = e^{i(\mu_1+\mu_2)t - (\sigma_1^2+\sigma_2^2)t^2/2}$$

so the characteristic function of $X_1 + X_2$ is that of a Gaussian random variable, with expectation $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. A similar computation shows that $c_1 X_1 + c_2 X_2$ is also Gaussian for every $c_1, c_2 \in \mathbb{R}$, which proves the claim. □

Of course, random vectors composed of independent Gaussian random variables are not the only examples of Gaussian random vectors, as we shall see below. Nevertheless, it does *not always* hold that random vectors composed of arbitrary Gaussian random variables are Gaussian random vectors. Here is a counter-example, which is actually the same as Example 6.2 from the previous section.

**Example 6.7.** Let again $X_1 \sim \mathcal{N}(0,1)$, $Z$ be independent of $X_1$ and such that $\mathbb{P}(\{Z = +1\}) = \mathbb{P}(\{Z = -1\}) = \frac{1}{2}$, and finally $X_2 = Z\,X_1$. We saw in the previous section that $X_2 \sim \mathcal{N}(0,1)$, so that both $X_1, X_2$ are Gaussian random variables. But we also saw that

$$\mathbb{P}(\{X_1 + X_2 = 0\}) = \frac{1}{2} > 0$$

which prevents $X_1 + X_2$ from being a continuous random variable, and by extension, a Gaussian random variable. According to the definition given above, $X = (X_1, X_2)$ is therefore not a Gaussian random vector, even though both $X_1, X_2$ are Gaussian random variables.

Now comes a well known property which says that the reciprocal statement of Proposition 6.4 holds for Gaussian random vectors. This property could be wrongly summarized as "uncorrelated Gaussian random variables are independent", because we will see below that in order for this statement to hold, the random variables need not only be Gaussian, but also be part of a Gaussian random vector.

**Proposition 6.8.** If $X = (X_1, \ldots, X_d)$ is a Gaussian random vector and its covariance matrix is diagonal, then $X_1, \ldots, X_d$ are independent random variables.

*Proof in the case $d = 2$.* We will use here the proposition from last lecture:

$$\text{If } \mathbb{E}(e^{it_1X_1+it_2X_2}) = \mathbb{E}(e^{it_1X_1})\,\mathbb{E}(e^{it_2X_2}) \text{ for every } t_1, t_2 \in \mathbb{R}, \quad \text{then } X_1 \perp\!\!\!\perp X_2.$$

To this end, let us fix $t_1, t_2 \in \mathbb{R}$, assume $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and define $Y = t_1X_1 + t_2X_2$. Because of the assumptions made, we have

$$\mathbb{E}(Y) = t_1\mu_1 + t_2\mu_2 \quad \text{and} \quad \mathrm{Var}(Y) = t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 0$$

Because $X = (X_1, X_2)$ is *assumed* to be a Gaussian random vector, we obtain that $Y$ is also Gaussian, more precisely, that $Y \sim \mathcal{N}(t_1\mu_1 + t_2\mu_2, t_1^2\sigma_1^2 + t_2^2\sigma_2^2)$. This implies that

$$\begin{aligned}\mathbb{E}(e^{it_1X_1+it_2X_2}) &= \mathbb{E}(e^{iY}) = e^{i(t_1\mu_1+t_2\mu_2)-(t_1^2\sigma_1^2+t_2^2\sigma_2^2)/2} \\ &= e^{it_1\mu_1-t_1^2\sigma_1^2/2}\,e^{it_2\mu_2-t_2^2\sigma_2^2/2} = \mathbb{E}(e^{it_1X_1})\,\mathbb{E}(e^{it_2X_2})\end{aligned}$$

which proves the independence of $X_1$ and $X_2$. □

More generally, if we only assume that $X_1$, $X_2$ are Gaussian random variables and $\mathrm{Cov}(X_1, X_2) = 0$, this does *not* necessarily imply that $X_1$ and $X_2$ are independent. The above example 6.7 provides again a proof of this claim: $X_1 \sim \mathcal{N}(0,1)$ and $X_2 = ZX_1 \sim \mathcal{N}(0,1)$ are clearly not independent (rather than stating a formal proof of this fact, let us observe simply that $|X_1| = |X_2|$), but

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1X_2) = \mathbb{E}(ZX_1^2) = \mathbb{E}(Z)\,\mathbb{E}(X_1^2) = 0 \cdot 1 = 0$$

where we have used the independence of $Z$ and $X_1$ in the third equality.

## 6.3 Joint distribution of Gaussian random vectors

Is it always the case that a Gaussian random vector $X$ admits a joint pdf $p_X$? The answer is no, as we allowed for constant random variables to be called "Gaussians", and these clearly do not admit a pdf, so if one of the components of our vector $X$ is such a constant random variable, then $X$ cannot admit a joint pdf. A less trivial example is $X = (X_1, -X_1)$, where $X_1 \sim \mathcal{N}(0,1)$: $X$ *is* a Gaussian random vector, because clearly, $c_1X_1 + c_2(-X_1) = (c_1 - c_2)X_1$ is a Gaussian random variable for all possible values of $c_1, c_2$, but the random vector $X$ only takes values in the diagonal $\Delta = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 = 0\}$, which has Lebesgue measure zero.

In order for a Gaussian random vector $X$ to admit a joint pdf $p_X$, its covariance matrix $\mathrm{Cov}(X)$ needs to be *positive definite*, i.e.,

$$\sum_{j,k=1}^{d} c_j, c_k \,\mathrm{Cov}(X_j, X_k) > 0 \quad \text{as soon as} \quad c_1, \ldots, c_d \in \mathbb{R} \quad \text{and} \quad (c_1, \ldots, c_d) \not\equiv 0$$

This turns out to be equivalent to asking (for positive semi-definite matrices) that $\det(\mathrm{Cov}(X)) > 0$, which is again equivalent to asking that the matrix $\mathrm{Cov}(X)$ is invertible. One can check that in the above example, the covariance matrix of $X = (X_1, -X_1)$ is given by

$$\mathrm{Cov}(X) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

whose determinant is given by $\det(\mathrm{Cov}(X)) = 1 \cdot 1 - (-1) \cdot (-1) = 1 - 1 = 0$, implying that $X$ does not admit a joint pdf.

Our aim now is to compute the joint pdf of a Gaussian random vector when it exists. In order to simplify the discussion, we will only deal with *centered* Gaussian random vectors, i.e., vectors for which $\mathbb{E}(X_j) = 0$ for all $j \in \{1, \ldots, d\}$ (the generalization to the non-centered case being relatively straightforward). We start with the simple case where the covariance matrix of $X$ is diagonal.

**Proposition 6.9.** Let $X$ be a centered Gaussian random vector with positive definite and diagonal covariance matrix $\text{Cov}(X) = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$. Then $X$ is a continuous random vector with joint pdf

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Lambda)}} \exp\left(-\frac{x^T \Lambda^{-1} x}{2}\right), \quad x \in \mathbb{R}^d$$

*Proof.* Because $\text{Cov}(X) = \Lambda$ is positive definite, all $\lambda_j > 0$ (so $\det(\Lambda) > 0$ and $\Lambda^{-1}$ exists), and because it is diagonal, Proposition 6.8 implies that $X_1, \ldots, X_d$ are independent, with each $X_j \sim \mathcal{N}(0, \lambda_j)$, so

$$\begin{aligned} p_X(x) = p_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = p_{X_1}(x_1) \cdots p_{X_d}(x_d) &= \prod_{j=1}^d \frac{1}{\sqrt{2\pi\lambda_j}} \exp\left(-\frac{x_j^2}{2\lambda_j}\right) \\ = \frac{1}{\sqrt{(2\pi)^d \prod_{j=1}^d \lambda_j}} \exp\left(-\sum_{j=1}^d \frac{x_j^2}{2\lambda_j}\right) &= \frac{1}{\sqrt{(2\pi)^d \det(\Lambda)}} \exp\left(-\frac{x^T \Lambda^{-1} x}{2}\right), \quad \text{for } x \in \mathbb{R}^d \end{aligned}$$

which completes the proof. □

In the general (i.e., non-diagonal) case, the expression for the joint pdf is actually the same!

**Proposition 6.10.** Let $X$ be a centered Gaussian random vector with positive definite covariance matrix $\text{Cov}(X) = A$. Then $X$ is a continuous random vector with joint pdf

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d \det(A)}} \exp\left(-\frac{x^T A^{-1} x}{2}\right), \quad x \in \mathbb{R}^d$$

*Proof.* Because $\text{Cov}(X) = A$ is a symmetric and positive semi-definite matrix, we saw at the end of Section 6.1 that $A$ can be decomposed into $A = V \Lambda V^T$, where $V$ is an orthogonal matrix (i.e., $VV^T = V^T V = I$) and $\lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix with non-negative entries. Because $A$ is assumed here to be also positive definite, all $\lambda_j > 0$. Also, $A^{-1}$ exists and can be decomposed into

$$A^{-1} = (V \Lambda V^T)^{-1} = (V^T)^{-1} \Lambda^{-1} V^{-1} = V \Lambda^{-1} V^T$$

as $V^{-1} = V^T$ (orthogonal matrix). Likewise, $\det(A) = \det(V \Lambda V^T) = \det(\Lambda V^T V) = \det(\Lambda) > 0$.

In order to deduce the above expression for the joint pdf of $X$, let us first consider the random vector defined as $Y = V^T X$, or component-wise, $Y_j = \sum_{\ell=1}^d V_{\ell j} X_\ell$, $j \in \{1, \ldots, d\}$. Because it is a linear transformation of a centered Gaussian random vector, $Y$ is also a centered Gaussian random vector and

$$\begin{aligned} \text{Cov}(Y_j, Y_k) &= \sum_{\ell,m=1}^d V_{\ell j} V_{mk} \underbrace{\text{Cov}(X_\ell, X_m)}_{=A_{\ell m}} = (V^T A V)_{jk} \\ &= (V^T V \Lambda V^T V)_{jk} = \Lambda_{jk} = \lambda_j \, \delta_{jk} \end{aligned}$$

so $Y$ is a centered Gaussian random vector with a diagonal and positive definite covariance matrix $\Lambda$. By Proposition 6.9 above, $Y$ is a continuous random vector with joint pdf

$$p_Y(y) = \frac{1}{\sqrt{(2\pi)^d \det(\Lambda)}} \exp\left(-\frac{y^T \Lambda^{-1} y}{2}\right), \quad y \in \mathbb{R}^d$$

Finally, because $Y = V^T X$ and $V$ is orthogonal, the Jacobian of this linear transformation is equal to 1, so

$$p_X(x) = p_Y(V^T x) = \frac{1}{\sqrt{(2\pi)^d \det(\Lambda)}} \exp\left(-\frac{x^T V \Lambda^{-1} V^T x}{2}\right) = \frac{1}{\sqrt{(2\pi)^d \det(A)}} \exp\left(-\frac{x^T A^{-1} x}{2}\right)$$

following the observations made at the beginning of this proof, which is now complete. □

As a by-product, this proof also shows that every centered Gaussian random vector $X$ can be written as $X = VY$, where $Y$ is a vector of independent Gaussian random variables (and note that this holds also in the general case where the covariance matrix $A$ is not invertible).

For the sake of completeness, we give below the general formula when the expectation $\mathbb{E}(X) = \mu$ is non-zero (and the covariance matrix $\text{Cov}(X) = A$ is positive definite):

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d \det(A)}} \exp\left(-\frac{(x-\mu)^T A^{-1}(x-\mu)}{2}\right), \quad x \in \mathbb{R}^d$$

# 7 Inequalities

We review below three important inequalities in probability.

**Cauchy-Schwarz's inequality.** If $X$, $Y$ are square-integrable random variables, then the product $XY$ is integrable and

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

**Remark.** The first inequality follows from the simple fact mentioned above that $|\mathbb{E}(Z)| \leq \mathbb{E}(|Z|)$ for any random variable $Z$.

Note that this inequality is a refinement of the inequality which is obtained using the pointwise inequality $|X(\omega)\,Y(\omega)| \leq \frac{X^2(\omega)+Y^2(\omega)}{2}$:

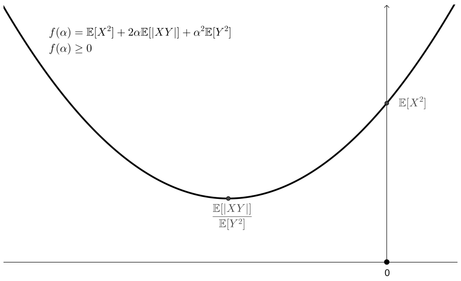
$$\mathbb{E}(|XY|) \leq \frac{\mathbb{E}(X^2) + \mathbb{E}(Y^2)}{2}$$

as we know by the arithmetic-geometric mean inequality that $\sqrt{\mathbb{E}(X^2)\,\mathbb{E}(Y^2)} \leq \frac{\mathbb{E}(X^2)+\mathbb{E}(Y^2)}{2}$.

*Proof.* Consider now the map $\alpha \mapsto f(\alpha) = \mathbb{E}((|X| + \alpha\,|Y|)^2)$. Note first that $f(\alpha) \leq 2\,\mathbb{E}(X^2 + \alpha^2\,Y^2) < +\infty$ by assumption, so $f$ is a well-defined map. This map clearly takes non-negative values for all values of $\alpha \in \mathbb{R}$. Moreover, using the basic properties of the expectation, we see that

$$f(\alpha) = \mathbb{E}(X^2) + 2\alpha\,\mathbb{E}(|XY|) + \alpha^2\,\mathbb{E}(Y^2)$$

The function $f$ is represented below:



As $f$ is a second-order polynomial with at most one root, we deduce that the discriminant of this polynomial is non-positive, i.e., that

$$\Delta = (\mathbb{E}(|XY|)^2 - \mathbb{E}(X^2)\,\mathbb{E}(Y^2) \leq 0$$

which implies the desired inequality:

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

□

**Remark.** An alternate proof of the inequality $|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\,\sqrt{\mathbb{E}(Y^2)}$ follows from the observation that the bilinear form $X, Y \mapsto \mathbb{E}(XY)$ is a (semi-)inner product on the space of square-integrable random variables (which further implies that $X \mapsto \sqrt{\mathbb{E}(X^2)}$ is a (semi-)norm on the same space). Indeed:

1) The fact that it is bilinear in $X, Y$ comes from the linearity of the expectation.

2) It is symmetric in $X, Y$ by definition (and by commutativity of the multiplication).

3) It is also positive, as $\mathbb{E}(X^2) \geq 0$ for every $X$.

The classical Cauchy-Schwarz inequality, valid for any (semi-)inner product, then implies that $|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\,\sqrt{\mathbb{E}(Y^2)}$. In order to obtain the inequality with absolute values inside the expectation, just apply the above inequality to $|X|$ and $|Y|$ instead of $X$ and $Y$.

**Jensen's inequality.** If $X$ is an integrable random variable and $\varphi : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, convex and such that $\mathbb{E}(|\varphi(X)|) < +\infty$, then

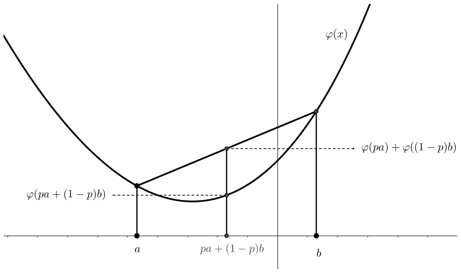
$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$$

In particular: $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$, $\mathbb{E}(X)^2 \leq \mathbb{E}(X^2)$, $\exp(\mathbb{E}(X)) \leq \mathbb{E}(\exp(X))$, $\exp(-\mathbb{E}(X)) \leq \mathbb{E}(\exp(-X))$ and for $X$ a positive random variable, $\log(\mathbb{E}(X)) \geq \mathbb{E}(\log(X))$ (as log is *concave*, $-\log$ is convex).

Also, if $X$ is such that $\mathbb{P}(\{X = a\}) = p \in ]0, 1[$ and $\mathbb{P}(\{X = b\}) = 1 - p$, then the above inequality says that

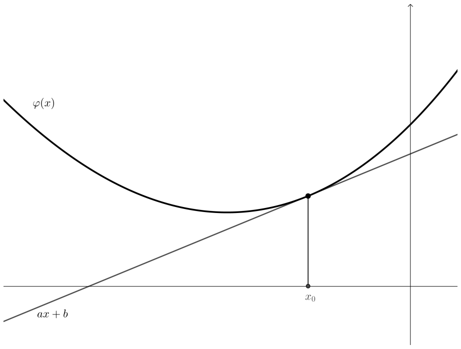
$$\varphi\,(pa + (1-p)b) \leq p\,\varphi(a) + (1-p)\,\varphi(b)$$

which matches actually the definition of convexity for $\varphi$:



*Proof.* The proof relies on the fact that for any convex function $\varphi : \mathbb{R} \to \mathbb{R}$ and any point $x_0 \in \mathbb{R}$, there exists an affine function $x \mapsto ax + b$ which is below $\varphi$ and is also tangent to $\varphi$ in $x_0$, i.e.,

$$\varphi(x) \geq ax + b, \quad \forall x \in \mathbb{R} \quad \text{and} \quad \varphi(x_0) = ax_0 + b$$



Therefore, choose $x_0 = \mathbb{E}(X)$. We obtain, by linearity of expectation:

$$\mathbb{E}(\varphi(X)) \geq \mathbb{E}(aX + b) = a\,\mathbb{E}(X) + b = ax_0 + b = \varphi(x_0) = \varphi(\mathbb{E}(X))$$

which is the desired inequality. □

Another consequence of Jensen's inequality is the following: the arithmetic mean of $n$ positive real numbers $x_1, \ldots, x_n$ is greater than or equal to their geometric mean, which is in turn greater than or equal to their harmonic mean, i.e.,

$$\frac{x_1 + \ldots + x_n}{n} \geq (x_1 \cdots x_n)^{1/n} \geq \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}$$

*Proof.* Consider $X$ taking values $x_1, \ldots, x_n$ with uniform probability. Then proving the above inequalities amounts to proving that

$$\mathbb{E}(X) \geq \exp(\mathbb{E}(\log(X))) \geq \frac{1}{\mathbb{E}(\frac{1}{X})}$$

Observe first that since log is a concave function, $-\log$ is convex, so by Jensen's inequality, we have $\mathbb{E}(\log(X)) \leq \log(\mathbb{E}(X))$. Taking exponentials, this proves the above left-hand side inequality. In order to prove the right-hand side inequality, observe that, again because $-\log$ is convex, we have

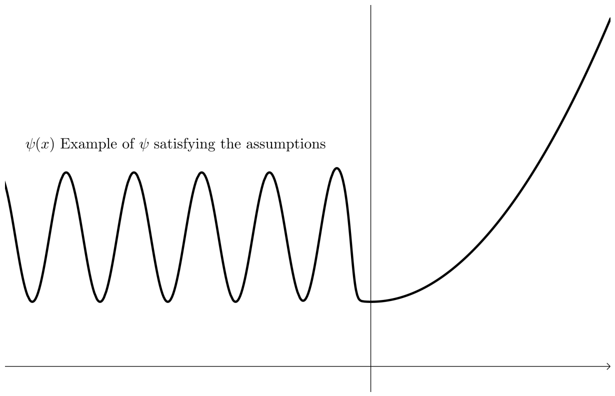
$$\exp(\mathbb{E}(\log(X)) = \exp(\mathbb{E}(-\log(\tfrac{1}{X}))) \geq \exp(-\log \mathbb{E}(\tfrac{1}{X})) = \tfrac{1}{\mathbb{E}(\frac{1}{X})}$$

proving the claim. □

**Chebyshev-Markov's inequality.** Let $X$ be a random variable and $t \in \mathbb{R}_+$. If $\psi : \mathbb{R} \to \mathbb{R}_+$ is a Borel-measurable function which is non-decreasing on $\mathbb{R}_+$ and such that $\psi(t) > 0$ and $\mathbb{E}(\psi(X)) < +\infty$, then

$$\mathbb{P}(\{X \geq t\}) \leq \frac{\mathbb{E}(\psi(X))}{\psi(t)}$$

In particular, if $X$ is square-integrable and $t > 0$, then taking $\psi(x) = x^2$ gives $\mathbb{P}(\{X \geq t\}) \leq \dfrac{\mathbb{E}(X^2)}{t^2}$. But please note that $\psi$ could also be a much more intricate function, such as:



*Proof.* Using successively the assumptions that $\psi$ takes values in $\mathbb{R}_+$, is non-decreasing on $\mathbb{R}_+$ and $t > 0$, we obtain

$$\mathbb{P}(\{X \geq t\}) = \mathbb{E}\left(1_{\{X \geq t\}}\right) \leq \mathbb{E}\left(\frac{\psi(X)}{\psi(t)}\, 1_{\{X \geq t\}}\right) = \frac{\mathbb{E}(\psi(X)\, 1_{\{X \geq t\}})}{\psi(t)} \leq \frac{\mathbb{E}(\psi(X))}{\psi(t)}$$

which is the desired inequality. □

# 8 Laws of large numbers

## 8.1 Preliminary: convergence of sequences of numbers

Let us recall that a sequence of real numbers $(a_n,\, n \ge 1)$ converges to a limit $a \in \mathbb{R}$ (and this is denoted as $a_n \underset{n\to\infty}{\to} a$) if and only if

$$\forall \varepsilon > 0,\ \exists N \ge 1 \text{ such that } \forall n \ge N,\ |a_n - a| < \varepsilon$$

Reciprocally, the sequence $(a_n,\, n \ge 1)$ does *not* converge to $a \in \mathbb{R}$ if and only if

$$\exists \varepsilon > 0 \text{ such that } \forall N \ge 1,\ \exists n \ge N \text{ such that } |a_n - a| \ge \varepsilon$$

This is still equivalent to saying that

$$\exists \varepsilon > 0 \text{ such that } |a_n - a| \ge \varepsilon \text{ for an infinite number of values of } n$$

In the previous sentence, "for an infinite number of values of $n$" may be abbreviated as "infinitely often".

## 8.2 Convergences of sequences of random variables

In order to extend the notion of convergence from sequences of numbers to sequences of random variables, there are quite a few possibilities. We have indeed seen in the previous lectures that random variables are functions. In a first year analysis course, one hears about various notions of convergence for sequences of functions, among which pointwise and uniform convergence. There are actually many others. We will see four of them that are most useful in the context of probability, and three of them in today's lecture.

Let $(X_n,\, n \ge 1)$ be a sequence of random variables and $X$ be another random variable, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

**1) Quadratic convergence.** Assume that all $X_n$, $X$ are square-integrable. The sequence $(X_n,\, n \ge 1)$ is said to *converge in quadratic mean* to $X$ (and this is denoted as $X_n \underset{n\to\infty}{\overset{L^2}{\to}} X$) if

$$\mathbb{E}(|X_n - X|^2) \underset{n\to\infty}{\to} 0$$

**2) Convergence in probability.** The sequence $(X_n,\, n \ge 1)$ is said to *converge in probability* to $X$ (and this is denoted as $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$) if

$$\forall \varepsilon > 0, \quad \mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \ge \varepsilon\}) \underset{n\to\infty}{\to} 0$$

**3) Almost sure convergence.** The sequence $(X_n,\, n \ge 1)$ is said to *converge almost surely* to $X$ (and this is denoted as $X_n \underset{n\to\infty}{\to} X$ a.s.) if

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

## 8.3 Relations between the three notions of convergence

**Quadratic convergence implies convergence in probability.** This is a direct consequence of Chebyshev's inequality. Assume indeed that $X_n \underset{n\to\infty}{\overset{L^2}{\to}} X$. Then we have for any fixed $\varepsilon > 0$:

$$\mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \ge \varepsilon\}) \le \frac{\mathbb{E}(|X_n - X|^2)}{\varepsilon^2} \underset{n\to\infty}{\to} 0 \quad \text{by assumption, so} \quad X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$$

□

**Convergence in probability does not imply quadratic convergence.** This is left for homework. Note that counter-examples exist even in the case where all $X_n$, $X$ are square-integrable. On the other hand, the following proposition holds [without proof].

**Proposition 8.1.** Let $(X_n,\, n \geq 1)$ be a sequence of random variables and $X$ be a random variable such that $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$. Assume moreover there exists $C > 0$ such that $|X_n(\omega)| \leq C$ for all $\omega \in \Omega$ and $n \geq 1$. Then $X_n \underset{n\to\infty}{\overset{L^2}{\to}} X$.

**Almost sure convergence implies convergence in probability.** In order to show this, we will the following lemma which provides an alternate characterization of almost sure convergence:

**Lemma 8.2.** (characterization of almost sure convergence)

$$X_n \underset{n\to\infty}{\to} X \text{ a.s.} \quad \text{if and only if} \quad \forall \varepsilon > 0, \quad \mathbb{P}\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ infinitely often}\}\right) = 0$$

*Proof.* We drop here the full notation with $\omega$'s and use the abbreviation "i.o." for "infinitely often" in order to lighten the writing. Based on what was said on the covergence of sequences of numbers, we obtain the following series of equivalences:

$$X_n \underset{n\to\infty}{\to} X \text{ a.s.} \quad \Longleftrightarrow \quad \mathbb{P}\left(\left\{\lim_{n\to\infty} X_n = X\right\}\right) = 1$$
$$\Longleftrightarrow \quad \mathbb{P}\left(\left\{\lim_{n\to\infty} X_n \neq X\right\}\right) = 0 \quad \Longleftrightarrow \quad \mathbb{P}\left(\{\exists \varepsilon > 0 \text{ such that } |X_n - X| \geq \varepsilon \text{ i.o.}\}\right) = 0$$
$$\Longleftrightarrow \quad \mathbb{P}\left(\left\{\exists M \geq 1 \text{ such that } |X_n - X| \geq \frac{1}{M} \text{ i.o.}\right\}\right) = 0 \quad \Longleftrightarrow \mathbb{P}\left(\bigcup_{M\geq 1} \left\{|X_n - X| \geq \frac{1}{M} \text{ i.o.}\right\}\right) = 0$$

This implies that

$$\forall M \geq 1, \quad \mathbb{P}\left(\left\{|X_n - X| \geq \frac{1}{M} \text{ i.o.}\right\}\right) = 0$$

which is in turn equivalent to saying that

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\{|X_n - X| \geq \varepsilon \text{ i.o.}\}\right) = 0$$

and completes the proof. Note the "hat trick" used here in order to replace an uncountable union over $\varepsilon$'s by a countable union over $M$'s. □

*Proof that almost sure convergence implies convergence in probability.* We have the following series of equivalences. By Lemma 8.2,

$$\begin{aligned} X_n \underset{n\to\infty}{\to} X \text{ a.s.} \quad &\Longleftrightarrow \quad \forall \varepsilon > 0, \quad \mathbb{P}\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ infinitely often}\}\right) = 0 \\ &\Longleftrightarrow \quad \forall \varepsilon > 0, \quad \mathbb{P}\left(\{\forall N \geq 1,\, \exists n \geq N \text{ such that } |X_n - X| \geq \varepsilon\}\right) = 0 \\ &\Longleftrightarrow \quad \forall \varepsilon > 0, \quad \mathbb{P}\Big(\cap_{N\geq 1} \underbrace{\cup_{n\geq N} \{|X_n - X| \geq \varepsilon\}}_{=B_N}\Big) = 0 \end{aligned}$$

Since $B_N \supset B_{N+1}$ for every $N \geq 1$, we obtain that

$$\forall \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\Big(\cup_{n\geq N} \{|X_n - X| \geq \varepsilon\}\Big) = 0$$

and this implies that $\forall \varepsilon > 0$, $\lim_{N\to\infty} \mathbb{P}\left(\{|X_N - X| \geq \varepsilon\}\right) = 0$. Said otherwise: $X_N \underset{N\to\infty}{\overset{\mathbb{P}}{\to}} X$. □

**Convergence in probability does not imply almost sure convergence,** as surprising as this may sound! Here is a counter-example: let us consider a sequence of independent and identically distributed (i.i.d.) heads (H) and tails (T). Out of this sequence, we construct another sequence of random variables:

$$X_1 = 1_H, \quad X_2 = 1_T, \quad X_3 = 1_{HH}, \quad X_4 = 1_{HT}, \quad X_5 = 1_{TH}, \quad X_6 = 1_{TT}, \quad X_7 = 1_{HHH}, \quad X_8 = 1_{HHT} \ldots$$

meaning "$X_1 = 1$ iff the first coin falls on heads", "$X_2 = 1$ iff the first coin falls on tails", "$X_3 = 1$ iff the first two coins fall on heads", etc. Note that this new sequence of random variables is everything but independent: there is indeed a strong dependency e.g. between $X_3, X_4, X_5$ and $X_6$, as only one of these random variables can take the value 1 for a given sequence of heads and tails.

