جامعة نيويورك أبوظبي

NYU | ABU DHABI

Advanced Probability

NYUAD

Pierre Youssef
yp27@nyu.edu
https://wp.nyu.edu/pyoussef/

Spring 2022

Contents

Chapter 1

Measurable spaces and Probability

1.1 Measurable spaces

Given an object, we would like to understand its size. For instance, we measure the area of a surface in $\mathbb{R}^2$ or the volume of a solid in $\mathbb{R}^3$ to get an idea to what extend these objects are big. This geometric intuition allows us to assert for example that the area of two disjoint surfaces is the sum of the two. On the one hand, we would like to generalize these notions of geometric measures to other types of measures which are not necessarily related to a geometric aspect, and on the other hand introduce all the necessary materials involved. To understand the importance of this, notice for example that our intuition may lead us to measure a set by summing the measure of its elements, which already causes a problem since $A = [0, 1]$ and $B = [0, 2]$ are in bijection while B is the "double" of A. Therefore, we could have taken A, decomposed it into a (uncountable) number of points then reassembled it to form B which is twice as large. This may seem weird and we may be led to believe that the root cause of our weird conclusion reside in the fact that the decomposition is uncountable. However, even when dropping this uncountable feature, the problem persists. More precisely, the *Banach-Tarski paradox* asserts that the Euclidean ball of radius 1 in $\mathbb{R}^3$ can be decomposed into a finite number of parts[1] which could be reassembled (after possible translations and rotations) to form two disjoint copies of the original Euclidean ball. Isn't this strange? This definitely messes with our intuition as we ended up with a set of size double the one we had while using finite parts of that object.

Then what is exactly the issue?

The five parts decomposing the ball in the Banach-Tarski paradox are actually constructed using the axiom of choice and are strange creatures. If we admit the axiom of choice, and that the volume is preserved under translation and rotation, then the only possible explanation to this paradox is that we can't simply measure everything! This leads us to the following natural questions:

- What does it mean for a set to be measurable?

- How do we define its measure?

- What are the properties of measurable sets and of measures?

[1] Five parts are actually sufficient and the construction relies on the axiom of choice.

- Do these general concepts correspond to "usual" measures in some context?

We will try in this chapter to address these questions and lay the foundation of measure theory which will make rigorous our study of Probability theory. Let us start by defining the notion of σ-algebra which will regroup the sets which we will be able to measure each time.

Definition 1.1 (*σ-algebra*)

Let S be a set. We say that a collection Σ of subsets of S is a σ-algebra if the following properties hold:

(i) $S \in \Sigma$.

(ii) $A \in \Sigma \Rightarrow A^c := S \backslash A \in \Sigma$.

(iii) $A_n \in \Sigma \ \forall n \in \Sigma \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \Sigma$.

Remark 1.2

- *Note that if Σ is a σ-algebra then $\varnothing \in \Sigma$ since it is the complement of S.*
- *If $A_n \in \Sigma \ \forall n \in \Sigma$ then $\bigcap_{n\in\mathbb{N}} A_n \in \Sigma$ since*

  $$\bigcap_{n\in\mathbb{N}} A_n = \Big(\bigcup_{n\in\mathbb{N}} A_n^c\Big)^c,$$

  and $A_n^c \in \Sigma$ by (ii) and the countable union belongs to Σ by (iii).
- *Note that if the countable union in (iii) is replaced by a finite union, then we say that Σ is an algebra. Therefore, a σ-algebra is an algebra while the reverse is not necessarily true.*
- *In some places, σ-algebras are also called σ-fields.*

Example 1.3

1. *Given a set S, we can define $\{\varnothing, S\}$ verifying all three properties on a σ-algebra. This is the smallest σ-algebra on S, and is called the trivial σ-algebra.*
2. *The set $\mathcal{P}(S)$ of all subsets of S is a σ-algebra. It is the largest σ-algebra on S. Note that any σ-algebra on S satisfies $\{\varnothing, S\} \subset \Sigma \subset \mathcal{P}(S)$.*

Definition 1.4 (*Mesurable space*)

We say that the pair (S, Σ) is a measurable space if S is a set and Σ is a σ-algebra on S. We say that the elements of Σ are the subsets of S which are Σ-measurable.

In linear algebra for example, in order to understand the elements of a vector space, it suffices to specify a basis and all elements of the space could be derived using this basis. We would like to do something of the same sort for a Σ-algebra and instead of specifying all its elements, it would suffice to recognize a generating system which would allow to obtain all other elements using the operations (union, intersection, complement) as these are the operations by which a Σ-algebra is stable.

Definition 1.5 (*Generated σ-algebra*)

Let $\mathcal{C}$ be a collection of subsets of S. The σ-algebra generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the smallest σ-algebra on S which contains $\mathcal{C}$. It is the intersection of all σ-algebras containing $\mathcal{C}$.

Remark 1.6

Note that the above definition assumes a fact which needs to be checked, that is: the intersection of σ-algebras is a σ-algebra. Try to check this by yourself.

Example 1.7

1. *Let $S = \{1,2,3,4\}$ and $\mathcal{C} = \{\{1\},\{2,3\}\}$. To form $\sigma(\mathcal{C})$, we start by adding the empty set, S then all elements of $\mathcal{C}$. Then we start using all possible operations and verify that the obtained set is indeed a σ-algebra. In this example, we should add $\{1\}^c = \{2,3,4\}$, $\{2,3\}^c = \{1,4\}$, $\{1,4\} \cap \{2,3,4\} = \{4\}$, and $\{4\}^c = \{1,2,3\}$ to get*

$$\sigma(\mathcal{C}) = \big\{\varnothing, S, \{1\}, \{2,3\}, \{1,2,3\}, \{4\}, \{1,4\}, \{2,3,4\}\big\}.$$

2. *An important example is the Borel σ-algebra. If S is a topological space, the Borel σ-algebra $\mathcal{B}(S)$ is the σ-algebra generated by the open sets of S. Thus $\mathcal{B} := \mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$ i.e. the σ-algebra generated by the open sets of $\mathbb{R}$.*

   Note that instead of working with all open sets of $\mathbb{R}$, we will select a generating system which is easier to work with. To this end, we define the class of intervals

$$I(\mathbb{R}) := \{(-\infty, x] : x \in \mathbb{R}\}$$

   and we will verify that it allows to obtain all other elements of the Borel σ-algebra on $\mathbb{R}$ i.e. we will show that $\mathcal{B} = \sigma\big(I(\mathbb{R})\big)$. We proceed by checking the double inclusion, by first noting that for any $x \in \mathbb{R}$ we have $(-\infty, x] = \bigcap_{n\in\mathbb{N}}(-\infty, x+1/n)$ which is a countable union of open sets and therefore $(-\infty, x] \in \mathcal{B}$. Since $\sigma\big(I(\mathbb{R})\big)$ is the smallest σ-algebra containing $I(\mathbb{R})$ then $\sigma\big(I(\mathbb{R})\big) \subseteq \mathcal{B}$.

