Advanced Probability and Applications (Part I)

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

February 8, 2019

Contents

1 σ-fields and random variables 3
  1.1 σ-fields ................................................................. 3
  1.2 σ-field generated by a collection of events .......................... 4
  1.3 Sub-σ-field ............................................................. 5
  1.4 Random variables ....................................................... 5
  1.5 σ-field generated by a collection of random variables ............. 7

2 Probability measures and distributions 8
  2.1 Probability measures .................................................... 8
  2.2 Distribution of a random variable ..................................... 10
  2.3 Cumulative distribution function ..................................... 10
  2.4 Two important classes of random variables .......................... 11
  2.5 The Cantor set and the devil’s staircase ............................ 13

3 Independence 15
  3.1 Independence of two events ......................................... 15
  3.2 Independence of two random variables ............................... 16
  3.3 Independence of two sub-σ-fields .................................... 17
  3.4 Independence of more sub-σ-fields ................................... 17
  3.5 Do independent random variables really exist? ........................ 18

4 Expectation 19
  4.1 Discrete non-negative random variables ............................. 19
  4.2 General definition ....................................................... 20
  4.3 Inequalities ............................................................... 23
5 Laws of large numbers 26

5.1 Preliminary: convergence of sequences of numbers ............................. 26
5.2 Convergences of sequences of random variables ............................... 26
5.3 Relations between the three notions of convergence ............................ 26
5.4 The Borel-Cantelli lemma .............................................................. 28
5.5 Laws of large numbers .................................................................. 29
5.6 Application: convergence of the empirical distribution ......................... 31
5.7 Kolmogorov's 0-1 law .................................................................. 31
5.8 Extension of the weak law: an example ............................................. 32

6 The central limit theorem 34

6.1 Convergence in distribution ............................................................... 34
6.2 Equivalent criterion for convergence in distribution ......................... 35
6.3 The central limit theorem ................................................................. 36

7 Concentration inequalities 38

7.1 Hoeffding’s inequality ................................................................. 38
7.2 Large deviations principle .............................................................. 40

A Appendix 43

A.1 Characteristic functions and their basic properties ............................... 43
A.2 An alternate proof of the central limit theorem .................................. 44
A.3 An intriguing fact ....................................................................... 45
A.4 Moments and Carleman’s theorem .................................................. 46

Basic terminology and conventions
- A discrete set means either a finite set (in bijection with \{1, \ldots, N\} for some \(N \geq 1\) or a countable set (in bijection with \(\mathbb{N}\)).

- 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 \(x_1 < x_2\).

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 σ-fields and random variables

1.1 σ-fields

In probability, the fundamental set Ω 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 Ω = {1, ..., 6} representing the outcomes of a die roll, or Ω = [0, 1] representing the outcomes of a concentration measurement of some chemical product. Notice moreover that the set Ω need not be composed of numbers exclusively; it would be for example perfectly valid to consider the set Ω = {banana, apple, orange}.

Given a fundamental set Ω, 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 σ-field, which is defined below. Notice 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 Ω be a set. A σ-field (or σ-algebra) on Ω is a collection $\mathcal{F}$ of subsets of Ω (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 $A_1, ..., A_n \in \mathcal{F}$, then $\bigcup_{j=1}^n A_j \in \mathcal{F}$.

(iii') If $(A_n, n \geq 1)$ is a sequence of subsets of Ω 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 $A_1, ..., A_n \in \mathcal{F}$, then $\bigcap_{j=1}^n A_j \in \mathcal{F}$.

(iv') If $(A_n, n \geq 1)$ is a sequence of subsets of Ω 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 $B \setminus A = B \cap A^c \in \mathcal{F}$.

**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 Ω) does the realization of the experiment $\omega$ belong.

**Example 1.2.** For a generic set Ω, the following are always σ-fields:

$\mathcal{F}_0 = \{\emptyset, \Omega\}$ (= trivial σ-field).

$\mathcal{P}(\Omega) = \{\text{all subsets of } \Omega\}$ (= complete σ-field).

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

$\mathcal{F}_1 = \{\emptyset, \{1\}, \{2, ..., 6\}, \Omega\}$.

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

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

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

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

In the discrete setting (that is, in a σ-field with a finite or countable number of elements), the smallest elements contained in a σ-field are called the atoms of the σ-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 $I_1, ..., I_n$. Notice moreover that a σ-field with $n$ atoms has $2^n$ elements, so that the number of elements of a finite σ-field is always a power of 2.
1.2 σ-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 \(A = \{A_i, i \in I\}\) be a collection of subsets of \(\Omega\) (where \(I\) need not be finite nor countable). The \(\sigma\)-field generated by \(A\) is the smallest \(\sigma\)-field on \(\Omega\) containing all the events \(A_i\). It is denoted as \(\sigma(A)\).

**Remark.** A natural question is whether such a vague definition makes sense. Observe first that there is always at least one \(\sigma\)-field containing \(A\): it is \(P(\Omega)\). 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 \(\sigma(A)\): it is the intersection of all \(\sigma\)-fields containing the collection \(A\), which is certainly a well-defined object.

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

- Let \(A_1 = \{\{1\}\}\). Then \(\sigma(A_1) = F_1\).
- Let \(A_2 = \{\{1, 3, 5\}\}\). Then \(\sigma(A_2) = F_2\).
- Let \(A_3 = \{\{1, 2, 3\}\}\). Then \(\sigma(A_3) = F_3\).
- Let \(A = \{\{1\}, \ldots, \{6\}\}\). Then \(\sigma(A) = P(\Omega)\).

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

**Example.** Let \(\Omega = [0, 1]\) and let \(A_4 = \{I_1, \ldots, I_n\}\) (cf. Example 1.4). Then \(\sigma(A_4) = F_4\). 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:

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

is the Borel \(\sigma\)-field on \([0, 1]\) and elements of \(B([0, 1])\) are called the Borel subsets of \([0, 1]\). As surprising as it may be, it turns out that \(B([0, 1]) \neq P([0, 1])\) [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 \(B([0, 1])\). Notice indeed that

a) All singletons belong to \(B([0, 1])\). Indeed, for any \(0 < x < 1\), \(\{x\} = \bigcap_{n \geq 1} [x - \frac{1}{n}, x + \frac{1}{n}]\) belongs to \(B([0, 1])\), 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 \(B([0, 1])\).

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

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 \(B([0, 1])\), 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 \(\Omega = \{1, \ldots, 6\}\).

**Remark.** It can be easily checked that the atoms of \(B([0, 1])\) are the singletons \(\{x\}, x \in [0, 1]\). Nevertheless, one can check that \(B([0, 1])\) 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 σ-field on $\mathbb{R}$ and $\mathbb{R}^2$.

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

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

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

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

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

Notice that even though $B(\mathbb{R}^2)$ 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 σ-field generated by a collection of events is much larger than the collection of events itself.

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

### 1.3 Sub-σ-field

One may have more or less information about a system. In mathematical terms, this translates into the fact that a σ-field contains more or less elements. It is therefore convenient to introduce a (partial) ordering on the ensemble of existing σ-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 σ-field on $\Omega$. A sub-σ-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 σ-field.

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

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

**Example.** Let $\Omega = \{1, \ldots, 6\}$ (cf. Example 1.3). Notice that $\mathcal{F}_1$ is not a sub-σ-field of $\mathcal{F}_2$ (even though $\{1\} \subset \{1, 3, 5\}$), nor is $\mathcal{F}_2$ a sub-σ-field of $\mathcal{F}_1$. In general, notice 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-σ-field of $B([0, 1])$. 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 \(X : \Omega \rightarrow \mathbb{R}\) satisfying
\[
\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R})
\] (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 \(X : \Omega \rightarrow \mathbb{R}\) 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 \(X : \Omega \rightarrow \mathbb{R}\) 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 \(\Omega \rightarrow \mathbb{R}\) 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 notice 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 \((\Omega, \mathcal{F})\). Notice 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 \(g : \mathbb{R} \rightarrow \mathbb{R}\) 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 \(g : \mathbb{R} \rightarrow \mathbb{R}\) 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 \(g : \mathbb{R} \rightarrow \mathbb{R}\) 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. \qed

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.** Notice 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]

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

**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 $\sigma(X_4) = \mathcal{F}_4$.

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 $X$.

Notice 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 \( \mathbb{P} : \mathcal{F} \rightarrow [0, 1] \) 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 \( \mathbb{P}(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} \mathbb{P}(A_j) \).

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

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

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

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

(v) 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 \( \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \mathbb{P}(A_n) \).

(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 \( \mathbb{P}(\bigcap_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \mathbb{P}(A_n) \).

**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-\( \sigma \)-additivity, of probability 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

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

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

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

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] Caratheodory’s extension theorem 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 $\mathbb{P}(B) = |B|$. Notice that it corresponds also to the uniform distribution on $[0,1]$.

Example. Let $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}(\mathbb{R})$. One defines 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 again be uniquely extended to all Borel subsets of $\mathbb{R}$: it is called the (normalized) Gaussian measure on $\mathbb{R}$.

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.

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 $([0,1], \mathcal{B}([0,1]), \mathbb{P} =$ 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,

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

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

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

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.** The knowledge of \(F_X\) is equivalent to the knowledge of \(\mu_X\) [without proof].

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. \(F_X(s) \leq F_X(t)\) for all \(s < t\).

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

Indeed:

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

(ii) If \(s \leq t\), then \(\{X \leq s\} \subset \{X \leq t\}\), so \(F_X(s) = \mathbb{P}(\{X \leq s\}) \leq \mathbb{P}(\{X \leq t\}) = F_X(t)\).
(iii) For any $t \in \mathbb{R}$, we have $\lim_{\varepsilon \downarrow 0} F_X(t + \varepsilon) = \lim_{n \to \infty} F_X(t + \frac{1}{n}) = \lim_{n \to \infty} \mathbb{P}(\{X \leq t + \frac{1}{n}\}) = \mathbb{P}(\{X \leq t\}) = F_X(t)$, as again the sequence $\{X \leq t + \frac{1}{n}\}$ is a decreasing sequence of events with $\bigcap_{n \geq 1} \{X \leq t + \frac{1}{n}\} = \{X \leq t\}$.