On the one hand, $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} 0$. Indeed, one can check that for any $\varepsilon > 0$, the probability

$$\mathbb{P}\left(\{|X_n - 0| \geq \varepsilon\}\right) = O\left(\frac{1}{n}\right)$$

It therefore converges to 0 as $n \to \infty$.

On the other hand, $X_n \not\to 0$ a.s. Indeed, for a given realization $\omega$ of heads and tails, such as e.g. HHTHTT..., the sequence $(X_n(\omega),\, n \geq 1)$ is equal to 10100001000000... Note that as you explore further the sequence, you encounter less and less 1's; nevertheless, you *always* encounter a 1 after a sufficiently large number of steps. So the sequence $(X_n(\omega),\, n \geq 1)$ is an alternating sequence of 0's and 1's, that therefore does not converge to 0 (according to the first definition of today's lecture), and this is true for any realization $\omega$. In conclusion,

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = 0\right\}\right) = 0$$

which is the complete opposite of the definition of almost sure convergence.

**Remark.** As we know that $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} 0$, the sequence cannot converge to anything else almost surely. Indeed, as almost sure convergence implies convergence in probability, if $X_n$ were to converge a.s. to another limit $X \neq 0$, this would imply that $X_n$ should also converge in probability towards this same limit $X$. But we know already that $X_n$ converges in probability to 0, so $X$ cannot be different from 0 (up to a set of probability 0). This is formalized by the proposition below.

**Proposition 8.3.** Let $(X_n,\, n \geq 1)$ be a sequence of random variables and $X, Y$ be two random variables such that $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$ and $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} Y$. Then $X = Y$ a.s.

**Note.** For two random variables $X, Y$, there is therefore no such thing as "$X = Y$ in probability" or "$X = Y$ in $L^2$".

*Proof.* We have for any fixed $\varepsilon > 0$:

$$\begin{aligned}
&\mathbb{P}(\{|X - Y| \geq \varepsilon\}) = \mathbb{P}(\{|X - X_n + X_n - Y| \geq \varepsilon\}) \leq \mathbb{P}(\{|X - X_n| + |X_n - Y| \geq \varepsilon\}) \\
&\leq \mathbb{P}(\{|X - X_n| \geq \frac{\varepsilon}{2}\} \cup \{|X_n - Y| \geq \frac{\varepsilon}{2}\}) \leq \mathbb{P}(\{|X_n - X| \geq \frac{\varepsilon}{2}\}) + \mathbb{P}(\{|X_n - Y| \geq \frac{\varepsilon}{2}\}) \underset{n\to\infty}{\to} 0
\end{aligned}$$

by the assumptions made. So $\mathbb{P}(\{|X - Y| \geq \varepsilon\}) = 0$ for any $\varepsilon > 0$, therefore $\mathbb{P}(\{X = Y\}) = 1$, which proves the claim. □

## 8.4 The Borel-Cantelli lemma

As one might guess from the previous pages, proving convergence in probability (using e.g. Chebyshev's inequality) is in general much easier than proving almost sure convergence. Still, these two notions are not equivalent, as the previous counter-example shows. So it would be convenient to have a criterion saying that if both convergence in probability and another easy-to-check condition hold, then almost sure convergence holds. This criterion is the (first) Borel-Cantelli lemma.

**Reminder.** Let $(a_n,\, n \geq 1)$ be a sequence of non-negative numbers. Then writing that $\sum_{n\geq 1} a_n < +\infty$ exactly means that

$$\lim_{N\to\infty} \sum_{n\geq N} a_n = 0$$

which is stronger than $\lim_{n\to\infty} a_n = 0$. This last condition alone does indeed *not* guarantee that $\sum_{n\geq 1} a_n < +\infty$. A famous counter-example is the harmonic series $a_n = \frac{1}{n}$.

**Lemma 8.4.** (Borel-Cantelli)
Let $(A_n,\, n \geq 1)$ be a sequence of events in $\mathcal{F}$ such that $\sum_{n\geq 1} \mathbb{P}(A_n) < +\infty$. Then

$$\mathbb{P}\left(\{\omega \in \Omega : \omega \in A_n \text{ infinitely often}\}\right) = 0$$

Before proving this lemma, let us see how it can be applied to the convergence of sequences of random variables. Let $(X_n,\, n \geq 1)$ be a sequence of random variables.

a) If for all $\varepsilon > 0$, $\mathbb{P}(\{|X_n - X| \geq \varepsilon\}) \underset{n\to\infty}{\to} 0$, then $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$, by definition.

b) If for all $\varepsilon > 0$, $\sum_{n\geq 1} \mathbb{P}(\{|X_n - X| \geq \varepsilon\}) < \infty$, then $X_n \underset{n\to\infty}{\to} X$ a.s. Indeed, by the Borel-Cantelli lemma, the condition on the sum implies that for all $\varepsilon > 0$,

$$\mathbb{P}\left(\{|X_n - X| \geq \varepsilon \text{ infinitely often}\}\right) = 0$$

which is exactly the characterization of almost sure convergence given in Lemma 8.2.

So we see that if one can prove that $\mathbb{P}(\{|X_n - X| \geq \varepsilon\}) = O(\frac{1}{n})$ for all $\varepsilon > 0$, this guarantees convergence in probability, but not almost sure convergence, as the condition on the sum is not necessarily satisfied (cf. the example in the previous section).

*Proof of the Borel-Cantelli lemma.* Let us first rewrite

$$\begin{aligned}
\mathbb{P}\left(\{\omega \in \Omega : \omega \in A_n \text{ infinitely often}\}\right) &= \mathbb{P}\left(\{\forall N \geq 1,\ \exists n \geq N \text{ such that } \omega \in A_n\}\right) \\
&= \mathbb{P}\left(\bigcap_{N\geq 1} \bigcup_{n\geq N} A_n\right) = \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{n\geq N} A_n\right)
\end{aligned}$$

where we have used Fact 2 for the last equality. Using finally the union bound, we obtain:

$$\mathbb{P}\left(\{\omega \in \Omega : \omega \in A_n \text{ infinitely often}\}\right) \leq \lim_{N\to\infty} \sum_{n\geq N} \mathbb{P}(A_n) = 0$$

by the assumption made on the sum (and the above reminder). This completes the proof. □

## 8.5 Laws of large numbers

We state below the law of large numbers. This law justifies the notion of theoretical expectation, as it shows that the average of a large number of independent and identically distributed (i.i.d.) random variables converges to this expectation (in probability and almost surely).

**Theorem 8.5.** Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. random variables such that $\mathbb{E}(X_1^2) < +\infty$, and let $S_n = X_1 + \ldots + X_n$. Then:

**a)** $\dfrac{S_n}{n} \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \mathbb{E}(X_1)$ (weak law).

**b)** $\dfrac{S_n}{n} \underset{n\to\infty}{\to} \mathbb{E}(X_1)$ almost surely (strong law).

**Remarks.** - Because the random variables $X$'s are i.i.d., the assumption $\mathbb{E}(X_1^2) < +\infty$ is an assumption on the whole sequence and not only on the random variable $X_1$.

- Both laws above hold under the weaker assumption that $\mathbb{E}(|X_1|) < +\infty$, but in this case, the proof of the theorem becomes significantly longer! Restricting ourselves to the assumption that $\mathbb{E}(X_1^2) < +\infty$ allows in particular to use $\mathrm{Var}(X_1)$ in the proof.

- One may wonder why should one state both a weak and a strong law, as the latter is obviously a stronger result than the former. A first simple reason is that the weak law was found historically by Jacob Bernoulli around year 1700, much before the strong law, obtained by Borel and Cantelli around year 1900. Besides, both the weak and the strong law can be generalized to different sets of assumptions on the random variables $X$'s. But more generalizations are possible for the weak law than for the strong law, as we shall see.

*Proof.* **a)** For all $\varepsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq \varepsilon\right\}\right) &= \mathbb{P}\left(\{|S_n - n\,\mathbb{E}(X_1)| \geq n\varepsilon\}\right) = \mathbb{P}\left(\{|S_n - \mathbb{E}(S_n)| \geq n\varepsilon\}\right) \\ &\leq \frac{\mathbb{E}((S_n - \mathbb{E}(S_n))^2)}{n^2\varepsilon^2} = \frac{\mathrm{Var}(S_n)}{n^2\,\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\,\varepsilon^2} \underset{n\to\infty}{\to} 0\end{aligned}$$

where we have used Chebyshev's inequality and the fact that the variance of a sum of independent variables is the sum of the variances of these random variables. This implies that $\dfrac{S_n}{n} \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \mathbb{E}(X_1)$ and therefore proves the weak law of large numbers.

**b)** Note that the former proof does not allow to conclude here, because we only showed that

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq \varepsilon\right\}\right) = O\left(\frac{1}{n}\right)$$

so we cannot apply Borel-Cantelli's lemma in this case, as mentioned already in the last lecture. There is nevertheless an elegant solution to this problem, as described in the sequel.

- Observe first that we may simply replace $n$ by $n^2$ in the previous equality, so as to obtain:

$$\mathbb{P}\left(\left\{\left|\frac{S_{n^2}}{n^2} - \mathbb{E}(X_1)\right| \geq \varepsilon\right\}\right) = O\left(\frac{1}{n^2}\right)$$

Using the Borel-Cantelli lemma and the fact that $\sum_{n\geq 1} \frac{1}{n^2} < \infty$, we obtain that $\dfrac{S_{n^2}}{n^2} \underset{n\to\infty}{\to} \mathbb{E}(X_1)$ almost surely. This alone of course does not prove the result, but...

- Assume for now that $X_n \geq 0$ for all $n \geq 1$ and consider an integer $m$ such that $n^2 \leq m \leq (n+1)^2$. Because the $X$'s are positive, the sequence $S_n$ is non-decreasing, so we obtain in this case

$$\frac{S_{n^2}}{(n+1)^2} \leq \frac{S_m}{m} \leq \frac{S_{(n+1)^2}}{n^2}$$

Note that by what was just shown above and by the fact that $\frac{(n+1)^2}{n^2} \underset{n\to\infty}{\to} 1$, both the left-most and the right-most terms converge almost surely to $\mathbb{E}(X_1)$ as $n \to \infty$. As $\dfrac{S_m}{m}$ is lower and upper bounded by these two terms, respectively, we deduce that $\dfrac{S_m}{m}$ also converges almost surely to $\mathbb{E}(X_1)$ as $m \to \infty$.

- Finally, we need to address the case where the $X$'s are not necessarily non-negative. Let us define in this case $X_n^+ = \max(X_n, 0)$ and $X_n^- = \max(-X_n, 0)$, so that $X_n = X_n^+ - X_n^-$. Similarly, let

$$S_n^+ = \sum_{j=1}^n X_j^+ \quad \text{and} \quad S_n^- = \sum_{j=1}^n X_j^-, \quad \text{so that} \quad S_n = S_n^+ - S_n^-$$

Note that it is not necessarily the case that $S_n^+ = \max(S_n, 0)$ and $S_n^- = \max(-S_n, 0)$, but this does not matter here. What matters is that both $S_n^+$ and $S_n^-$ are sums of i.i.d. non-negative random variables, so that the previous result applies to both:

$$\frac{S_n^+}{n} \underset{n\to\infty}{\to} \mathbb{E}(X_1^+) \quad \text{a.s.} \quad \text{and} \quad \frac{S_n^-}{n} \underset{n\to\infty}{\to} \mathbb{E}(X_1^-) \quad \text{a.s.}$$

which in turn implies that

$$\frac{S_n}{n} = \frac{S_n^+}{n} - \frac{S_n^-}{n} \underset{n\to\infty}{\to} \mathbb{E}(X_1^+) - \mathbb{E}(X_1^-) = \mathbb{E}(X_1) \quad \text{a.s.}$$

and therefore completes the proof. □

**Remark.** With the weaker assumption that $\mathbb{E}(|X_1|) < +\infty$, one cannot mimic the above proof using Chebyshev's inequality with $\psi(x) = |x|$ instead of $\psi(x) = x^2$. Indeed, in part a), one would get

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq \varepsilon\right\}\right) \leq \frac{\mathbb{E}(|S_n - \mathbb{E}(S_n)|)}{n\,\varepsilon}$$

but then how to upper bound $\mathbb{E}(|S_n - \mathbb{E}(S_n)|$? Using the triangle inequality, one would get

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq \varepsilon\right\}\right) \leq \frac{n\,\mathbb{E}(|X_1 - \mathbb{E}(X_1)|)}{n\,\varepsilon} = \frac{\mathbb{E}(|X_1 - \mathbb{E}(X_1)|)}{\varepsilon}$$

which does not decrease to 0 as $n$ increases...

## 8.6 Application: convergence of the empirical distribution

Let again $(X_n,\, n \geq 1)$ be a sequence of i.i.d. random variables (but without any assumption on their integrability), and let $F$ denote their common cdf ($F(t) = \mathbb{P}(\{X_1 \leq t\})$, $t \in \mathbb{R}$.).

Let now $F_n(t) = \frac{1}{n}\, \sharp\{1 \leq j \leq n : X_j \leq t\}$ for $t \in \mathbb{R}$; $F_n$ is the empirical distribution (or cdf) of the first $n$ random variables $X_1, \ldots, X_n$. Note that for fixed $n$, it is a discrete distribution (i.e., the cdf is a staircase function). As well as the law of large numbers provides a justification for the notion of expectation, the statement below provides a justification for the notion of distribution.

**Theorem 8.6.** For every $t \in \mathbb{R}$,

$$F_n(t) \underset{n\to\infty}{\to} F(t) \quad \text{almost surely}$$

*Proof.* Fix $t \in \mathbb{R}$ and let $Y_j = 1_{\{X_j \leq t\}}$. As the $X$'s are i.i.d., so are the $Y$'s. On top of that, the $Y$'s are square-integrable, as they are Bernoulli random variables, taking values in $\{0, 1\}$ only. Also, $F_n(t)$ may be rewritten as

$$F_n(t) = \frac{1}{n}\sum_{j=1}^n Y_j \underset{n\to\infty}{\to} \mathbb{E}(Y_1) \quad \text{almost surely}$$

by the strong law of large numbers. Noticing finally that $\mathbb{E}(Y_1) = \mathbb{P}(\{X_1 \leq t\}) = F(t)$ completes the proof. □

## 8.7 Extension of the strong law: Kolmogorov's 0-1 law

The strong law *cannot* be extended beyond the $\mathbb{E}(|X_1|) < +\infty$ assumption. One can show actually the following more precise statement:

- If $\mathbb{E}(|X_1|) < +\infty$, then $\lim_{n\to\infty} \dfrac{S_n}{n} = \mathbb{E}(X_1)$ a.s.

- If $\mathbb{E}(|X_1|) = +\infty$, then $\limsup_{n\to\infty} \left|\dfrac{S_n}{n}\right| = +\infty$ a.s., i.e., $\dfrac{S_n}{n}$ diverges a.s.

The aim here is not to give a full proof of the above statements, but rather to explain why $\dfrac{S_n}{n}$ can only converge or diverge a.s.

Let $(X_n,\, n \geq 1)$ be a sequence of random variables, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For $n \geq 1$, define

$$\mathcal{G}_n = \sigma(X_n, X_{n+1}, X_{n+2}, \ldots) \quad \text{and} \quad \mathcal{T} = \bigcap_{n\geq 1} \mathcal{G}_n$$

$\mathcal{T}$ is called the *tail $\sigma$-field* of the sequence $(X_n,\, n \geq 1)$. In words, it is the information related to the asymptotic behaviour of the sequence when $n \to \infty$ (which in general might contain lots of information!). Here is an example of event in $\mathcal{T}$:

$$A_1 = \left\{ \omega \in \Omega \,:\, \sum_{n\geq 1} X_n(\omega) \text{ converges} \right\}$$

Indeed, note that for *every* $N \geq 1$, we have

$$A_1 = \left\{ \omega \in \Omega \,:\, \sum_{n\geq N} X_n(\omega) \text{ converges} \right\}$$

so $A_1 \in \mathcal{G}_N$ for every $N \geq 1$. It therefore also belongs to $\mathcal{T} = \cap_{N\geq 1} \mathcal{G}_N$.

Along the same lines, here is another example of event in $\mathcal{T}$:

$$A_2 = \left\{ \omega \in \Omega \,:\, \lim_{n\to\infty} \frac{S_n(\omega)}{n} \text{ exists} \right\}$$

which is of direct interest to us in the sequel.

**Theorem 8.7.** (Kolmogorov's 0-1 law)
If the sequence $(X_n,\, n \geq 1)$ is independent and $A \in \mathcal{T}$, then $\mathbb{P}(A) \in \{0, 1\}$ (so in this case, $\mathcal{T}$ is essentially a trivial $\sigma$-field).

**Remark.** Please pay attention that $\mathbb{P}(A) \in \{0, 1\}$ means that either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, not that $0 \leq \mathbb{P}(A) \leq 1$, which is always true.

**Consequence.** Because the event $A_2$ above belongs to $\mathcal{T}$, we can therefore conclude that either the sequence $\frac{S_n}{n}$ converges a.s., or it diverges a.s., but it cannot be that convergence takes place with a probability which is strictly between 0 and 1. It turns out [without proof] that a.s. convergence takes place if and only if $\mathbb{E}(|X_1|) < +\infty$, and correspondingly that a.s. divergence takes place if and only if $\mathbb{E}(|X_1|) = +\infty$.

*Proof of Theorem 8.7.* The strategy for the proof of the above theorem is to show that when the $X_n$'s are independent, any event $A \in \mathcal{T}$ is independent of itself! So that $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$, implying $\mathbb{P}(A) \in \{0, 1\}$. First note that because of the independence of the $X$'s, for every $n \geq 1$, the $\sigma$-fields

$$\mathcal{F}_n = \sigma(X_1, \ldots, X_n) \quad \text{and} \quad \mathcal{G}_{n+1} = \sigma(X_{n+1}, X_{n+2}, \ldots)$$

are independent. As $\mathcal{T} \subset \mathcal{G}_{n+1}$ for every $n \geq 1$, this also implies that $\mathcal{T}$ is independent of $\mathcal{F}_n$ for every $n \geq 1$. But this implie that $\mathcal{T}$ is also independent of $\sigma(X_1, X_2, \ldots, X_n, \ldots)$[7], which is the $\sigma$-field generated by all the $\mathcal{F}_n$'s. Finally, observe that $\mathcal{T} \subset \sigma(X_1, X_2, \ldots, X_n, \ldots)$, which implies that $\mathcal{T}$ is independent of itself! So that any event $A \in \mathcal{T}$ has probability 0 or 1, as mentioned above. □

[7] Note that we skip here the measure-theoretic argument allowing to prove this.

## 8.8 Extension of the weak law: an example

As just seen, the strong law of large numbers does not extend beyond the assumption $\mathbb{E}(|X_1|) < +\infty$. The situation is different for the weak law. Let us consider an example of random variable with infinite expectation, which can be described through the following game: toss a (fair) coin until it falls on "heads" and call $T$ the total number of tosses; your gain is then $G = 2^T$ francs. What is your expected reward?

$$\mathbb{P}(\{G = 2^k\}) = \mathbb{P}(\{T = k\}) = \frac{1}{2^k}, \quad \text{so} \quad \mathbb{E}(G) = \sum_{k \geq 1} 2^k \, \mathbb{P}(\{G = 2^k\}) = \sum_{k \geq 1} 1 = +\infty$$

Nevertheless, note that it is not encouraged to bet a too large amount of money on such a game, as you might be disappointed with the result... This is called the *St-Petersburg paradox*.

Consider now that you are allowed to play this game multiple times, say $n$ times, with $n$ large. We assume that the $n$ realizations $G_1, \ldots, G_n$ of the game are independent. How much would you agree to pay to be allowed to play these $n$ games? If these games were "reasonable" games with finite expectation $\mu$, a plausible answer would be $n\mu$, according to the law of large numbers seen above, telling you that the sum of the $n$ games would then be of the order of $n\mu$ for large $n$. But as we saw, $\mathbb{E}(G_1) = +\infty$. So by the previous paragraph, defining $S_n = G_1 + \ldots + G_n$, we obtain that $\frac{S_n}{n}$ diverges almost surely (to $+\infty$ in this case). This again does not sound like a reasonable answer.

A more reasonable answer can be obtained via the following proposition, which is an extension of the weak law.

**Proposition 8.8.** Under the above assumptions, it holds that $\dfrac{S_n}{n \log_2 n} \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} 1$, i.e., that for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n \log_2 n} - 1\right| \geq \varepsilon\right\}\right) \underset{n\to\infty}{\to} 0$$

So for a large number of games $n$, investing $\log_2 n$ francs per game (that is, $n \log_2 n$ francs in total) is the right thing to do. We provide below the proof of this proposition, skipping some technical details.

*Proof.* A classical way to handle random variables with infinite expectation is to cut them off, that is, to consider, for a *fixed* value of $n$: $H_j = G_j \, 1_{\{G_j \leq n \log_2 n\}}$, $1 \leq j \leq n$, as well as $S'_n = H_1 + \ldots + H_n$ (the value of the cutoff, $n \log_2 n$, is of course chosen in hindsight to make everything work well, as we shall see below). Let us now write

$$\begin{aligned}
\mathbb{P}\left(\left\{\left|\frac{S_n}{n \log_2 n} - 1\right| \geq \varepsilon\right\}\right) &= \mathbb{P}(\{|S_n - n \log_2 n| \geq \varepsilon n \log_2 n\}) \\
&= \mathbb{P}(\{|S_n - n \log_2 n| \geq \varepsilon n \log_2 n, \ G_j \leq n \log_2 n, \ \forall 1 \leq j \leq n\}) \\
&\quad + \mathbb{P}(\{|S_n - n \log_2 n| \geq \varepsilon n \log_2 n, \ \exists 1 \leq j \leq n \text{ such that } G_j > n \log_2 n\})
\end{aligned}$$

Noticing that $G_j = H_j$ when $G_j \leq n \log_2 n$ and using the simple fact that $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ or $\mathbb{P}(B)$, the above expression can be upperbounded by

$$\mathbb{P}(\{|S'_n - n \log_2 n| \geq \varepsilon n \log_2 n\}) + \mathbb{P}(\{\exists 1 \leq j \leq n \text{ such that } G_j > n \log_2 n\}) \tag{6}$$

$S'_n = H_1 + \ldots + H_n$ is a sum of bounded random variables, and one can check that

$$\mathbb{E}(H_1) = \sum_{k \leq \log_2(n \log_2 n)} 2^k \, \frac{1}{2^k} = \sum_{k \leq \log_2(n \log_2 n)} 1 \simeq \log_2 n$$

so that $\mathbb{E}(S'_n) \simeq n \log_2 n$ and that, using Chebyshev's inequality, we obtain:

$$\mathbb{P}(\{|S'_n - n \log_2 n| \geq \varepsilon n \log_2 n\}) \leq \frac{\mathrm{Var}(S'_n)}{(\varepsilon n \log_2 n)^2} = \frac{n \, \mathrm{Var}(H_1)}{(\varepsilon n \log_2 n)^2}$$

where

$$\mathrm{Var}(H_1) \le \mathbb{E}(H_1^2) = \sum_{k \le \log_2(n \log_2 n)} 2^{2k} \, \frac{1}{2^k} = \sum_{k \le \log_2(n \log_2 n)} 2^k \simeq 2n \log_2 n$$

Therefore,

$$\mathbb{P}(\{|S'_n - n \log_2 n| \ge \varepsilon n \log_2 n\}) \le \frac{2n^2 \log_2 n}{(\varepsilon n \log_2 n)^2} = \frac{2}{\varepsilon^2 \, \log_2 n} \underset{n\to\infty}{\to} 0$$

The second term in (6) can be upperbounded using the union bound:

$$\mathbb{P}(\{\exists 1 \le j \le n \text{ such that } G_j > n \log_2 n\}) \le n \, \mathbb{P}(\{G_1 > n \log_2 n\}) = n \sum_{k > \log_2(n \log_2 n)} \frac{1}{2^k} \simeq \frac{1}{\log_2 n} \underset{n\to\infty}{\to} 0$$

which completes the proof. □

**Remark.** From the above proof, we deduce that the result cannot be extended to become a strong law type of result. Indeed, the decay of the probability $\mathbb{P}\left(\left\{\left|\frac{S_n}{n \log_2 n} - 1\right| \ge \varepsilon\right\}\right)$ is only $O\left(\frac{1}{\log_2(n)}\right)$, far too slow to apply the Borel-Cantelli lemma (and also too slow to apply the same trick as in our proof of the strong law; this last observation is left as exercise).

# 9 The central limit theorem

## 9.1 Convergence in distribution

Convergence in distribution is a key tool in probability, allowing notably to state the central limit theorem.

**Definition 9.1.** Let $(X_n,\, n \ge 1)$ be a sequence of random variables, *not necessarily* defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence $(X_n,\, n \ge 1)$ is said to *converge in distribution* to a limiting random variable $X$ (and this is denoted as $X_n \underset{n\to\infty}{\overset{d}{\to}} X$) if

$$F_{X_n}(t) = \mathbb{P}(\{X_n \le t\}) \underset{n\to\infty}{\to} F_X(t) = \mathbb{P}(\{X \le t\})$$

for every $t \in \mathbb{R}$ continuity point of the limiting cdf $F_X$.

**Remark 9.2.** Why asking only for convergence in continuity points of $F_X$ and not in all $t \in \mathbb{R}$? There are two main reasons for this:

- The fact is, it may happen that the limit cdf $F_X$ is itself discontinuous (if it is e.g. the cdf of a discrete random variable, in which case we recall that $F_X$ is a staircase function). In this case, it would be asking for too much to have a sequence of functions converging to $F_X$ in *every* $t \in \mathbb{R}$, including in points where the function $F_X$ makes a jump; one can at least imagine easily examples of sequences of functions that converge everywhere except in these points.

- Besides, as we know that the limit $F_X$ is a right-continuous function, by definition, this implies that even if there is no convergence in a point where $F_X$ makes a jump, it is always possible to reconstruct $F_X$ in this point by taking the limit from the right.

**Remark 9.3.** We have already seen an instance of convergence in distribution in Section 8.6: the convergence of the *empirical distribution* of a sequence of i.i.d. random variables. Note however that this example is a bit more complicated, as the empirical cdfs are also *random* in this case!

The proposition below shows that convergence in distribution is the weakest of the four notions of convergence we have seen so far.

**Proposition 9.4.** Let $(X_n,\, n \ge 1)$ be a sequence of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $X$ be another random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. If $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$, then $X_n \underset{n\to\infty}{\overset{d}{\to}} X$.

**Example 9.5.** The weak law of large numbers states that $\frac{S_n}{n} \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} \mu = \mathbb{E}(X_1)$ if $S_n = X_1 + \ldots + X_n$ and the $X$'s are i.i.d. random variables with finite expectation. The above proposition implies therefore that

$$F_{S_n/n}(t) = \mathbb{P}(\{S_n/n \le t\}) \underset{n\to\infty}{\to} \begin{cases} 1, & \text{if } t > \mu \\ 0, & \text{if } t < \mu \end{cases}$$

*Proof.* Recall that $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$ means that for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(\{|X_n - X| \ge \varepsilon\}) = 0$. We show below that this implies that $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ for every $t \in \mathbb{R}$ continuity point of $F_X$.