   For the other inclusion, it is enough to show that any open interval in $\mathbb{R}$ can be generated using elements of $I(\mathbb{R})$. Let $a < b$, and note that since $(a,b) = \bigcup_{n\in\mathbb{N}}(a, b-1/n]$ it is enough to show that an interval of the form $(a,u]$ can be generated using the elements of $I(\mathbb{R})$. Now observe that

$$(a,u] = (-\infty, u] \cap (-\infty, a]^c \in \sigma\big(I(\mathbb{R})\big),$$

   which finishes the proof.

1.2 Measure and Probability

We are now ready to introduce the notion of measure. The work done previously consisted in defining the domain on which the measure will act, that is the sets which will be measurable, and making sure for example that if two sets are measurable, then their union is measurable, etc.

Definition 1.8 (*Measure*)

Let (S, Σ) be a measurable space. A mesure μ on (S, Σ) is a function

$$\mu : \Sigma \to [0, \infty]$$

which satisfies the following properties:

(i) $\mu(\varnothing) = 0$.

(ii) If $(A_n)_{n\in\mathbb{N}}$ is a sequence of disjoint sets in Σ then $\mu\big(\bigcup_{n\in\mathbb{N}} A_n\big) = \sum_{n\in\mathbb{N}} \mu(A_n)$.

We then say that (S, Σ, μ) is a measure space.
If in addition $\mu(S) = 1$ then we say that μ is a probability measure and that (S, Σ, μ) is a probability space.

Remark 1.9

1. *Property (ii) above is known as the σ-additivity of the mesure. If the countable union is replaced by a finite union, then we talk about additivity.*
2. *Note that the fact that Σ is a σ-algebra is consistent with properties (i) and (ii) since $\varnothing \in \Sigma$ and we can consider $\mu(\varnothing)$ and similarly, if $(A_n)_{n\in\mathbb{N}}$ are in Σ then their union also belongs to Σ and we can consider $\mu\big(\bigcup_{n\in\mathbb{N}} A_n\big)$.*
3. *We say that μ is finite if $\mu(S) < \infty$. Therefore, a probability measure is in particular a finite measure.*
4. *We say that μ is σ-finite if there exists a sequence $(S_n)_{n\in\mathbb{N}}$ of elements of Σ such that $\bigcup_{n\in\mathbb{N}} S_n = S$ and $\mu(S_n) < \infty$ for every $n \in \mathbb{N}$.*

Example 1.10

1. *Let (S, Σ) be a measurable space. $\mu : \Sigma \to [0, \infty]$ defined by $\mu(A) = 0$ for every $A \in \Sigma$ is a measure.*
2. *The cardinality of a set is a measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$.*
3. *Let (S, Σ) be a measurable space and $x \in S$. We define $\delta_x : \Sigma \to [0, \infty]$ by $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. We verify that δ_x is a probability measure which is called the Dirac measure.*
4. *A two-point probability space can be used to model the fair coin flipping experiment. Let $\Omega = \{H, T\}$ and $\Sigma = \mathcal{P}(\Omega)$, and take $\mathbb{P}$ to be the probability measure on Ω which assigns values $1/2$ to both $\{H\}$ and $\{T\}$.*

Here are some of the elementary properties of measurs.

Proposition 1.11

Let (S, Σ, μ) be a measure space. Then we have

1. *$\mu(A \cup B) \leqslant \mu(A) + \mu(B)$ for any $A, B \in \Sigma$.*
2. *$\mu\big(\bigcup_{i \leqslant n} A_i\big) \leqslant \sum_{i \leqslant n} \mu(A_i)$ for any $A_1, \ldots, A_n \in \Sigma$.*
3. *If $\mu(S) < \infty$ then for any $A, B \in \Sigma$ we have*

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B).$$

More generally, we have the following inclusion-exclusion formula

$$\mu\big(\bigcup_{i \leqslant n} A_i\big) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} \mu\big(\bigcap_{j \leqslant k} A_{i_j}\big),$$

for any $A_1, \ldots, A_n \in \Sigma$.

Proof Exercise. □

When one defines a function on $\mathbb{R}$ for example, an important property is its continuity which one of its consequences asserts that $\lim_n f(x_n) = f(\lim_n x_n)$. We would like to obtain a similar continuity property for measures, but we will need for this to make sense of limits for sets (since a measure takes sets as imput).

Therefore, for a sequence $(A_n)_{n \in \mathbb{N}} \in \Sigma$, we will write $A_n \uparrow A$ if $A_n \subseteq A_{n+1}$ for any $n \in \mathbb{N}$ and $\bigcup_{n \in \mathbb{N}} A_n = A$. Similarly, we write $A_n \downarrow A$ if $A_{n+1} \subseteq A_n$ for every $n \in \mathbb{N}$ and $\bigcap_{n \in \mathbb{N}} A_n = A$.

Lemme 1.12 (*Monotone convergence lemma*)

Let (S, Σ, μ) be a measure space and $(A_n)_{n \in \mathbb{N}} \in \Sigma$.