**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 $t \in \mathbb{R}$, $\lim_{\varepsilon \downarrow 0} F_X(t - \varepsilon) = \lim_{n \to \infty} F_X(t - \frac{1}{n}) = \lim_{n \to \infty} \mathbb{P}(\{X \leq t - \frac{1}{n}\}) = \mathbb{P}(\{X < t\})$, as $\{X \leq t - \frac{1}{n}\}$ is an increasing sequence of events with $\bigcup_{n \geq 1} \{X \leq t - \frac{1}{n}\} = \{X < t\}$. But $\mathbb{P}(\{X < t\}) \neq F_X(t)$ 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 $C$ of $\mathbb{R}$, that is, $X(\omega) \in C$ 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 $x \in C$. Notice that $0 \leq p_x \leq 1$ for all $x \in C$ and that $\sum_{x \in C} p_x = \mathbb{P}(\{X \in C\}) = 1$. The sequence of numbers $(p_x, x \in C)$ 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:

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

and

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

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 $X \sim Bi(n, p)$) 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 $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ 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 $|\{x\}| = 0 \ \forall x \in \mathbb{R}$).

**Fact.** [without proof] 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 $F_X'(t) = p_X(t)$), as illustrated below:

![Plot of probability density function and cumulative distribution function](image)

**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

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

---

\(^1\)This fact is known as the Radon-Nikodym theorem. It has many different formulations and important applications in probability theory.
so in particular, 

\[ p_X(\mu) = \frac{1}{\sqrt{2\pi} \sigma^2} > 1 \text{ if } \sigma < \frac{1}{\sqrt{2\pi}}. \]

**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 \( Y \)? Not much in general, but assume for example that \( f \) is increasing, i.e., that \( f(x_1) < f(x_2) \) as soon as \( x_1 < x_2 \). Then \( f \) is invertible, so

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

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

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

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 \( x_1 < x_2 \)), notice 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

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

be the set of (open) intervals removed at stage \( n \). In particular, \( A_1 = \left] \frac{1}{3}, \frac{2}{3} \right[ \), \( A_2 = \left] \frac{1}{9}, \frac{2}{9} \right[ \cup \left] \frac{5}{9}, \frac{8}{9} \right[ \), etc. Notice also that \( A_n \) and \( A_m \) are *disjoint* for every \( n \neq m \). The Cantor set \( C \) is then defined as

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

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

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

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:

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

$$x = \sum_{n \geq 1} \frac{a_n}{3^n}, \quad \text{where} \quad a_n \in \{0, 2\}$$  \hspace{1cm} (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 $b_n = 0 \leftrightarrow a_n = 0$ and $b_n = 1 \leftrightarrow a_n = 2$), 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:

![Devil's staircase graph]

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

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

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].

Notice 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 $t \in \bigcup_{n \geq 1} A_n$, 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...

### 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 $\sigma$-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):

$$
\mathbb{P}(A_1 \cap A_2^c) = \mathbb{P}(A_1 \setminus (A_1 \cap A_2)) = \mathbb{P}(A_1) - \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) - \mathbb{P}(A_1) \mathbb{P}(A_2) = \mathbb{P}(A_1)(1 - \mathbb{P}(A_2)) = \mathbb{P}(A_1) \mathbb{P}(A_2^c)
$$

Notice 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

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

Notation. $X_1 \perp X_2$.

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 $(\Omega, \mathcal{F}, \mathbb{P})$.

The above general definition with Borel sets has an interesting direct consequence.

Proposition 3.4. Let $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R}$ 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 $B_1, B_2 \in \mathcal{B}(\mathbb{R})$:

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

$$= \mathbb{P}(\{X_1 \in f_1^{-1}(B_1)\}) \mathbb{P}(\{X_2 \in f_2^{-1}(B_2)\}) = \mathbb{P}(\{f_1(X_1) \in B_1\}) \mathbb{P}(\{f_2(X_2) \in B_2\})$$

Notice again 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 preimage of $B_i$.

Checking the independence relation for all Borel sets $B_1$ and $B_2$ can be painful... Luckily, the following proposition helps [without proof].

Proposition 3.5. $X_1, X_2$ are independent if and only if

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

The above welcome simplification is connected to the fact that the knowledge of the distribution $\mu_X$ of a random variable $X$ is equivalent to that of its cdf $F_X$.

Further simplifications of this definition occur in the two following situations [again, without proofs]:

- Assume $X_1, X_2$ are two discrete random variables, taking values in a common countable set $C$. Then $X_1, X_2$ are independent if and only if

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

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 $p_{X_1, X_2} : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ (= 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:

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

This can always be assumed for discrete random variables: indeed, if $X_1 \in C_1$ and $X_2 \in C_2$, then simply consider $C = C_1 \cup C_2$, which is also countable.
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

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

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

One can readily check (using in particular Proposition 3.2 for the first line) that
- $A_1, A_2$ are independent according to Definition 3.3 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 3.4 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 $A_i^c$, $i \in \{1, \ldots, n\}$. Notice 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

$$\mathbb{P}\left(\bigcap_{j \in S} A_j\right) = \prod_{j \in S} \mathbb{P}(A_j), \quad \forall S \subset \{1, \ldots, n\}$$

Notice 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 subcollection $\{\mathcal{G}_{i_1}, \ldots, \mathcal{G}_{i_n}\}$ 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.
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 and identically distributed (i.i.d.) 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 i.i.d. random variables $X_1$ and $X_2$, each distributed according to the same distribution $\mu$ on $\mathbb{R}$, say. 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

$$[a_1, b_1] \times [a_2, b_2], \text{ with } a_1 < b_1 \text{ and } a_2 < b_2$$

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

$$\mathbb{P}([a_1, b_1] \times [a_2, b_2]) = \mu([a_1, b_1]) \mu([a_2, b_2])$$

on the rectangles. Caratheodory’s extension theorem then ensures 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

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

while

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

and similarly for $B_2$, 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:

$$[a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] \times \mathbb{R} \times \mathbb{R} \times \ldots$$

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

$$\mathbb{P}([a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] \times \mathbb{R} \times \mathbb{R} \times \ldots) = \mu([a_1, b_1]) \cdot \mu([a_2, b_2]) \cdots \mu([a_n, b_n])$$

and Caratheodory’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.
4 Expectation

4.1 Discrete non-negative random variables

Let us first define what the expectation (or mean) of a random variable is in the “simple” case where $X$ is a discrete non-negative random variable.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X$ be a discrete non-negative random variable defined on this space with values in a countable set $C$ (notice that $C \subseteq \mathbb{R}_+$, as $X$ is non-negative by assumption). Let also $(p_x, x \in C)$ denote its pmf (recall that $p_x = \mathbb{P}(\{X = x\})$, so that $0 \leq p_x \leq 1$ for every $x \in C$ and $\sum_{x \in C} p_x = 1$).

**Definition 4.1.** The expectation or expected value of $X$ is defined as follows:

$$E(X) = \sum_{x \in C} x \mathbb{P}(\{X = x\}) = \sum_{x \in C} x p_x$$

Being a countable sum of non-negative terms (as all $x \in C$ are non-negative), the above expectation can therefore take values in the interval $[0, +\infty]$, with $+\infty$ included (in case the series diverges). In order to illustrate this, let us consider some particular cases.

**Example 4.2.** A simple example first. Let $X$ be a Bernoulli random variable with parameter $0 < p < 1$, that is, $\mathbb{P}(\{X = 1\}) = p = 1 - \mathbb{P}(\{X = 0\})$. Then

$$E(X) = p \cdot 1 + (1 - p) \cdot 0 = p$$

So the expected value of $X$ is $p$. But what does that exactly mean, as $X$ never takes the value $p$? The answer will come in a few chapters.

**Example 4.3** (known as Saint-Petersburg’s paradox). Here is a more puzzling example. Let us consider an i.i.d. sequence of random variables $(X_n, n \geq 1)$ with $\mathbb{P}(\{X_1 = +1\}) = \mathbb{P}(\{X_1 = -1\}) = \frac{1}{2}$, and define the random variables $T = \inf\{n \geq 1 : X_n = +1\}$ and $G = 2^T$

This models the following game: you toss a coin multiple times, until “heads” ($X_n = +1$) comes out for the first time. Denote this time $T$. Then your gain is $2^T$. How much money would you be ready to pay to be allowed to play such a game? In general, the answer to such a question is the expected gain of the game. Let us compute this expected gain in the present case:

$$E(G) = E(2^T) = \sum_{n \geq 1} 2^n \mathbb{P}(\{T = n\}) = \sum_{n \geq 1} 2^n \mathbb{P}(\{X_1 = \ldots = X_{n-1} = -1, X_n = +1\}) = \sum_{n \geq 1} 2^n \frac{1}{2^n} = +\infty$$

where we have used the independence of the $X$’s. The expected gain being infinite, this is potentially saying that you would be ready to pay any finite amount of money to play such a game, and still believe you would be on the winning side, as you would pay less than your expected gain in any case. But would you, really? Probably no...

There are two ways to resolve this paradox:

1) A first idea is the following: the expected gain is infinite, because one assumes that you are ready to potentially toss a coin indefinitely, which is of course not the case in practice. If we modify the game by saying that there is a time horizon $N$ after which, assuming you had only obtained “tails” until then, you are out of the game and do not gain anything, then your expected gain becomes:

$$E(G) = \sum_{n=1}^{N} 2^n \frac{1}{2^n} + 0 = N$$
This seems already like a much more reasonable amount to invest on this game! Still, it is not clear that one would be willing to invest an amount of one million in order to get the right to toss a coin a million times, as most of the times, “heads” would come up after 1, 2, 3 or 4 tosses.