- Let us first compute, for a given fixed $\varepsilon > 0$:

$$\begin{aligned} F_{X_n}(t) &= \mathbb{P}(\{X_n \le t\}) = \mathbb{P}(\{X_n \le t,\, X \le t+\varepsilon\}) + \mathbb{P}(\{X_n \le t,\, X > t+\varepsilon\}) \\ &\le \mathbb{P}(\{X \le t+\varepsilon\}) + \mathbb{P}(\{|X_n - X| \ge \varepsilon\}) \end{aligned}$$

Because of the assumption made, this implies that $\limsup_{n\to\infty} F_{X_n}(t) \le F_X(t+\varepsilon) + 0$.

- Let us then compute, again for a given fixed $\varepsilon > 0$:

$$\begin{aligned} F_X(t-\varepsilon) &= \mathbb{P}(\{X \le t-\varepsilon\}) = \mathbb{P}(\{X \le t-\varepsilon,\, X_n \le t\}) + \mathbb{P}(\{X \le t-\varepsilon,\, X_n > t\}) \\ &\le \mathbb{P}(\{X_n \le t\}) + \mathbb{P}(\{|X_n - X| \ge \varepsilon\}) \end{aligned}$$

Again, because of the assumption made, this implies that $F_X(t-\varepsilon) \le \liminf_{n\to\infty} F_{X_n}(t) + 0$.

- In conclusion, we obtain for any given $\varepsilon > 0$:

$$F_X(t-\varepsilon) \le \liminf_{n\to\infty} F_{X_n}(t) \le \limsup_{n\to\infty} F_{X_n}(t) \le F_X(t+\varepsilon)$$

Assuming $t \in \mathbb{R}$ is a continuity point of $F_X$, we have $\lim_{\varepsilon\downarrow 0} F_X(t-\varepsilon) = \lim_{\varepsilon\downarrow 0} F_X(t+\varepsilon) = F_X(t)$, so by the above inequalities,

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$$

which proves the claim. □

**Remark 9.6.** Let us insist here that convergence in distribution is a *much weaker* notion than convergence in probability. For example, a sequence $(X_n,\, n \ge 1)$ of i.i.d. random variables does never converge in probability (unless these random variables are all constants), but it converges in distribution, because all cdfs are the same, so that the sequence $(F_{X_n}(t),\, n \ge 1)$ necessarily converges for every $t \in \mathbb{R}$.

## 9.2 Application: the Curie-Weiss model

In the following, we give an example, coming from statistical physics, of a sequence of random variables $(X_n, n \ge 1)$ taking values in $\{-1, +1\}$ such that the sequence $\frac{1}{n}(X_1 + \ldots + X_n)$ does not converge almost surely, nor in probability, but only in distribution. Actually, the sequence of random variables $(X_n, n \ge 1)$ we will consider is not an "orthodox" one, because the interaction between the first $n$ random variables $X_1, \ldots, X_n$ depends on the value of $n$. For each $n \ge 1$, we will therefore need a *different* probability space to describe the $n$-tuple $X_1, \ldots, X_n$. Let us see what this concretely means.

For a given $n$, the probability space on which we define the $n$ random variables $X_1, \ldots, X_n$ is the following: $\Omega_n = \{-1, +1\}^n$, $\mathcal{F}_n = \mathcal{P}(\Omega_n)$ and

$$\mathbb{P}_n(\{\omega\}) = \frac{1}{C_n} \exp\left(\frac{\beta}{n} \sum_{\substack{j,k=1 \\ j<k}}^{n} \omega_j \omega_k\right), \quad \omega \in \Omega_n$$

where $\beta > 0$ is a fixed parameter and the normalization constant $C_n$ is given by

$$C_n = \sum_{\omega \in \Omega_n} \exp\left( \frac{\beta}{n} \sum_{\substack{j,k=1 \\ j<k}}^{n} \omega_j \omega_k \right)$$

so that $\sum_{\omega \in \Omega_n} \mathbb{P}_n(\{\omega\}) = 1$. The random variables $X_1, \ldots, X_n$ are then defined as

$$X_j(\omega) = \omega_j, \quad j \in \{1, \ldots, n\}$$

We are interested below in the asymptotic behaviour of the sequence of random variables

$$M_n(\omega) = \frac{1}{n} \left( X_1(\omega) + \ldots + X_n(\omega) \right), \quad n \geq 1$$

under this model. As the probability space $\Omega_n$ changes for each value of $n$, we can only hope to obtain convergence *in distribution* for this sequence.

**Interpretation in statistical physics.** The random variables $X_1, \ldots, X_n$ represent *spins* (i.e., small magnetic moments) interacting together. The interaction here is *ferromagnetic*, which means that the spins are positively correlated: if $X_j = +1$, then this encourages[8] other spins $X_k$ to also take the value $+1$. The model is also a *mean field model*, in the sense that all pairs of spins interact together, with the same interaction strength for each pair. The parameter $\beta > 0$ models the inverse of the temperature of the system: when $\beta$ is small, the probability $\mathbb{P}_n$ tends to be uniform over the whole space $\Omega_n$, so the spins are highly fluctuating; on the contrary, when $\beta$ is large, the probability $\mathbb{P}_n$ concentrates on certain parts of $\Omega_n$ and the spins tend to freeze in these configurations. Finally, the quantity $M_n$ represents the *magnetization* of the system, which is the macroscopic quantity of interest that can actually be measured in a physical experiment (please remember that in a physical system, the number of particles (spins) is of the order of the Avogadro number, which is roughly $6 \cdot 10^{23}$, so it *really* makes sense to consider large $n$ limits!).

Contrary to many other unsolved problems of the same type, the Curie-Weiss model can be exactly analyzed in the large $n$ limit. To this end, let us first observe that for a given value of $n$,

$$\sum_{\substack{j,k=1 \\ j<k}}^{n} \omega_j \omega_k = \frac{1}{2} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} \omega_j \omega_k = \frac{1}{2} \left( \left( \sum_{j=1}^{n} \omega_j \right)^2 - n \right)$$

so that the probability $\mathbb{P}_n$ may be rewritten as

$$\mathbb{P}_n(\{\omega\}) = \frac{1}{\widetilde{C}_n} \exp\left( \frac{\beta}{2n} \left( \sum_{j=1}^{n} \omega_j \right)^2 \right), \quad \omega \in \Omega_n$$

where

$$\widetilde{C}_n = \sum_{\omega \in \Omega_n} \exp\left( \frac{\beta}{2n} \left( \sum_{j=1}^{n} \omega_j \right)^2 \right)$$

Observe also that for a given value of $n$, the magnetization $M_n$, being the average of $n$ random variables taking values in $\{-1, +1\}$, can take the following values:

$$-1, \quad -1 + \frac{2}{n}, \quad -1 + \frac{4}{n}, \quad \ldots \quad 1 - \frac{4}{n}, \quad 1 - \frac{2}{n}, \quad 1$$

[8] i.e., increases the probability of

Let $m$ be one such value and let us compute

$$\mathbb{P}_n(\{\omega \in \Omega \,:\, M_n(\omega) = m\}) = \sum_{\omega \in \Omega_n \,:\, \frac{1}{n}\,(\omega_1 + \ldots + \omega_n) = m} \mathbb{P}_n(\{\omega\})$$

Using the above formula for $\mathbb{P}_n(\{\omega\})$, this can be rewritten as

$$\begin{aligned}\mathbb{P}_n(\{\omega \in \Omega \,:\, M_n(\omega) = m\}) &= \sum_{\omega \in \Omega_n \,:\, \frac{1}{n}\,(\omega_1 + \ldots + \omega_n) = m} \frac{1}{\widetilde{C}_n} \exp\left(\frac{n\beta}{2} m^2\right) \\ &= \frac{1}{\widetilde{C}_n} \exp\left(\frac{n\beta}{2} m^2\right) \#\{\omega \in \Omega_n \,:\, \frac{1}{n}\,(\omega_1 + \ldots + \omega_n) = m\}\end{aligned}$$

where $\#A$ denotes the cardinality of the set $A$. Some combinatorics are required now, which we will skip, as it is not our main focus here. These combinatorics lead to the following approximation:

$$\#\{\omega \in \Omega_n \,:\, \frac{1}{n}\,(\omega_1 + \ldots + \omega_n) = m\} \underset{n\to\infty}{\simeq} \exp\left(nh\left(\frac{1+m}{2}\right)\right)$$

where $h(p) = -p\log(p) - (1-p)\log(1-p)$ is the classical entropy function, defined for $0 \le p \le 1$. Gathering the last two computations together, we obtain

$$\mathbb{P}_n(\{\omega \in \Omega \,:\, M_n(\omega) = m\}) \underset{n\to\infty}{\simeq} \frac{1}{\widetilde{C}_n} \exp\left(n\left(\frac{\beta m^2}{2} + h\left(\frac{1+m}{2}\right)\right)\right)$$

From the above expression, we infer[9] that the possible values taken by the magnetization $M_n$ in the large $n$ limit are the ones maximizing the following function:

$$f(m) = \frac{\beta m^2}{2} + h\left(\frac{1+m}{2}\right), \quad \text{where } -1 \le m \le +1$$

This function $f$ has an interesting behaviour: when $\beta \le 1$ (i.e., in the so-called "high temperature regime"), $f$ has a unique maximum in $m = 0$ and it can be shown that

$$M_n \underset{n\to\infty}{\overset{d}{\to}} 0$$

but note again that convergence only takes place in distribution here, so this is not a "law of large numbers" result. More interestingly, when $\beta > 1$ (i.e., in the so-called "low temperature regime"), the function $f$ has two maxima at the same level in $m = +m^*(\beta) > 0$ and $m = -m^*(\beta) < 0$, and it can be shown in this case that

$$M_n \underset{n\to\infty}{\overset{d}{\to}} M$$

where the limiting random variable $M$ takes values $+m^*(\beta)$ and $-m^*(\beta)$ with probability $\frac{1}{2}$ each.

The physical interpretation of the above result is the following: at high temperature ($\beta \le 1$), there is lots of agitation in the system, so the spins are disordered and the resulting magnetization is zero. On the other hand, at low temperature ($\beta > 1$), the positive interaction between the spins wins over the agitation, so that the spins get aligned in one direction ($+m^*(\beta)$) or the other ($-m^*(\beta)$).

## 9.3 Equivalent criterion for convergence in distribution

Observe first that if two random variables $X, Y$ share the same distribution, then

$$\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$$

for any continuous and bounded function $g : \mathbb{R} \to \mathbb{R}$. It turns out that the reciprocal statement holds, and even better: the following theorem, also known as (part of) the Portemanteau theorem, gives an equivalent criterion for convergence in distribution.

[9] Disclaimer: this needs to be seriously proven.

**Theorem 9.7.** Let $(X_n,\, n \geq 1)$ be a sequence of random variables and $X$ be another random variable. Then $X_n \underset{n\to\infty}{\overset{d}{\to}} X$ if and only if

$$\mathbb{E}(g(X_n)) \underset{n\to\infty}{\to} \mathbb{E}(g(X))$$

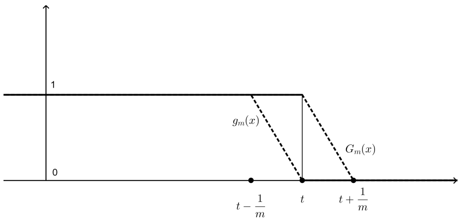
for every continuous and bounded function $g : \mathbb{R} \to \mathbb{R}$.

*Proof of the "if" part.* Observe first that Definition 9.1 is equivalent to saying that

$$\mathbb{E}(h_t(X_n)) \underset{n\to\infty}{\to} \mathbb{E}(h_t(X)),$$

for every function $h_t : \mathbb{R} \to \mathbb{R}$ of the form $h_t(x) = 1_{\{x\leq t\}}$, where $t$ is a continuity point of $F$. Our aim is to show that for fixed $t \in \mathbb{R}$, there is a way to approximate from above and from below the step function $h_t$ with continuous and bounded functions. To this end, assume without loss of generality that $t > 0$ and define for $m \geq 1$:

$$g_m(x) = \begin{cases} 1 & \text{if } x \leq t\left(1 - \frac{1}{m}\right) \\ m\left(1 - \frac{x}{t}\right) & \text{if } t\left(1 - \frac{1}{m}\right) < x \leq t \\ 0 & \text{if } x > t \end{cases} \quad \text{and} \quad G_m(x) = \begin{cases} 1 & \text{if } x \leq t \\ m\left(1 + \frac{1}{m} - \frac{x}{t}\right) & \text{if } t < x \leq t\left(1 + \frac{1}{m}\right) \\ 0 & \text{if } x > t\left(1 + \frac{1}{m}\right) \end{cases}$$



One deduces the following facts from the above figure (valid for any $m \geq 1$):

- the functions $g_m$ and $G_m$ are continuous and bounded;
- $g_m(x) \leq h_t(x) \leq G_m(x)$, $g_m(x) \leq g_{m+1}(x)$ and $G_m(x) \geq G_{m+1}(x)$ for every $x \in \mathbb{R}$;
- $\lim_{m\to\infty} g_m(x) = 1_{\{x<t\}}$ and $\lim_{m\to\infty} G_m(x) = 1_{\{x\leq t\}}$.

From the assumption made, we also obtain that for every $m \geq 1$:

$$\mathbb{E}(g_m(X)) = \lim_{n\to\infty} \mathbb{E}(g_m(X_n)) \leq \liminf_{n\to\infty} \mathbb{E}(h_t(X_n)) \leq \limsup_{n\to\infty} \mathbb{E}(h_t(X_n)) \leq \lim_{n\to\infty} \mathbb{E}(G_m(X_n)) = \mathbb{E}(G_m(X))$$

Besides, the monotone convergence theorem (not seen in this course) implies that

$$\lim_{m\to\infty} \mathbb{E}(g_m(X)) = \mathbb{E}(\lim_{n\to\infty} g_m(X_n)) = \mathbb{E}(1_{\{X<t\}}) = \mathbb{P}(\{X < t\})$$
$$\lim_{m\to\infty} \mathbb{E}(G_m(X)) = \mathbb{E}(\lim_{n\to\infty} G_m(X_n)) = \mathbb{E}(1_{\{X\leq t\}}) = \mathbb{P}(\{X \leq t\})$$

so

$$\mathbb{P}(\{X < t\}) \leq \liminf_{n\to\infty} \mathbb{E}(h_t(X_n)) \leq \limsup_{n\to\infty} \mathbb{E}(h_t(X_n)) \leq \mathbb{P}(\{X \leq t\})$$

which in turn implies that the limit exists (and is equal to what we want) when $t$ is a continuity point of $F_X$ (i.e., when $\mathbb{P}(\{X = t\}) = 0$). □

**Remark 9.8.** For $k \in \mathbb{N}$, let $C_b^k(\mathbb{R})$ denote the space of $k$ times continuously differentiable functions $g : \mathbb{R} \to \mathbb{R}$, which are bounded and whose all $k$ derivatives are also bounded. Replacing the above functions $g_m$ and $G_m$ by regular cubic splines, one can show the following improvement of the above theorem:

$$\text{If } \mathbb{E}(g(X_n)) \underset{n\to\infty}{\to} \mathbb{E}(g(X)) \text{ for every function } g \in C_b^3(\mathbb{R}), \text{ then } X_n \underset{n\to\infty}{\overset{d}{\to}} X$$

This remark will be useful below.

## 9.4 The central limit theorem

Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. square-integrable random variables. Let also $\mu = \mathbb{E}(X_1)$, $\sigma^2 = \mathrm{Var}(X_1)$ and $S_n = X_1 + \ldots + X_n$. Using basic properties of expectation and variance, we obtain the following:

$$\mathbb{E}(S_n) = n\,\mathbb{E}(X_1) = n\,\mu \quad \text{and} \quad \mathrm{Var}(S_n) = n\,\mathrm{Var}(X_1) = n\,\sigma^2$$

Please watch out that independence of the $X$'s is needed for the second computation, but not for the first one. These two equalities may be restated in a single one:

$$S_n = n\,\mu + \sqrt{n}\,\sigma\,\widetilde{S}_n,$$

where $\widetilde{S}_n$ is a random variable with $\mathbb{E}(\widetilde{S}_n) = 0$ and $\mathrm{Var}(\widetilde{S}_n) = \mathbb{E}(\widetilde{S}_n^2) = 1$. The central limit theorem states that as $n$ grows large, $\widetilde{S}_n$ converges in distribution to a Gaussian random variable with zero mean and unit variance. This result is therefore *universal*: the Gaussian distribution appears in the limit, independently of the distribution chosen for the $X$'s. Slightly more formally, we have the following.

**Theorem 9.9.** Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. random variables such that $\mathbb{E}(|X_1|^3) < +\infty$. defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let also $\mu = \mathbb{E}(X_1)$, $\sigma^2 = \mathrm{Var}(X_1)$ (the latter being assumed to be strictly positive), $S_n = X_1 + \ldots + X_n$ and

$$\widetilde{S}_n = \frac{S_n - n\,\mu}{\sqrt{n}\,\sigma}, \quad n \geq 1$$

Then $\widetilde{S}_n \underset{n\to\infty}{\overset{d}{\to}} Z \sim \mathcal{N}(0,1)$, i.e.,

$$\mathbb{P}(\{\widetilde{S}_n \leq t\}) \underset{n\to\infty}{\to} \mathbb{P}(\{Z \leq t\}) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx, \quad \forall t \in \mathbb{R}$$

Note that convergence has to take place here for *every* $t \in \mathbb{R}$, as the limiting cdf is continuous.

**Remark 9.10.** The above condition $\mathbb{E}(|X_1|^3) < +\infty$ is not needed for the conclusion of the theorem to hold; with extra effort, it can be proven under the weaker assumption that $X_1$ is only square-integrable (and this last assumption is clearly needed in order for $\sigma$ to be well defined).

The proof of the theorem relies on the following sequence of two lemmas, which together are known as *Lindeberg's principle* that is of interest in its own right. This principle can be informally stated as follows: in a sum of independent and small random variables, one may replace each small random variable by another one having the same mean and variance, without changing asymptotically the distribution of the sum.

**Lemma 9.11.** Let $g \in C_b^3(\mathbb{R})$ and $X$, $Y$, $Z$ be random variables such that $X$ is independent of both $Y$ and $Z$, $\mathbb{E}(Y) = \mathbb{E}(Z)$, $\mathbb{E}(Y^2) = \mathbb{E}(Z^2)$ and $\mathbb{E}(|Y|^3) < +\infty$, $\mathbb{E}(|Z|^3) < +\infty$. Then

$$|\mathbb{E}(g(X+Y)) - \mathbb{E}(g(X+Z))| \leq \frac{C}{6}\left(\mathbb{E}(|Y|^3) + \mathbb{E}(|Z|^3)\right)$$

where $C = \sup_{x\in\mathbb{R}} |g'''(x)|$.

What this lemma is essentially saying is that provided both $Y$ and $Z$ are "small" random variables, one may trade $Y$ for $Z$ in the expression $\mathbb{E}(g(X+Y))$ without changing much the value of the expectation.

*Proof.* By Taylor's expansion, we obtain for real numbers $x, y$:

$$g(x+y) = g(x) + y\,g'(x) + \frac{y^2}{2} g''(x) + \frac{y^3}{6} g'''(u) \quad \text{for some } u \text{ such that } |u - x| \leq |y|$$

The independence of $X$ and $Y$ then implies that

$$\mathbb{E}(g(X+Y)) = \mathbb{E}(g(X)) + \mathbb{E}(Y)\,\mathbb{E}(g'(X)) + \frac{1}{2}\,\mathbb{E}(Y^2)\,\mathbb{E}(g''(X)) + \frac{1}{6}\,\mathbb{E}(Y^3\,g'''(U))$$

where $U$ is a random variable satisfying $|U-X| \leq |Y|$. Similarly, one may write

$$\mathbb{E}(g(X+Z)) = \mathbb{E}(g(X)) + \mathbb{E}(Z)\,\mathbb{E}(g'(X)) + \frac{1}{2}\,\mathbb{E}(Z^2)\,\mathbb{E}(g''(X)) + \frac{1}{6}\,\mathbb{E}(Z^3\,g'''(V))$$

where $V$ is a random variable satisfying $|V-X| \leq |Z|$. By the assumptions made, we obtain

$$\mathbb{E}(g(X+Y)) - \mathbb{E}(g(X+Z)) = \frac{1}{6}\left(\mathbb{E}(Y^3\,g'''(U)) - \mathbb{E}(Z^3 g'''(V))\right)$$

so

$$|\mathbb{E}(g(X+Y)) - \mathbb{E}(g(X+Z))| \leq \frac{C}{6}\left(\mathbb{E}(|Y|^3) + \mathbb{E}(|Z|^3)\right)$$

which completes the proof. □

The above lemma generalizes to the case of sums of multiple random variables, as shown below.

**Lemma 9.12.** Let $g \in C_b^3(\mathbb{R})$ and $Y_1,\ldots,Y_n$, $Z_1,\ldots,Z_n$ be random variables, all independent and such that $\mathbb{E}(Y_i) = \mathbb{E}(Z_i)$, $\mathbb{E}(Y_i^2) = \mathbb{E}(Z_i^2)$ and $\mathbb{E}(|Y_i|^3) < +\infty$, $\mathbb{E}(|Z_i|^3) < +\infty$ for all $i \in \{1,\ldots,n\}$. Then

$$|\mathbb{E}(g(Y_1+\ldots+Y_n)) - \mathbb{E}(g(Z_1+\ldots+Z_n))| \leq \frac{C}{6}\sum_{i=1}^{n}\left(\mathbb{E}(|Y_i|^3) + \mathbb{E}(|Z_i|^3)\right)$$

where $C = \sup_{x\in\mathbb{R}} |g'''(x)|$.

*Proof.* Define

$$\begin{aligned} &X_1 = Z_2+\ldots+Z_n, \qquad X_n = Y_1+\ldots+Y_{n-1}, \qquad \text{and} \\ &X_i = Y_1+\ldots+Y_{i-1}+Z_{i+1}+\ldots+Z_n \quad \text{for } i \in \{2,\ldots,n-1\} \end{aligned}$$

Observe then that

$$\begin{aligned} &Y_1+\ldots+Y_n = X_n+Y_n, \qquad Z_1+\ldots+Z_n = X_1+Z_1, \qquad \text{and} \\ &X_i+Y_i = Y_1+\ldots+Y_i+Z_{i+1}+\ldots+Z_n = X_{i+1}+Z_{i+1}, \quad \text{for } i \in \{1,\ldots,n-1\} \end{aligned}$$

so that

$$\begin{aligned} &|\mathbb{E}(g(Y_1+\ldots+Y_n)) - \mathbb{E}(g(Z_1+\ldots+Zn))| = |\mathbb{E}(g(X_n+Y_n)) - \mathbb{E}(g(X_1+Z_1))| \\ &= |\mathbb{E}(g(X_n+Y_n)) - \mathbb{E}(g(X_n+Z_n)) + \mathbb{E}(g(X_{n-1}+Y_{n-1})) + \ldots \\ &\qquad \ldots - \mathbb{E}(g(X_2+Z_2)) + \mathbb{E}(g(X_1+Y_1)) - \mathbb{E}(g(X_1+Z_1))| \\ &\leq \sum_{i=1}^{n} |\mathbb{E}(g(X_i+Y_i)) - \mathbb{E}(g(X_i+Z_i))| \leq \frac{C}{6}\sum_{i=1}^{n}\left(\mathbb{E}(|Y_i|^3) + \mathbb{E}(|Z_i|^3)\right) \end{aligned}$$

by repeated uses of Lemma 9.11. The proof is complete. □

As the above lemma is valid for any function $g \in C_b^3(\mathbb{R})$, this says that if a random variable is the sum of multiple small independent components, then essentially only the first and second moments of these components matter for the computation of the distribution of the random variable itself. We are now in position to prove Theorem 9.9.

*Proof of Theorem 9.9.* Let us first estimate $\mathbb{E}(g(\widetilde{S}_n))$ for $g \in C_b^3(\mathbb{R})$ and $n$ large. Defining $Y_i = \frac{X_i - \mu}{\sqrt{n}\,\sigma}$, we may rewrite

$$\widetilde{S}_n = \frac{S_n - n\,\mu}{\sqrt{n}\,\sigma} = \sum_{i=1}^n \frac{X_i - \mu}{\sqrt{n}\,\sigma} = \sum_{i=1}^n Y_i$$

The random variables $Y_i$ are i.i.d. with $\mathbb{E}(Y_i) = 0$, $\mathbb{E}(Y_i^2) = \frac{1}{n}$ and $\mathbb{E}(|Y_i|^3) = O(n^{-3/2})$.

Let now $Z_1, \ldots, Z_n$ be i.i.d. $\mathcal{N}(0, \frac{1}{n})$ random variables, independent of $X_1, \ldots, X_n$ (and therefore also of $Y_1, \ldots, Y_n$). Clearly, it is also the case $\mathbb{E}(Z_i) = 0$, $\mathbb{E}(Z_i^2) = \frac{1}{n}$ and $\mathbb{E}(|Z_i|^3) = O(n^{-3/2})$. By Lemma 9.12, we have

$$\begin{aligned} |\mathbb{E}(g(Y_1 + \ldots + Y_n)) - \mathbb{E}(g(Z_1 + \ldots + Z_n))| &\leq \frac{C}{6} \sum_{i=1}^n \left(\mathbb{E}(|Y_i|^3) + \mathbb{E}(|Z_i|^3)\right) \\ &\leq \frac{C}{6} \sum_{i=1}^n O(n^{-3/2}) = O(n^{-1/2}) \underset{n\to\infty}{\to} 0 \end{aligned}$$

Observe next that as $\mathrm{Var}(Z_1 + \ldots + Z_n) = \frac{1}{n} + \ldots + \frac{1}{n} = 1$ for all $n \geq 1$, the random variables $Z_1 + \ldots + Z_n$ all share the same distribution $\mathcal{N}(0,1)$, so $\mathbb{E}(g(Z_1 + \ldots + Z_n)) = \mathbb{E}(g(Z))$, for all $n \geq 1$, where $Z \sim \mathcal{N}(0,1)$. By Theorem 9.7 (or more precisely Remark 9.8 following it), we finally deduce that

$$\widetilde{S}_n = Y_1 + \ldots + Y_n \underset{n\to\infty}{\overset{d}{\to}} Z$$

which completes the proof of the theorem. □.

## 9.5 An alternate proof of the central limit theorem

The proof of the central limit theorem 9.9 provided below relies on the use of characteristic functions and more specifically on the following interesting characterization of convergence in distribution (whose proof is omitted here, but is related to the Portemanteau theorem seen in Section 9.3).

**Proposition 9.13.** Let $(X_n,\, n \geq 1)$ be a sequence of random variables and $X$ be another random variable. Then

$$X_n \underset{n\to\infty}{\overset{d}{\to}} X \quad \text{if and only if} \quad \phi_{X_n}(t) \underset{n\to\infty}{\to} \phi_X(t), \quad \forall t \in \mathbb{R}$$

Note that such a proposition is made possible because a characteristic function is fully characterizing its corresponding distribution, as guaranteed by the inversion formula (see Section 5).

*Alternate proof of Theorem 9.9.* As already seen above, we have

$$\widetilde{S}_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \sum_{j=1}^n Y_j$$

where $Y_j = \frac{X_j - \mu}{\sqrt{n}\,\sigma}$ are i.i.d. random variables with $\mathbb{E}(Y_1) = 0$, $\mathbb{E}(Y_1^2) = \frac{1}{n}$ and $\mathbb{E}(|Y_1|^3) = O(n^{-3/2})$) (using the assumption that $\mathbb{E}(|X_1|^3) < +\infty$).