1. *If $A_n \uparrow A$ then $\mu(A_n) \uparrow \mu(A)$.*
2. *If $A_n \downarrow A$ and $\mu(A_k) < \infty$ for a certain $k \in \mathbb{N}$, then $\mu(A_n) \downarrow \mu(A)$.*

Proof

1. Let $B_1 = A_1$, $B_2 = A_2 \backslash A_1$ and more generally $B_n = A_n \backslash A_{n-1}$ for every $n \geqslant 2$. We remark that the B_n's are disjoint and that $A_n = \bigcup_{i=1}^{n} B_i$. Therefore

$$\mu(A_n) = \mu(B_1 \cup \ldots \cup B_n) = \sum_{i=1}^{n} \mu(B_i) \underset{n \to \infty}{\longrightarrow} \sum_{i=1}^{\infty} \mu(B_i) = \mu(A).$$

2. Let $B_n = A_k \backslash A_{k+n}$ for every $n \in \mathbb{N}$. Therefore $B_n \uparrow A_k \backslash \bigcap_{\ell \in \mathbb{N}} A_\ell$ and by the first part

$$\mu(B_n) = \mu(A_k) - \mu(A_{k+n}) \underset{n \to \infty}{\longrightarrow} \mu\big(A_k \backslash \bigcap_{\ell \in \mathbb{N}} A_\ell\big) = \mu(A_k) - \mu\big(\bigcap_{\ell \in \mathbb{N}} A_\ell\big).$$

It remains to subtract $\mu(A_k)$ (which is finite) from both sides.

□

1.3 Extension of measures

A σ-algebra could be huge and very complicated which makes our task of understanding a measure on it difficult. The idea is to restrict our attention to a smaller part of the σ-algebra, and easier to manipulate. The goal is to prove that this is possible and that it would be enough to know the measure on a "special part" of the σ-algebra in order to define it everywhere. This special part is what we will call a π-system

Definition 1.13 (*π-system*)

Let S be a set. We say that a family $\mathcal{I}$ of subsets of S is a π-system if it is stable by finite intersection i.e.

$$A_1, A_2 \in \mathcal{I} \Rightarrow A_1 \cap A_2 \in \mathcal{I}.$$

Example 1.14

1. *A σ-algebra is in particular a π-system.*
2. *The class of intervals $\{(-\infty, t] : t \in \mathbb{R}\}$ is a π-system.*
3. *Let S be a set. Then $\mathcal{I} = \{\{x\} : x \in S\} \cup \{\varnothing\}$ is a π-system.*
4. *Let Ω be a set. Then $\mathcal{I} = \{A \times B : A, B \in \mathcal{P}(\Omega)\}$ is a π-system.*

To form a σ-algebra starting from a π-system, one is led to add complements of the sets we have and perform all operations needed to form a σ-algebra from a generating system. Therefore, the σ-algebra generated by the π-system could be much larger than the original system, which explains our will to work with the system instead of the much larger σ-algebra. The following lemma will help us prove that in many cases, we can proceed with such strategy. Let us first define monotone classes.

Definition 1.15 (*Monotone classes*)

Let S be a set. We say that a family $\mathcal{M}$ of subsets of S is a monotone class if

- *$S \in \mathcal{M}$.*
- *If $A, B \in \mathcal{M}$ and $A \subset B$, then $B \backslash A \in \mathcal{M}$.*
- *If $A_n \in \mathcal{M}$ and $A_n \subset A_{n+1}$ then $\bigcup_n A_n \in \mathcal{M}$.*

Note that a σ-algebra is in particular a monotone class.

Theorem 1.16 (*Monotone classes lemma*)

Let S be a set and $\mathcal{I}$ be a π-system. Then $\mathcal{M}(\mathcal{I}) = \sigma(\mathcal{I})$ where

$$\mathcal{M}(\mathcal{I}) = \bigcap_{\mathcal{M} \text{ classe monotone}, \mathcal{I} \subset \mathcal{M}} \mathcal{M}.$$

Proof Since $\sigma(\mathcal{I})$ is a monotone class (since it is a σ-algebra), then $\mathcal{M}(\mathcal{I}) \subset \sigma(\mathcal{I})$. To prove the

other inclusion, it is enough to show that $\mathcal{M}(\mathcal{I})$ is a σ-algebra since $\sigma(\mathcal{I})$ is the smallest σ-algebra containing $\mathcal{I}$.
We say that $S \in \mathcal{M}(\mathcal{I})$ by definition. Let $A \in \mathcal{M}(\mathcal{I})$ then $A \in \mathcal{M}$ for every monotone class $\mathcal{M}$ containing $\mathcal{I}$. Since $\mathcal{M}$ is a monotone class and $A \subset S$ belong to $\mathcal{M}$, then the same holds for $A^c = S \backslash A$.

It remains then to show that $\mathcal{M}(\mathcal{I})$ is stable by countable union. By property (iii) of monotone classes, it is enough to show that $\mathcal{M}(\mathcal{I})$ is stable by finite union. Since $\mathcal{M}(\mathcal{I})$ is stable by taking the complement, it is enough to verify that $\mathcal{M}(\mathcal{I})$ is stable by finite intersection. To this aim, we need to show that if $A, B \in \mathcal{M}(\mathcal{I})$ then $A \cap B \in \mathcal{M}(\mathcal{I})$.

Suppose first that $A \in \mathcal{I}$. Let $\mathcal{M}_A = \{B \in \mathcal{M}(\mathcal{I}) : A \cap B \in \mathcal{M}(\mathcal{I})\}$. Since $\mathcal{I}$ is a π-system then $\mathcal{I} \subset \mathcal{M}_A$. We will verify that $\mathcal{M}_A$ is a monotone class to deduce that $\mathcal{M}(\mathcal{I}) \subset \mathcal{M}_A$. We have

(i) $S \cap A = A \in \mathcal{I} \subset \mathcal{M}(\mathcal{I})$ therefore $S \in \mathcal{M}_A$.

(ii) Let $B, B' \in \mathcal{M}_A$ with $B \subset B'$. Then $(B' \backslash B) \cap A = (B' \cap A) \backslash (B \cap A) \in \mathcal{M}(\mathcal{I})$ since $B' \cap A \in \mathcal{M}(\mathcal{I})$, $B \cap A \in \mathcal{M}(\mathcal{I})$ and $\mathcal{M}(\mathcal{I})$ is a monotone class.