2) Another (and better) way out is the following: considering the expected gain as the fair price to pay for being allowed to play the game does not reflect one’s aversion to risk. Rather than paying $x = \mathbb{E}(G)$, many people would probably be ready to pay the amount $y$ satisfying

$$\sqrt{y} = \mathbb{E}({\sqrt{G}})$$

or even more risk-averse people would only consider paying the amount $z$ satisfying

$$\log_2(z) = \mathbb{E}({\log_2 G})$$

One can check that $y = \frac{1}{3-2\sqrt{2}} \approx 5.83$ and that $z = 4$. The square root and log functions are called utility functions. These are generally taken to be increasing and concave, with their particular shape reflecting one’s risk aversion.

Remark. Why did we restrict ourselves here to non-negative random variables? The advantage of this restriction is that in this case, the expectation is always well defined, even if possibly infinite. In the general case, the sum $\sum_{x \in C} x p_x$ may contain positive and negative terms, and may therefore “truly” diverge, making it impossible to define what $\mathbb{E}(X)$ is. More on this in the next section.

4.2 General definition

From the point of view of measure theory, random variables are maps from $\Omega$ to $\mathbb{R}$. Correspondingly, the expectation of a random variable $X$ is the Lebesgue integral of the map $X$, that is, the “area under the curve $\omega \mapsto X(\omega)$”, where the horizontal axis is measured with the probability measure $P$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X$ be a random variable defined on this probability space. The expectation of $X$, denoted as $\mathbb{E}(X)$, will be defined in three steps.

Step 1. Assume first that $X$ is a discrete non-negative (and by definition, $\mathcal{F}$-measurable) random variable. We already saw the definition of $\mathbb{E}(X)$ above, but let us observe here that such a random variable $X$ may be written as

$$X(\omega) = \sum_{i \geq 1} x_i 1_{A_i}(\omega), \quad \omega \in \Omega$$

where $x_i \geq 0$ are distinct and $A_i = \{\omega \in \Omega : X(\omega) = x_i\} \in \mathcal{F}$ are disjoint, as illustrated below:

![Diagram](image)

The expectation of $X$ is then defined as

$$\mathbb{E}(X) = \sum_{i \geq 1} x_i P(A_i) = \sum_{i \geq 1} x_i P(\{X = x_i\})$$

20
which corresponds to the definition seen above. It is therefore again the case that $E(X) \in [0, +\infty]$. Notice moreover that in the particular case where $X = 1_A$ with $A \in \mathcal{F}$ (which is nothing but a Bernoulli random variable), one has $E(X) = P(A)$.

**Step 2.** Assume now that $X$ is a generic $\mathcal{F}$-measurable non-negative random variable, i.e., $X(\omega) \geq 0$, $\forall \omega \in \Omega$. Let us define the following sequence of discrete random variables:

$$X_n(\omega) = \sum_{i \geq 1} \frac{i - 1}{2^n} 1_{A_i}^{(n)}(\omega)$$

where $A_i^{(n)} = \left\{ \omega \in \Omega : \frac{i - 1}{2^n} < X(\omega) \leq \frac{i}{2^n} \right\}$

Notice that $x_i = \frac{i - 1}{2^n} \geq 0$ and that $A_i^{(n)} \in \mathcal{F}$, since $X$ is $\mathcal{F}$-measurable. So according to Step 1, one has for each $n$

$$E(X_n) = \sum_{i \geq 1} \frac{i - 1}{2^n} P \left( A_i^{(n)} \right) = \sum_{i \geq 1} \frac{i - 1}{2^n} P \left( \left\{ \frac{i - 1}{2^n} < X \leq \frac{i}{2^n} \right\} \right) \in [0, +\infty].$$

It should be observed that $(X_n, n \in \mathbb{N})$ is actually a non-decreasing sequence of non-negative “staircases”, that is,

$$0 \leq X_n(\omega) \leq X_{n+1}(\omega), \quad \forall n$$

The staircase gets indeed refined at each step, as the size of the steps is divided by 2 from $n$ to $n+1$, as illustrated below:

![Staircase Diagram](image-url)

Likewise, one easily sees that $E(X_n) \leq E(X_{n+1})$ for all $n$, so $(E(X_n), n \in \mathbb{N})$ is an non-decreasing sequence, that therefore converges (possibly to $+\infty$). One defines

$$E(X) = \lim_{n \to \infty} E(X_n) = \lim_{n \to \infty} \sum_{i \geq 1} \frac{i - 1}{2^n} \cdot P \left( \left\{ \frac{i - 1}{2^n} < X \leq \frac{i}{2^n} \right\} \right) \in [0, +\infty]$$

**Step 3.** Finally, consider a generic $\mathcal{F}$-measurable random variable $X$. One defines its *positive and negative parts:*

$$X^+(\omega) = \max(0, X(\omega)), \quad X^-(\omega) = \max(0, -X(\omega))$$

Notice that both $X^+(\omega) \geq 0$ and $X^-(\omega) \geq 0$, and that

$$X^+(\omega) - X^-(\omega) = X(\omega), \quad X^+(\omega) + X^-(\omega) = |X(\omega)|$$

In measure theory, one does not want to deal with ill-defined quantities such as $\infty - \infty$. One therefore defines $E(X)$ only when $E(|X|) = E(X^+) + E(X^-) < +\infty$, using the formula:

$$E(X) = E(X^+) - E(X^-)$$
Simplified expressions in two important particular cases. [without proofs]
Let \( X \) be a random variable and \( g : \mathbb{R} \to \mathbb{R} \) be a Borel-measurable function.
- If \( X \) is a discrete random variable with values in \( C \) and \( \mathbb{E}(|g(X)|) = \sum_{x \in C} |g(x)| \mathbb{P}({X = x}) < +\infty \), then
  \[
  \mathbb{E}(g(X)) = \sum_{x \in C} g(x) \mathbb{P}({X = x})
  \]
- If \( X \) is a continuous random variable with pdf \( p_X \) and \( \mathbb{E}(|g(X)|) = \int_{\mathbb{R}} |g(x)| p_X(x) \, dx < +\infty \), then
  \[
  \mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) p_X(x) \, dx
  \]

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 \( c > 0 \) such that \( |X(\omega)| \leq c, \forall \omega \in \Omega \), then \( X \) is said to be a bounded random variable.
- If \( \mathbb{E}(X) = 0 \), then \( X \) is said to be a centered random variable.

One has the following series of implications:

\[
\begin{align*}
X \text{ is bounded} & \Rightarrow X \text{ is square-integrable} \Rightarrow X \text{ is integrable} \\
X \text{ is integrable and } Y \text{ is bounded} & \Rightarrow XY \text{ is integrable} \\
X, Y \text{ are both square-integrable} & \Rightarrow XY \text{ is integrable}
\end{align*}
\]

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 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 \( |X(\omega)Y(\omega)| \leq C \cdot |X(\omega)| \), so \( XY \) is also integrable. The other implications follow from Cauchy-Schwarz’ inequality (see next section).

Basic properties. [without proofs]

Linearity. If \( c \in \mathbb{R} \) is a constant and \( X, Y \) are integrable, then
\[
\mathbb{E}(cX) = c \mathbb{E}(X) \quad \text{and} \quad \mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)
\]

Remark. As simple as the above statement looks, it is actually not so easy to prove, essentially because there are many different “simple” representations of the random variable \( X + Y \).

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 \( \mathbb{E}(X) \geq \mathbb{E}(Y) \).

Variance, covariance and independence.

Definition 4.4. Let \( X, Y \) be two square-integrable random variables. The variance of \( X \) is defined as
\[
\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0
\]
and the covariance of \( X \) and \( Y \) is defined as
\[
\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
\]
so that \( \text{Cov}(X, X) = \text{Var}(X) \).
**Terminology.** If 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) \(\text{Var}(cX) = c^2 \text{Var}(X)\).

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

In addition, if \(X, Y\) are independent, then

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

d) \(\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)\).

### 4.3 Inequalities

**Cauchy-Schwarz’s inequality.** If \(X, Y\) are square-integrable random variables, then the product \(XY\) is integrable and

\[
|E(XY)| \leq E(|XY|) \leq \sqrt{E(X^2)} \sqrt{E(Y^2)}
\]

In particular, considering \(Y = 1\) shows that if \(X\) is square-integrable, then it is also integrable.