Let $\phi_n$ be the characteristic function of $\widetilde{S}_n$. Using Proposition 9.13, it is sufficient to prove that

$$\phi_n(t) \underset{n\to\infty}{\to} \phi(t) = e^{-t^2/2}, \quad \forall t \in \mathbb{R}$$

as $\phi$ is the characteristic function of $Z \sim \mathcal{N}(0,1)$. To this end, making use of the factorization property of characteristic functions, we compute

$$\phi_n(t) = \mathbb{E}\left(\exp\left(it\widetilde{S}_n\right)\right) = \prod_{j=1}^n \mathbb{E}(\exp(itY_j)) = (\mathbb{E}(\exp(itY_1)))^n$$

Using Taylor's expansion, we further obtain

$$\mathbb{E}(\exp(itY_1)) = 1 + (it)\,\mathbb{E}(Y_1) + \frac{(it)^2}{2}\,\mathbb{E}(Y_1^2) + \frac{(it)^3}{6}\,O(\mathbb{E}(|Y_1|^3)) = 1 + 0 - \frac{t^2}{2n} + O(n^{-3/2})$$

so

$$\phi_n(t) = \left(1 - \frac{t^2}{2n} + O(n^{-3/2})\right)^n \underset{n\to\infty}{\to} e^{-t^2/2}$$

which actually needs some careful verification here, as the $O(n^{-3/2})$ is complex-valued. This completes our second proof of the central limit theorem. □

## 9.6 Application: the coupon collector problem

We describe below a "convergence in distribution" type of result occurring in the context of a famous problem in probability theory: the coupon collector problem. Interestingly, the limiting distribution is not Gaussian in this case. Another famous example of that kind is the birthday problem (see exercises).

**Problem formulation.** The question is simple: suppose that $m$ balls are thrown independently and uniformly at random into $n$ bins. How large need $m$ be in order to ensure that the $n$ bins are all occupied by at least one ball? This problem was first studied by Laplace in 1812. Since then, many variations of the problem have been studied, with interesting applications in various areas (including the one of estimating how much money you need to spend, once every four years, in order to complete an empty book with 682 stickers[10]).

**Expected behaviour.** For $k \in \{1, \ldots, n\}$, define $T_k$ to be the first time (i.e., the smallest value of $m$) such that $k$ bins are occupied. We can compute the expected value of $T_n$ by the following reasoning. Let $X_1 = T_1 = 1$ and $X_k = T_k - T_{k-1}$ for $k \in \{2, \ldots, n\}$ (in words, $X_k$ is the waiting time for a new bin to be reached after $k-1$ of them have already been reached). Then

$$\mathbb{P}(\{X_k = \ell\}) = \left(\frac{k-1}{n}\right)^{\ell-1} \left(1 - \frac{k-1}{n}\right) \quad \text{for } \ell \geq 1$$

i.e., $X_k$ is a geometric random variable of parameter $p_k = 1 - \frac{k-1}{n} = \frac{n-k+1}{n}$. Such a random variable has expectation

$$\mathbb{E}(X_k) = \frac{1}{p_k} = \frac{n}{n-k+1}$$

logically translating the fact that the average time duration for a new bin to be reached increases with $k$. We finally obtain

$$\mathbb{E}(T_n) = \sum_{k=1}^n \mathbb{E}(X_k) = n \sum_{k=1}^n \frac{1}{n-k+1} = n \sum_{k=1}^n \frac{1}{k} \simeq n \log n$$

as $n$ gets large (we spare you here the additional Euler constant that refines this approximation).

**Threshold phenomenon.** The result below (due to Erdős and Rényi) shows that $m = n \log n$ is not only an average behaviour, but also the precise threshold before which the coupon collection is not complete with high probability and after which it is complete with high probability.

**Proposition.** For $t \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbb{P}(\{T_n \leq n \log n + nt\}) = \exp(-e^{-t})$$

[10] For those interested, the answer was around 1'000 Swiss Francs for the last edition!

In other words, this is saying that the sequence of random variables $(G_n,\, n \geq 1)$ defined as

$$G_n = \frac{T_n - n\log n}{n}$$

converges in distribution as $n \to \infty$ towards the random variable $G$ with cdf

$$F_G(t) = \exp(-e^{-t}), \quad t \in \mathbb{R}$$

also known as the standard Gumbel distribution.

**Remark.** The implications of the above result are the following:

- If $t$ is large and positive, then $\mathbb{P}(\{T_n \leq n\log n + nt\}) \simeq \exp(-e^{-t}) \simeq \exp(-0) \simeq 1$.

- If $t$ is large and negative, then $\mathbb{P}(\{T_n \leq n\log n + nt\}) \simeq \exp(-e^{-t}) \simeq \exp(-\infty) \simeq 0$.

which proves the claim made at the beginning of this section.

*Proof sketch.* We do not provide below a complete proof of the above proposition, but just an approximation argument. For $m \geq 1$ and $i \in \{1, \ldots, n\}$, define

$$E_{im} = \{\text{bin } i \text{ is still empty after } m \text{ throws}\}$$

Considering $m = \lceil n\log n + tn \rceil$ with $t \in \mathbb{R}$ fixed, we obtain as $n$ gets large:

$$\mathbb{P}(E_{im}) = \left(1 - \frac{1}{n}\right)^m \simeq \exp\left(-\frac{m}{n}\right) \simeq \exp(-\log n - t) = \frac{e^{-t}}{n}, \quad \forall i \in \{1, \ldots, n\} \tag{7}$$

Besides, the events $E_{1m}, \ldots, E_{nm}$ are "approximately independent" in the following sense[11]: for every $i \in \{1, \ldots, m\}$ and $J \subset \{1, \ldots, n\}$ such that $J \cap \{i\} = \emptyset$ and $|J| = k$ with $k$ fixed, we have

$$\mathbb{P}(E_{im} \,|\, \cap_{j \in J} E_{jm}) \simeq \mathbb{P}(E_{im}) \quad \text{as } n \to \infty \text{ (and } m = \lceil n\log n + tn \rceil) \tag{8}$$

Indeed:

$$\begin{aligned} \mathbb{P}(E_{im} \,|\, \cap_{j \in J} E_{jm}) &= \frac{\mathbb{P}(E_{im} \cap (\cap_{j \in J} E_{jm}))}{\mathbb{P}(\cap_{j \in J} E_{jm})} = \frac{(1 - (k+1)/n)^m}{(1 - k/n)^m} \\ &= \left(\frac{n-k-1}{n-k}\right)^m = \left(1 - \frac{1}{n-k}\right)^m \\ &\simeq \exp\left(-\frac{m}{n-k}\right) \simeq \exp\left(-\frac{m}{n}\right) \simeq \mathbb{P}(E_{im}) \end{aligned}$$

for $k$ fixed and $n$ large (with $m = \lceil n\log n + tn \rceil$). Using the above approximations (7) and (8), we obtain

$$\begin{aligned} \mathbb{P}(\{T_n \leq m\}) &= \mathbb{P}(\{\text{all bins are occupied after } m \text{ throws}\}) \\ &= \mathbb{P}\left(\bigcap_{i=1}^n E_{im}^c\right) \simeq \prod_{i=1}^n \mathbb{P}(E_{im}^c) \simeq \left(1 - \frac{e^{-t}}{n}\right)^n \simeq \exp(-e^{-t}) \end{aligned}$$

which "proves" the claim (for a full proof, the inclusion-exclusion principle is needed, as well as the Bonferroni inequalities). □

[11] Note that they cannot be all independent, because if for example the first $n-1$ bins are empty after $m$ throws, then the last one can't be.

# 10 Moments and Carleman's theorem

**Definition 10.1.** Let $X$ be a random variable and $k \geq 0$. If $\mathbb{E}(\big(|X|^k\big) < +\infty$, we say that the *moment of order* $k$ of the random variable $X$ is finite and defined as

$$m_k = \mathbb{E}\left(X^k\right)$$

Note that if a random variable has a finite moment of order $k$, then all moments of order $0 \leq j \leq k$ are also finite. Indeed, using Jensen's inequality along with the fact that $f(x) = x^{j/k}$ is concave for $x \geq 0$ and $j \leq k$, we obtain

$$\mathbb{E}\left(|X|^j\right) = \mathbb{E}\left(f\left(|X|^k\right)\right) \leq f\left(\mathbb{E}\left(|X|^k\right)\right) < +\infty$$

Note also that the moment of order 0 is always finite and equal to $m_0 = 1$.

Furthermore, if all the moments $(m_k,\, k \geq 0)$ of a given random variable $X$ are finite, then the infinite matrix $M$ whose entries are given by $M_{jk} = m_{j+k},\ j,k \geq 0$ is symmetric and positive semi-definite, that is,

$$\sum_{j,k=0}^{n} c_j\, c_k\, M_{jk} \geq 0, \quad \forall n \geq 1,\ c_0, ..., c_n \in \mathbb{R}$$

Indeed,

$$\sum_{j,k=0}^{n} c_j\, c_k\, m_{j+k} = \sum_{j,k=0}^{n} c_j\, c_k\, \mathbb{E}\left(X^{j+k}\right) = \mathbb{E}\left(\left(\sum_{j=1}^{n} c_j\, X^j\right)^2\right) \geq 0$$

And the reciprocal statement also holds (this is known as the solution of the *Hamburger moment problem*): if $(m_n, n \geq 0)$ is a sequence of real numbers such that $m_0 = 1$ and the infinite matrix $M$ defined as $M_{jk} = m_{j+k},\ j,k \geq 0$ is symmetric and positive semi-definite, then $(m_n, n \geq 0)$ is the sequence of moment of some random variable $X$. It is however not clear whether the distribution of the random variable $X$ is fully characterized by the sequence of moments $(m_n, n \geq 0)$, even if these are all finite (if they are not, then the answer is clearly no). It turns out that the general answer to this question depends on the *growth* of the moments $m_k$ as a function of $k$. Here are first two particular cases:

1. If $X$ is a bounded random variable (i.e., there exists $C > 0$ such that $|X(\omega)| \leq C$ for every $\omega \in \Omega$), then
$$|m_k| = |\mathbb{E}(X^k)| \leq \mathbb{E}(|X|^k) \leq C^k, \quad \forall k \geq 0$$
and in this case, it can be proven that the knowledge of the sequence $(m_k,\, k \geq 0)$ fully characterizes the distribution of $X$.

2. Consider now a random variable $X$ whose moments are given by $m_k = \exp(k^2/2),\ k \geq 0$. In this case, it is impossible to tell whether a) $X$ is a continuous log-normal random variable:
$$X = e^Z, \quad \text{where } Z \sim \mathcal{N}(0,1)$$
or b) $X$ is a *discrete* random variable with pmf:
$$\mathbb{P}(\{X = e^j\}) = \frac{1}{C}\, \exp(-j^2/2), \quad j \in \mathbb{Z}, \quad \text{where } C = \sum_{j \in \mathbb{Z}} \exp(-j^2/2)$$

The most precise answer to the above question is given by Carleman's theorem, stated below.

**Theorem 10.2.** If $X$ is a random variable whose moments are all finite and satisfy

$$\sum_{k \geq 1} m_{2k}^{-\frac{1}{2k}} = +\infty \tag{9}$$

then the distribution of the random variable $X$ is uniquely determined by its sequence of moments $(m_k,\, k \geq 0)$.

**Remarks.** - Even though this is far from obvious, condition (9) is actually a limit on the growth of the sequence $(m_k,\, k \geq 0)$, which can be translated approximately as

$$|m_k| \leq \exp(ck \log k), \quad \text{where } c \leq 1$$

In order to better understand what condition (9) means, let us see what it gives in our two examples above:

1. When $|m_k| \leq C^k$, we obtain that
$$m_{2k}^{-\frac{1}{2k}} \geq C^{-1}, \quad \text{so that indeed,} \quad \sum_{k \geq 1} m_{2k}^{-\frac{1}{2k}} \geq \sum_{k \geq 1} C^{-1} = +\infty$$

2. When $m_k = \exp(k^2/2)$, we obtain
$$m_{2k}^{-\frac{1}{2k}} = \exp\left(-\frac{(2k)^2}{2(2k)}\right) = \exp(-k), \quad \text{so} \quad \sum_{k \geq 1} m_{2k}^{-\frac{1}{2k}} = \sum_{k \geq 1} \exp(-k) < +\infty$$

- One may wonder why Carleman's condition (9) only involves the even moments $m_{2k}$ and not the odd ones. This is because the growth of the odd moments is controlled by that of the even moments. Cauchy-Schwarz' inequality indeed guarantees that

$$\left(\mathbb{E}\left(|X|^{2k+1}\right)\right)^2 = \left(\mathbb{E}\left(|X|^k\,|X|^{k+1}\right)\right)^2 \leq \mathbb{E}\left(X^{2k}\right)\,\mathbb{E}\left(X^{2k+2}\right) = m_{2k}\, m_{2k+2}, \quad \forall k \geq 0$$

**Yet another characterization of convergence in distribution.**

The above analysis allows us to state the following nice criterion guaranteeing convergence in distribution [without proof].

**Theorem 10.3.** Let $(m_k,\, k \geq 0)$ be a sequence of real numbers satisfying Carleman's condition (9). If $(X_n,\, n \geq 1)$ is a sequence of random variables such that for all $k \geq 0$,

$$\mathbb{E}(X_n^k) \longrightarrow m_k, \quad \text{as } n \to \infty$$

then $X_n$ converges in distribution towards a random variable $X$, whose moments are $(m_k,\, k \geq 0)$.

Note that this characterization offers another way to prove the central limit theorem, but is of course useful in numerous other contexts also.

# 11 Concentration inequalities

The weak law of large numbers states that $\dfrac{S_n}{n}$ converges in probability to $\mathbb{E}(X_1)$, when $S_n = X_1 + \ldots + X_n$ and the $X$'s are i.i.d. random variables. This is exactly saying that for every fixed $t > 0$,

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq t\right\}\right) \underset{n \to \infty}{\to} 0$$

However, this law does not say anything about the *speed* of convergence to 0 of this probability. The answer to this question is provided by concentration inequalities.

## 11.1 Hoeffding's inequality

**Theorem 11.1.** Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. and integrable random variables, defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and such that $|X_n(\omega) - \mathbb{E}(X_n)| \leq 1$, for all $n \geq 1$ and $\omega \in \Omega$. Let also $S_n = X_1 + \ldots + X_n$. Then

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq t\right\}\right) \leq 2\, \exp\left(-\frac{nt^2}{2}\right), \quad \forall t > 0,\, n \geq 1$$

Before proving this theorem, let us make a few observations.

**Remarks.** - From this result, we easily recover the strong law of large numbers. Indeed, for all $t > 0$, we have

$$\sum_{n\geq 1} \mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq t\right\}\right) \leq 2 \sum_{n \geq 1} \exp\left(-\frac{nt^2}{2}\right) < \infty$$

which implies by the Borel-Cantelli lemma that

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq t \quad \text{infinitely often}\right\}\right) = 0$$

This therefore says that $\dfrac{S_n}{n} \underset{n\to\infty}{\to} \mathbb{E}(X_1)$ almost surely.

- Note the universal character of this result (as already observed for the central limit theorem): the upper bound on the probability does not depend on the distribution of the $X$'s (except for the fact that these are bounded random variables by assumption).

- Note also the following: replacing $t$ by $\frac{u}{\sqrt{n}}$ in the above statement, we obtain

$$\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq \frac{u}{\sqrt{n}}\right\}\right) \leq 2\, \exp\left(-\frac{u^2}{2}\right)$$

Recalling that the cdf of a $\mathcal{N}(0,1)$ random variable behaves as

$$F(u) \sim 1 - \exp\left(-\frac{u^2}{2}\right) \quad \text{when } u \text{ is large}$$

we observe here another analogy with the central limit theorem.

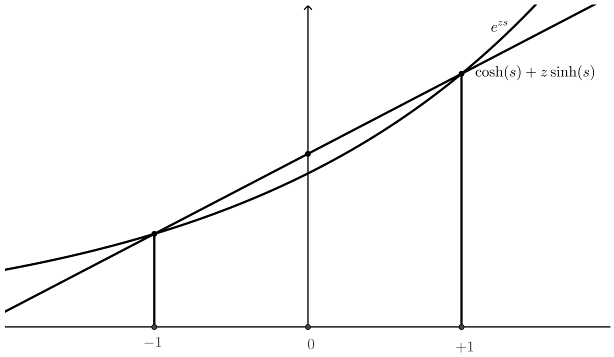
The following lemma is the key to the proof of Theorem 11.1.

**Lemma 11.2.** Let $Z$ be a random variable such that $|Z(\omega)| \leq 1$ for all $\omega \in \Omega$ and $\mathbb{E}(Z) = 0$. Then

$$\mathbb{E}\left(e^{sZ}\right) \leq \exp\left(\frac{s^2}{2}\right), \quad \forall s \in \mathbb{R}$$

*Proof.* First observe that the mapping $z \mapsto e^{sz}$ is convex for any $s \in \mathbb{R}$, so for any $z \in [-1, 1]$, we have

$$e^{sz} \leq \frac{e^s + e^{-s}}{2} + z\, \left(\frac{e^s - e^{-s}}{2}\right) = \cosh(s) + z\, \sinh(s)$$



Therefore, as $|Z(\omega)| \leq 1$ for all $\omega \in \Omega$ and $\mathbb{E}(Z) = 0$, we obtain

$$\mathbb{E}\left(e^{sZ}\right) \leq \cosh(s) + \mathbb{E}(Z)\,\sinh(s) = \cosh(s)$$

The rest of the proof is calculus. Let $f(s) = \log\cosh(s)$; then

$$f'(s) = \frac{\sinh(s)}{\cosh(s)} = \tanh(s) \quad \text{and} \quad f''(s) = 1 - (\tanh(s))^2 \leq 1$$

So for $s \geq 0$, $f'(s) = f'(0) + \int_0^s f''(t)\,dt \leq 0 + \int_0^s dt = s$. Similarly,

$$f(s) = f(0) + \int_0^s f'(t)\,dt \leq 0 + \int_0^s t\,dt = \frac{s^2}{2}$$

This implies that $\cosh(s) \leq \exp\left(\frac{s^2}{2}\right)$. The same reasoning can be applied to the case $s \leq 0$, leading to the same conclusion. This proves the lemma. □

*Proof of Theorem 11.1.* Let us compute

$$\begin{aligned}\mathbb{P}\left(\left\{\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq t\right\}\right) &= \mathbb{P}\left(\{|S_n - n\,\mathbb{E}(X_1)| \geq nt\}\right) = \mathbb{P}\left(\{|S_n - \mathbb{E}(S_n)| \geq nt\}\right) \\ &= \mathbb{P}\left(\{S_n - \mathbb{E}(S_n) \geq nt\}\right) + \mathbb{P}\left(\{S_n - \mathbb{E}(S_n) \leq -nt\}\right) \qquad (10)\end{aligned}$$

Let us focus on the first term, as the second can be handled exactly in the same way. By Chebyshev's inequality (using $\varphi(x) = e^{sx}$ with $s \geq 0$), we obtain

$$\begin{aligned}\mathbb{P}\left(\{S_n - \mathbb{E}(S_n) \geq nt\}\right) &\leq \frac{\mathbb{E}\left(e^{s\,(S_n - \mathbb{E}(S_n))}\right)}{e^{snt}} = e^{-nst}\,\mathbb{E}\left(\prod_{j=1}^n e^{s\,(X_j - \mathbb{E}(X_j))}\right) \\ &= e^{-nst}\,\left(\mathbb{E}\left(e^{s(X_1 - \mathbb{E}(X_1))}\right)\right)^n\end{aligned}$$

By the assumptions made, the random variable $Z = X_1 - \mathbb{E}(X_1)$ satisfies the assumptions of Lemma 11.2, so $\mathbb{E}\left(e^{s(X_1 - \mathbb{E}(X_1))}\right) \leq e^{s^2/2}$. This implies finally that

$$\mathbb{P}\left(\{S_n - \mathbb{E}(S_n) \geq nt\}\right) \leq e^{-nst + ns^2/2} = e^{n(s^2/2 - st)}$$

As $s \geq 0$ is a freely chosen parameter, we deduce that

$$\mathbb{P}\left(\{S_n - \mathbb{E}(S_n) \geq nt\}\right) \leq \min_{s \geq 0} e^{n(s^2/2 - st)} = e^{-n \max_{s \geq 0}(st - s^2/2)}$$

A simple derivation shows that the maximum (i.e., the tightest upper bound) is reached in $s^* = t$. This gives $\mathbb{P}\left(\{S_n - \mathbb{E}(S_n) \geq nt\}\right) \leq e^{-nt^2/2}$. As mentioned above, a similar reasoning gives the same upper bound on the second term in (10), and this concludes the proof. □

**Generalization.** (This is actually Hoeffding's original statement.)

Let $(X_n,\, n \geq 1)$ be a sequence of independent and integrable random variables (so not necessarily i.i.d.) such that $X_n(\omega) \in [a_n, b_n]$ for all $n \geq 1$ and $\omega \in \Omega$. Let also $S_n = X_1 + \ldots + X_n$. Then

$$\mathbb{P}\left(\{|S_n - \mathbb{E}(S_n)| \geq nt\}\right) \leq 2\, \exp\left(-\frac{2n^2t^2}{\sum_{j=1}^n (b_j - a_j)^2}\right), \quad \forall t > 0,\, n \geq 1$$

The proof is strictly speaking the same as above, but note that in this general case, $\dfrac{S_n}{n}$ need not converge to a limit as $n \to \infty$.

## 11.2 Large deviations principle

Large deviations estimates lead to a refinement of Hoeffding's inequality. Rather than stating the result from the beginning, let us discover it together!

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 such that $\mathbb{E}(e^{sX_1}) < +\infty$ for all $|s| < s_0$, for some $s_0 > 0$ [12]. Let also $t > \mathbb{E}(X_1)$ and $S_n = X_1 + \ldots + X_n$. Using then again Chebyshev's inequality (with $\varphi(x) = e^{sx}$ and $s \geq 0$), we obtain

$$\begin{aligned} \mathbb{P}\left(\left\{\frac{S_n}{n} \geq t\right\}\right) &= \mathbb{P}(\{S_n > nt\}) \leq \frac{\mathbb{E}\left(e^{sS_n}\right)}{e^{snt}} = e^{-nst}\, \left(\mathbb{E}\left(e^{sX_1}\right)\right)^n \\ &= e^{-nst}\, \exp\left(n \log \mathbb{E}\left(e^{sX_1}\right)\right) = \exp\left(-n\left(st - \log \mathbb{E}\left(e^{sX_1}\right)\right)\right) \end{aligned}$$

Optimizing this upper bound over $s \geq 0$, we obtain

$$\mathbb{P}\left(\left\{\frac{S_n}{n} \geq t\right\}\right) \leq \exp\left(-n\, \max_{s \geq 0} \left(st - \log \mathbb{E}\left(e^{sX_1}\right)\right)\right), \quad \forall t > \mathbb{E}(X_1)$$

Let us make a slightly technical observation at this point. First, the function $st - \log \mathbb{E}\left(e^{sX_1}\right)$ takes the value 0 in $s = 0$, so the above maximum is greater than or equal to 0. Second, for all $s < 0$, we obtain, using Jensen's inequality:

$$st - \log \mathbb{E}\left(e^{sX_1}\right) \leq s\left(t - \mathbb{E}(X_1)\right) < 0$$

as $s < 0$ and $t - \mathbb{E}(X_1) > 0$ by assumption. In the above inequality, we may therefore replace the maximum over $s \geq 0$ by the maximum over all $s \in \mathbb{R}$, leading to:

$$\mathbb{P}\left(\left\{\frac{S_n}{n} \geq t\right\}\right) \leq \exp\left(-n\, \max_{s \in \mathbb{R}} \left(st - \log \mathbb{E}\left(e^{sX_1}\right)\right)\right), \quad \forall t > \mathbb{E}(X_1)$$

Let us now define the function $\Lambda(s) = \log \mathbb{E}\left(e^{sX_1}\right)$, for $s \in \mathbb{R}$. This function might take the value $+\infty$ for some values of $s$ above $s_0$, but this is not a problem here.

Let us also define what is called the *Legendre transform* of $\Lambda$: $\Lambda^*(t) = \max_{s \in \mathbb{R}} \left(st - \Lambda(s)\right)$. It is a non-negative and convex function, which of course depends on the distribution of $X_1$. By the above inequality, we have:

$$\mathbb{P}\left(\left\{\frac{S_n}{n} \geq t\right\}\right) \leq \exp(-n\, \Lambda^*(t)), \quad \forall t > \mathbb{E}(X_1) \tag{11}$$

This is our first large deviations estimate, which is more precise than Hoeffding's inequality. This is normal, as we take into account here the specificity of the distribution; we are not after a universal upper bound. Note also that the only inequality in the above derivation comes from the use of Chebyshev's inequality at the beginning. All the rest are equalities. Moreover, we optimize our choice over a large set of functions ($\varphi(x) = e^{sx}$) while using Chebyshev's inequality, so this upper bound is hopefully tight.

[12] One can show that this condition is equivalent to saying that there exists $c > 0$ such that $\mathbb{P}(\{|X_1| \geq x\}) \overset{\sim}{\leq} \exp(-cx)$ as $x \to \infty$.

Likewise, for $t < \mathbb{E}(X_1)$, we obtain for $s \geq 0$:

$$\begin{aligned}\mathbb{P}\left(\left\{\frac{S_n}{n} \leq t\right\}\right) &= \mathbb{P}(\{S_n \leq nt\}) = \mathbb{P}(\{-S_n \geq -nt\}) \leq \frac{\mathbb{E}\left(e^{-sS_n}\right)}{e^{-snt}} = e^{nst}\,\left(\mathbb{E}\left(e^{-sX_1}\right)\right)^n \\ &= e^{nst}\,\exp\left(n\log\mathbb{E}\left(e^{-sX_1}\right)\right) = \exp\left(-n\left(-st - \log\mathbb{E}\left(e^{-sX_1}\right)\right)\right)\end{aligned}$$

Optimizing over $s \geq 0$, we further obtain

$$\mathbb{P}\left(\left\{\frac{S_n}{n} \leq t\right\}\right) \leq \exp\left(-n\max_{s\geq 0}\left(-st - \log\mathbb{E}\left(e^{-sX_1}\right)\right)\right), \quad \forall t < \mathbb{E}(X_1)$$

and for similar reasons as before, the maximum can be turned into a maximum over $\mathbb{R}$, so that

$$\begin{aligned}\mathbb{P}\left(\left\{\frac{S_n}{n} \leq t\right\}\right) &\leq \exp\left(-n\max_{s\in\mathbb{R}}\left(-st - \log\mathbb{E}\left(e^{-sX_1}\right)\right)\right) \\ &= \exp\left(-n\max_{s\in\mathbb{R}}\left(st - \log\mathbb{E}\left(e^{sX_1}\right)\right)\right) = \exp\left(-n\Lambda^*(t)\right), \quad \forall t < \mathbb{E}(X_1) \qquad (12)\end{aligned}$$

What do these two equations (11) and (12) actually mean?

One can check that $\Lambda^*(t) = 0$ if and only if $t = \mathbb{E}(X_1)$, so we see that in both cases ($t > \mathbb{E}(X_1)$ and $t < \mathbb{E}(X_1)$), the upper bound on the probability is decreasing exponentially in $n$, as it was the case with Hoeffding's inequality. What changes here is the multiplicative factor $\Lambda^*(t)$ which differs from (and is generally larger than) $t^2/2$, as we will see in the examples below.