(iii) If B_n is an increasing sequence of $\mathcal{M}_A$, then $B_n \cap A$ is an increasing sequence of $\mathcal{M}(\mathcal{I})$. Since $\mathcal{M}(\mathcal{I})$ is a monotone class, then $\bigcup (B_n \cap A) \in \mathcal{M}(\mathcal{I})$. Since $\bigcup (B_n \cap A) = A \cap \bigcup B_n$ we deduce that $\bigcup B_n \in \mathcal{M}_A$.

Let now

$$\mathcal{M}_1 = \{A \in \mathcal{M}(\mathcal{I}) : \forall B \in \mathcal{M}(\mathcal{I}), A \cap B \in \mathcal{M}(\mathcal{I})\} = \{A \in \mathcal{M}(\mathcal{I}) : \mathcal{M}(\mathcal{I}) \subset \mathcal{M}_A\}.$$

We just showed that $\mathcal{I} \subset \mathcal{M}_1$. We verify as previously that $\mathcal{M}_1$ is a monotone class, which implies that $\mathcal{M}(\mathcal{I}) \subset \mathcal{M}_1$. This shows that for any $A, B \in \mathcal{M}(\mathcal{I})$, we have $A \cap B \in \mathcal{M}(\mathcal{I})$ and we finish the proof. □

To illustrate the importance of what we previously mentioned, let us state the following corollary.

Corollary 1.17 (*Uniqueness of a measure*)

Let S be a set, $\mathcal{I}$ a π-system on S and $\Sigma = \sigma(\mathcal{I})$. If two finite measures μ_1, μ_2 on (S, Σ) coincide on $\mathcal{I}$ (i.e. $\mu_1 = \mu_2$ on $\mathcal{I}$) and $\mu_1(S) = \mu_2(S)$ then they coincide on Σ.

Proof We will prove $\mathcal{F} = \{A \in \Sigma : \mu_1(A) = \mu_2(A)\}$ is a monotone class. Indeed this would imply by the monotone classes Lemma that $\sigma(\mathcal{I}) = \mathcal{M}(\mathcal{I}) \subset \mathcal{F}$ and that $\mathcal{F} = \Sigma$ finishing the proof.
To verify that it is a monotone class, we first note that by hypothesis we have $\mu_1(S) = \mu_2(S)$.
Next we show that $\mathcal{F}$ is stable by increasing union. Let $A_n \in \mathcal{F}$, $n \in \mathbb{N}$, with $A_n \subset A_{n+1}$. By the monotone convergence lemma we have

$$\mu_1\big(\bigcup_n A_n\big) = \lim_n \mu_1(A_n) = \lim_n \mu_2(A_n) = \mu_2\big(\bigcup_n A_n\big).$$

Now if $A, B \in \mathcal{F}$ and $A \subset B$ then $\mu_1(B \backslash A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \backslash A)$. Note that we used here that the measures are finite. □

This result gains more importance when combined with the following extension theorem.

Theorem 1.18 (*Caratheodory extension theorem*)

Let S be a set, Σ_0 be a π-system on S which also satisfies that if $A, B \in \Sigma_0$ then $A \backslash B \in \Sigma_0$ and

let $\Sigma = \sigma(\Sigma_0)$. If $\mu_0 : \Sigma_0 \to [0, \infty]$ is σ-additive and $\mu_0(\varnothing) = 0$ then there exists an extension of μ_0 to Σ i.e. there exists a measure μ on (S, Σ) such that $\mu = \mu_0$ on Σ_0.
Moreover, if μ_0 is σ-finite then this extension is unique.

The moral behind this result is that one doesn't need to define the measure on the whole σ-algebra but only on some π-system with an additional property, and the previous theorem ensures that one can extend the measure to the whole σ-algebra. Moreover, it can be done in a unique way when the corresponding measure is σ-finite.

1.4 Lebesgue measure

We choose not to prove Theorem 1.18 in its full generality but only focus on verifying part of it for an important particular example: *the Lebesgue measure* on $\mathbb{R}$.
Consider Σ_0 to be the collection of all sets which can be written as a countable union of intervals of the form

$$(a_i, b_i],\ (a_i, b_i),\ [a_i, b_i],\ [a_i, b_i),$$

with $a_i \leqslant b_i$. It is not difficult to check that Σ_0 is a π-system which satisfies the assumption in Theorem 1.18. We will only to define the measure λ on Σ_0 and we do so in the most natural way as

$$\lambda(I) = \sum_{i=1}^{\infty} |b_i - a_i|,$$

for every $I = \bigcup_{i=1}^{\infty}(a_i, b_i]$ (or any other interval form). Note that $\lambda(I)$ could be infinite. One can check that λ is σ-additive on Σ_0. The previous theorem ensures that there exists an extension of λ to $\sigma(\Sigma_0) = \mathcal{B}(\mathbb{R})$, this measure is called *the Lebesgue measure* on $\mathbb{R}$. To verify this claim, we will explicitly construct such extension as follows. Given an open set $A \in \mathcal{B}(\mathbb{R})$, define

$$\lambda^*(A) = \inf_{\bigcup_i (a_i, b_i) \supset A} \sum_i |b_i - a_i|.$$

λ^* is called the *Lebesgue outer measure.* Clearly, λ^* and λ coincide on Σ_0 and we only need to show that λ^* is indeed a measure. It is possible to check that λ^* satisfies the properties of outer measures

- $\lambda^*(\varnothing) = 0$.
- If $A \subset B \subset \mathbb{R}$, then $\lambda^*(A) \leqslant \lambda^*(B)$.
- We have the sub-additivity property $\lambda^*(\bigcup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{\infty} \lambda^*(A_i)$.

Let us check that if A and B are disjoint then $\lambda^*(A \cup B) = \lambda^*(A) + \lambda^*(B)$. Using the above properties, we have

$$\lambda^*(A \cup B) \leqslant \lambda^*(A) + \lambda^*(B).$$

To prove the reverse inequality, we may assume that $\lambda^*(A \cup B)$ is finite. For any $\varepsilon > 0$, let $\{(a_i, b_i)\}_i$ be a cover of $A \cup B$ such that

$$\sum_{i=1}^{\infty} |a_i - b_i| \leqslant \lambda^*(A \cup B) + \varepsilon.$$

These intervals (a_i, b_i) can be split into several sub-intervals in such a way that we obtain two disjoint covers of A and B respectively. Therefore, we get that

$$\lambda^*(A) + \lambda^*(B) \leqslant \sum_{i=1}^{\infty} |a_i - b_i| \leqslant \lambda^*(A \cup B) + \varepsilon.$$

Since this is true for any $\varepsilon > 0$, we get the claim.