**Proof.** Notice that the first inequality follows either from the definition of expectation or from Jensen’s inequality (see below). Consider now the map \(\alpha \mapsto f(\alpha) = E((|X| + \alpha |Y|)^2)\). Notice first that \(f(\alpha) \leq 2E(X^2 + \alpha^2 Y^2) < +\infty\) 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

\[
f(\alpha) = E(X^2) + 2\alpha E(|XY|) + \alpha^2 E(Y^2)
\]

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

\[
\Delta = (E(|XY|)^2 - E(X^2) E(Y^2)) \leq 0
\]

which implies the desired inequality:

\[
E(|XY|) \leq \sqrt{E(X^2)} \sqrt{E(Y^2)}
\]

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

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

$$\varphi(E(X)) \leq E(\varphi(X))$$

In particular: $|E(X)| \leq E(|X|), E(X)^2 \leq E(X^2), \exp(E(X)) \leq E(\exp(X)), \exp(-E(X)) \leq E(\exp(-X))$ and for $X$ a positive random variable, $\log(E(X)) \geq E(\log(X))$ (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 $\varphi$:

\[\begin{array}{c}
\varphi(pa + (1-p)b) \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot 
\end{array}\]

\[\begin{array}{c}
\varphi(a) \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
\cdot 
\end{array}\]

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 $x_0 = E(X)$. We obtain, by linearity of expectation:

$$E(\varphi(X)) \geq E(aX + b) = aE(X) + b = ax_0 + b = \varphi(x_0) = \varphi(E(X))$$

which is the desired inequality. \qed
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.,

\[
\frac{x_1 + \ldots + x_n}{n} \geq (x_1 \cdot \ldots \cdot x_n)^{1/n} \geq \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}}.
\]

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

\[
E(X) \geq \exp(E(\log(X))) \geq \frac{1}{E(\frac{1}{X})}
\]

The inequality on the left-hand side follows from Jensen’s inequality with \( \phi(x) = -\log(x) \), while the one on the right-hand side is equivalent to (writing \( X = \exp(Y) \)):

\[
\exp(E(Y)) \geq \frac{1}{E(\exp(-Y))} \quad \text{or} \quad E(\exp(-Y)) \geq \exp(-E(Y))
\]

which is Jensen’s inequality with \( \phi(x) = \exp(-x) \) this time. \( \square \)

**Chebyshev’s inequality.** If \( X \) is a random variable and \( \psi : \mathbb{R} \to \mathbb{R}_+ \) is Borel-measurable, non-decreasing on \( \mathbb{R}_+ \) and such that \( E(\psi(X)) < +\infty \), then for any \( a \geq 0 \) such that \( \psi(a) > 0 \), one has

\[
P(\{X \geq a\}) \leq \frac{E(\psi(X))}{\psi(a)}
\]

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

\[
\psi(x) \text{ Example of } \psi \text{ satisfying the assumptions}
\]

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

\[
P(\{X \geq a\}) = E(1_{\{X \geq a\}}) \leq E\left( \frac{\psi(X)}{\psi(a)} 1_{\{X \geq a\}} \right) = \frac{E(\psi(X) 1_{\{X \geq a\}})}{\psi(a)} \leq \frac{E(\psi(X))}{\psi(a)}
\]

which is the desired inequality. \( \square \)
5 Laws of large numbers

5.1 Preliminary: convergence of sequences of numbers

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

\[
\forall \varepsilon > 0, \exists N \geq 1 \text{ such that } \forall n \geq N, |a_n - a| \leq \varepsilon
\]

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

\[
\exists \varepsilon > 0 \text{ such that } \forall N \geq 1, \exists n \geq N \text{ such that } |a_n - a| > \varepsilon
\]

This is still equivalent to saying that

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

In the previous sentence, “for an infinite number of values of \(n\)” may be abbreviated as “infinitely often”.

5.2 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 \((X_n, n \geq 1)\) 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 \((X_n, n \geq 1)\) is said to converge in quadratic mean to \(X\) (and this is denoted as \(X_n \xrightarrow{L^2} X\)) if

\[
\mathbb{E}(|X_n - X|^2) \rightarrow_{n \rightarrow \infty} 0
\]

2) Convergence in probability. The sequence \((X_n, n \geq 1)\) is said to converge in probability to \(X\) (and this is denoted as \(X_n \xrightarrow{P} X\)) if

\[
\forall \varepsilon > 0, \quad \mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) \rightarrow_{n \rightarrow \infty} 0
\]

3) Almost sure convergence. The sequence \((X_n, n \geq 1)\) is said to converge almost surely to \(X\) (and this is denoted as \(X_n \rightarrow X\) a.s.) if

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

5.3 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 \(X_n \xrightarrow{L^2} X\). Then we have for any fixed \(\varepsilon > 0\):

\[
\mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) \leq \frac{\mathbb{E}(|X_n - X|^2)}{\varepsilon^2} \rightarrow_{n \rightarrow \infty} 0
\]

by assumption, so \(X_n \xrightarrow{P} X\).
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 5.1.** Let $(X_n, n \geq 1)$ be a sequence of random variables and $X$ be a random variable such that $X_n \xrightarrow{p} X$. Assume moreover there exists $C > 0$ such that $|X_n(\omega)| \leq C$ for all $\omega \in \Omega$ and $n \geq 1$. Then $X_n \xrightarrow{L^2} X$.

Almost sure convergence implies convergence in probability. In order to show this, we will need two facts that can be easily deduced from the basic axioms on probability measures, as well as a lemma that provides an alternate characterization of almost sure convergence.

**Fact 1.** $\mathbb{P}(A_M) = 0$ for all $M \geq 1$ if and only if $\mathbb{P}(\bigcup_{M \geq 1} A_M) = 0$.

**Fact 2.** $\mathbb{P} (\bigcap_{N \geq 1} A_N) = \lim_{N \to \infty} \mathbb{P}(A_N)$ if $A_N \supset A_{N+1}$ for all $N \geq 1$.

**Lemma 5.2.** (characterization of almost sure convergence)

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

*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 convergence of sequences of numbers, we obtain the following series of equivalences:

$$
X_n \xrightarrow{n \to \infty} X \quad \text{a.s.} \iff \mathbb{P} \left( \left\{ \lim_{n \to \infty} X_n = X \right\} \right) = 1 \iff \mathbb{P} \left( \left\{ \lim_{n \to \infty} X_n \neq X \right\} \right) = 0
$$

$$
\iff \mathbb{P} (\{ \exists \varepsilon > 0 \text{ such that } |X_n - X| > \varepsilon \text{ i.o.} \}) = 0
$$

$$
\iff \mathbb{P} \left( \exists M \geq 1 \text{ such that } |X_n - X| > \frac{1}{M} \text{ i.o.} \right) = 0
$$

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

By Fact 1, this last assertion is equivalent to saying that

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

which is in turn equivalent to saying that

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

and completes the proof. Notice 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 5.2,

$$
X_n \xrightarrow{n \to \infty} X \quad \text{a.s.} \iff \forall \varepsilon > 0, \quad \mathbb{P} (\{ \omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon \text{ infinitely often} \}) = 0
$$

$$
\iff \forall \varepsilon > 0, \quad \mathbb{P} (\{ \forall N \geq 1, \exists n \geq N \text{ such that } |X_n - X| > \varepsilon \}) = 0
$$

$$
\iff \forall \varepsilon > 0, \quad \mathbb{P} \left( \bigcap_{N \geq 1} \bigcup_{n \geq N} \{|X_n - X| > \varepsilon\} \right) = 0
$$

27
By Fact 2, this is equivalent to saying that \( \forall \varepsilon > 0, \lim_{N \to \infty} P \left( \bigcup_{n \geq N} \{ |X_n - X| > \varepsilon \} \right) = 0 \), and this last assertion implies that \( \forall \varepsilon > 0, \lim_{N \to \infty} P (\{|X_N - X| > \varepsilon\}) = 0 \). Said otherwise: \( X_N \xrightarrow{P} X \). \( \square \)

**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:

$$X_1 = 1_H, \quad X_2 = 1_T, \quad X_3 = 1_H H, \quad X_4 = 1_H T, \quad X_5 = 1_T H, \quad X_6 = 1_T T, \quad X_7 = 1_H H H, \quad X_8 = 1_H H T, \ldots$$

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, \( X_n \xrightarrow{P} 0 \). Indeed, one can check that for any \( \varepsilon > 0 \), the probability

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

It therefore converges to 0 as \( n \to \infty \).

On the other hand, \( X_n \not\xrightarrow{a.s.} 0 \). 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 1010001000000... 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,

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

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

**Remark.** As we know that \( X_n \xrightarrow{P} 0 \), 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 5.3.** Let \( (X_n, n \geq 1) \) be a sequence of random variables and \( X, Y \) be two random variables such that \( X_n \xrightarrow{P} X \) and \( X_n \xrightarrow{P} Y \). Then \( X = Y \) a.s.

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

$$P(|X - Y| > \varepsilon) = P(|X - X_n + X_n - Y| > \varepsilon) \leq P(|X_n - X| > \frac{\varepsilon}{2}) + P(|X_n - Y| > \frac{\varepsilon}{2}) \xrightarrow{n \to \infty} 0$$

by the assumptions made. So \( P(|X - Y| > \varepsilon) = 0 \) for any \( \varepsilon > 0 \), therefore \( P(|X - Y| = 0) = 1 \), which proves the claim. \( \square \)

### 5.4 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 positive 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 5.4. (Borel-Cantelli)
Let \((A_n, n \geq 1)\) be a sequence of events in \(\mathcal{F}\) such that \(\sum_{n \geq 1} P(A_n) < \infty\). Then
\[
P(\{\omega \in \Omega : \omega \in A_n \text{ infinitely often}\}) = 0
\]
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 \(\varepsilon > 0\), \(\mathbb{P}(\{|X_n - X| > \varepsilon\}) \to 0\) as \(n \to \infty\), then \(X_n \xrightarrow{p} X\), by definition.

b) If for all \(\varepsilon > 0\), \(\sum_{n \geq 1} \mathbb{P}(\{|X_n - X| > \varepsilon\}) < \infty\), then \(X_n \xrightarrow{a.s.} X\). Indeed, by the Borel-Cantelli lemma, the condition on the sum implies that for all \(\varepsilon > 0\),
\[
P(\{|X_n - X| > \varepsilon \text{ infinitely often}\}) = 0
\]
which is exactly the characterization of almost sure convergence given in Lemma 5.2.