**Generalization.** Before that, let us mention the generalization of the above result, also known as Cramér's theorem. Let $A$ be a "nice" subset of $\mathbb{R}$ (think e.g. of an interval). Then

$$\mathbb{P}\left(\left\{\frac{S_n}{n} \in A\right\}\right) \underset{n\to\infty}{\simeq} \exp\left(-n\inf_{t\in A}\Lambda^*(t)\right)$$

Note therefore that the above probability is decreasing exponentially in $n$ if and only if $\mathbb{E}(X_1) \notin A$.

In the particular cases where $A$ is either the interval $]-\infty, t[$ with $t < \mathbb{E}(X_1)$, or the interval $]t, +\infty[$ with $t > \mathbb{E}(X_1)$, one recovers the above equations (11) and (12). Indeed, one can check that the infimum of $\Lambda^*$ on $]-\infty, t[$ is achieved in $t$ when $t < \mathbb{E}(X_1)$, and likewise for the interval on the positive axis.

Finally, let us mention that Cramér's theorem not only provides an *upper bound* on the probability, but also a corresponding *lower bound* which is matching the upper bound in some asymptotic sense (see Appendix A.6 for a proof of this fact). This is therefore a quite remarkable and complete result.

**Examples.** - Let $X_1 \sim \mathcal{N}(0,1)$. Note that $X_1$ is an unbounded random variable, so Hoeffding's inequality does not apply here. Let us compute

$$\Lambda(s) = \log\mathbb{E}\left(e^{sX_1}\right) = \log\left(\int_{\mathbb{R}} e^{sx}\,\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx\right) = \log\left(e^{s^2/2}\right) = \frac{s^2}{2}$$

and

$$\Lambda^*(t) = \max_{s\in\mathbb{R}}\left(st - \frac{s^2}{2}\right) = \frac{t^2}{2}, \quad \text{attained in } s^* = t$$

Also, $\mathbb{E}(X_1) = 0$, so $\mathbb{P}\left(\left\{\frac{S_n}{n} \geq t\right\}\right) \leq \exp\left(-\frac{nt^2}{2}\right)$, for all $t > 0$. Surprisingly perhaps, this gives exactly the same upper bound as the one derived by Hoeffding (even though the random variables $X$'s are unbounded here).

- Let $X_1$ be such that $\mathbb{P}(\{X_1 = +1\}) = \mathbb{P}(\{X_1 = -1\}) = \frac{1}{2}$. In this case,

$$\Lambda(s) = \log\left(\frac{e^s + e^{-s}}{2}\right) = -s + \log(1 + e^{2s}) - \log(2)$$

and

$$\Lambda^*(t) = \max_{s\in\mathbb{R}}(st - \Lambda(s)) = \max_{s\in\mathbb{R}} \big(s\,(t+1) - \log(1+e^{2s}) + \log(2)\big)$$

Looking for the value where the maximum is attained, we obtain $s^* = \frac{1}{2}\log\left(\frac{1+t}{1-t}\right)$ (note that it does not make sense to consider values of $t$ such that $|t| > 1$, as it is always the case here that $|S_n/n| \le 1$). Correspondingly, after some computations:

$$\Lambda^*(t) = \frac{1+t}{2}\log\left(\frac{1+t}{2}\right) + \frac{1-t}{2}\log\left(\frac{1-t}{2}\right) + \log(2) \quad \text{for } |t| \le 1$$

Also $\mathbb{E}(X_1) = 0$, so $\mathbb{P}\left(\left\{\frac{S_n}{n} \ge t\right\}\right) \le \exp\left(-n\,\Lambda^*(t)\right)$ for all $t > 0$. Let us compare this result with Hoeffding's inequality, which reads in this case:

$$\mathbb{P}\left(\left\{\frac{S_n}{n} \ge t\right\}\right) \le \exp\left(-\frac{nt^2}{2}\right), \quad \forall t > 0$$

It can be observed that for $t > 0$, the above function $\Lambda^*(t)$ dominates the function $t^2/2$ obtained via Hoeffding's inequality (in particular, $\Lambda^*(\pm 1) = \log(2) > 1/2$, and $\Lambda^*(t) \simeq t^2$ around $t = 0$, which is greater than $t^2/2$).

# 12 Conditional expectation

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

## 12.1 Conditioning with respect to an event $B \in \mathcal{F}$

The conditional probability of an event $A \in \mathcal{F}$ given another event $B \in \mathcal{F}$ is defined as

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

Note that if $A$ and $B$ are independent, then $\mathbb{P}(A|B) = \mathbb{P}(A)$; the conditioning does not affect the probability. This fact remains true in more generality (see below).

In a similar manner, the conditional expectation of an integrable random variable $X$ given $B \in \mathcal{F}$ is defined as

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

## 12.2 Conditioning with respect to a discrete random variable $Y$

Let us assume that the random variable $Y$ (is $\mathcal{F}$-measurable and) takes values in a discrete set $D$.

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

If $X$ is also a discrete random variable with values in $D$, then

$$\mathbb{E}(X|Y) = \psi(Y), \text{ where } \psi(y) = \frac{\mathbb{E}(X\,1_{\{Y=y\}})}{\mathbb{P}(\{Y=y\})} = \sum_{x\in D} x\,\frac{\mathbb{E}(1_{\{X=x\}\cap\{Y=y\}})}{\mathbb{P}(\{Y=y\})} = \sum_{x\in D} x\,\mathbb{P}(\{X=x\}|\{Y=y\})$$

**Important remark.** $\varphi(y)$ and $\psi(y)$ are functions, while $\varphi(Y) = \mathbb{P}(A|Y)$ and $\psi(Y) = \mathbb{E}(X|Y)$ are *random variables*. They both are functions of the outcome of the random variable $Y$, that is, they are $\sigma(Y)$-measurable random variables.

**Example.** Let $X_1$, $X_2$ be two independent die rolls and let us compute $\mathbb{E}(X_1 + X_2|X_2) = \psi(X_2)$, where

$$\begin{aligned}
\psi(y) &= \mathbb{E}(X_1 + X_2|\{X_2 = y\}) = \frac{\mathbb{E}((X_1 + X_2)\, 1_{\{X_2=y\}})}{\mathbb{P}(\{X_2 = y\})} \\
&= \frac{\mathbb{E}(X_1\, 1_{\{X_2=y\}}) + \mathbb{E}(X_2\, 1_{\{X_2=y\}})}{\mathbb{P}(\{X_2 = y\})} \overset{(a)}{=} \frac{\mathbb{E}(X_1)\, \mathbb{E}(1_{\{X_2=y\}}) + \mathbb{E}(y\, 1_{\{X_2=y\}})}{\mathbb{P}(\{X_2 = y\})} \\
&= \frac{\mathbb{E}(X_1)\, \mathbb{P}(\{X_2 = y\}) + y\, \mathbb{P}(\{X_2 = y\})}{\mathbb{P}(\{X_2 = y\})} = \mathbb{E}(X_1) + y
\end{aligned}$$

where the independence assumption between $X_1$ and $X_2$ has been used in equality (a). So finally (as one would expect), $\mathbb{E}(X_1 + X_2|X_2) = \mathbb{E}(X_1) + X_2$, which can be explained intuitively as follows: the expectation of $X_1$ conditioned on $X_2$ is nothing but the expectation of $X_1$, as the outcome of $X_2$ provides no information on the outcome of $X_1$ ($X_1$ and $X_2$ being independent); on the other hand, the expectation of $X_2$ conditioned on $X_2$ is exactly $X_2$, as the outcome of $X_2$ is known.

## 12.3 Conditioning with respect to a continuous random variable $Y$?

In this case, one faces the following problem: if $Y$ is a continuous random variable, $\mathbb{P}(\{Y = y\}) = 0$ for all $y \in \mathbb{R}$. So a direct generalization of the above formulas to the continuous case is impossible at first sight. A possible solution to this problem is to replace the event $\{Y = y\}$ by $\{y \le Y < y + \varepsilon\}$ and to take the limit $\varepsilon \to 0$ for the definition of conditional expectation. This actually works, but also leads to a paradox in the multidimensional setting (known as Borel's paradox). In addition, some random variables are neither discrete, nor continuous. It turns out that the cleanest way to define conditional expectation in the general case is through $\sigma$-fields.

## 12.4 Conditioning with respect to a sub-$\sigma$-field $\mathcal{G}$

In order to define the conditional expectation in the general case, one needs the following proposition[13].

**Proposition 12.1.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}$ be a sub-$\sigma$-field of $\mathcal{F}$ and $X$ be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. There exists then an integrable random variable $Z$ such that

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

(ii) $\mathbb{E}(ZU) = \mathbb{E}(XU)$ for any random variable $U$ $\mathcal{G}$-measurable and bounded.

Moreover, if $Z_1$, $Z_2$ are two integrable random variables satisfying (i) and (ii), then $Z_1 = Z_2$ a.s.

**Definition 12.2.** The above random variable $Z$ is called the *conditional expectation of $X$ given $\mathcal{G}$* and is denoted as $\mathbb{E}(X|\mathcal{G})$. Because of the last part of the above proposition, it is defined up to a negligible set.

**Definition 12.3.** One further *defines* $\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(1_A|\mathcal{G})$ for $A \in \mathcal{F}$.

**Remark.** Note that as before, both $\mathbb{P}(A|\mathcal{G})$ and $\mathbb{E}(X|\mathcal{G})$ are ($\mathcal{G}$-measurable) random variables.

**Properties.** The above definition does not give a computation rule for the conditional expectation; it is only an existence theorem. The properties listed below will therefore be of help for computing conditional expectations. The proofs of the first two are omitted, while the next five are left as (important!) exercises.

- Linearity. $\mathbb{E}(c\, X + Y|\mathcal{G}) = c\, \mathbb{E}(X|\mathcal{G}) + \mathbb{E}(Y|\mathcal{G})$ a.s.

- Monotonicity. If $X \ge Y$ a.s., then $\mathbb{E}(X|\mathcal{G}) \ge \mathbb{E}(Y|\mathcal{G})$ a.s. (so if $X \ge 0$ a.s., then $\mathbb{E}(X|\mathcal{G}) \ge 0$ a.s.)

[13] We do not prove here this proposition: let us just mention that it is a consequence of the Radon-Nikodym theorem.

- $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.

- If $X$ is independent of $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ a.s.

- If $X$ is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a.s.

- If $Y$ is $\mathcal{G}$-measurable and bounded (or if $Y$ is $\mathcal{G}$-measurable and both $X$ and $Y$ are square-integrable; what actually matters here is that the random variable $XY$ is integrable), then $\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})\, Y$ a.s.

- If $\mathcal{H}$ is a sub-$\sigma$-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. (in other words, the smallest $\sigma$-field always "wins": this property is also known as the"towering property" of conditional expectation)

Some of the above properties are illustrated below with an example.

**Example.** Let $\Omega = \{1, \ldots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$ for $\omega = 1, \ldots, 6$ (the probability space of the die roll). Let also $X(\omega) = \omega$ be the outcome of the die roll and consider the two sub-$\sigma$-fields:

$$\mathcal{G} = \sigma(\{1,3\}, \{2\}, \{5\}, \{4,6\}) \quad \text{and} \quad \mathcal{H} = \sigma(\{1,3,5\}, \{2,4,6\})$$

Then $\mathbb{E}(X) = 3.5$,

$$\mathbb{E}(X|\mathcal{G})(\omega) = \begin{cases} 2 & \text{if } \omega \in \{1,3\} \text{ or } \omega = 2 \\ 5 & \text{if } \omega \in \{4,6\} \text{ or } \omega = 5 \end{cases} \quad \text{and} \quad \mathbb{E}(X|\mathcal{H})(\omega) = \begin{cases} 3 & \text{if } \omega \in \{1,3,5\} \\ 4 & \text{if } \omega \in \{2,4,6\} \end{cases}$$

So $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})) = \mathbb{E}(X)$. Moreover,

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})(\omega) = \begin{cases} \frac{1}{3}(2+2+5) = 3 & \text{if } \omega \in \{1,3,5\} \\ \frac{1}{3}(2+5+5) = 4 & \text{if } \omega \in \{2,4,6\} \end{cases} = \mathbb{E}(X|\mathcal{H})(\omega)$$

and

$$\mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G})(\omega) = \begin{cases} 3 & \text{if } \omega \in \{1,3\} \text{ or } \omega = 5 \\ 4 & \text{if } \omega \in \{4,6\} \text{ or } \omega = 2 \end{cases} = \mathbb{E}(X|\mathcal{H})(\omega)$$

The proposition below (given here without proof) is an extension of some of the above properties.

**Proposition 12.4.** Let $\mathcal{G}$ be a sub-$\sigma$-field of $\mathcal{F}$, $X$, $Y$ be two random variables such that $X$ is independent of $\mathcal{G}$ and $Y$ is $\mathcal{G}$-measurable, and let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ be a Borel-measurable function such that $\mathbb{E}(|\varphi(X,Y)|) < +\infty$. Then

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

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

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

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

Finally, the proposition below (given again without proof) shows that Jensen's inequality also holds for conditional expectation.

**Proposition 12.5.** Let $X$ be a random variable, $\mathcal{G}$ be a sub-$\sigma$-field of $\mathcal{F}$ and $\psi : \mathbb{R} \to \mathbb{R}$ be Borel-measurable, convex and such that $\mathbb{E}(|\psi(X)|) < +\infty$. Then

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

In particular, $|\mathbb{E}(X|\mathcal{G})| \le \mathbb{E}(|X||\mathcal{G})$ a.s.

## 12.5 Conditioning with respect to a random variable $Y$

Once the definition of conditional expectation with respect to a $\sigma$-field is set, it is natural to define it for a generic random variable $Y$:

$$\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y)) \quad \text{and} \quad \mathbb{P}(A|Y) = \mathbb{P}(A|\sigma(Y))$$

**Remark.** Since any $\sigma(Y)$-measurable random variable may be written as $g(Y)$, where $g$ is a Borel-measurable function, the definition of $\mathbb{E}(X|Y)$ may be rephrased as follows.

**Definition 12.6.** *$E(X|Y) = \psi(Y)$, where $\psi : \mathbb{R} \to \mathbb{R}$ is the unique Borel-measurable function such that $\mathbb{E}(\psi(Y)\, g(Y)) = \mathbb{E}(Xg(Y))$ for any function $g : \mathbb{R} \to \mathbb{R}$ Borel-measurable and bounded.*

In two particular cases, the function $\psi$ can be made explicit, which allows for concrete computations.

- If $X$, $Y$ are two discrete random variables with values in a set $D$, then

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

which matches the formula given in Section 12.2. The proof that it also matches the theoretical definition of conditional expectation is left as an exercise.

- If $X, Y$ are two jointly continuous random variables with joint pdf $p_{X,Y}$, then

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

and $p_Y$ is the marginal pdf of $Y$ given by $p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x,y)\, dx$, assumed here to be strictly positive (but this assumption is not needed, actually). Let us check that the random variable $\psi(Y)$ is indeed the conditional expectation of $X$ given $Y$ according to Definition 12.6: for any function $g : \mathbb{R} \to \mathbb{R}$ Borel-measurable and bounded, one has

$$\begin{aligned}
\mathbb{E}(\psi(Y)\, g(Y)) &= \int_{\mathbb{R}} \psi(y)\, g(y)\, p_Y(y)\, dy \\
&= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} x\, \frac{p_{X,Y}(x,y)}{p_Y(y)}\, dx \right) g(y)\, p_Y(y)\, dy \\
&= \iint_{\mathbb{R}^2} x\, g(y)\, p_{X,Y}(x,y)\, dx\, dy = \mathbb{E}(Xg(Y))
\end{aligned}$$

□

**Remark.** The function $\psi(y) = \int_{\mathbb{R}} x\, \frac{p_{X,Y}(x,y)}{p_Y(y)}\, dx$ is sometimes dangerously denoted as "$\mathbb{E}(X|\{Y = y\})$".

## 12.6 Geometric interpretation

Finally, let us go back to the general case and mention a geometric interpretation of conditional expectation when $X$ is a square-integrable random variable. Observe first that in this case, according to definition 12.2, $Z = \mathbb{E}(X|\mathcal{G})$ is the unique[14] square-integrable random variable such that

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

(ii) $\mathbb{E}(ZU) = \mathbb{E}(XU)$ for any random variable $U$ $\mathcal{G}$-measurable and square-integrable.

Indeed, as $X$ is square-integrable, $Z = \mathbb{E}(X|\mathcal{G})$ also is (use Jensen's inequality), so the above equality may be extended from bounded $U$'s to square-integrable ones (ensuring that both $XU$ and $ZU$ are integrable).

[14] up to set of probability 0

Denote now by $L^2(\Omega)$ the Hilbert space[15] of square-integrable random variables equipped with the scalar product:

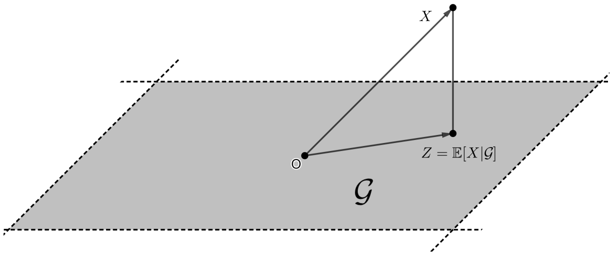
$$\langle X, Y\rangle_2 = \mathbb{E}(XY) \quad \text{and the corresponding norm} \quad \|X\|_2 = \sqrt{\mathbb{E}(X^2)}$$

and denote by $G$ the linear subspace of $L^2(\Omega)$ comprising $\mathcal{G}$-measurable and square-integrable random variables. With these two definitions, the above definition can be rephrased as: $Z = \mathbb{E}(X|\mathcal{G})$ is the unique[11] random variable such that

(i) $Z \in G$;

(ii) $\mathbb{E}((Z - X)\, U) = 0$ for every $U \in G$ (i.e., $Z - X \perp G$).

In geometric terms, this is nothing but saying that $Z$ is the orthogonal projection of $X$ onto the linear subspace $G$, as illustrated on the following page.



We provide below an alternate characterization of this orthogonal projection, as well as the proof of the equivalence between the two characterizations, for the sake of completeness.

**Proposition 12.7.** Let $X$ be a square-integrable random variable, $\mathcal{G}$ be a sub-$\sigma$-field of $\mathcal{F}$ and $G$ be the linear subspace of square-integrable and $\mathcal{G}$-measurable random variables. Then $Z \in G$ satisfies

$$\mathbb{E}((Z - X)\, U) = 0 \quad \forall U \in G \tag{13}$$

if and only if

$$\mathbb{E}((Z - X)^2) \leq \mathbb{E}((Z' - X)^2) \quad \forall Z' \in G \tag{14}$$

*Proof.* - Assume first that $Z \in G$ satisfies (13). Then for every $Z' \in G$, we have

$$\begin{aligned}\mathbb{E}((Z' - X)^2) &= \mathbb{E}((Z' - Z + Z - X)^2) = \mathbb{E}((Z' - Z)^2) + 2\mathbb{E}((Z' - Z)(Z - X)) + \mathbb{E}((Z - X)^2)\\ &= \mathbb{E}((Z' - Z)^2) + \mathbb{E}((Z - X)^2)\end{aligned}$$

as by assumption, $\mathbb{E}((Z' - Z)(Z - X)) = 0$ since $Z' - Z \in G$. From this, we deduce that

$$\mathbb{E}((Z' - X)^2) \geq \mathbb{E}((Z - X)^2) \quad \forall Z' \in G$$

implying (14).

- Assume now that $Z \in G$ satisfies (14). Fix $U \in G$ and for $\alpha \in \mathbb{R}$, define

$$F(\alpha) = \mathbb{E}((Z - X + \alpha U)^2) = \mathbb{E}((Z - X)^2) + 2\alpha\, \mathbb{E}((Z - X)\, U) + \alpha^2\, \mathbb{E}(U^2)$$

By the assumption made, we know that $F$ has a global minimum in $\alpha = 0$. So $0 = F'(0) = 2\,\mathbb{E}((Z-X)\, U)$, implying (13), as $U \in G$ is arbitrary. □

[15] Strictly speaking, we should consider here $L^2(\Omega)$ to be the space of equivalence classes of square-integrable random variables, two random variables $X, Y$ being equivalent if $X = Y$ a.s.

# 13 Martingales

## 13.1 Basic definitions

**Definition 13.1.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *filtration* is a sequence $(\mathcal{F}_n,\, n \in \mathbb{N})$ of sub-$\sigma$-fields of $\mathcal{F}$ such that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, $\forall n \in \mathbb{N}$.

**Example.** Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$, $X_n(\omega) = n^{th}$ decimal of $\omega$, for $n \geq 1$. Let also $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. Then $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, $\forall n \in \mathbb{N}$.

**Definitions 13.2.** - A discrete-time process $(X_n,\, n \in \mathbb{N})$ is said to be *adapted* to the filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$ if $X_n$ is $\mathcal{F}_n$-measurable $\forall n \in \mathbb{N}$.

- The *natural filtration* of a process $(X_n,\, n \in \mathbb{N})$ is defined as $\mathcal{F}_n^X = \sigma(X_0, \dots, X_n)$, $n \in \mathbb{N}$. It represents the available amount of information about the process at time $n$.

**Remark.** A process is adapted to its natural filtration, by definition.

Let now $(\mathcal{F}_n,\, n \in \mathbb{N})$ be a given filtration.

**Definition 13.3.** A discrete-time process $(M_n,\, n \in \mathbb{N})$ is a *martingale* with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ if

(i) $\mathbb{E}(|M_n|) < +\infty$, $\forall n \in \mathbb{N}$.

(ii) $M_n$ is $\mathcal{F}_n$-measurable, $\forall n \in \mathbb{N}$ (i.e., $(M_n,\, n \in \mathbb{N})$ is adapted to $(\mathcal{F}_n,\, n \in \mathbb{N})$).

(iii) $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n$ a.s., $\forall n \in \mathbb{N}$.

A martingale is therefore a fair game: the expectation of the process at time $n+1$ given the information at time $n$ is equal to the value of the process at time $n$.

**Remark.** Conditions (ii) and (iii) are actually redundant, as (iii) implies (ii).

**Properties.** If $(M_n,\, n \in \mathbb{N})$ is a martingale, then

- $\mathbb{E}(M_{n+1}) = \mathbb{E}(M_n)\, (= \ldots = \mathbb{E}(M_0))$, $\forall n \in \mathbb{N}$ (by the first property of conditional expectation).

- $\mathbb{E}(M_{n+1} - M_n|\mathcal{F}_n) = 0$ a.s. (nearly by definition).

- $\mathbb{E}(M_{n+m}|\mathcal{F}_n) = M_n$ a.s., $\forall n, m \in \mathbb{N}$.

This last property is important, as it says that the martingale property propagates over time. Here is a short proof, which uses the towering property of conditional expectation:

$$\mathbb{E}(M_{n+m}|\mathcal{F}_n) = \mathbb{E}(\mathbb{E}(M_{n+m}|\mathcal{F}_{n+m-1})|\mathcal{F}_n) = \mathbb{E}(M_{n+m-1}|\mathcal{F}_n) = \ldots = \mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n \quad \text{a.s.}$$

**Example: the simple symmetric random walk.**

Let $(S_n,\, n \in \mathbb{N})$ be the simple symmetric random walk : $S_0 = 0$, $S_n = X_1 + \ldots + X_n$, where the $X_n$ are i.i.d. and $\mathbb{P}(\{X_1 = +1\}) = \mathbb{P}(\{X_1 = -1\}) = 1/2$.

Let us define the following filtration: $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, $n \geq 1$. Then $(S_n,\, n \in \mathbb{N})$ is a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$. Indeed:

(i) $\mathbb{E}(|S_n|) \leq \mathbb{E}(|X_1|) + \ldots + \mathbb{E}(|X_n|) = 1 + \ldots + 1 = n < +\infty$, $\forall n \in \mathbb{N}$.

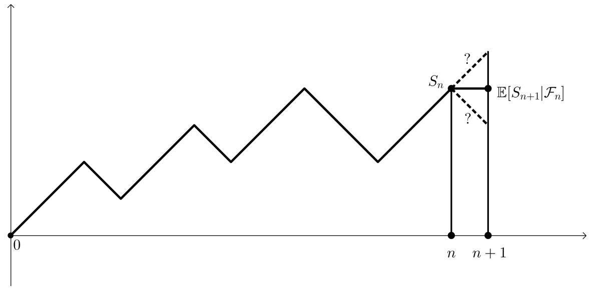
(ii) $S_n = X_1 + \ldots + X_n$ is a function of $(X_1, \ldots, X_n)$, i.e., is $\sigma(X_1, \ldots, X_n) = \mathcal{F}_n$-measurable.

(iii) We have

$$\begin{aligned} \mathbb{E}(S_{n+1}|\mathcal{F}_n) &= \mathbb{E}(S_n + X_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n|\mathcal{F}_n) + \mathbb{E}(X_{n+1}|\mathcal{F}_n) \\ &= S_n + \mathbb{E}(X_{n+1}) = S_n + 0 = S_n \quad \text{a.s.} \end{aligned}$$

The first equality on the second line follows from the fact that $S_n$ is $\mathcal{F}_n$-measurable and that $X_{n+1}$ is independent of $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. □

Here is an additional illustration of the martingale property of the simple symmetric random walk:



**Remark.** Even though one uses generally the same letter "M" for both martingales and Markov process, these are a priori completely different processes! A possible way to state the Markov property is to say that

$$\mathbb{E}(g(M_{n+1})|\mathcal{F}_n) = \mathbb{E}(g(M_{n+1})|X_n) \quad \text{a.s.} \quad \text{for any } g : \mathbb{R} \to \mathbb{R} \text{ continuous and bounded}$$

which is clearly different from the above stated martingale property. Beyond the use of the same letter "M", the confusion between the two notions comes also from the fact that the simple symmetric random walk is usually taken a paradigm example for *both* martingales and Markov processes.

**Generalization.** If the random variables $X_n$ are i.i.d. and such that $\mathbb{E}(|X_1|) < +\infty$ and $\mathbb{E}(X_1) = 0$, then $(S_n,\, n \in \mathbb{N})$ is also a martingale (in particular, $X_1 \sim \mathcal{N}(0,1)$ works).

**Definition 13.4.** Let $(\mathcal{F}_n,\, n \in \mathbb{N})$ be a filtration. A process $(M_n,\, n \in \mathbb{N})$ is a *submartingale* (resp. a *supermartingale*) with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ if

(i) $\mathbb{E}(|M_n|) < +\infty$, $\forall n \in \mathbb{N}$.

(ii) $M_n$ is $\mathcal{F}_n$-measurable, $\forall n \in \mathbb{N}$.

(iii) $\mathbb{E}(M_{n+1}|\mathcal{F}_n) \geq M_n$ a.s., $\forall n \in \mathbb{N}$ (resp. $\mathbb{E}(M_{n+1}|\mathcal{F}_n) \leq M_n$ a.s., $\forall n \in \mathbb{N}$).

**Remarks.** - Not every process is either a sub- or a supermartingale!

- The appellations sub- and supermartingale are counter-intuitive. They are due to historical reasons.

- Condition (ii) is now necessary in itself, as (iii) does not imply it.

- If $(M_n,\, n \in \mathbb{N})$ is both a submartingale and a supermartingale, then it is a martingale.

**Example: the simple asymmetric random walk.**

- If $\mathbb{P}(\{X_1 = +1\}) = p = 1 - \mathbb{P}(\{X_1 = -1\})$ with $p \geq 1/2$, then $S_n = X_1 + \ldots + X_n$ is a submartingale.

- More generally, $S_n = X_1 + \ldots + X_n$ is a submartingale if $\mathbb{E}(X_1) \geq 0$.

**Proposition 13.5.** If $(M_n,\, n \in \mathbb{N})$ is a martingale with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable and convex function such that $\mathbb{E}(|\varphi(M_n)|) < +\infty$, $\forall n \in \mathbb{N}$, then $(\varphi(M_n),\, n \in \mathbb{N})$ is a submartingale.