Remark 1.19

- *Let us note that the Lebesgue measure can be defined beyond Borel sets. In fact the above measure is usually referred to as Borel measure (since it is defined on the Borel σ-algebra) and the Lebesgue measure extends it to the σ-algebra of Lebesgue measurable sets.*

  A set A is said to be Lebesgue measurable if for every $\varepsilon > 0$ there exists an open set U such that $\lambda^*(U \backslash A) \leqslant \varepsilon$. Note that not every Lebesgue measurable set is a Borelian set so that the σ-algebra of Lebesgue measurable sets is strictly larger than $\mathcal{B}(\mathbb{R})$. See here for more details.

- *Note that there are sets which are not Lebesgue measurable. An example are Vitali sets (see here).*

- *The definition of the Lebesgue measure extends in the obvious way to $\mathbb{R}^d$.*

- *Note that when restricted to $\mathcal{B}([0,1])$, the Lebesgue measure is a Probability measure which coincide with the uniform measure.*

- *Note that the Lebesgue measure of any countable set is equal to zero. However, there exists an uncountable set whose Lebesgue measure is zero. Such example is given by the Cantor set.*

1.5 Events

We now turn our focus to probability spaces. Recall that a triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is called a probability space if $(\Omega, \mathcal{A})$ is a measurable space $\mathcal{A}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure.

In this setting, we refer to Ω as the universe, an element $\omega \in \Omega$ as a sample and $\mathcal{A}$ the family of events. The randomness here stems from sampling an element ω from Ω randomly according to the distribution $\mathbb{P}$ i.e. for any $A \in \mathcal{A}$, the probability that we sampled an element belonging to A is given by $\mathbb{P}(A)$.

Example 1.20

1. *We toss a coin twice. The result of our experiment consists of all pairs of head and tail. We set*

   $$\Omega = \{HT, HH, TH, TT\}$$

   and take $\mathcal{A} = \mathcal{P}(\Omega)$ the collection of all subsets of Ω. We can define $\mathbb{P}$ as the uniform probability on Ω. Therefore, the event "obtaining at least one head" is given by the subset $\{HH, HT, TH\}$ which has probability $3/4$.

2. *If our experiment consisted of choosing at random a number between 0 and 1. Then the*

universe would be $\Omega = [0,1]$ which we equip with the Borel σ-algebra $\mathcal{A} = \mathcal{B}([0,1])$ and of the Lebesgue measure.

3. *Consider a card deck with 52 cards. Each order of cards in the deck can be represented as an element of the permutation group S_{52}. Let $\Omega := S_{52}$, $\Sigma = \mathcal{P}(\Omega)$ and $\mathbb{P}$ be a uniform probability measure which assigns value $1/52!$ to each permutation. This probability space can be used to model a perfect card shuffling.*

Definition 1.21 (*Almost surely*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We say that a statement Γ holds $\mathbb{P}$-almost-surely *(or simply a.s.) if $C = \{\omega \in \Omega : \Gamma(\omega) \text{ holds}\} \in \mathcal{A}$ and $\mathbb{P}(C) = 1$.*

Example 1.22

1. *Taking the coin toss. We make the statement Γ:"We obtain at least one head or one tail". We see that $C = \{\omega \in \Omega : \Gamma(\omega) \text{ holds}\} = \{HH, HT, TH, TT\} = \Omega \in \mathcal{A}$ and that $\mathbb{P}(C) = 1$. Thus Γ holds a.s.*

2. *Taking the example of picking a number in $[0,1]$. We make the statement Γ:"The number we obtain is different than $0, 1/2, 1$". We see that $C = \{\omega \in [0,1] : \Gamma(\omega) \text{ holds}\} = (0,1)\backslash\{1/2\} \in \mathcal{B}([0,1])$ and that its Lebesgue measure is 1. Thus Γ holds a.s.*

Proposition 1.23

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. If $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}$ are such that $\mathbb{P}(A_n) = 1$ for any $n \in \mathbb{N}$, then $\mathbb{P}\big(\bigcap_{n\in\mathbb{N}} A_n\big) = 1$. In other words, the countable intersection of almost sure events is an almost sure event.

Proof We have $\mathbb{P}(A_n^c) = 0$ for every $n \in \mathbb{N}$ and therefore $\mathbb{P}\big(\bigcup_{k=1}^n A_k^c\big) = 0$. But $\bigcup_{k=1}^n A_k^c \uparrow \bigcup_{\ell\in\mathbb{N}} A_\ell^c$, thus $\mathbb{P}\big(\bigcup_{\ell\in\mathbb{N}} A_\ell\big) = \lim_{n\to\infty} \mathbb{P}\big(\bigcup_{k=1}^n A_k^c\big) = 0$ and therefore $\mathbb{P}\big(\bigcap_{n\in\mathbb{N}} A_n\big) = 1$. □

In the same manner that we defined the limit of sets, we will now define the lim sup and lim inf. Let us first recall these definitions for a sequence of reals: given a sequence of real numbers $(x_n)_{n\in\mathbb{N}}$, the inner and outer limits are given by

$$\liminf x_n := \sup_m \inf_{n\geqslant m} x_n = \lim_m \big[\inf_{n\geqslant m} x_n\big] \quad \text{and} \quad \limsup x_n := \inf_m \sup_{n\geqslant m} x_n = \lim_m \big[\sup_{n\geqslant m} x_n\big].$$

We say that the limit of x_n exist when the two previous quantities coincide. Doing the analogy between inf and $\cap$ on one hand, and between sup and $\cup$ on the other, we define the inner and outer limits of sets as follows.