So we see that if one can prove that \(\mathbb{P}(\{|X_n - X| > \varepsilon\}) = O(\frac{1}{n})\) 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
\[
\mathbb{P}(\{\omega \in \Omega : \omega \in A_n \text{ infinitely often}\}) = \mathbb{P}(\{\forall N \geq 1, \exists n \geq N \text{ such that } \omega \in A_n\})
\]
\[
= \mathbb{P}\left(\bigcap_{N \geq 1} \bigcup_{n \geq N} A_n\right) = \lim_{N \to \infty} \mathbb{P}\left(\bigcup_{n \geq N} A_n\right)
\]
by Fact 2. Using the union bound, this implies that
\[
\mathbb{P}(\{\omega \in \Omega : \omega \in A_n \text{ infinitely often}\}) \leq \lim_{N \to \infty} \sum_{n \geq N} \mathbb{P}(A_n) = 0
\]
by the assumption made on the sum. This proves the Borel-Cantelli lemma. \(\Box\)

5.5 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 5.5. 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) \(\frac{S_n}{n} \xrightarrow{p} \mathbb{E}(X_1)\) (weak law).

b) \(\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}(X_1)\) almost surely (strong law).
Remarks. - 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 \( \text{Var}(X_1) \) 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.

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

\[
\mathbb{P} \left( \left\{ \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > \varepsilon \right\} \right) = \mathbb{P} \left( \left\{ |S_n - n\mathbb{E}(X_1)| > n\varepsilon \right\} \right) = \mathbb{P} \left( \left\{ |S_n - \mathbb{E}(S_n)| > n\varepsilon \right\} \right)
\leq \frac{\mathbb{E}((S_n - \mathbb{E}(S_n))^2)}{n^2\varepsilon^2} = \frac{\text{Var}(S_n)}{n^2\varepsilon^2} = \frac{\text{Var}(X_1)}{n\varepsilon^2} \xrightarrow{n \to \infty} 0
\]

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 \( \frac{S_n}{n} \xrightarrow{n \to \infty} \mathbb{E}(X_1) \) and therefore proves the weak law of large numbers.

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

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

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:

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

Using the Borel-Cantelli lemma and the fact that \( \sum_{n \geq 1} \frac{1}{n^2} < \infty \), we obtain that \( \frac{S_{n^2}}{n^2} \xrightarrow{n \to \infty} \mathbb{E}(X_1) \) 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

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

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

\[
S_n^+ = \sum_{j=1}^{n} X_j^+ \quad \text{and} \quad S_n^- = \sum_{j=1}^{n} X_j^- , \quad \text{so that} \quad S_n = S_n^+ - S_n^-
\]
Notice 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:

\[
\frac{S_n^+}{n} \to_n \mathbb{E}(X_1^+) \quad \text{a.s. and} \quad \frac{S_n^-}{n} \to_n \mathbb{E}(X_1^-) \quad \text{a.s.}
\]

which in turn implies that

\[
\frac{S_n}{n} = \frac{S_n^+}{n} - \frac{S_n^-}{n} \to_n \mathbb{E}(X_1^+) - \mathbb{E}(X_1^-) = \mathbb{E}(X_1) \quad \text{a.s.}
\]

and therefore completes the proof. \( \square \)

### 5.6 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 \(F_n(t) = \frac{1}{n} \sum_{1 \leq j \leq n : X_j \leq t} 1\) for \(t \in \mathbb{R} \); \(F_n\) is the empirical distribution (or cdf) of the first \(n\) random variables \(X_1, \ldots, X_n\). Notice 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 5.6.** For every \(t \in \mathbb{R}\),

\[
F_n(t) \to_n F(t) \quad \text{almost surely}
\]

**Proof.** Fix \(t \in \mathbb{R}\) and let \(Y_j = \mathbb{1}_{\{X_j \leq t\}}\). 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

\[
F_n(t) = \frac{1}{n} \sum_{j=1}^{n} Y_j \to_n \mathbb{E}(Y_1) \quad \text{almost surely}
\]

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. \( \square \)

### 5.7 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 \(\lim_{n \to \infty} \frac{S_n}{n} = \mathbb{E}(X_1)\) a.s.

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

The aim here is not to give a full proof of the above statements, but rather to explain why \(\frac{S_n}{n}\) 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 behavior of the sequence when \( n \to \infty \). 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, notice 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 \( \mathcal{T} = \bigcap_{N \geq 1} \mathcal{G}_N \).

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 5.7.** (Kolmogorov’s 0-1 law)
If the sequence \((X_n, n \geq 1)\) is independent and \( A \in \mathcal{T} \), then \( \mathbb{P}(A) \in \{0, 1\} \).

**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 5.7.** 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 notice 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 implies (by definition) that \( \mathcal{T} \) is also independent of \( \sigma(X_1, X_2, \ldots, X_n, \ldots) \), 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.

\[\square\]

### 5.8 Extension of the weak law: an example

Let us consider again the Saint-Petersburg game, whose gain distribution is given by \( \mathbb{P}(\{G_1 = 2^k\}) = \frac{1}{2^k} \), for \( k \geq 1 \). Assume now you are allowed to play \( n \) times this game, 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 reasonable answer can still be obtained via the following proposition.
Proposition 5.8. Under the above assumptions, it holds that \( \frac{S_n}{n \log_2 n} \xrightarrow{p} 1 \), i.e., that for any \( \varepsilon > 0 \),

\[
P \left( \left\{ \left| \frac{S_n}{n \log_2 n} - 1 \right| > \varepsilon \right\} \right) \xrightarrow{n \to \infty} 0
\]

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. Of course, how much does one deviate at finite \( n \) from this average behaviour is still an open question (to be addressed later). We now give the proof of the 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 \( n \): \( H_j = G_j \mathbf{1}_{\{G_j \leq n \log_2 n\}} \), \( 1 \leq j \leq n \), 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

\[
P \left( \left| \frac{S_n}{n \log_2 n} - 1 \right| > \varepsilon \right) = P(\{|S_n - n \log_2 n| > \varepsilon n \log_2 n\})
\]

= \( P(\{|S_n - n \log_2 n| > \varepsilon n \log_2 n, \ G_j \leq n \log_2 n, \ \forall 1 \leq j \leq n\}) + P(\{|S_n - n \log_2 n| > \varepsilon n \log_2 n, \ \exists 1 \leq j \leq n \text{ such that } G_j > n \log_2 n\}) \)

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

\[
P(\{|S_n' - n \log_2 n| > \varepsilon n \log_2 n\}) + P(\{\exists 1 \leq j \leq n \text{ such that } G_j > n \log_2 n\})
\]

(4)

\( S_n' = H_1 + \ldots + H_n \) is a sum of bounded random variables, and one can check that

\[
E(H_1) = \sum_{k \leq \log_2(n \log_2 n)} 2^k \frac{1}{2^k} = \sum_{k \leq \log_2(n \log_2 n)} 1 \approx \log_2 n
\]

so that \( E(S_n') \approx n \log_2 n \) and that, using Chebyshev's inequality, we obtain:

\[
P(\{|S_n' - n \log_2 n| > \varepsilon n \log_2 n\}) \leq \frac{\text{Var}(S_n')}{(\varepsilon n \log_2 n)^2} = \frac{n \text{Var}(H_1)}{(\varepsilon n \log_2 n)^2}
\]

where

\[
\text{Var}(H_1) = E(H_1^2) = \sum_{k \leq \log_2(n \log_2 n)} 2^{2k} \frac{1}{2^k} = \sum_{k \leq \log_2(n \log_2 n)} 2^k \approx 2n \log_2 n
\]

Therefore,

\[
P(\{|S_n' - n \log_2 n| > \varepsilon n \log_2 n\}) \leq \frac{2n^2 \log_2 n}{(\varepsilon n \log_2 n)^2} = \frac{2}{\varepsilon^2 \log_2 n} \xrightarrow{n \to \infty} 0
\]

The second term in (4) can be upperbounded using the union bound:

\[
P(\{\exists 1 \leq j \leq n \text{ such that } G_j > n \log_2 n\}) \leq n P(\{G_1 > n \log_2 n\}) = n \sum_{k > \log_2(n \log_2 n)} \frac{1}{2^k} \approx \frac{1}{\log_2 n} \xrightarrow{n \to \infty} 0
\]

which completes the proof. \(\square\)
6 The central limit theorem

6.1 Convergence in distribution

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

Definition 6.1. Let \((X_n, n \geq 1)\) be a sequence of random variables, **not necessarily** defined on the same probability space \((\Omega, \mathcal{F}, \mathbb{P})\). The sequence \((X_n, n \geq 1)\) is said to **converge in distribution** to a limiting random variable \(X\) (and this is denoted as \(X_n \xrightarrow{d} X\)) if

\[
F_{X_n}(t) = \mathbb{P}(\{X_n \leq t\}) \xrightarrow{n \to \infty} F_X(t) = \mathbb{P}(\{X \leq t\})
\]

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

Remark 6.2. 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.

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

Proposition 6.3. Let \((X_n, n \geq 1)\) 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 \(X_n \xrightarrow{d} X\), then \(X_n \xrightarrow{d} X\).

**Proof.** Recall that \(X_n \xrightarrow{d} X\) means that for all \(\varepsilon > 0\), \(\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0\). 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 \(\varepsilon > 0\):

\[
F_{X_n}(t) = \mathbb{P}(\{X_n \leq t\}) = \mathbb{P}(\{X_n \leq t, X \leq t + \varepsilon\}) + \mathbb{P}(\{X_n \leq t, X > t + \varepsilon\})
\]

\[
\leq \mathbb{P}(\{X \leq t + \varepsilon\}) + \mathbb{P}(\{|X_n - X| > \varepsilon\})
\]

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

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

\[
F_X(t - \varepsilon) = \mathbb{P}(\{X \leq t - \varepsilon\}) = \mathbb{P}(\{X \leq t - \varepsilon, X_n \leq t\}) + \mathbb{P}(\{X \leq t - \varepsilon, X_n > t\})
\]

\[
\leq \mathbb{P}(\{X_n \leq t\}) + \mathbb{P}(\{|X_n - X| > \varepsilon\})
\]

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

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

\[
F_X(t - \varepsilon) \leq \liminf_{n \to \infty} F_{X_n}(t) \leq \limsup_{n \to \infty} F_{X_n}(t) \leq F_X(t + \varepsilon)
\]

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. \(\square\)
**Remark 6.4.** Let us insist here that convergence in distribution is a much weaker notion than convergence in probability. For example, a sequence \( (X_n, n \geq 1) \) 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 \( (F_{X_n}(t), n \geq 1) \) necessarily converges for every \( t \in \mathbb{R} \).