*Proof.* (i) $\mathbb{E}(|\varphi(M_n)|) < +\infty$ by assumption.

(ii) $\varphi(M_n)$ is $\mathcal{F}_n$-measurable as $M_n$ is (and $\varphi$ is Borel-measurable).

(iii) $\mathbb{E}(\varphi(M_{n+1})|\mathcal{F}_n) \geq \varphi(\mathbb{E}(M_{n+1}|\mathcal{F}_n)) = \varphi(M_n)$ a.s.

In (iii), the first inequality follows from Jensen's inequality and the second follows from the fact that $M$ is a martingale. □

**Example.** If $(M_n,\, n \in \mathbb{N})$ is a square-integrable martingale (i.e., $\mathbb{E}(M_n^2) < +\infty$, $\forall n \in \mathbb{N}$), then the process $(M_n^2,\, n \in \mathbb{N})$ is a submartingale (as $x \mapsto x^2$ is convex).

## 13.2 Stopping times

**Definitions 13.6.** - A *random time* is a random variable $T$ with values in $\mathbb{N} \cup \{+\infty\}$. It is said to be *finite* if $T(\omega) < +\infty$ for every $\omega \in \Omega$ and *bounded* if there exists moreover an integer $N$ such that $T(\omega) \leq N$ for every $\omega \in \Omega$ (note that a finite random time is not necessarily bounded).

- Let $(X_n,\, n \in \mathbb{N})$ be a stochastic process and assume $T$ is finite. One then defines $X_T(\omega) = X_{T(\omega)}(\omega) = \sum_{n\in\mathbb{N}} X_n(\omega)\, 1_{\{T=n\}}(\omega)$.

- A *stopping time* with respect to a filtration $(\mathcal{F}_n, n \in \mathbb{N})$ is a random time $T$ such that $\{T = n\} \in \mathcal{F}_n$, $\forall n \in \mathbb{N}$.

**Example.** Let $(X_n,\, n \in \mathbb{N})$ be a process adapted to $(\mathcal{F}_n,\, n \in \mathbb{N})$ and $a > 0$. Then $T_a = \inf\{n \in \mathbb{N} : |X_n| \geq a\}$ is a stopping time with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$. Indeed:

$$\begin{aligned} \{T_a = n\} &= \{|X_k| < a,\ \forall 0 \leq k \leq n-1 \text{ and } |X_n| \geq a\} \\ &= \bigcap_{k=0}^{n-1} \underbrace{\{|X_k| < a\}}_{\in\mathcal{F}_k\subset\mathcal{F}_n} \cap \{|X_n| \geq a\} \in \mathcal{F}_n, \quad \forall n \in \mathbb{N} \end{aligned}$$

**Definition 13.7.** Let $T$ be a stopping time with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$. One defines the information one possesses at time $T$ as the following $\sigma$-field:

$$\mathcal{F}_T = \{A \in \mathcal{F} \,:\, A \cap \{T = n\} \in \mathcal{F}_n,\, \forall n \in \mathbb{N}\}$$

**Facts.**

- If $T(\omega) = N\ \forall \omega \in \Omega$, then $\mathcal{F}_T = \mathcal{F}_N$. This is obvious from the definition.

- If $T_1$, $T_2$ are stopping times such that $T_1(\omega) \leq T_2(\omega)\ \forall \omega \in \Omega$, then $\mathcal{F}_{T_1} \subset \mathcal{F}_{T_2}$. Indeed, if $T_1(\omega) \leq T_2(\omega)\ \forall \omega \in \Omega$ and $A \in \mathcal{F}_{T_1}$, then for all $n \in \mathbb{N}$, we have:

$$A \cap \{T_2 = n\} = A \cap (\cup_{k=1}^n \{T_1 = k\}) \cap \{T_2 = n\} = \Big( \cup_{k=1}^n \underbrace{A \cap \{T_1 = k\}}_{\in\mathcal{F}_k\subset\mathcal{F}_n} \Big) \cap \{T_2 = n\} \in \mathcal{F}_n$$

so $A \in \mathcal{F}_{T_2}$. By the way, here is an example of stopping times $T_1$, $T_2$ such that $T_1(\omega) \leq T_2(\omega)\ \forall \omega \in \Omega$: let $0 < a < b$ and consider $T_1 = \inf\{n \in \mathbb{N} : |X_n| \geq a\}$ and $T_2 = \inf\{n \in \mathbb{N} : |X_n| \geq b\}$.

- A random variable $Y$ is $\mathcal{F}_T$-measurable if and only if $Y\, 1_{\{T=n\}}$ is $\mathcal{F}_n$-measurable, $\forall n \in \mathbb{N}$. As a consequence: if $(X_n,\, n \in \mathbb{N})$ is adapted to $(\mathcal{F}_n,\, n \in \mathbb{N})$, then $X_T$ is $\mathcal{F}_T$-measurable.

## 13.3 Doob's optional stopping theorem, version 1

Let $(M_n,\, n \in \mathbb{N})$ be a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$, $N \in NN$ be fixed and $T_1$, $T_2$ be two stopping times such that $0 \leq T_1(\omega) \leq T_2(\omega) \leq N < +\infty$, $\forall \omega \in \Omega$. Then

$$\mathbb{E}(M_{T_2}|\mathcal{F}_{T_1}) = M_{T_1} \text{ a.s.}$$

In particular, $\mathbb{E}(M_{T_2}) = \mathbb{E}(M_{T_1})$.

In particular, if $T$ is a stopping time such that $0 \leq T(\omega) \leq N < +\infty$, $\forall \omega \in \Omega$, then $\mathbb{E}(M_T) = \mathbb{E}(M_0)$.

**Remarks.** - The above theorem says that the martingale property holds even if one is given the option to stop at any (bounded) stopping time.

- The theorem also holds for sub- and supermartingales (i.e., if $M$ is a submartingale, then $\mathbb{E}(M_{T_2}|\mathcal{F}_{T_1}) \geq M_{T_1}$ a.s.).

*Proof.* - We first show that if $T$ is a stopping time such that $0 \le T(\omega) \le N$, $\forall \omega \in \Omega$, then

$$\mathbb{E}(M_N|\mathcal{F}_T) = M_T \tag{15}$$

Indeed, let $Z = M_T = \sum_{n=0}^N M_n \, 1_{\{T=n\}}$. We check below that $Z$ is the conditional expectation of $M_N$ given $\mathcal{F}_T$:

(i) $Z$ is $\mathcal{F}_T$-measurable: $Z \, 1_{\{T=n\}} = M_n \, 1_{\{T=n\}}$ is $\mathcal{F}_n$-measurable $\forall n$, so $Z$ is $\mathcal{F}_T$-measurable.

(ii) $\mathbb{E}(ZU) = \mathbb{E}(M_N U)$, $\forall U$ $\mathcal{F}_T$-measurable and bounded:

$$\mathbb{E}(ZU) = \sum_{n=0}^N \mathbb{E}(M_n 1_{\{T=n\}} U) = \sum_{n=0}^N \mathbb{E}(\mathbb{E}(M_N|\mathcal{F}_n) \underbrace{1_{\{T=n\}} U}_{\mathcal{F}_n-\text{measurable}}) = \sum_{n=0}^N \mathbb{E}(M_N 1_{\{T=n\}} U) = \mathbb{E}(M_N U)$$

- Second, let us check that $\mathbb{E}(M_{T_2}|\mathcal{F}_{T_1}) = M_{T_1}$:

$$M_{T_1} \underset{(15) \text{ with } T=T_1}{=} \mathbb{E}(M_N|\mathcal{F}_{T_1}) \underset{\mathcal{F}_{T_1} \subset \mathcal{F}_{T_2}}{=} \mathbb{E}(\mathbb{E}(M_N|\mathcal{F}_{T_2})|\mathcal{F}_{T_1}) \underset{(15) \text{ with } T=T_2}{=} \mathbb{E}(M_{T_2}|\mathcal{F}_{T_1})$$

This concludes the proof of the theorem. □

## 13.4 The reflection principle

Let $(S_n, \, n \in \mathbb{N})$ be the simple symmetric random walk and

$$T = \inf\{n \ge 1 \; : \; S_n = +1 \quad \text{or} \quad n = N\}$$

As $S$ is a martingale and $T$ is a bounded stopping time (indeed, $T(\omega) \le N$ for every $\omega \in \Omega$), the optional stopping theorem applies here, so it holds that $\mathbb{E}(S_T) = \mathbb{E}(S_0) = 0$. But what is the distribution of the random variable $S_T$? Intuitively, for $N$ large, $S_T$ will be $+1$ with high probability, but in case it does not reach this value, what is the average loss we should expect? More precisely, we are asking here for the value of

$$\mathbb{E}\left( S_T \,\middle|\, \left\{ \max_{0 \le n \le N} S_n \le 0 \right\} \right) = \mathbb{E}\left( S_N \,\middle|\, \left\{ \max_{0 \le n \le N} S_n \le 0 \right\} \right) = \frac{\mathbb{E}\left( S_N \, 1_{\{\max_{0 \le n \le N} S_n \le 0\}} \right)}{\mathbb{P}\left( \{\max_{0 \le n \le N} S_n \le 0\} \right)} \tag{16}$$

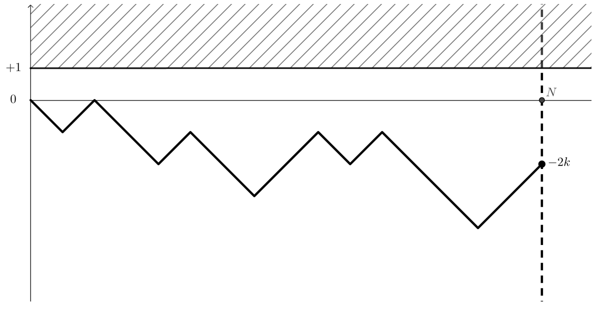
Let us first compute the denominator in (16), assuming that $N$ is even to simplify notations:

$$\mathbb{P}\left( \left\{ \max_{0 \le n \le N} S_n \le 0 \right\} \right) = \mathbb{P}(\{S_n \le 0, \forall 0 \le n \le N\}) = \sum_{k \ge 0} \mathbb{P}(\{S_N = -2k, S_n \le 0, \forall 0 \le n \le N-1\})$$

noticing that $S_N$ can only take even values (because $N$ itself is even) and that we are asking here that $S_N \le 0$. Let us now consider a *fixed* value of $k \ge 0$. In order to compute the probability

$$\mathbb{P}(\{S_N = -2k, S_n \le 0, \forall 0 \le n \le N-1\})$$

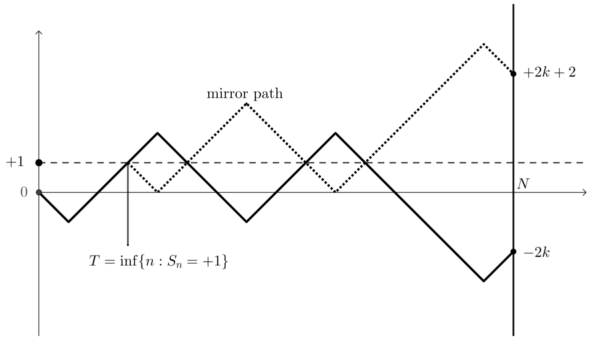
we should enumerate all paths of the following form:



but this is rather complicated combinatorics. In order to avoid such a computation, first observe that

$$\mathbb{P}(\{S_N = -2k, S_n \leq 0, \forall 0 \leq n \leq N-1, \}) = \mathbb{P}(\{S_N = -2k\}) - \mathbb{P}(\{S_N = -2k, \, \exists 1 \leq n \leq N_1 \text{ with } S_n = +1\})$$

A second important observation, which is at the heart of the reflection principle, is that to each path going from 0 (at time 0) to $-2k$ (at time $N$) "via" $+1$ corresponds a mirror path that goes from 0 to $2k+2$, also "via" $+1$, as illustrated below:



so that in total:

$$\mathbb{P}(\{S_N = -2k, \, \exists 1 \leq n \leq N_1 \text{ with } S_n = +1\}) = \mathbb{P}(\{S_N = 2k+2, \, \exists 1 \leq n \leq N_1 \text{ with } S_n = +1\})$$

A third observation is that for any $k \geq 0$, there is no way to go from 0 to $2k+2$ without crossing the $+1$ line, so that

$$\mathbb{P}(\{S_N = 2k+2, \, \exists 1 \leq n \leq N_1 \text{ with } S_n = +1\}) = \mathbb{P}(\{S_N = 2k+2\})$$

Finally, we obtain

$$\mathbb{P}\left(\left\{\max_{0 \leq n \leq N} S_n \leq 0\right\}\right) = \sum_{k \geq 0} (\mathbb{P}(\{S_N = -2k\}) - \mathbb{P}(\{S_N = 2k+2\})) = \sum_{k \geq 0} (\mathbb{P}(\{S_N = 2k\}) - \mathbb{P}(\{S_N = 2k+2\}))$$

by symmetry. But this is a telescopic sum, and we know that for finite $N$, it ends before $k = +\infty$. At the end, we therefore obtain:

$$\mathbb{P}\left(\left\{\max_{0 \leq n \leq N} S_n \leq 0\right\}\right) = \mathbb{P}(\{S_N = 0\})$$

which can be computed via simple combinatorics (writing here $N = 2M$):

$$\mathbb{P}(\{S_{2M} = 0\}) = \frac{1}{2^{2M}} \binom{2M}{M} = \frac{1}{2^{2M}} \frac{(2M)!}{(M!)^2}$$

which gives for large $M$, using Stirling's formula: $M! \simeq M^M e^{-M} \sqrt{2\pi M}$:

$$\mathbb{P}(\{S_{2M} = 0\}) \simeq \frac{1}{2^{2M}} \frac{(2M)^{2M} \, e^{-2M} \, \sqrt{4\pi M}}{(M \, e^{-M} \sqrt{2\pi M})^2} = \frac{1}{\sqrt{\pi M}}$$

This leads to the approximation for large $N$:

$$\mathbb{P}\left(\left\{\max_{0 \leq n \leq N} S_n \leq 0\right\}\right) \simeq \sqrt{\frac{2}{\pi N}}$$

Finally, the optional stopping theorem spares us the direct computation of the numerator in (16), since

$$0 = \mathbb{E}(S_T) = 1 \cdot \mathbb{P}\left(\left\{\max_{0 \leq n \leq N} S_n \geq +1\right\}\right) + \mathbb{E}\left(S_N \, 1_{\{\max_{0 \leq n \leq N} S_n \leq 0\}}\right)$$

so

$$\mathbb{E}\left(S_N\, 1_{\{\max_{0\le n\le N} S_n \le 0\}}\right) = -1 + \mathbb{P}\left(\left\{\max_{0\le n\le N} S_n \le 0\right\}\right) \simeq -1 + \sqrt{\frac{2}{\pi N}}$$

for large $N$, and finally

$$\mathbb{E}\left(S_T \,\middle|\, \left\{\max_{0\le n\le N} S_n \le 0\right\}\right) \simeq \frac{-1+\sqrt{\frac{2}{\pi N}}}{\sqrt{\frac{2}{\pi N}}} = 1 - \sqrt{\frac{\pi N}{2}}$$

for large $N$. In conclusion, in case $S$ does not reach the value $+1$ during the time interval $\{0, \ldots N\}$, we should expect a loss of order $-\sqrt{N}$.

## 13.5 Martingale transforms

**Definition 13.8.** A process $(H_n,\, n \in \mathbb{N})$ is said to be *predictable* with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$ if $H_0 = 0$ and $H_n$ is $\mathcal{F}_{n-1}$-measurable $\forall n \ge 1$.

**Remark.** If a process is predictable, then it is adapted.

Let now $(\mathcal{F}_n,\, n \in \mathbb{N})$ be a filtration, $(H_n,\, n \in \mathbb{N})$ be a predictable process with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ and $(M_n,\, n \in \mathbb{N})$ be a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$.

**Definition 13.9.** The process $G$ defined as

$$G_0 = 0, \quad G_n = (H \cdot M)_n = \sum_{i=1}^n H_i (M_i - M_{i-1}), \quad n \ge 1$$

is called the *martingale transform* of $M$ through $H$.

**Remark.** This process is the discrete version of the stochastic integral. It represents the gain obtained by applying the strategy $H$ to the game $M$:

- $H_i$ = amount bet on day $i$ ($\mathcal{F}_{i-1}$-measurable).

- $M_i - M_{i-1}$ = increment of the process $M$ on day $i$.

- $G_n$ = gain on day $n$.

**Proposition 13.10.** If $H_n$ is a bounded random variable for each $n$ (i.e., $|H_n(\omega)| \le K_n \ \forall \omega \in \Omega$), then the process $G$ is a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$.

In other words, one cannot win on a martingale!

*Proof.* (i) $\mathbb{E}(|G_n|) \le \sum_{i=1}^n \mathbb{E}(|H_i|\,|M_i - M_{i-1}|) \le \sum_{i=1}^n K_i \left(\mathbb{E}(|M_i|) + \mathbb{E}(|M_{i-1}|)\right) < +\infty$.

(ii) $G_n$ is $\mathcal{F}_n$-measurable by construction.

(iii) $\mathbb{E}(G_{n+1}|\mathcal{F}_n) = \mathbb{E}(G_n + H_{n+1}\,(M_{n+1} - M_n)|\mathcal{F}_n) = G_n + H_{n+1}\,\mathbb{E}(M_{n+1} - M_n|\mathcal{F}_n) = G_n + 0 = G_n$. □

**Example: "the" martingale.**

Let $(M_n,\, n \in \mathbb{N})$ be the simple symmetric random walk ($M_n = X_1 + \ldots + X_n$) and consider the following strategy:

$$H_0 = 0,\ H_1 = 1,\ H_{n+1} = \begin{cases} 2H_n, & \text{if } \xi_1 = \ldots = \xi_n = -1 \\ 0, & \text{otherwise} \end{cases}$$

Note that all the $H_n$ are bounded random variables. Then by the above proposition, the process $G$ defined as

$$G_0 = 0, \quad G_n = \sum_{i=1}^n H_i\,(M_i - M_{i-1}) = \sum_{i=1}^n H_i\,X_i, \quad n \ge 1$$

is a martingale. So $\mathbb{E}(G_n) = \mathbb{E}(G_0) = 0$, $\forall n \in \mathbb{N}$. Let now

$$T = \inf\{n \ge 1 \,:\, X_n = +1\}$$

$T$ is a stopping time and it is easily seen that $G_T = +1$. But then $\mathbb{E}(G_T) = 1 \ne 0 = \mathbb{E}(G_0)$? Is there a contradiction? Actually no. The optional stopping theorem does not apply here, because the time $T$ is unbounded: $\mathbb{P}(T = n) = 2^{-n}$, $\forall n \in \mathbb{N}$, i.e., there does not exist $N$ fixed such that $T(\omega) \le N$, $\forall \omega \in \Omega$.

## 13.6 Doob's decomposition theorem

**Theorem 13.11.** Let $(X_n,\, n \in \mathbb{N})$ be a submartingale with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$. Then there exists a martingale $(M_n,\, n \in \mathbb{N})$ with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ and a process $(A_n,\, n \in \mathbb{N})$ predictable with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ and increasing (i.e., $A_n \le A_{n+1}$ $\forall n \in \mathbb{N}$) such that $A_0 = 0$ and $X_n = M_n + A_n$, $\forall n \in \mathbb{N}$. Moreover, this decomposition of the process $X$ is unique.

*Proof.* (main idea)
$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \ge X_n$, so a natural candidate for the process $A$ is to set $A_0 = 0$ and $A_{n+1} = A_n + \mathbb{E}(X_{n+1}|\mathcal{F}_n) - X_n$ $(\ge A_n)$, which is a predictable and increasing process. Then, $M_0 = X_0$ and $M_{n+1} - M_n = X_{n+1} - X_n - (A_{n+1} - A_n) = X_{n+1} - \mathbb{E}(X_{n+1}|\mathcal{F}_n)$ is indeed a martingale, as $\mathbb{E}(M_{n+1} - M_n|\mathcal{F}_n) = 0$. □

# 14 Martingale convergence theorems

## 14.1 Preliminary: Doob's martingale

**Proposition 14.1.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n,\, n \in \mathbb{N})$ be a filtration and $X : \Omega \to \mathbb{R}$ be an $\mathcal{F}$-measurable and integrable random variable. Then the process $(M_n,\, n \in \mathbb{N})$ defined as

$$M_n = \mathbb{E}(X|\mathcal{F}_n), \quad n \in \mathbb{N}$$

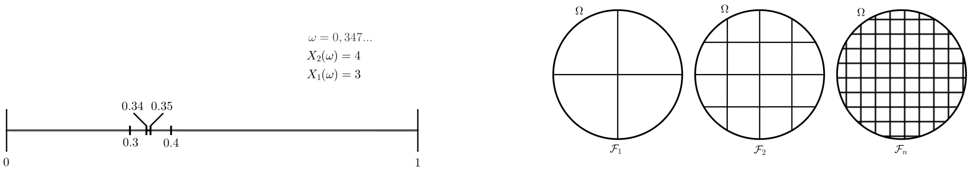
is a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$.

*Proof.* (i) $\mathbb{E}(|M_n|) = \mathbb{E}(|\mathbb{E}(X|\mathcal{F}_n)|) \le \mathbb{E}(\mathbb{E}(|X|\,|\mathcal{F}_n)) = \mathbb{E}(|X|) < +\infty$, for all $n \in \mathbb{N}$.

(ii) By the definition of conditional expectation, $M_n = \mathbb{E}(X|\mathcal{F}_n)$ is $\mathcal{F}_n$-measurable, for all $n \in \mathbb{N}$.

(iii) $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_{n+1})|\mathcal{F}_n) = \mathbb{E}(X|\mathcal{F}_n) = M_n$, for all $n \in \mathbb{N}$. □

**Remarks. -** This process describes the situation where one acquires more and more information about a random variable. Think e.g. at the case where $X$ is a number drawn uniformly at random between 0 and 1, and one reads this number from left to right: while reading, one obtains more and more information about the number, as illustrated on the left-hand side of the figure below:



- On the right-hand side of the figure is another illustration of a Doob martingale: as time goes by, one gets more and more information about where to locate oneself in the space $\Omega$.

- Are Doob's martingales a very particular type of martingales? No! As the following paragraph shows, there are quite many such martingales!

## 14.2 The martingale convergence theorem: first version

**Theorem 14.2.** Let $(M_n,\, n \in \mathbb{N})$ be a square-integrable martingale (i.e., a martingale such that $\mathbb{E}(M_n^2) < +\infty$ for all $n \in \mathbb{N}$) with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$. Under the additional assumption that

$$\sup_{n\in\mathbb{N}} \mathbb{E}(M_n^2) < +\infty \tag{17}$$

there exists a limiting random variable $M_\infty$ such that

(i) $M_n \underset{n\to\infty}{\to} M_\infty$ almost surely.

(ii) $\lim_{n\to\infty} \mathbb{E}\left((M_n - M_\infty)^2\right) = 0$ (quadratic convergence).

(iii) $M_n = \mathbb{E}(M_\infty|\mathcal{F}_n)$, for all $n \in \mathbb{N}$ (this last property is referred to as the martingale $M$ being "closed at infinity").

**Remarks.** - Condition (17) is of course much stronger than just asking that $\mathbb{E}(M_n^2) < +\infty$ for every $n$. Think for example at the simple symmetric random walk $S_n$: $\mathbb{E}(S_n^2) = n < +\infty$ for every $n$, but the supremum is infinite.

- By conclusion (iii) in the theorem, any square-integrable martingale satisfying condition (17) is actually a Doob martingale (take $X = M_\infty$)!

- A priori, one could think that all the conclusions of the theorem hold true if one replaces all the squares by absolute values in the above statement (such as e.g. replacing condition (17) by $\sup_{n\in\mathbb{N}} \mathbb{E}(|M_n|) < +\infty$, etc.). This is wrong, and we will see interesting counter-examples later.

- A stronger condition than (17) (leading therefore to the same conclusion) is the following:

$$\sup_{n\in\mathbb{N}} \sup_{\omega\in\Omega} |M_n(\omega)| < +\infty. \tag{18}$$

Martingales satisfying this stronger condition are called *bounded martingales.*

**Example 14.3.** Let $M_0 = x$, where $x \in [0, 1]$ is a fixed number, and let us define recursively:

$$M_{n+1} = \begin{cases} M_n^2, & \text{with probability } \frac{1}{2} \\ 2M_n - M_n^2, & \text{with probability } \frac{1}{2} \end{cases}$$

The process $M$ is a bounded martingale. Indeed:

(i) By induction, if $M_n \in [0, 1]$, then $M_{n+1} \in [0, 1]$, for every $n \in \mathbb{N}$, so as $M_0 = x \in [0, 1]$, we obtain

$$\sup_{n\in\mathbb{N}} \sup_{\omega\in\Omega} |M_n(\omega)| \leq 1 < +\infty$$

(ii) $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \frac{1}{2}\, M_n^2 + \frac{1}{2}\,(2M_n - M_n^2) = M_n$, for every $n \in \mathbb{N}$.

By the theorem, there exists therefore a random variable $M_\infty$ such that the three conclusions of the theorem hold. In addition, it can be shown by contradiction that $M_\infty$ takes values in the binary set $\{0, 1\}$ only, so that

$$x = \mathbb{E}(M_0) = \mathbb{E}(M_\infty) = \mathbb{P}(\{M_\infty = 1\})$$

## 14.3 Consequences of the theorem

Before diving into the proof of the above important theorem, let us first explore a few of its interesting consequences.

**Optional stopping theorem, version 2.** Let $(\mathcal{F}_n,\, n \in \mathbb{N})$ be a filtration, let $(M_n,\, n \in \mathbb{N})$ be a square-integrable martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ which satisfies condition (17) and let $0 \leq T_1 \leq T_2 \leq +\infty$ be two stopping times with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$. Then

$$\mathbb{E}(M_{T_2}|\mathcal{F}_{T_1}) = M_{T_1} \text{ a.s.} \quad \text{and} \quad \mathbb{E}(M_{T_2}) = \mathbb{E}(M_{T_1})$$

*Proof.* Simply replace $N$ by $\infty$ in the proof of the first version and use the fact that $M$ is a closed martingale by the convergence theorem. □

**Stopped martingale.** Let $(M_n,\, n \in \mathbb{N})$ be a martingale and $T$ be a stopping time with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$, without any further assumption. Let us also define the *stopped process*

$$(M_{T\wedge n},\, n \in \mathbb{N})$$

where $T \wedge n = \min\{T, n\}$ by definition. Then this stopped process is also a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ (we skip the proof here).

**Optional stopping theorem, version 3.** Let $(M_n,\, n \in \mathbb{N})$ be a martingale with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ such that there exists $c > 0$ with $|M_{n+1}(\omega) - M_n(\omega)| \leq c$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$ (this assumption ensures that the martingale does not make jumps of uncontrolled size: the simple symmetric random walk $S_n$ satisfies in particular this assumption). Let also $a, b > 0$ and

$$T = \inf\{n \in \mathbb{N} \,:\, M_n \leq -a \text{ or } M_n \geq b\}$$

Observe that $T$ is a stopping time with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ and that $-a - c \leq M_{T\wedge n}(\omega) \leq b + c$ for all $\omega \in \Omega$ and $n \in \mathbb{N}$. In particular,

$$\sup_{n\in\mathbb{N}} \mathbb{E}(M_{T\wedge n}^2) < +\infty$$

so the stopped process $(M_{T\wedge n},\, n \in \mathbb{N})$ satisfies the assumptions of the first version of the martingale convergence theorem. By the conclusion of this theorem, the stopped martingale $(M_{T\wedge n},\, n \in \mathbb{N})$ is closed, i.e. it admits a limit $M_{T\wedge\infty} = M_T$ and

$$\mathbb{E}(M_T) = \mathbb{E}(M_{T\wedge\infty}) = \mathbb{E}(M_{T\wedge 0}) = \mathbb{E}(M_0)$$

**Application.** Let $(S_n,\, n \in \mathbb{N})$ be the simple symmetric random walk (which satisfies the above assumptions with $c = 1$) and $T$ be the above stopping time (with $a, b$ positive integers). Then $\mathbb{E}(S_T) = \mathbb{E}(S_0) = 0$. Given that $S_T \in \{-a, +b\}$, we obtain

$$0 = \mathbb{E}(S_T) = (+b)\,\mathbb{P}(\{S_T = +b\}) + (-a)\,\mathbb{P}(\{S_T = -a\}) = bp - a(1-p), \quad \text{where } p = \mathbb{P}(\{S_T = +b\})$$

From this, we deduce that $\mathbb{P}(\{S_T = +b\}) = p = \frac{a}{a+b}$.