Definition 1.24

Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets. The outer limit of A_n, denoted $\limsup A_n$, is given by

$$\limsup A_n = \bigcap_m \bigcup_{n\geqslant m} A_n.$$

Similarly, the inner limit of A_n, denoted by $\liminf A_n$, is given by

$$\liminf A_n = \bigcup_m \bigcap_{n\geqslant m} A_n.$$

Remark 1.25

1. *It is important to understand the meaning of* $\limsup A_n$*. Saying that* $\omega \in \limsup A_n$ *is equivalent to saying that*
$$\forall m, \exists n \geqslant m \text{ such that } \omega \in A_n.$$
Therefore ω *belongs to an infinity of sets* A_n*. The reverse is also true and we should keep in mind the following*
$$\omega \in \limsup A_n \Leftrightarrow \omega \text{ belongs to an infinity of } A_n.$$

2. *Saying that* $\omega \in \liminf A_n$ *is equivalent to saying that*
$$\exists m, \forall n \geqslant m \text{ we have } \omega \in A_n.$$
Therefore
$$\omega \in \liminf A_n \Leftrightarrow \text{starting from some threshold, } \omega \text{ belongs to all } A_n.$$

3. *Exercise: Verify that* $\liminf A_n \subseteq \limsup A_n$*.*

4. *Exercise: Show that*
$$\mathbf{1}_{\limsup A_n} = \limsup \mathbf{1}_{A_n} \quad \text{and} \quad \mathbf{1}_{\liminf A_n} = \liminf \mathbf{1}_{A_n},$$
where $\mathbf{1}_A$ *denote the indicator function of* A*.*

Proposition 1.26

Let $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space. Let* $(A_n)_{n\in\mathbb{N}}$ *be a sequence of events, we then have:*

(i) (First Borel-Cantelli lemma) If $\sum_n \mathbb{P}(A_n) < \infty$*, then*
$$\mathbb{P}(\limsup A_n) = 0.$$

(ii) (Fatou Lemma)
$$\mathbb{P}(\liminf A_n) \leqslant \liminf \mathbb{P}(A_n).$$

(iii) (Inverse Fatou Lemma)
$$\mathbb{P}(\limsup A_n) \geqslant \limsup \mathbb{P}(A_n).$$

Proof

(i) Set $B = \limsup A_n$ and $B_m = \bigcup_{n\geqslant m} A_n$. By the hypothesis $\mathbb{P}(B_m) < \infty$ for every $m \in \mathbb{N}$. Moreover, the B_m's are non-increasing and $\bigcap_m B_m = B$, therefore
$$\mathbb{P}(B) = \mathbb{P}\Big(\bigcap_m B_m\Big) = \lim_m \mathbb{P}(B_m) \leqslant \lim_m \sum_{n\geqslant m} \mathbb{P}(A_n) = 0.$$

(ii) Let $C = \liminf A_n$ and $C_m = \bigcap_{n\geqslant m} A_n$. The C_m's being non-decreasing and $\bigcup_m C_m = C$, we have

$$\mathbb{P}(C) = \lim_m \mathbb{P}(C_m).$$

But $\mathbb{P}(C_m) \leqslant \mathbb{P}(A_n)$ for every $n \geqslant m$ since $\bigcap_{n \geqslant m} A_n \subseteq A_n$. Thus

$$\mathbb{P}(C) = \lim_m \mathbb{P}(C_m) \leqslant \lim_m \inf_n \mathbb{P}(A_n) = \liminf \mathbb{P}(A_n).$$

(iii) Set $B = \limsup A_n$ and $B_m = \bigcup_{n \geqslant m} A_n$. The B_m's are non-increasing and $\mathbb{P}$ is a finite measure, thus

$$\mathbb{P}(A) = \mathbb{P}\big(\bigcap_m B_m\big) = \lim_m \mathbb{P}(B_m).$$

But $B_m \supseteq A_n$ for every $n \geqslant m$, thus $\mathbb{P}(B_m) \geqslant \mathbb{P}(A_n)$ for every $n \geqslant m$. Therefore

$$\mathbb{P}(A) = \mathbb{P}\big(\bigcap_m B_m\big) = \lim_m \mathbb{P}(B_m) \geqslant \lim_m \sup_{n \geqslant m} \mathbb{P}(A_n) = \limsup \mathbb{P}(A_n).$$

□

Chapter 2

Integration and product measures

2.1 Measurable maps

In the previous chapter, we started with an arbitrary set and equipped it with a σ-algebra to form a measurable space. In Algebra for example, we start with a set and equip it with an addition or product rule. In topology, we identify the open and closed sets of the space or we equip the space with a metric. In all these fields, the next step is to define maps between those different spaces formed. The map should of course take into account the underlying structure. For example, in Algebra, we talk about group morphism which respects the addition/product rule by asserting that the image of the sum (or product) is the sum (or product) of the two images obtained by this map. In Topology, we speak about continuous maps which allow to transfer the topology from the domain space to the range one. We ask that the reciprocal of an open set in the range is an open set in the domain.

This last example hints at what we would be imposing on maps defined between measurable spaces. For simplicity, let us look at maps which land in $\mathbb{R}$.

Definition 2.1 (*Measurable map*)

Let (S, Σ) be a measurable space. We say that $h : S \to \mathbb{R}$ is Σ-measurable if for every $A \in \mathcal{B}(\mathbb{R})$, we have $h^{-1}(A) := \{x \in S : h(x) \in A\} \in \Sigma$.

Remark 2.2

1. *More generally, given another measurable space (S', Σ') and a map $h : S \to S'$, we say that h is (Σ, Σ')-measurable if for any $A \in \Sigma'$ we have $h^{-1}(A) \in \Sigma$. Therefore, a map is measurable if the reciprocal image of any measurable set is measurable.*

2. *When the start and end space are both topoligical spaces equipped with their Borel σ-algebra, then the definition of a measurable map coincide with of a continuous map. In this case, we also say that the map is Borelian.*

3. *In the probabilistic language, a measurable map is called a random variable. We will revisit this notion in the next chapter.*

Example 2.3

1. *Let $c \in \mathbb{R}$ and $h : S \to \mathbb{R}$ defined by $h(x) = c$ for every $x \in S$. Then h is Σ-measurable. Indeed, let A be a Borel set in $\mathbb{R}$: if $c \in A$ then $h^{-1}(A) = S \in \Sigma$ and if $c \notin A$ then $h^{-1}(A) = \varnothing \in \Sigma$.*