### 6.2 Equivalent criterion for convergence in distribution

The following theorem, also known as (part of) the “Portemanteau theorem”, gives an equivalent criterion for convergence in distribution.

**Theorem 6.5.** Let \( (X_n, n \geq 1) \) be a sequence of random variables and \( X \) be another random variable. Then \( X_n \xrightarrow{d} X \) if and only if

\[
\mathbb{E}(g(X_n)) \xrightarrow{n \to \infty} \mathbb{E}(g(X))
\]

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

**Proof sketch of the “if” part.** Observe first that Definition 6.1 is equivalent to saying that

\[
\mathbb{E}(h_t(X_n)) \xrightarrow{n \to \infty} \mathbb{E}(h_t(X)),
\]

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, let us define for \( m \geq 1 \):

\[
g_m(x) = \begin{cases} 
1 & \text{if } x \leq t \left(1 - \frac{1}{m}\right) \\
1 - \frac{x - t}{m} & \text{if } \left(1 - \frac{1}{m}\right) < x \leq t \\
0 & \text{if } x > t
\end{cases}
\]

and

\[
G_m(x) = \begin{cases} 
1 & \text{if } x \leq t \\
1 - \frac{x - t}{m} & \text{if } t < x \leq (1 + \frac{1}{m}) \\
0 & \text{if } x > (1 + \frac{1}{m})
\end{cases}
\]

As one easily sees from the figure above, the functions \( g_m \) and \( G_m \) are continuous and bounded and \( g_m(x) \leq h_t(x) \leq G_m(x) \) for every \( x \in \mathbb{R} \) and every \( m \geq 1 \). Also, \( \lim_{m \to \infty} g_m(x) = \lim_{m \to \infty} G_m(x) = h_t(x) \) for every \( x \neq t \). From the assumption made, we obtain that for every \( m \geq 1 \):

\[
\mathbb{E}(g_m(X)) = \lim_{n \to \infty} \mathbb{E}(g_m(X_n)) \leq \lim_{n \to \infty} \mathbb{E}(h_t(X_n)) \leq \lim_{n \to \infty} \mathbb{E}(G_m(X_n)) = \mathbb{E}(G_m(X))
\]

A last technical but important detail (not shown here) allows to conclude that for \( t \) a continuity point of \( F_X \), both \( \mathbb{E}(g_m(X)) \xrightarrow{m \to \infty} \mathbb{E}(h_t(X)) \) and \( \mathbb{E}(G_m(X)) \xrightarrow{m \to \infty} \mathbb{E}(h_t(X)) \), implying the result. \( \square \)

**Remark 6.6.** For \( k \in \mathbb{N} \), let \( C^k_b(\mathbb{R}) \) 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:

If \( \mathbb{E}(g(X_n)) \xrightarrow{n \to \infty} \mathbb{E}(g(X)) \) for every function \( g \in C^3_b(\mathbb{R}) \), then \( X_n \xrightarrow{d} X \)

This remark will be useful below.
6.3 The central limit theorem

Let \((X_n, n \geq 1)\) be a sequence of i.i.d. square-integrable random variables. Let also \(\mu = \mathbb{E}(X_1), \sigma^2 = \text{Var}(X_1)\) and \(S_n = X_1 + \ldots + X_n\). Using basic properties of expectation and variance, we obtain the following:

\[
\mathbb{E}(S_n) = n \mathbb{E}(X_1) = n \mu \quad \text{and} \quad \text{Var}(S_n) = n \text{Var}(X_1) = n \sigma^2
\]

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:

\[
S_n = n \mu + \sqrt{n} \sigma \tilde{S}_n,
\]

where \(\tilde{S}_n\) is a random variable with \(\mathbb{E}(\tilde{S}_n) = 0\) and \(\text{Var}(\tilde{S}_n) = \mathbb{E}(\tilde{S}_n^2) = 1\). The central limit theorem states that as \(n\) grows large, \(\tilde{S}_n\) converges in distribution to a Gaussian random variable with zero mean and unit variance. This result is therefore \textit{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 6.7.** 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 \(\mu = \mathbb{E}(X_1), \sigma^2 = \text{Var}(X_1), S_n = X_1 + \ldots + X_n\) and

\[
\tilde{S}_n = \frac{S_n - n \mu}{\sqrt{n} \sigma}, \quad n \geq 1
\]

Then \(\tilde{S}_n \xrightarrow{d} n \to \infty Z \sim \mathcal{N}(0, 1)\), i.e.,

\[
\mathbb{P}(\{\tilde{S}_n \leq t\}) \xrightarrow[n \to \infty]{} \mathbb{P}(\{Z \leq t\}) = \int_{-\infty}^{t} \frac{1}{\sqrt{2 \pi}} e^{-x^2/2} dx, \quad \forall t \in \mathbb{R}
\]

Note that convergence has to take place here for \textit{every} \(t \in \mathbb{R}\), as the limiting cdf is continuous.

**Remark 6.8.** 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.

**Lemma 6.9.** Let \(g \in C^3_b(\mathbb{R})\) 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

$$E(g(X + Y)) - E(g(X + Z)) = \frac{1}{6} \left( E(Y^3 g'''(U)) - E(Z^3 g'''(V)) \right)$$

so

$$|E(g(X + Y)) - E(g(X + Z))| \leq \frac{C}{6} \left( E(|Y|^3) + E(|Z|^3) \right)$$

which completes the proof.

The above lemma generalizes to the case of multiple random variables, as shown below.

**Lemma 6.10.** Let $g \in C^3_b(\mathbb{R})$ and $Y_1, \ldots, Y_n, Z_1, \ldots, Z_n$ be random variables, all independent and such that $E(Y_i) = E(Z_i)$, $E(Y_i^2) = E(Z_i^2)$ and $E(|Y_i|^3) < +\infty, E(|Z_i|^3) < +\infty$ for all $i \in \{1, \ldots, n\}$. Then

$$|E(g(Y_1 + \ldots + Y_n)) - E(g(Z_1 + \ldots + Z_n))| \leq \frac{C}{6} \sum_{i=1}^n (E(|Y_i|^3) + E(|Z_i|^3))$$

where $C = \sup_{x \in \mathbb{R}} |g'''(x)|$.

**Proof.** Define

$$X_1 = Z_2 + \ldots + Z_n, \quad X_n = Y_1 + \ldots + Y_{n-1}, \quad \text{and} \quad X_i = Y_1 + \ldots + Y_{i-1} + Z_{i+1} + \ldots + Z_n \quad \text{for } i \in \{2, \ldots, n-1\}$$

Observe then that

$$Y_1 + \ldots + Y_n = X_n + Y_n, \quad Z_1 + \ldots + Z_n = X_1 + Z_1, \quad \text{and} \quad X_i + Y_i = Y_1 + \ldots + Y_i + Z_{i+1} + \ldots + Z_n = X_{i+1} + Z_{i+1}, \quad \text{for } i \in \{1, \ldots, n-1\}$$

so that

$$|E(g(Y_1 + \ldots + Y_n)) - E(g(Z_1 + \ldots + Z_n))| = |E(g(X_n + Y_n)) - E(g(X_1 + Z_1))| = |E(g(X_n + Y_n)) - E(g(X_n + Z_n)) + E(g(X_{n-1} + Y_{n-1})) + \ldots$$

$$\ldots - E(g(X_2 + Z_2)) + E(g(X_1 + Y_1)) - E(g(X_1 + Z_1))|$$

$$\leq \sum_{i=1}^n |E(g(X_i + Y_i)) - E(g(X_i + Z_i))| \leq \frac{C}{6} \sum_{i=1}^n (E(|Y_i|^3) + E(|Z_i|^3))$$

by repeated uses of Lemma 6.9. The proof is complete.

As the above lemma is valid for any function $g \in C^3_b(\mathbb{R})$, 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 6.7.

**Proof of Theorem 6.7.** Let us first estimate $E(g(\tilde{S}_n))$ for $g \in C^3_b(\mathbb{R})$ and $n$ large. Defining $Y_i = \frac{X_i - \mu}{\sqrt{n} \sigma}$, we may rewrite

$$\tilde{S}_n = \frac{S_n - n \mu}{\sqrt{n} \sigma} = \sum_{i=1}^n \frac{X_i - \mu}{\sqrt{n} \sigma} = \sum_{i=1}^n Y_i$$

The random variables $Y_i$ are i.i.d. with $E(Y_i) = 0$, $E(Y_i^2) = \frac{1}{n}$ and $E(|Y_i|^3) = O(n^{-3/2})$. 37
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 6.10, we have

$$|\mathbb{E}(g(Y_1 + \ldots + Y_n)) - \mathbb{E}(g(Z_1 + \ldots + Z_n))| \leq \frac{C}{6} \sum_{i=1}^{n} \left( \mathbb{E}(|Y_i|^3) + \mathbb{E}(|Z_i|^3) \right)$$

$$\leq \frac{C}{6} \sum_{i=1}^{n} O(n^{-3/2}) = O(n^{-1/2}) \to 0$$

Observe next that as $\text{Var}(Z_1 + \ldots + Z_n) = \frac{1}{n} + \ldots + \frac{1}{n} = 1$ 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 6.5 (or more precisely Remark 6.6 following it), we finally deduce that

$$S_n = Y_1 + \ldots + Y_n \overset{d}{\to} Z \quad n \to \infty$$

which completes the proof of the theorem. □.