**Remark.** Note that the same reasoning does not hold if we replace the stopping time $T$ by a stopping time of the form

$$T' = \inf\{n \in \mathbb{N} \,:\, M_n \geq b\}$$

There is indeed no guarantee in this case that the stopped martingale $(M_{T'\wedge n},\, n \in \mathbb{N})$ is bounded (from below).

## 14.4 Proof of the theorem

**A key ingredient for the proof: the maximal inequality.** The following inequality, apart from being useful for the proof of the martingale convergence theorem, is interesting in itself. Let $(M_n,\, n \in \mathbb{N})$ be a square-integrable martingale. Then for every $N \in \mathbb{N}$ and $x > 0$,

$$\mathbb{P}\left(\left\{\max_{0\leq n\leq N} |M_n| \geq x\right\}\right) \leq \frac{\mathbb{E}(M_N^2)}{x^2}$$

**Remark.** This inequality resembles Chebyshev's inequality, but it is actually much stronger. In particular, note the remarkable fact that deviation probability of the maximum value of the martingale *over the whole time interval* $\{0, \ldots, N\}$ is controlled by the second moment of the martingale at the final instant $N$ alone.

*Proof.* - First, let $x > 0$ and let $T_x = \inf\{n \in \mathbb{N} \,:\, |M_n| \geq x\}$: $T_x$ is a stopping time and note that

$$\{T_x \leq N\} = \left\{\max_{0\leq n\leq N} |M_n| \geq x\right\}$$

So what we need actually to prove is that $\mathbb{P}(\{T_x \leq N\}) \leq \frac{\mathbb{E}(M_N^2)}{x^2}$.

- Second, observe that as $M$ is a martingale, $M^2$ is a submartingale. So by the optional stopping theorem, we obtain

$$\begin{aligned} \mathbb{E}(M_N^2) &\geq \mathbb{E}\left(M_{T_x\wedge N}^2\right) \geq \mathbb{E}\left(M_{T_x\wedge N}^2\, 1_{\{T_x\leq N\}}\right) \\ &= \mathbb{E}\left(M_{T_x}^2\, 1_{\{T_x\leq N\}}\right) \geq \mathbb{E}\left(x^2\, 1_{\{T_x\leq N\}}\right) = x^2\, \mathbb{P}(\{T_x \leq N\}) \end{aligned}$$

where the last inequality comes from the fact that $|M_{T_x}| \geq x$, by definition of $T_x$. This proves the claim. □

*Proof of Theorem 14.2.* - We first prove conclusion (i), namely that the sequence $(M_n,\, n \in \mathbb{N})$ converges almost surely to some limit. This proof is divided in two parts.

*Part 1.* We first show that for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left\{\sup_{n\in\mathbb{N}} |M_{n+m} - M_m| \geq \varepsilon\right\}\right) \underset{m\to\infty}{\to} 0 \tag{19}$$

This is saying that for every $\varepsilon > 0$, the probability that the martingale $M$ deviates by more than $\varepsilon$ after a given time $m$ can be made arbitrarily small by taking $m$ large enough. This essentially says that the fluctuations of the martingale decay with time, i.e. that the martingale ultimately converges! Of course, this is just an intuition and needs a formal proof, which will be done in the second part of the proof. For now, let us focus on proving (19).

a) Let $m \in \mathbb{N}$ be fixed and define the process $(Y_n,\, n \in \mathbb{N})$ by $Y_n = M_{n+m} - M_m$, for $n \in \mathbb{N}$. $Y$ is a square-integrable martingale, so by the maximal inequality, we have for every $N \in \mathbb{N}$ and every $\varepsilon > 0$:

$$\mathbb{P}\left(\left\{\max_{0\leq n\leq N} |Y_n| \geq \varepsilon\right\}\right) \leq \frac{\mathbb{E}(Y_N^2)}{\varepsilon^2}$$

b) Let us now prove that

$$\mathbb{E}(Y_N^2) = \mathbb{E}(M_{m+N}^2) - \mathbb{E}(M_m^2).$$

This equality follows from the orthogonality of the increments of $M$. Here is a detailed proof:

$$\begin{aligned} \mathbb{E}(Y_N^2) &= \mathbb{E}((M_{m+N} - M_m)^2) = \mathbb{E}(M_{m+N}^2) - 2\,\mathbb{E}(M_{m+N}\, M_m) + \mathbb{E}(M_m^2) \\ &= \mathbb{E}(M_{m+N}^2) - 2\,\mathbb{E}(\mathbb{E}(M_{m+N}\, M_m|\mathcal{F}_m)) + \mathbb{E}(M_m^2) \\ &= \mathbb{E}(M_{m+N}^2) - 2\,\mathbb{E}(\mathbb{E}(M_{m+N}|\mathcal{F}_m)\, M_m) + \mathbb{E}(M_m^2) \\ &= \mathbb{E}(M_{m+N}^2) - 2\,\mathbb{E}(M_m^2) + \mathbb{E}(M_m^2) = \mathbb{E}(M_{m+N}^2) - \mathbb{E}(M_m^2) \end{aligned}$$

Gathering a) and b) together, we obtain for every $m, N \in \mathbb{N}$ and every $\varepsilon > 0$:

$$\mathbb{P}\left(\left\{\max_{0\leq n\leq N} |M_{m+n} - M_m| \geq \varepsilon\right\}\right) \leq \frac{\mathbb{E}(M_{m+N}^2) - \mathbb{E}(M_m^2)}{\varepsilon^2}.$$

c) Assumption (17) states that $\sup_{n\in\mathbb{N}} \mathbb{E}(M_n^2) < +\infty$. As the sequence $(\mathbb{E}(M_n^2),\, n \in \mathbb{N})$ is increasing (since $M^2$ is a submartingale), this also says that the sequence has a limit: $\lim_{n\to\infty} \mathbb{E}(M_n^2) = K < +\infty$. Therefore, for every $m \in \mathbb{N}$ and $\varepsilon > 0$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\left\{\sup_{n\in\mathbb{N}} |M_{m+n} - M_m| \geq \varepsilon\right\}\right) &= \lim_{N\to\infty} \mathbb{P}\left(\left\{\max_{0\leq n\leq N} |M_{m+n} - M_m| \geq \varepsilon\right\}\right) \\
&\leq \lim_{N\to\infty} \frac{\mathbb{E}(M_{m+N}^2) - \mathbb{E}(M_m^2)}{\varepsilon^2} = \frac{K - \mathbb{E}(M_m^2)}{\varepsilon^2}
\end{aligned}$$

Taking now $m$ to infinity, we further obtain

$$\mathbb{P}\left(\left\{\sup_{n\in\mathbb{N}} |M_{m+n} - M_m| \geq \varepsilon\right\}\right) \leq \frac{K - \mathbb{E}(M_m^2)}{\varepsilon^2} \underset{m\to\infty}{\to} \frac{K-K}{\varepsilon^2} = 0$$

for every $\varepsilon > 0$. This proves (19) and concludes therefore the first part of the proof.

*Part 2.* Let $C = \{\omega \in \Omega \,:\, \lim_{n\to\infty} M_n(\omega) \text{ exists}\}$. In this second part, we prove that $\mathbb{P}(C) = 1$, which is conclusion (i).

Here is what we have proven so far. For $m \in \mathbb{N}$ and $\varepsilon > 0$, define $A_m(\varepsilon) = \{\sup_{n\in\mathbb{N}} |M_{m+n} - M_m| \geq \varepsilon\}$. Then (19) says that for every fixed $\varepsilon > 0$, $\lim_{m\to\infty} \mathbb{P}(A_m(\varepsilon)) = 0$. This implies in particular that

$$\forall \varepsilon > 0, \quad \mathbb{P}(\cap_{m\in\mathbb{N}} A_m(\varepsilon)) = 0$$

We then have the following (long!) series of equivalent statements:

$$\begin{aligned}
&\forall \varepsilon > 0, \quad \mathbb{P}(\cap_{m\in\mathbb{N}} A_m(\varepsilon)) = 0 \quad \Longleftrightarrow \quad \forall M \geq 1, \quad \mathbb{P}(\cap_{m\in\mathbb{N}} A_m(\tfrac{1}{M})) = 0 \\
\Longleftrightarrow \quad & \mathbb{P}(\cup_{M\geq 1} \cap_{m\in\mathbb{N}} A_m(\tfrac{1}{M})) = 0 \quad \Longleftrightarrow \quad \mathbb{P}(\cup_{\varepsilon>0} \cap_{m\in\mathbb{N}} A_m(\varepsilon)) = 0 \\
\Longleftrightarrow \quad & \mathbb{P}(\{\exists \varepsilon > 0 \text{ s.t. } \forall m \in \mathbb{N},\ \sup_{n\in\mathbb{N}} |M_{m+n} - M_m| \geq \varepsilon\}) = 0 \\
\Longleftrightarrow \quad & \mathbb{P}(\{\forall \varepsilon > 0,\ \exists m \in \mathbb{N} \text{ s.t. } \sup_{n\in\mathbb{N}} |M_{m+n} - M_m| < \varepsilon\}) = 1 \\
\Longleftrightarrow \quad & \mathbb{P}(\{\forall \varepsilon > 0,\ \exists m \in \mathbb{N} \text{ s.t. } |M_{m+n} - M_m| < \varepsilon,\ \forall n \in \mathbb{N}\}) = 1 \\
\Longleftrightarrow \quad & \mathbb{P}(\{\forall \varepsilon > 0,\ \exists m \in \mathbb{N} \text{ s.t. } |M_{m+n} - M_{m+p}| < \varepsilon,\ \forall n, p \in \mathbb{N}\}) = 1 \\
\Longleftrightarrow \quad & \mathbb{P}(\{\text{the sequence } (M_n,\, n \in \mathbb{N}) \text{ is a Cauchy sequence}\}) = 1 \Longleftrightarrow \mathbb{P}(C) = 1
\end{aligned}$$

as every Cauchy sequence in $\mathbb{R}$ converges. This completes the proof of conclusion (i) in the theorem.

- In order to prove conclusion (ii) (quadratic convergence), let us recall that from what was shown above

$$\mathbb{E}((M_n - M_m)^2) = \mathbb{E}(M_n^2) - \mathbb{E}(M_m^2), \quad \forall n \geq m \geq 0$$

This, together with the fact that $\lim_{n\to\infty} \mathbb{E}(M_n^2) = K$, implies that $M_n$ is a Cauchy sequence in $L^2$: it therefore converges to some limit, as the space of square-integrable random variables is complete. Let us call this limit $\widetilde{M}_\infty$. But does it hold that $\widetilde{M}_\infty = M_\infty$, the a.s. limit of part (i)? Yes, as both quadratic convergence and a.s. convergence imply convergence in probability, and we have seen in part I (Theorem 5.3) that if a sequence of random variables converges in probability to two possible limits, then these two limits are equal almost surely.

- Conclusion (iii) then follows from the following reasoning. We need to prove that $M_n = \mathbb{E}(M_\infty|\mathcal{F}_n)$ for every (fixed) $n \in \mathbb{N}$ (where $M_\infty$ is the limit found in parts (i) and (ii)). To this end, let us go back to the very definition of conditional expectation and simply check that

(i) $M_n$ is $\mathcal{F}_n$-measurable: this is by definition.

(ii) $\mathbb{E}(M_\infty U) = \mathbb{E}(M_n U)$ for every random variable $U$ $\mathcal{F}_n$-measurable and bounded. This follows from the following observation:

$$\mathbb{E}(M_n U) = \mathbb{E}(M_N U), \quad \forall N \geq n$$

This equality together with the Cauchy-Schwarz inequality imply that for every $N \geq n$:

$$|\mathbb{E}(M_\infty U) - \mathbb{E}(M_n U)| = |\mathbb{E}(M_\infty U) - \mathbb{E}(M_N U)| = |\mathbb{E}((M_\infty - M_N)\, U)| \leq \sqrt{\mathbb{E}((M_\infty - M_N)^2)}\, \sqrt{\mathbb{E}(U^2)} \underset{N\to\infty}{\to} 0$$

by quadratic convergence (conclusion (ii)). So we obtain that necessarily, $\mathbb{E}(M_\infty U) = \mathbb{E}(M_n U)$ (remember that $n$ is fixed here). This completes the proof of Theorem 14.2. □

## 14.5 The martingale convergence theorem: second version

**Theorem 14.4.** Let $(M_n,\, n \in \mathbb{N})$ be a martingale such that

$$\sup_{n\in\mathbb{N}} \mathbb{E}(|M_n|) < +\infty \tag{20}$$

Then there exists a limiting random variable $M_\infty$ such that $M_n \underset{n\to\infty}{\to} M_\infty$ almost surely.

We shall not go through the proof of this second version of the martingale convergence theorem[16], whose order of difficulty resembles that of the first one. Let us just make a few remarks and also exhibit an interesting example below.

**Remarks.** - Contrary to what one could perhaps expect, it does not necessarily hold in this case that $\lim_{n\to\infty} \mathbb{E}(|M_n - M_\infty|) = 0$, nor that $\mathbb{E}(M_\infty | \mathcal{F}_n) = M_n$ for every $n \in \mathbb{N}$.

- By the Cauchy-Schwarz inequality, we see that condition (20) is weaker than condition (17).

- On the other hand, condition (20) is of course stronger than just asking $\mathbb{E}(|M_n|) < +\infty$ for all $n \in \mathbb{N}$ (this last condition is by the way satisfied by every martingale, by definition). It is also stronger than asking $\sup_{n\in\mathbb{N}} \mathbb{E}(M_n) < +\infty$. Why? Simply because for every martingale, $\mathbb{E}(M_n) = \mathbb{E}(M_0)$ for every $n \in \mathbb{N}$, so the supremum is always finite! The same does not hold when one adds absolute values: the process $(|M_n|,\, n \in \mathbb{N})$ is a submartingale, so the sequence $(\mathbb{E}(|M_n|),\, n \in \mathbb{N})$ is non-decreasing, possibly growing to infinity.

- If $M$ is a non-negative martingale, then $|M_n| = M_n$ for every $n \in \mathbb{N}$ and by what was just said above, condition (20) is satisfied! So non-negative martingales *always* converge to a limit almost surely! But they might not be closed at infinity.

**A puzzling example.** Let $(S_n,\, n \in \mathbb{N})$ be the simple symmetric random walk and $(M_n,\, n \in \mathbb{N})$ be the process defined as

$$M_n = \exp(S_n - cn), \quad n \in \mathbb{N}$$

where $c = \log\left(\frac{e+e^{-1}}{2}\right) > 0$ is such that $M$ is a martingale, with $\mathbb{E}(M_n) = \mathbb{E}(M_0) = 1$ for every $n \in \mathbb{N}$. On top of that, $M$ is a positive martingale, so by the previous remark, there exists a random variable $M_\infty$ such that $M_n \underset{n\to\infty}{\to} M_\infty$ almost surely. So far so good. Let us now consider some more puzzling facts:

- A simple computation shows that $\sup_{n\in\mathbb{N}} \mathbb{E}(M_n^2) = \sup_{n\in\mathbb{N}} \mathbb{E}(\exp(2S_n - 2cn)) = +\infty$, so we cannot conclude that (ii) and (iii) in Theorem 14.2 hold. Actually, these conclusions do not hold, as we will see below.

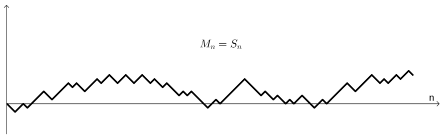
- What can the random variable $M_\infty$ be? It can be shown that $S_n - cn \underset{n\to\infty}{\to} -\infty$ almost surely, from which we deduce that $M_n = \exp(S_n - cn) \underset{n\to\infty}{\to} 0$ almost surely, i.e. $M_\infty = 0$!

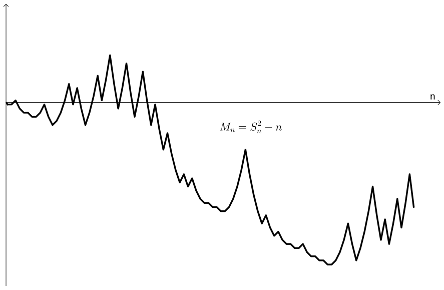
[16] It is sometimes called the first version in the literature!

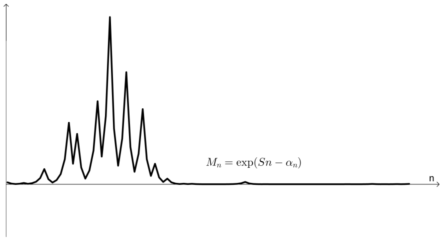
- It is therefore impossible that $\mathbb{E}(M_\infty|\mathcal{F}_n) = M_n$, as the left-hand side is 0, while the right-hand side is not. Likewise, quadratic convergence to 0 does not hold (this would mean that $\lim_{n\to\infty} \mathbb{E}(M_n^2) = 0$, which does not hold).

- On the contrary, we just said above that $\mathrm{Var}(M_n) = \mathbb{E}(M_n^2) - (\mathbb{E}(M_n))^2 = \mathbb{E}(M_n^2) - 1$ grows to infinity as $n$ goes to infinity. Still, $M_n$ converges to 0 almost surely. If this sounds puzzling to you, be reassured that you are not alone!

For illustration purposes, here are below three well known martingales: $(S_n,\, n \in \mathbb{N})$, $(S_n^2 - n,\, n \in \mathbb{N})$ and $(M_n = \exp(S_n - cn,\, n \in \mathbb{N})$ just seen above:







We see again here that even though theses three processes are all contant mean processes, they do exhibit very different behaviours!

## 14.6 Generalization to sub- and supermartingales

We state below the generalization of the two convergence theorems to sub- and supermartingales.

**Theorem 14.5.** (Generalization of Theorem 14.2)
Let $(M_n,\, n \in \mathbb{N})$ be a square-integrable submartingale (resp., supermartingale) with respect to a filtration $(\mathcal{F}_n,\, n \in \mathbb{N})$. Under the additional assumption that

$$\sup_{n\in\mathbb{N}} \mathbb{E}(M_n^2) < +\infty \tag{21}$$

there exists a limiting random variable $M_\infty$ such that

(i) $M_n \underset{n\to\infty}{\to} M_\infty$ almost surely.

(ii) $\lim_{n\to\infty} \mathbb{E}\left(|M_n - M_\infty|\right) = 0$ ($L^1$ convergence).

(iii) $M_n \leq \mathbb{E}(M_\infty|\mathcal{F}_n)$ (resp., $M_n \geq \mathbb{E}(M_\infty|\mathcal{F}_n)$), for all $n \in \mathbb{N}$.

Note that $L^2$ convergence does not hold in general when $M$ is a sub- or a supermartingale, but $L^1$ convergence does, and the sub- or supermartingale is also closed at infinity in this case (with the equality $M_n = \mathbb{E}(M_\infty|\mathcal{F}_n)$ being replaced by an inequality, of course).

**Theorem 14.6.** (Generalization of Theorem 14.4)
Let $(M_n,\, n \in \mathbb{N})$ be a submartingale (resp., supermartingale) such that

$$\sup_{n\in\mathbb{N}} \mathbb{E}(M_n^+) < +\infty \quad \text{(resp., } \sup_{n\in\mathbb{N}} \mathbb{E}(M_n^-) < +\infty) \tag{22}$$

where recall here that $M_n+ = \max(M_n, 0)$ and $M_n^- = \max(-M_n, 0)$. Then there exists a limiting random variable $M_\infty$ such that $M_n \underset{n\to\infty}{\to} M_\infty$ almost surely.

As one can see, not much changes in the assumptions and conclusions of both theorems! Let us mention some interesting consequences.

- From Theorem 14.5, it holds that if $M$ is a sub- or a supermartingale satisfying condition (21), then $M_n$ converges both almost surely and in $L^1$ to some limit $M_\infty$. In case where $M$ is a (non-trivial) martingale, we saw previously that the limit $M_\infty$ cannot be equal to 0, as this would lead to a contradiction, because of the third part of the conclusion stating that $M_n = \mathbb{E}(M_\infty|\mathcal{F}_n) = 0$ for all $n$. In the case of a sub- or supermartingale, this third part only says that $M_n \le \mathbb{E}(M_\infty|\mathcal{F}_n) = 0$ or $M_n \ge \mathbb{E}(M_\infty|\mathcal{F}_n) = 0$, which is not necessarily a contradiction.

- From Theorem 14.6, one deduces that any positive supermartingale admits an almost sure limit at infinity. But the same conclusion cannot be drawn for a positive submartingale (think simply of $M_n = n$: this very particular positive submartingale does not converge). From the same theorem, one deduces also that any *negative* submartingale admits an almost sure limit at infinity.

## 14.7 Azuma's and McDiarmid's inequalities

**Theorem 14.7.** (Azuma's inequality)
Let $(M_n,\, n \in \mathbb{N})$ be a martingale such that $|M_n(\omega) - M_{n-1}(\omega)| \le 1$ for every $n \ge 1$ and $\omega \in \Omega$. Such a martingale is said to have *bounded differences*. Assume also that $M_0$ is constant. Then for every $n \ge 1$ and $t > 0$, we have

$$\mathbb{P}(\{|M_n - M_0| \ge nt\}) \le 2\,\exp\left(-\frac{nt^2}{2}\right)$$

**Remark.** This statement resembles that of Hoeffding's inequality! The difference here is that a martingale is not necessarily a sum of i.i.d. random variables.

*Proof.* Let $X_n = M_n - M_{n-1}$ for $n \ge 1$. Then, by the assumptions made, $M_n - M_0 = \sum_{j=1}^n X_j$, with $|X_j(\omega)| \le 1$ for every $j \ge 1$ and $\omega \in \Omega$, but as mentioned above, the $X_j$'s are not necessarily i.i.d.: we only know that $\mathbb{E}(X_j|\mathcal{F}_{j-1}) = 0$ for every $j \ge 1$. We need to bound

$$\mathbb{P}\left(\left\{\left|\sum_{j=1}^n X_j\right| \ge nt\right\}\right) = \mathbb{P}\left(\left\{\sum_{j=1}^n X_j \ge nt\right\}\right) + \mathbb{P}\left(\left\{\sum_{j=1}^n X_j \le -nt\right\}\right)$$

By Chebyshev's inequality with $\varphi(x) = e^{sx}$ and $s > 0$, we obtain

$$\begin{aligned}\mathbb{P}\left(\left\{\sum_{j=1}^n X_j \ge nt\right\}\right) &\le \frac{\mathbb{E}\left(\exp\left(s\sum_{j=1}^n X_j\right)\right)}{\exp(snt)} = e^{-snt}\,\mathbb{E}\left(\mathbb{E}\left(\exp\left(s\sum_{j=1}^n X_j\right)|\mathcal{F}_{n-1}\right)\right)\\ &= e^{-snt}\,\mathbb{E}\left(\exp\left(s\sum_{j=1}^{n-1} X_j\right)\mathbb{E}\left(\exp\left(sX_n\right)|\mathcal{F}_{n-1}\right)\right)\end{aligned}$$

As $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = 0$ and $|X_n(\omega)| \leq 1$ for every $\omega \in \Omega$, we can apply the same lemma as in the proof of Hoeffding's inequality to conclude that

$$\mathbb{E}(\exp(sX_n)|\mathcal{F}_{n-1}) \leq e^{s^2/2}$$

So

$$\mathbb{P}\left(\left\{\sum_{j=1}^n X_j \geq nt\right\}\right) \leq e^{-snt}\,\mathbb{E}\left(\exp\left(s\sum_{j=1}^{n-1} X_j\right)\right)\, e^{s^2/2}$$

and working backwards, we finally obtain the upper bound

$$\mathbb{P}\left(\left\{\sum_{j=1}^n X_j \geq nt\right\}\right) \leq e^{-snt+ns^2/2}$$

which is again minimum for $s^* = t$ and equal then to $\exp(-nt^2/2)$. By symmetry, the same bound is obtained for the other term:

$$\mathbb{P}\left(\left\{\sum_{j=1}^n X_j \leq -nt\right\}\right) \leq \exp(-nt^2/2)$$

which completes the proof. □

**Generalization.** Exactly like Hoeffding's inequality, Azuma's inequality can be generalized as follows. Let $M$ be a martingale such that $M_n(\omega) - M_{n-1}(\omega) \in [a_n, b_n]$ for every $n \geq 1$ and every $\omega \in \Omega$. Then

$$\mathbb{P}(\{|M_n - M_0| \geq nt\}) \leq 2\,\exp\left(-\frac{2n^2t^2}{\sum_{j=1}^n (b_j - a_j)^2}\right)$$

**Application 1.** Consider the martingale transform of Section 13.5 defined as follows. Let $(X_n,\, n \geq 1)$ be a sequence of i.i.d. random variables such that $\mathbb{P}(\{X_1 = +1\}) = \mathbb{P}(\{X_1 = -1\}) = \frac{1}{2}$. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ for $n \geq 1$. Let $(H_n,\, n \in \mathbb{N})$ be a predictable process with respect to $(\mathcal{F}_n,\, n \in \mathbb{N})$ such that $|H_n(\omega)| \leq K_n$ for every $n \in \mathbb{N}$ and $\omega \in \Omega$. Let finally $G_0 = 0$, $G_n = \sum_{j=1}^n H_j X_j$, $n \geq 1$. Then

$$\mathbb{P}(\{|G_n - G_0| \geq nt\}) \leq 2\,\exp\left(-\frac{n^2t^2}{2\sum_{j=1}^n K_j^2}\right)$$

In the case where $K_n = K$ for every $n \in \mathbb{N}$, this says that

$$\mathbb{P}(\{|G_n - G_0| \geq nt\}) \leq 2\,\exp\left(-\frac{nt^2}{2K^2}\right)$$

We had obtained the same conclusion earlier for the random walk, but here, the increments of $G$ are in general far from being independent.