2. *Let $F \in \Sigma$, then $h := \mathbf{1}_F$ is Σ-measurable. Indeed, let A be a Borel set in $\mathbb{R}$: if 0 and 1 both belong to A, then $h^{-1}(A) = S \in \Sigma$; if $0 \in A$ and $1 \notin A$ then $h^{-1}(A) = F^c \in \Sigma$; if $0 \notin A$ and $1 \in A$ then $h^{-1}(A) = F \in \Sigma$; and finally if 0 and 1 do not belong to A then $h^{-1}(A) = \varnothing \in \Sigma$.*

3. *Let $S = \mathbb{R}$ and $\Sigma = \{\varnothing, \mathbb{R}\}$, and define h by $h(x) = 0$ if $x \geqslant 0$ and $h(x) = 3$ otherwise. Then h is not Σ-measurable since $h^{-1}(\{0\}) = [0, \infty) \notin \Sigma$.*

As for continuous functions where we had a kit of basic continuous functions (such as polynomials, rational functions, exponential, etc) and made sure that basic operations such as the sum, product, composition preserved the continuity, we will do the same for measurable maps. To this aim, let us first recall some basic properties of the reciprocal image.

Proposition 2.4

(i) We have $h^{-1}\big(\bigcup_n A_n\big) = \bigcup_n h^{-1}(A_n)$, $h^{-1}(A^c) = \big(h^{-1}(A)\big)^c$, $h^{-1}(\bigcap A_n) = \bigcap h^{-1}(A_n)$. In other words, the reciprocal image preserves common operations performed on sets.

(ii) Let $\mathcal{C} \subset \mathcal{B}(\mathbb{R})$ be a generating system i.e. $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$. If $h^{-1}(A) \in \Sigma$ for every $A \in \mathcal{C}$ then h is Σ-measurable.

(iii) If for every $c \in \mathbb{R}$ we have

$$\{h \leqslant c\} := h^{-1}\big((-\infty, c]\big) = \{x \in S : h(x) \leqslant c\} \in \Sigma,$$

then h is Σ-measurable.

Proof

(i) Exercise.

(ii) To obtain all elements of $\mathcal{B}(\mathbb{R})$ using the generating system, it is enough to perform the operations countable union/intersection and taking complements. From (i), the reciprocal image preserves all these operations and since $h^{-1}(A) \in \Sigma$ for every $A \in \mathcal{C}$ we get that $A \in \mathcal{B}(\mathbb{R})$ which proves that h is Σ-measurable.

(iii) We have seen that the class of intervals of the form $(-\infty, x]$ is a generating system of $\mathcal{B}(\mathbb{R})$. It remains to use (ii).

□

Proposition 2.5

(i) Let $h, h_1, h_2 : S \to \mathbb{R}$ be Σ-measurable maps and $\lambda \in \mathbb{R}$. Then λh, $h_1 + h_2$, $h_1 h_2$ are Σ-measurable.

(ii) Let $h : S \to \mathbb{R}$ Σ-measurable and $f : \mathbb{R} \to \mathbb{R}$ $\mathcal{B}(\mathbb{R})$-measurable. Then $f \circ h$ est Σ-measurable.

(iii) Let $(h_n)_{n\in\mathbb{N}}$ be a sequence of Σ-measurable maps. Then $\inf h_n$ (resp. $\sup h_n$) and $\liminf h_n$

(resp. $\limsup h_n$) are Σ-measurables (note that these maps take values in $[-\infty, \infty]$ and not in $\mathbb{R}$, thus one has to work with the Borel σ-algebra of $\bar{\mathbb{R}}$).

Proof

(i) Suppose for simplicity that $\lambda \geqslant 0$ (the other case can be treated in a similar manner) then $\{\lambda h \geqslant c\} = \{h \geqslant c/\lambda\}$ for every $c \in \mathbb{R}$. Since h is Σ-measurable then $\{h \geqslant c/\lambda\} \in \Sigma$. Thus $\{\lambda h \geqslant c\} \in \Sigma$ for every $c \in \mathbb{R}$ and by the previous proposition we conclude that λh is Σ-measurable.

Note that for every $c \in \mathbb{R}$ we have

$$\{h_1 + h_2 \geqslant c\} = \bigcup_{q \in \mathbb{Q}} \big(\{h_1 \geqslant q\} \bigcap \{h_2 \geqslant c - q\}\big).$$

This set is therefore measurable since it is the countable union of intersections of measurable sets. We proceed the same way for the product.

(ii) Let $c \in \mathbb{R}$ and note that $\{x : f \circ h(x) \leqslant c\} = h^{-1} \circ f^{-1}\big((-\infty, c]\big)$. Since f is $\mathcal{B}(\mathbb{R})$-measurable then $f^{-1}\big((-\infty, c]\big) \in \mathcal{B}(\mathbb{R})$ and since h is Σ-measurable then $h^{-1} \circ f^{-1}\big((-\infty, c]\big) \in \Sigma$ which shows that $f \circ h$ is Σ-measurable.

(iii) For the first point, it is enough to note that for every $c \in \mathbb{R}$, we have $\{\inf_n h_n \geqslant c\} = \bigcap_n \{h_n \geqslant c\}$. Since h_n is Σ-measurable then $\{h_n \geqslant c\} \in \Sigma$ and $\{\inf_n h_n \geqslant c\} \in \Sigma$ since it is the countable intersection of mesurable sets.

For the second point, note that $\liminf h_n = \lim_{r \to \infty} \inf_{n \geqslant r} h_n$. Similar to above, we have that $\inf_{n \geqslant r} h_n$ is Σ-measurable. Note also that $\inf_{n \geqslant r} h_n$ is non-decreasing in r. Thus, it is enough to prove that the limit of a non-decreasing sequence of measurable maps $(g_r)_{r \in \mathbb{N}}$ is measurable. To this aim, we note that $\lim_r g_r = \sup_r g_r$ and that by the previous argument this map is Σ-measurable.

□

2.2 Integration of nonnegative functions

As before, the strategy is to define the integral over "simple" objects then extend the definition to a general setting.