# 7 Concentration inequalities

The weak law of large numbers states that $\frac{S_n}{n}$ 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 $t > 0$,

$$\mathbb{P} \left( \left\{ \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > t \right\} \right) \to 0$$

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.

## 7.1 Hoeffding’s inequality

**Theorem 7.1.** 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

$$\mathbb{P} \left( \left\{ \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > t \right\} \right) \leq 2 \exp \left( -\frac{nt^2}{2} \right), \quad \forall t > 0, n \geq 1$$

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

$$\sum_{n \geq 1} \mathbb{P} \left( \left\{ \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > t \right\} \right) \leq 2 \sum_{n \geq 1} \exp \left( -\frac{nt^2}{2} \right) < \infty$$

which implies by the Borel-Cantelli lemma that

$$\mathbb{P} \left( \left\{ \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > t \text{ infinitely often} \right\} \right) = 0$$

This therefore says that $\frac{S_n}{n} \overset{n \to \infty}{\to} \mathbb{E}(X_1)$ 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

$$\Pr \left( \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > \frac{u}{\sqrt{n}} \right) \leq 2 \exp \left( -\frac{u^2}{2} \right)$$

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 7.1.

**Lemma 7.2.** Let $Z$ be a random variable such that $|Z(\omega)| \leq 1$ for all $\omega \in \Omega$ and $\mathbb{E}(Z) = 0$. Then

$$\mathbb{E}(e^{sZ}) \leq \exp \left( \frac{s^2}{2} \right), \quad \forall s \in \mathbb{R}$$

**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

$$\mathbb{E}(e^{sZ}) \leq \cosh(s) + \mathbb{E}(Z) \sinh(s) = \cosh(s)$$

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 7.1. Let us compute

\[
\mathbb{P} \left( \left\{ \left| \frac{S_n}{n} - \mathbb{E}(X_1) \right| > t \right\} \right) = \mathbb{P} \left( \left\{ |S_n - n \mathbb{E}(X_1)| > nt \right\} \right) = \mathbb{P} \left( \left\{ |S_n - \mathbb{E}(S_n)| > nt \right\} \right)
\]

\[
= \mathbb{P} \left( \{S_n - \mathbb{E}(S_n) > nt\} \right) + \mathbb{P} \left( \{S_n - \mathbb{E}(S_n) < -nt\} \right)
\]

(5)

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

\[
\mathbb{P} \left( \{S_n - \mathbb{E}(S_n) > nt\} \right) \leq \frac{\mathbb{E}(e^{s(S_n - \mathbb{E}(S_n))})}{e^{nst}} = e^{-nst} \mathbb{E} \left( \prod_{j=1}^{n} e^{s(X_j - \mathbb{E}(X_j))} \right) = e^{-nst} \left( \mathbb{E} \left( e^{s(X_1 - \mathbb{E}(X_1))} \right) \right)^n
\]

By the assumptions made, the random variable \(Z = X_1 - \mathbb{E}(X_1)\) satisfies the assumptions of Lemma 7.2, so \(\mathbb{E} \left( e^{s(X_1 - \mathbb{E}(X_1))} \right) \leq e^{s^2/2}\). This implies finally that

\[
\mathbb{P} \left( \{S_n - \mathbb{E}(S_n) > nt\} \right) \leq e^{-nst + ns^2/2} = e^{n(s^2/2 - st)}
\]

As \(s \geq 0\) is a freely chosen parameter, we deduce that

\[
\mathbb{P} \left( \{S_n - \mathbb{E}(S_n) > nt\} \right) \leq \min_{s \geq 0} e^{n(s^2/2 - st)} = e^{-n \max_{s \geq 0} (st - s^2/2)}
\]

A simple derivation shows that the maximum (i.e., the tightest upper bound) is reached in \(s^* = t\). This gives \(\mathbb{P} \left( \{S_n - \mathbb{E}(S_n) > nt\} \right) \leq e^{-nt^2/2}\). As mentioned above, a similar reasoning gives the same upper bound on the second term in (5), and this concludes the proof. \qed

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

\[
\mathbb{P} \left( \left\{ \left| S_n - \mathbb{E}(S_n) \right| > nt \right\} \right) \leq 2 \exp \left( - \frac{2n^2 t^2}{\sum_{j=1}^{n} (b_j - a_j)^2} \right), \quad \forall t > 0, \ n \geq 1
\]

The proof is strictly speaking the same as above, but note that in this general case, \(\frac{S_n}{n}\) need not converge to a limit as \(n \to \infty\).

7.2 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 \(s_0 > 0\). 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

\[
\mathbb{P} \left( \left\{ \frac{S_n}{n} > t \right\} \right) = \mathbb{P} \left( \{S_n > nt\} \right) \leq \frac{\mathbb{E}(e^{S_n})}{e^{nst}} = e^{-nst} (\mathbb{E}(e^{sX_1}))^n
\]

\[
= e^{-nst} \exp \left( n \log (e^{sX_1}) \right) = \exp \left( -n (st - \log (\mathbb{E}(e^{sX_1})) \right)
\]

\(^3\text{One can show that this condition is equivalent to saying that there exists } c > 0 \text{ such that } \mathbb{P}((|X_1| > x)) \sim \exp(-cx) \text{ as } x \to \infty.\)
Optimizing this upper bound over \( s \geq 0 \), we obtain
\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} > t \right\} \right) \leq \exp\left( -n \max_{s \geq 0} \left( st - \log \mathbb{E} \left( e^{sX_1} \right) \right) \right), \quad \forall t > \mathbb{E}(X_1)
\]

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:

\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} > t \right\} \right) \leq \exp\left( -n \max_{s \in \mathbb{R}} \left( st - \log \mathbb{E} \left( e^{sX_1} \right) \right) \right), \quad \forall t > \mathbb{E}(X_1)
\]

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:

\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} > t \right\} \right) \leq \exp(-n \Lambda^*(t)), \quad \forall t > \mathbb{E}(X_1)
\] (6)

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.

Likewise, for \( t < \mathbb{E}(X_1) \), we obtain for \( s \geq 0 \):

\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} < t \right\} \right) = \mathbb{P}(\{S_n < nt\}) = \mathbb{P}(\{-S_n > -nt\}) \leq \frac{\mathbb{E}(e^{-sS_n})}{e^{-snt}} = e^{nst} \left( \mathbb{E}(e^{-sX_1}) \right)^n
\]

\[= e^{nts} \exp\left( n \log \mathbb{E} \left( e^{-sX_1} \right) \right) = \exp\left( -n \left( -st - \log \mathbb{E} \left( e^{-sX_1} \right) \right) \right)\]

Optimizing over \( s \geq 0 \), we further obtain

\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} < t \right\} \right) \leq \exp\left( -n \max_{s \geq 0} \left( -st - \log \mathbb{E} \left( e^{-sX_1} \right) \right) \right), \quad \forall t < \mathbb{E}(X_1)
\]

and for similar reasons as before, the maximum can be turned into a maximum over \( \mathbb{R} \), so that

\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} < t \right\} \right) \leq \exp\left( -n \max_{s \in \mathbb{R}} \left( -st - \log \mathbb{E} \left( e^{-sX_1} \right) \right) \right) = \exp\left( -n \Lambda^*(t) \right), \quad \forall t < \mathbb{E}(X_1)
\] (7)

What do these two equations (6) and (7) 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

\[
\mathbb{P}\left( \left\{ \frac{S_n}{n} \in A \right\} \right) \overset{n \to \infty}{\simeq} \exp\left( -n \inf_{t \in A} \Lambda^*(t) \right)
\]
Notice therefore that the above probability is decreasing exponentially in $n$ if and only if $E(X_1) \notin A$.

In the particular cases where $A$ is either the interval $]-\infty, t[$ with $t < E(X_1)$, or the interval $]t, +\infty[$ with $t > E(X_1)$, one recovers the above equations (6) and (7). Indeed, one can check that the infimum of $\Lambda^*$ on $]-\infty, t[$ is achieved in $t$ when $t < E(X_1)$, 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. 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

$$\Lambda(s) = \log E(e^{sX_1}) = \log \left( \int_{\mathbb{R}} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \right) = \log \left( e^{s^2/2} \right) = \frac{s^2}{2}$$

and

$$\Lambda^*(t) = \max_{s \in \mathbb{R}} \left( st - \frac{s^2}{2} \right) = \frac{t^2}{2}, \text{ attained in } s^* = t$$

Also, $E(X_1) = 0$, so $P\left( \{ \frac{S_n}{n} > t \} \right) \leq \exp \left( -\frac{nt^2}{2} \right)$, 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 $P(\{ X_1 = +1 \}) = P(\{ X_1 = -1 \}) = \frac{1}{2}$. In this case,

$$\Lambda(s) = \log \left( \frac{e^s + e^{-s}}{2} \right) = -s + \log(1 + e^{2s}) - \log(2)$$

and

$$\Lambda^*(t) = \max_{s \in \mathbb{R}} (st - \Lambda(s)) = \max_{s \in \mathbb{R}} (s(t + 1) - \log(1 + e^{2s}) + \log 2)$$

Looking for the value where the maximum is attained, we obtain $s^* = \frac{1}{2} \log \left( \frac{1+t}{1-t} \right)$ and correspondingly, after some computations:

$$\Lambda^*(t) = \frac{1}{2} \left( (1 + t) \log(1 + t) + (1 - t) \log(1 - t) \right)$$

Also $E(X_1) = 0$, so $P\left( \{ \frac{S_n}{n} > t \} \right) \leq \exp \left( -n \Lambda^*(t) \right)$ for all $t > 0$. Let us compare this result with Hoeffding’s inequality, which reads in this case:

$$P\left( \left\{ \frac{S_n}{n} > t \right\} \right) \leq \exp \left( -\frac{nt^2}{2} \right), \quad \forall t > 0$$

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 $\Lambda^*(t) \approx t^2$ around $t = 0$, which is greater than $t^2/2$).
A Appendix

In this section, we review some topics that were not discussed during the course, but which are of central importance in probability theory.