**Application 2: McDiarmid's inequality.** Let $n \geq 1$ be fixed, let $X_1, \ldots, X_n$ be i.i.d. random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ be a Borel-measurable function such that

$$|f(x_1, \ldots, x_j, \ldots, x_n) - f(x_1, \ldots, x_j', \ldots, x_n)| \leq K_j, \quad \forall x_1, \ldots, x_j, x_j', \ldots, x_n \in \mathbb{R},\ 1 \leq j \leq n$$

Then

$$\mathbb{P}(\{|f(X_1, \ldots, X_n) - \mathbb{E}(f(X_1, \ldots, X_n))| \geq nt\}) \leq 2\,\exp\left(-\frac{n^2t^2}{2\sum_{j=1}^n K_j^2}\right)$$

*Proof.* Define $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_j = \sigma(X_1, \ldots, X_j)$ and $M_j = \mathbb{E}(F(X_1, \ldots, X_n)|\mathcal{F}_j)$ for $j \in \{0, \ldots, n\}$. By definition, $M$ is a martingale and observe that

$$M_n = f(X_1, \ldots, X_n) \quad \text{and} \quad M_0 = \mathbb{E}(f(X_1, \ldots, X_n))$$

Moreover,

$$|M_j - M_{j-1}| = |\mathbb{E}(f(X_1,\ldots,X_n)|\mathcal{F}_j) - \mathbb{E}(f(X_1,\ldots,X_n)|\mathcal{F}_{j-1})| = |g(X_1,\ldots,X_j) - h(X_1,\ldots,X_{j-1})|$$

where $g(x_1,\ldots,x_j) = \mathbb{E}(f(x_1,\ldots,x_j,X_{j+1},\ldots,X_n))$ and $h(x_1,\ldots,x_{j-1}) = \mathbb{E}(f(x_1,\ldots,x_{j-1},X_j,\ldots,X_n))$. By the assumption made, we find that for every $x_1,\ldots,x_j \in \mathbb{R}$,

$$|g(x_1,\ldots,x_j) - h(x_1,\ldots,x_{j-1})| \leq \mathbb{E}(|f(x_1,\ldots,x_j,X_{j+1},\ldots,X_n) - f(x_1,\ldots,x_{j-1},X_j,\ldots,X_n)|) \leq K_j$$

so Azuma's inequality applies. This completes the proof. □

# A Appendix

## A.1 Carathéodory's extension theorem

**Definition A.1.** Let $\Omega$ be a set. A *ring* on $\Omega$ is a collection $\mathcal{R}$ of subsets of $\Omega$ such that

(i) $\emptyset, \Omega \in \mathcal{R}$.

(ii) if $A, B \in \mathcal{R}$, then $A \cup B \in \mathcal{R}$.

(iii) if $A, B \in \mathcal{R}$, then $A \backslash B \in \mathcal{R}$.

**Example.** One easily checks that the collection $\mathcal{R}_1$ of *finite* unions of pairwise disjoint and half-open intervals in $\mathbb{R}$, such as

$$A = \bigcup_{j=1}^{n} [a_j, b_j[ \quad \text{with } [a_j, b_j[ \text{ disjoint}$$

is a ring.

**Definition A.2.** A *pre-probability measure* on $(\Omega, \mathcal{R})$ is a mapping $\mathbb{P} : \mathcal{R} \to [0,1]$ such that

(i) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$.

(ii) If $(A_n,\, n \geq 1)$ is a sequence of disjoint subsets of $\Omega$ such that $A_n \in \mathcal{R}$ for all $n \geq 1$ *and* $\cup_{n\geq 1} A_n \in \mathcal{R}$, then $\mathbb{P}(\cup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$.

**Example.** Let $p : \mathbb{R} \to \mathbb{R}$ be a given pdf on $\mathbb{R}$ (take your favorite, e.g. $p(x) = 1_{[0,1]}(x)$ or $p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$). Let now $\mathbb{P}_1(A) = \int_A p(x)\, dx$ for $A \in \mathcal{R}_1$ defined above. As $A$ is a finite union of pairwise disjoint intervals, $\mathbb{P}_1(A)$ is well defined and one can also check that $\mathbb{P}_1$ is a pre-probablity measure on $\mathcal{R}_1$.

**Theorem A.3.** (Carathéodory's extension theorem, without proof)
Let $\mathcal{R}$ be a ring on $\Omega$ and $\mathbb{P}$ be a pre-probability measure on $(\Omega, \mathcal{R})$. Then there exists a unique probability measure $\widetilde{\mathbb{P}}$ defined on the $\sigma$-field $\mathcal{F} = \sigma(\mathcal{R})$ such that $\widetilde{\mathbb{P}}(A) = \mathbb{P}(A)$ for every $A \in \mathcal{R}$.

**Remark.** This theorem has of course little to do with the other well-known Carathéodory theorem stating that in a convex body $K$ in $\mathbb{R}^d$, any interior point can be written as a convex combination of at most $d+1$ extremal points of $K$.

## A.2 Distances between random variables associated to various convergences

For quadratic convergence, there is of course a natural distance related to it:

$$d_2(X, Y) = \|X - Y\|_2, \quad \text{where } \|X - Y\|_2^2 = \mathbb{E}((X - Y)^2)$$

whereas there is *no* distance associated to almost sure convergence. Below, we describe distances associated to convergence in probability and convergence in distribution.

### Distances for convergence in probability: the Ky-Fan distances

Recall the definition: $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$ if $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(\{|X_n - X| \geq \varepsilon\}) = 0$ (applicable only to random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$).

It can be shown that for any of the following three distances (attributed to Ky Fan):

$$d(X,Y) = \inf\{\varepsilon > 0 \,:\, \mathbb{P}(\{|X-Y| \geq \varepsilon\}) \leq \varepsilon\}$$

$$d(X,Y) = \mathbb{E}(\min\{|X-Y|, 1\}) \quad \text{or} \quad d(X,Y) = \mathbb{E}\left(\frac{|X-Y|}{1+|X-Y|}\right)$$

it holds that $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X$ if and only if $d(X_n, X) \underset{n\to\infty}{\to} 0$.

**Distances for convergence in distribution**

Recall the definition: $X_n \underset{n\to\infty}{\overset{d}{\to}} X$ if $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$, $\forall t \in \mathbb{R}$ continuity point of $F_X$ (and we do not need to assume that the random variables $X_n$ are defined on a common probability space).

**The Lévy distance**

The Lévy distance between two random variables $X, Y$ (or more precisely, between the distribution of $X$ and that of $Y$) is defined as:

$$d_L(X,Y) = \inf\{\varepsilon > 0 \,:\, F_X(t-\varepsilon) - \varepsilon \leq F_Y(t) \leq F_X(t+\varepsilon) + \varepsilon, \quad \forall t \in \mathbb{R}\}$$

It can be shown that $X_n \underset{n\to\infty}{\overset{d}{\to}} X$ if and only if $d_L(X_n, X) \underset{n\to\infty}{\to} 0$, i.e., convergence in distribution is equivalent to convergence in the Lévy distance.

The next three distances defined below are stronger than the Lévy distance, in the sense that if convergence in any of these distance takes place, then convergence in distribution takes place (but the reciprocal statement does not hold).

**The Kantorovich ($L^1$) distance**

$$d_{K,1}(X,Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)|\, dt$$

Observe that this distance is not a proper distance as it might take the value $+\infty$ for a given pair of random variables $X, Y$ (this is the case for example if $X, Y$ are two Cauchy random variables with different parameters). Nevertheless, it can be shown that $X_n \underset{n\to\infty}{\overset{d}{\to}} X$ if $d_{K,1}(X_n, X) \underset{n\to\infty}{\to} 0$.

To gain some intuition on what this distance means, observe also that if $Y = X + c$, where c is a constant, then $d_{K,1}(X,Y) = c$. Moreover, when both $X$ and $Y$ are *bounded* random variables (implying that $d_{K,1}(X,Y) < +\infty$ in this case), this distance is equal to the *Wasserstein distance*:

$$d_W(X,Y) = \sup\{|\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| \,:\, f : \mathbb{R} \to \mathbb{R} \text{ is such that } |f(x) - f(y)| \leq |x-y|, \ \forall x, y \in \mathbb{R}\}$$

**The Kolmogorov ($L^\infty$) distance**

$$d_{K,\infty}(X,Y) = \sup_{t\in\mathbb{R}} |F_X(t) - F_Y(t)|$$

It can be shown (heathy exercise!) that $d_{K,\infty}(X,Y) \geq d_L(X,Y)$ for every pair of random variables $X, Y$, implying that convergence in the Kolmogorov distance implies convergence in the Levy distance (and therefore also convergence in distribution).

**The Radon or total variation (TV) distance**

$$d_{\text{TV}}(X,Y) = \sup_{B\in\mathcal{B}(\mathbb{R})} |\mu_X(B) - \mu_Y(B)|$$

It is immediate that $d_{\text{TV}}(X,Y) \geq d_K(X,Y)$ (as $d_K$ is a supremum over a restricted number of sets). It is therefore the strongest of all norms above, so strong that $d_{\text{TV}}(X,Y) = 1$ as soon as $X$ is discrete and $Y$ is continuous, for example. The total variation distance can also be rewritten as

$$d_{\text{TV}}(X,Y) = \sup\{|\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| \,:\, f : \mathbb{R} \to \mathbb{R} \text{ is a continuous function with values in } [-1, +1]\}$$

## A.3 Two useful facts about convergences of sequences of random variables

**Continuous mapping theorem**

Consider a sequence $(X_n,\, n \geq 1)$ of random variables, another random variable $X$ and a function $g : \mathbb{R} \to \mathbb{R}$ continuous on a domain $D$ such that $\mathbb{P}(\{X \in D\}) = 1$. Then it is straightforward to check that if $X_n \underset{n\to\infty}{\to} X$ almost surely, then $g(X_n) \underset{n\to\infty}{\to} g(X)$ almost surely. The more surprising fact (stated here without proof) is that the same holds true both for convergence in probability and convergence in distribution, namely:

$$\text{if } X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} X, \text{ then } g(X_n) \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} g(X) \qquad\qquad \text{if } X_n \underset{n\to\infty}{\overset{d}{\to}} X, \text{ then } g(X_n) \underset{n\to\infty}{\overset{d}{\to}} g(X)$$

This fact does not hold true for $L^2$ convergence, as $X$ being square-integrable does not necessarily imply that $g(X)$ also is.

**Cauchy sequences**

It is straightforward to check that a sequence $(X_n,\, n \geq 1)$ of random variables converges almost surely if and only if it is almost surely a Cauchy sequence, i.e.

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

A similar statement holds true for quadratic convergence (as the space $L^2(\Omega)$ equipped with the scalar product $\langle X, Y\rangle_2 = \mathbb{E}(XY)$ is a Hilbert space[17]), as well as for convergence in probability: i.e., the sequence $(X_n,\, n \geq 1)$ converges in probability if and only if

$$\forall \varepsilon > 0,\ \mathbb{P}(\{|X_n - X_m| \geq \varepsilon\}) \underset{n,m\to\infty}{\to} 0$$

but this does *not* hold true for convergence in distribution. Here is a counter-example: the sequence $(X_n,\, n \geq 1)$ where $X_n \sim \mathcal{U}([-n, +n])$ does satisfy

$$|F_{X_n}(t) - F_{X_m}(t)| \underset{n,m\to\infty}{\to} 0, \quad \forall t \in \mathbb{R}$$

but does not converge in distribution.

## A.4 An intriguing fact about convergence in distribution

The above counter-example is related to an intriguing fact that has two different versions:

1. Let $(F_n,\, n \geq 1)$ be a sequence of cdfs such that there exists a function $F : \mathbb{R} \to [0, 1]$ with $\lim_{n\to\infty} F_n(t) = F(t)$ for every $t \in \mathbb{R}$. Then it is not necessarily the case that $F$ itself is the cdf of a random variable.
2. Let $(\phi_n,\, n \geq 1)$ be a sequence of characteristic functions such that there exists a function $\phi : \mathbb{R} \to \mathbb{C}$ with $\lim_{n\to\infty} \phi_n(t) = \phi(t)$ for every $t \in \mathbb{R}$. Then it is not necessarily the case that $\phi$ is itself the characteristic function of a random variable.

Here is an example illustrating this. Consider $(X_n,\, n \geq 1)$ a sequence of random variables with $X_n \sim \mathcal{U}([-n, +n])$ for every $n \geq 1$. Then their pdfs are given by

$$p_{X_n}(x) = \frac{1}{2n}\, 1_{[-n,n]}(x), \quad x \in \mathbb{R}$$

[17] As already mentioned before, we should consider here $L^2(\Omega)$ to be the space of equivalence classes of square-integrable random variables, two random variables $X, Y$ being equivalent if $X = Y$ a.s.

and correspondingly,

$$F_{X_n}(t) = \frac{1}{2n} \int_{-\infty}^{t} 1_{[-n,n]}(x)\, dx = \begin{cases} 0, & \text{if } t \leq -n \\ \frac{1}{2}(1 + \frac{t}{n}), & \text{if } -n < t \leq n \\ 1, & \text{if } t > n \end{cases}$$

One can check easily that for every fixed $t \in \mathbb{R}$, $F(t) = \lim_{n\to\infty} F_{X_n}(t) = \frac{1}{2}$, which is not a cdf, as $\lim_{t\to\pm\infty} F(t) = \frac{1}{2} \neq 0$ or 1.

Likewise,

$$\phi_{X_n}(t) = \frac{1}{2n} \int_{-n}^{n} e^{itx}\, dx = \begin{cases} \frac{e^{itn} - e^{-itn}}{2itn} = \frac{\sin(tn)}{tn}, & \text{if } t \neq 0 \\ 1, & \text{if } t = 0 \end{cases}$$

so

$$\phi(t) = \lim_{n\to\infty} \phi_{X_n}(t) = \begin{cases} 0, & \text{if } t \neq 0 \\ 1, & \text{if } t = 0 \end{cases}$$

but $\phi$ is not a characteristic function, as it is not continuous in $t = 0$.

From either of the above observations, we conclude that the sequence of random variables $(X_n,\, n \geq 1)$ does not converge in distribution, even though both their cdfs and characteristic functions converge to a limit. What is the problem?

The intuition, first: the random variable $X_n$ is uniformly distributed on $[-n, +n]$, so as $n \to \infty$, $X_n$ should converge to a random variable $X$ "uniformly distributed on $\mathbb{R}$", but we know that such a random variable does not exist. What is happening here is that all the mass of the distribution of $X_n$ is escaping to $\pm\infty$ as $n$ increases. In order to prevent such a thing from happening, we could impose that the sequence $X_n$ remains bounded as $n$ increases, i.e., that there exists $C > 0$ such that

$$\sup_{n\geq 1} \sup_{\omega\in\Omega} |X_n(\omega)| \leq C$$

This however is a strong assumption, ruling out nicely behaved random variables such as Gaussian or Poisson random variables. It can be relaxed to the following: for every $\varepsilon > 0$, there exists $C > 0$ such that

$$\inf_{n\geq 1} \mathbb{P}(\{|X_n| \leq C\}) \geq 1 - \varepsilon \tag{23}$$

Such a sequence of random variables is said to be *tight*. This assumption guarantees that most of the mass of the distribution of the random variables $X_n$ remains in a bounded interval, even for large values of $n$ (more precisely, that for every $\varepsilon > 0$, there exists $C > 0$ such that at least a fraction $1 - \varepsilon$ of the mass of all random variables $X_n$ remains in the interval $[-C, +C]$). Note that this assumption is natural, as for a single real-valued random variable $X$, it always holds by definition that for every $\varepsilon > 0$, there exists $C > 0$ such that

$$\mathbb{P}(\{|X| \leq C\}) \geq 1 - \varepsilon$$

If we therefore expect the sequence $X_n$ to converge in distribution as $n$ gets large, condition (23) should hold. Reversely, one can show that when condition (23) is met, pointwise convergence of either sequence $(F_{X_n},\, n \geq 1)$ or $(\phi_{X_n},\, n \geq 1)$ guarantees convergence in distribution of the sequence $(X_n,\, n \geq 1)$.

## A.5 Stein's method

Stein's method is (yet) another method allowing to prove the central limit theorem. It is based on the following equality (seen in the exercises and known as *Stein's lemma*):

$$\mathbb{E}(Z\, f(Z)) = \mathbb{E}(f'(Z)) \tag{24}$$

if $Z \sim \mathcal{N}(0, 1)$ and $f$ is a continuously differentiable function such that both $|f(x)|$ and $|f'(x)|$ do not grow faster than polynomially when $|x| \to \infty$. More interestingly, one can show that that if $Z$ is a random variable satisfying (24) for every $f \in C_b^1(\mathbb{R})$, then the distribution of $Z$ is necessarily $\mathcal{N}(0,1)$. Even more interestingly, if $(Y_n,\, n \geq 1)$ is a sequence of random variables such that

$$\mathbb{E}(Y_n\, f(Y_n) - f'(Y_n)) \underset{n\to\infty}{\to} 0, \quad \forall f \in C_b^1(\mathbb{R}) \tag{25}$$

then $Y_n \underset{n\to\infty}{\overset{d}{\to}} Z \sim \mathcal{N}(0,1)$.

*Proof.* We show below that condition (25) implies that

$$\mathbb{E}(g(Y_n)) \underset{n\to\infty}{\to} \mathbb{E}(g(Z)), \quad \forall g \in C_b(\mathbb{R})$$

implying in turn convergence in distribution. To this end, let $g$ be a given function in $C_b(\mathbb{R})$ and $f$ be the solution of the following differential equation:

$$x\, f(x) - f'(x) = g(x) - \mathbb{E}(g(Z)), \quad x \in \mathbb{R} \tag{26}$$

It turns out that the solution $f$ of this equation belongs to $C_b^1(\mathbb{R})$ [TO BE CHECKED] and can be written explicitly as

$$f(x) = e^{x^2/2} \int_{-\infty}^{x} e^{-t^2/2}\, (g(t) - \mathbb{E}(g(Z))\, dt$$

[This can be checked by simply computing the derivative of $f$]. Because $f$ satisfies equation (26), we can evaluate this equation in $x = Y_n$ and take expectations on both sides to conclude that

$$\mathbb{E}(g(Y_n)) - \mathbb{E}(g(Z)) = \mathbb{E}(Y_n\, f(Y_n) - f'(Y_n)) \underset{n\to\infty}{\to} 0$$

by assumption. □

The interest of Stein's method is twofold:

1. It allows to obtain precise estimates on how "close" the distribution of a random variable is from the Gaussian distribution, by estimating either the Wasserstein distance between the two distributions or their Kolmogorov distance (= max difference between their cdfs).
2. It allows to generalize the statement of the central limit theorem from sums of i.i.d. random variables to expressions of the form $F(X_1, \ldots, X_n)$, where the random variables

$$W_j = F(X_1, \ldots, X_j', \ldots, X_n) - F(X_1, \ldots, X_j, \ldots, X_n)$$
$$\text{(with } X_j' \text{ an independent copy of } X_j\text{)}$$

   are "small" and "approximately independent" as $n$ gets large. This extension is therefore significant (of the same flavor as McDiarmid's inequality for concentration).

We present below the method for the classical setup: assume $(X_n,\, n \geq 1)$ are i.i.d. random variables with mean 0 and variance 1 (for simplicity) and define, for $n \geq 1$, $S_n = X_1 + \ldots + X_n$ and $Y_n = S_n/\sqrt{n}$. The central limit theorem states in this case that $Y_n \underset{n\to\infty}{\overset{d}{\to}} Z \sim \mathcal{N}(0,1)$.

We now prove (leaving out some technical details) that the sequence $(Y_n,\, n \geq 1)$ satisfies condition (25) above. To this end, let us compute, for a given $f \in C_b^1(\mathbb{R})$:

$$\mathbb{E}(Y_n\, f(Y_n)) = \frac{1}{\sqrt{n}}\, \mathbb{E}(S_n\, f(Y_n)) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbb{E}(X_j\, f(Y_n)) \tag{27}$$

For $j \in \{1, \ldots, n\}$, define $Y_n^{(j)} = Y_n - \frac{X_j}{\sqrt{n}}$. Since $f \in C_b^1(\mathbb{R})$, we can write, using Taylor's expansion:

$$f(Y_n^{(j)}) \simeq f(Y_n) + (Y_n^{(j)} - Y_n)\, f'(Y_n) = f(Y_n) - \frac{X_j}{\sqrt{n}}\, f'(Y_n)$$

or equivalenty:

$$f(Y_n) \simeq f(Y_n^{(j)}) + \frac{X_j}{\sqrt{n}} f'(Y_n)$$

Introducing this in (27), we obtain

$$\mathbb{E}(Y_n\, f(Y_n)) \simeq \frac{1}{\sqrt{n}} \sum_{j=1}^n \left\{ \mathbb{E}\left( X_j \left( f(Y_n^{(j)}) + \frac{X_j}{\sqrt{n}} f'(Y_n) \right) \right) \right\}$$

The independence of the $X$'s implies that $X_j \perp\!\!\!\perp Y_n^{(j)}$ for every $j \in \{1, \ldots, n\}$, so

$$\begin{aligned} \mathbb{E}(Y_n\, f(Y_n)) &\simeq \frac{1}{\sqrt{n}} \sum_{j=1}^n \underbrace{\mathbb{E}(X_j)}_{=0} \mathbb{E}(f(Y_n^{(j)})) + \frac{1}{n} \sum_{j=1}^n \mathbb{E}(X_j^2\, f'(Y_n)) \\ &\simeq \mathbb{E}\left( \frac{1}{n} \sum_{j=1}^n X_j^2\, f'(Y_n) \right) \end{aligned}$$

The last two ingredients are: the law of large numbers which says that

$$\frac{1}{n} \sum_{j=1}^n X_j^2 \underset{n\to\infty}{\to} \mathbb{E}(X_1^2) = 1 \quad \text{almost surely}$$

and another “technical detail” (i.e., the dominated convergence theorem), which allow to conclude that

$$\mathbb{E}(Y_n\, f(Y_n)) \simeq \mathbb{E}(f'(Y_n))$$

as $n$ gets large, therefore proving (25). □

## A.6 Lower bound on large deviations estimates

In Section 11.2, we proved the following upper bound on the probability of a large deviation (at a given value of $n \geq 1$):

$$\mathbb{P}\left( \left\{ \frac{S_n}{n} \geq t \right\} \right) \leq \exp(-n\, \Lambda^*(t)), \quad \forall t > \mu$$

where we recall that $S_n = X_1 + \ldots + X_n$, $\mu = \mathbb{E}(X_1)$, $\Lambda^*(t) = \max_{s\in\mathbb{R}}(st - \Lambda(s))$ and $\Lambda(s) = \log \mathbb{E}(e^{sX_1})$. This implies in particular that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left( \left\{ \frac{S_n}{n} \geq t \right\} \right) \leq -\Lambda^*(t), \quad \forall t > \mu$$

In the following, we prove the matching lower bound

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left( \left\{ \frac{S_n}{n} \geq t \right\} \right) \geq -\Lambda^*(t), \quad \forall t > \mu$$

implying that the limit exists and is equal to $-\Lambda^*(t)$, and therefore that the upper bound is tight[18].

*Proof.* In order to obtain this lower bound, we will do something strange: for a given $n \geq 1$ and $t > \mu$, the event $\{S_n \geq nt\}$ is a rare event (this is what the upper bound says). Let us now *change the probability measure* $\mathbb{P}$ so as to make this event become *frequent.* To this end, define first for $s \geq 0$ and $A \in \mathcal{F}$:

$$\widetilde{\mathbb{P}}_s(A) = \mathbb{E}(1_A \cdot \exp(sX_1 - \Lambda(s)))$$

[18] Note that a similar result holds for $t < \mu$

(observing that $\widetilde{\mathbb{P}}_0 = \mathbb{P}$). All these new probability measures $\widetilde{\mathbb{P}}_s$ are candidates for transforming the rare event $\{S_n \geq nt\}$ into a frequent event, but this will work for a unique value of $s$, as we shall see. Let us first check that for every $s \geq 0$, $\widetilde{\mathbb{P}}_s$ is indeed a probability measure: clearly, $\widetilde{\mathbb{P}}_s(\emptyset) = 0$ and the $\sigma$-additivity property follows from the fact that $\mathbb{P}$ itself is a probability measure. What remains to be checked is that $\widetilde{\mathbb{P}}_s(\Omega) = 1$:

$$\widetilde{\mathbb{P}}_s(\Omega) = \mathbb{E}(\exp(sX_1 - \Lambda(s))) = \mathbb{E}(\exp(sX_1)) \cdot \exp(-\Lambda(s)) = \frac{\mathbb{E}(\exp(sX_1))}{\mathbb{E}(\exp(sX_1))} = 1$$

Moreover, under this new probability measure, we have:

$$\widetilde{\mathbb{E}}_s(X_1) = \mathbb{E}(X_1 \cdot \exp(sX_1 - \Lambda(s))) = \frac{\mathbb{E}(X_1 \, \exp(sX_1))}{\mathbb{E}(\exp(sX_1))} = \frac{\frac{\partial}{\partial s}\mathbb{E}(\exp(sX_1))}{\mathbb{E}(\exp(sX_1))} = \frac{\partial}{\partial s} \log \mathbb{E}(\exp(sX_1)) = \frac{\partial}{\partial s}\Lambda(s)$$

By carefully choosing the value of $s$, we can therefore tune the value of $\widetilde{\mathbb{E}}_s(X_1)$ so as to make it close to $t$ and not $\mu$ (thereby making the event $\{S_n \geq nt\}$ frequent).

This is the main idea behind the change of probability measure, but the story is actually a little bit more complicated than that, as we need to define a change of probability measure involving all the random variables $X_1, \ldots, X_n$ [19] and not only $X_1$. To this end, let us define for $s \geq 0$ and $A \in \mathcal{F}$: (overriding the above definition with a single $X_1$):

$$\widetilde{\mathbb{P}}_s(A) = \mathbb{E}\left(1_A \cdot \prod_{j=1}^{n} \exp(sX_j - \Lambda(s))\right)$$

Using the very same argument as above, we can show that $\widetilde{\mathbb{P}}_s$ is a probability measure for every $s \geq 0$. Moreover, observe that

$$\widetilde{\mathbb{P}}_s(A) = \mathbb{E}\left(1_A \cdot \exp\left(s\sum_{j=1}^{n} X_j - n\Lambda(s)\right)\right) = \mathbb{E}(1_A \, \exp(sS_n - n\Lambda(s))) \tag{28}$$

Following the same procedure as above, we also obtain the equality $\widetilde{\mathbb{E}}_s(X_1) = \frac{\partial}{\partial s}\Lambda(s)$. Let us not choose yet the value of $s$, but let us try instead to obtain a lower bound $\mathbb{P}(\{S_n \geq nt\})$ (which is our goal). To this end, observe first that the equality (28) can be rewritten as

$$\mathbb{P}(A) = \widetilde{\mathbb{E}}_s(1_A \, \exp(-sS_n + n\Lambda(s))) \quad \text{for } A \in \mathcal{F}$$

With this, we obtain (setting $\varepsilon > 0$):

$$\begin{aligned}
\mathbb{P}(\{S_n \geq nt\}) \geq \mathbb{P}(\{nt \leq S_n \leq n(t+\varepsilon)\}) &= \widetilde{\mathbb{E}}_s\left(1_{\{nt \leq S_n \leq n(t+\varepsilon)\}} \, \exp(-sS_n + n\Lambda(s))\right) \\
&= \exp(-n(st - \Lambda(s))) \, \widetilde{\mathbb{E}}_s\left(1_{\{nt \leq S_n \leq n(t+\varepsilon)\}} \, \exp(-s(S_n - nt))\right) \\
&\geq \exp(-n(st - \Lambda(s))) \, \exp(-sn\varepsilon) \, \widetilde{\mathbb{P}}_s(\{nt \leq S_n \leq n(t+\varepsilon)\})
\end{aligned}$$

Let us now choose $s \geq 0$ that maximizes $st - \Lambda(s)$, i.e., $\frac{\partial \Lambda(s)}{\partial s} = t$ and $st - \Lambda(s) = \Lambda^*(t)$. Under the probability measure $\widetilde{\mathbb{P}}_s$, we have seen above that $\widetilde{\mathbb{E}}_s(X_1) = \frac{\partial \Lambda(s)}{\partial s} = t$, so $S_n \simeq nt$ "on average" for large $n$. One can actually show more precisely that for any $\varepsilon > 0$

$$\widetilde{\mathbb{P}}_s(\{nt \leq S_n \leq n(t+\varepsilon)\}) \underset{n \to \infty}{\to} c > 0$$

Therefore, as $n$ gets large,

$$\mathbb{P}(\{S_n \geq nt\}) \overset{\sim}{\geq} \exp(-n\Lambda^*(t)) \, \exp(-sn\varepsilon) \, c$$

implying that

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\{S_n \geq nt\}) \geq -\Lambda^*(t) - s\varepsilon$$

implying in turn the result, as $\varepsilon > 0$ is arbitrary (and $s \geq 0$ is fixed). □

[19] Recall that $n$ is fixed here.