A.1 Characteristic functions and their basic properties

Definition A.1. The characteristic function (or Fourier transform) of a random variable $X$ is the mapping $\phi_X : \mathbb{R} \to \mathbb{C}$ defined as

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

Notice that as $|e^{itx}| = 1$ for all $x \in \mathbb{R}$, the characteristic function of a random variable is well defined for all $t \in \mathbb{R}$.

Example A.2. Let $X \sim \mathcal{U}([a,b])$. Then

$$\phi_X(t) = \frac{1}{b-a} \int_a^b e^{itx} \, dx = \frac{e^{itb} - e^{ita}}{it(b-a)}$$

Example A.3. 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

$$\phi_X(t) = \mathbb{E}\left(e^{it(\mu + \sigma Z)}\right) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right)$$

Notice that the above computation requires a bit of complex analysis, actually.

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

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

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

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

Remark. Properties (iv) and (v) are actually implied by properties (i) to (iii).

Proof of (iii)$^4$. For every $n \geq 1$, $c_1, \ldots, c_n \in \mathbb{C}$, $t_1, \ldots, t_n \in \mathbb{R}$, we have

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

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

Theorem A.4. Any function $\phi : \mathbb{R} \to \mathbb{C}$ satisfying properties (i) to (iii) is the characteristic function of a random variable $X$.

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

$^4$The other properties are quite straightforward, except for the continuity property, which requires the dominated convergence theorem (not seen in this course).
In particular, this says that one can identify a distribution by its characteristic function.

**Inversion formula.**

a) To a given characteristic function $\phi$ corresponds a unique cdf $F$ given by

$$F(b) - F(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi(t) \, dt$$

for all $a < b$ continuity points of $F$.

b) If in addition $\int_{\mathbb{R}} |\phi(t)| \, dt < \infty$, then the cdf $F$ admits a pdf $p$ given by

$$p(x) = F'(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \phi(t) \, dt$$

for all $x \in \mathbb{R}$.

The proposition below states a well-known and extremely useful property of characteristic functions.

**Proposition A.5.** If $X$, $Y$ are independent random variables, then $\phi_{X+Y}(t) = \phi_X(t) \phi_Y(t)$, for all $t \in \mathbb{R}$.

**Proof.** By independence $X$ and $Y$, we have independence of $e^{itX}$ and $e^{itY}$, and therefore

$$\mathbb{E}\left(e^{it(X+Y)}\right) = \mathbb{E}\left(e^{itX} e^{itY}\right) = \mathbb{E}\left(e^{itX}\right) \mathbb{E}\left(e^{itY}\right)$$

which completes the proof. \qed

**Another characterization of convergence in distribution.**

Thanks to characteristic functions, we have the following interesting characterization of convergence in distribution. Its proof, which is omitted here, makes use of the Portemanteau theorem from Section 6.2.

**Proposition A.6.** Let $(X_n, n \geq 1)$ be a sequence of random variables and $X$ be another random variable. Then

$$X_n \overset{d}{\to} X \text{ if and only if } \phi_{X_n}(t) \underset{n \to \infty}{\to} \phi_X(t), \text{ for every } t \in \mathbb{R}$$

Notice that such a proposition is made possible because a characteristic function is fully characterizing its corresponding distribution, as guaranteed by the above inversion formula.

**A.2 An alternate proof of the central limit theorem**

The last two propositions of the last section allow for an alternate proof of the central limit theorem of Section 6.3 (notice that we skip some technical details below).

**Proof of Theorem 6.7.** Remember that

$$\tilde{S}_n = \frac{S_n - n\mu}{\sqrt{n} \sigma} = \sum_{j=1}^{n} Y_j$$

where $Y_j = \frac{X_i - \mu}{\sqrt{n} \sigma}$ are i.i.d. random variables with $\mathbb{E}(Y_i) = 0$, $\mathbb{E}(Y_i^2) = \frac{1}{n}$ and $\mathbb{E}(|Y_i|^3) = O(n^{-3/2})$ (using the assumption that $\mathbb{E}(|X_1|^3) < +\infty$).

Let $\phi_n$ be the characteristic function of $\tilde{S}_n$. Using Proposition A.6, it is sufficient to prove that

$$\phi_n(t) \overset{n \to \infty}{\to} \phi(t) = e^{-t^2/2}, \quad \forall t \in \mathbb{R}$$
as $\phi$ is the characteristic function of $Z \sim \mathcal{N}(0, 1)$. To this end, making use of Proposition A.5, we compute

$$\phi_n(t) = \mathbb{E} \left( \exp \left( it \tilde{S}_n \right) \right) = \prod_{j=1}^{n} \mathbb{E} (\exp(itY_j)) = (\mathbb{E}(\exp(itY_1)))^n$$

Using Taylor’s expansion, we further obtain

$$\mathbb{E}(\exp(itY_1)) = 1 + (it) \mathbb{E}(Y_1) + \frac{(it)^2}{2} \mathbb{E}(Y_1^2) + \frac{(it)^3}{6} \mathbb{O}(\mathbb{E}(|Y_1|^3)) = 1 + 0 - \frac{t^2}{2n} + \mathbb{O}(n^{-3/2})$$

so

$$\phi_n(t) = \left( 1 - \frac{t^2}{2n} + \mathbb{O}(n^{-3/2}) \right)^n \xrightarrow{n \to \infty} e^{-t^2/2}$$

which actually needs some careful verification here, as the $\mathbb{O}(n^{-3/2})$ is complex-valued. This completes our second proof of the central limit theorem. □

A.3 An intriguing fact

This intriguing fact 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

$$p_{X_n}(x) = \frac{1}{2n} 1_{[-n, n]}(x), \quad x \in \mathbb{R}$$

and correspondingly,

$$F_{X_n}(t) = \frac{1}{2n} \int_{-\infty}^{t} 1_{[-n, n]}(x) \, dx = \begin{cases} 0, & \text{if } t \leq -n \\ \frac{1}{2} \left( 1 + \frac{t}{n} \right), & \text{if } -n < t \leq n \\ 1, & \text{if } t > n \end{cases}$$

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,

$$\phi_{X_n}(t) = \frac{1}{2n} \int_{-n}^{n} e^{itx} \, dx = \begin{cases} e^{itn - e^{itn}} \frac{\sin(tn)}{tn}, & \text{if } t \neq 0 \\ 1, & \text{if } t = 0 \end{cases}$$

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

\[
\inf_{n \geq 1} \mathbb{P}(\{|X| \leq C\}) \geq 1 - \varepsilon
\]

Such a sequence of random variables is said to be \textit{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 \([-C,+C]\)). Notice 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 (8) should hold. Reversely, one can show that when condition (8) 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.4 Moments and Carleman’s theorem

**Definition A.7.** Let \( X \) be a random variable and \( k \geq 0 \). If \( \mathbb{E}(|X|^k) < \infty \), we say that the moment of order \( k \) of the random variable \( X \) is finite and defined as

\[
m_k = \mathbb{E}(X^k)
\]

Notice 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

\[
\mathbb{E}(|X|^1) = \mathbb{E}(f(|X|^k)) \leq f(\mathbb{E}(|X|^k)) < \infty
\]

Notice 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 positive semi-definite, that is,

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

Indeed,

\[
\sum_{j,k=1}^{n} c_j \overline{c_k} m_{j+k} = \sum_{j,k=1}^{n} c_j \overline{c_k} \mathbb{E}(X^{j+k}) = \mathbb{E} \left( \left| \sum_{j=1}^{n} c_j X^j \right|^2 \right) \geq 0
\]

It is however not clear whether the distribution of a random variable \( X \) is fully characterized by its moments, 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 \textit{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
\[ |m_k| = |E(X^k)| \leq E(|X|^k) \leq C^k, \quad \forall k \geq 0 \]
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 A.8.** If $X$ is a random variable whose moments are all finite and satisfy
\[ \sum_{k \geq 1} \frac{m_k}{m_{2k}} = +\infty \tag{9} \]
then the distribution of the random variable $X$ is uniquely determined by its sequence of moments $(m_k, k \geq 0)$.

**Remarks.** - Even though it 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 get a better grasp on condition (9), let us see what it gives in our two examples above:

1. When $|m_k| \leq C^k$, we obtain that
\[ m_{2k}^{-1/2} \geq C^{-1}, \quad \text{so that indeed, } \sum_{k \geq 1} m_{2k}^{-1/2} \geq \sum_{k \geq 1} C^{-1} = +\infty \]
2. When $m_k = \exp(k^2/2)$, we obtain
\[ m_{2k}^{-1/2} = \exp\left(-\frac{(2k^2)^2}{2(2k)} \right) = \exp(-k), \quad \text{so } \sum_{k \geq 1} m_{2k}^{-1/2} = \sum_{k \geq 1} \exp(-k) < +\infty \]

- One may wonder why 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
\[ (E(|X|^{2k+1}))^2 = (E(|X|^k |X|^{k+1}))^2 \leq E(X^{2k}) E(X^{2k+2}) = m_{2k} m_{2k+2}, \quad \forall k \geq 0 \]

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 A.9.** Let $(m_k, k \geq 0)$ be a sequence of real numbers satisfying condition (9). If $(X_n, n \geq 1)$ is a sequence of random variables such that for all $k \geq 0$,
\[ E(X_n^k) \rightarrow m_k, \quad \text{as } n \rightarrow \infty \]
then $X_n$ converges in distribution towards a random variable $X$, whose moments are $(m_k, k \geq 0)$.

Notice that this characterization offers another way to prove the central limit theorem, but is of course useful in numerous other contexts also.