Definition 2.6 (*Step functions*)

Let $(S, \mathcal{A})$ be a measurable space and $f : S \to \mathbb{R}$ be a measurable function. We say that f is a staircase or step function if it takes a finite number of values. More precisely, if $a_1 < a_2 < \ldots < a_n$ are the values taken by f then we have

$$f(x) = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}(x),$$

where $A_i = f^{-1}(\{a_i\}) \in \mathcal{A}$ for every $i \in \{1, \ldots, n\}$.

Definition 2.7 (*Integral of nonnegative step functions*)

Let $(S, \mathcal{A}, \mu)$ be a measure space and let f be a nonnegative staircase function. The integral of f with respect to μ is defined by

$$\int f d\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

We make the convention that $0 \cdot \infty = 0$ when $a_i = 0$ and $\mu(A_i) = \infty$.

Remark 2.8

1. *If f was written with a different expression as $f = \sum_{i=1}^{m} b_i \mathbf{1}_{B_i}$ with the b_i's not necessarily distinct. Then, each set A_i is a disjoint union of the B_j's satisfying $b_j = a_i$. Since in such case, we have $\mu(A_i) = \sum_{\{j:\, b_j = a_i\}} \mu(B_j)$ then we obtain*

$$\sum_{j=1}^{m} b_j \mu(B_j) = \sum_{i=1}^{n} a_i \mu(A_i),$$

   and the definition stays consistent.

2. *In the probabilistic language, we speak of the expectation of a (discrete finite) random variable.*

3. *We say that a property on $(S, \mathcal{A}, \mu)$ holds true μ-almost surely if the set of $x \in S$ which do not satisfy it is of μ-measure zero. Clearly, if f and g are two nonnegative step functions such that $f = g$ μ-almost surely i.e. $\{x \in S : f(x) = g(x)\}$ is a μ-almost sure set, then $\int f\, d\mu = \int g\, d\mu$.*

Example 2.9

Consider $([0,1], \mathcal{B}([0,1]), \lambda)$ and define $f(x) = 1$ if $x \leqslant \frac{1}{3}$, $f(x) = 99$ if $x \in (\frac{1}{3}, \frac{2}{3}]$ and 0 otherwise. Clearly this is a step function since it only takes 3 values and we can write $f(x) = \mathbf{1}_{[0,\frac{1}{3}]} + 99\mathbf{1}_{(\frac{1}{3},\frac{2}{3}]}$. Therefore $\int f\, d\lambda = \lambda([0, \frac{1}{3}]) + 99\lambda((\frac{1}{3}, \frac{2}{3}]) = 33 + \frac{1}{3}$.

We easily verify the following properties.

Proposition 2.10

Let f and g be two nonnegative step functions.

(i) For all $\alpha, \beta \geqslant 0$, we have

$$\int (\alpha f + \beta g)\, d\mu = \alpha \int f\, d\mu + \beta \int g\, d\mu.$$

(ii) If $f \leqslant g$ then $\int f\, d\mu \leqslant \int g\, d\mu$.

Proof Exercise. □

The idea which we would like to exploit is that every nonnegative measurable function is the limit of nonnegative step functions. This will allow us to extend the definition of the integral to any nonnegative function. To this aim, we still need to address the continuity of integrals, and this is the goal of the monotone convergence theorem.

Proposition 2.11

Let f be a measurable function taking values in $[0, \infty]$. Then there exists a nondecreasing sequence of nonnegative step functions (f_n) such that $f = \lim_n f_n$.

Proof The idea is to truncate f at a high threshold and then proceed with dyadic approximation for the rest of the values. To this aim, we set for every $n \geqslant 1$

$$A_n = \{x \in S : f(x) \geqslant n\},$$

and for every $i \in \{0, 1, \ldots, n2^n - 1\}$ we define

$$B_{n,i} = \{x \in S : i2^{-n} \leqslant f(x) < (i+1)2^{-n}\}.$$

We then take

$$f_n = \sum_{i=0}^{n2^n-1} \frac{i}{2^n} \mathbf{1}_{B_{n,i}} + n\mathbf{1}_{A_n},$$

which is a step function. Let us now verify that $\lim_n f_n(x) = f(x)$ for every $x \in S$. Indeed, let $x \in S$ and note that if $f(x) = \infty$ then $x \in A_n$ and $f_n(x) \to \infty = f(x)$. Otherwise, for every $n > f(x)$ we have $x \in B_{n,i}$ with $i = \lfloor 2^n f(x) \rfloor$. Therefore, we have $f_n(x) = \frac{\lfloor 2^n f(x) \rfloor}{2^n}$ which converges to $f(x)$ when n goes to infinity. □

Definition 2.12 (*Integral of nonnegative measurable function*)

Let $(S, \mathcal{A}, \mu)$ be a measure space and $f : S \to [0, \infty]$ be a measurable function. We define the integral of f with respect to μ by

$$\int f \, d\mu = \sup_{\{h \text{ nonnegative step function}, h \leqslant f\}} \int h \, d\mu$$

Remark 2.13

1. *This definition coincide with the one we gave to step functions since if h is a step function and $h \leqslant f$ then $\int h \, d\mu \leqslant \int f \, d\mu$.*
2. *The previous property remains valid when we take the supremum over all step functions. Thus, if f and g are two nonnegative measurable functions and $f \leqslant g$ then $\int f \, d\mu \leqslant \int g \, d\mu$.*

Theorem 2.14 (*Monotone convergence theorem*)

Let $(f_n)_{n \in \mathbb{N}}$ be a nondecreasing sequence of measurable functions taking values in $[0, \infty]$. Then

$$\int (\lim_n f_n) \, d\mu = \lim_n \int f_n \, d\mu.$$

Proof Denote $f = \lim_n f_n$. Since f_n is a nondecreasing sequence then $f_n \leqslant f$ for every $n \in \mathbb{N}$. By the previous remark, we have $\int f_n \, d\mu \leqslant \int f \, d\mu$ for every $n \in \mathbb{N}$. Taking the limit, we get

$$\lim_n \int f_n \, d\mu \leqslant \int f \, d\mu.$$

It remains to prove the reverse inequality. Let $h = \sum_{i=1}^m a_i \mathbf{1}_{A_i}$ be a nonnegative step function with $h \leqslant f$. Let $\alpha < 1$ and

$$B_n = \{x \in S : \alpha h(x) \leqslant f_n(x)\}.$$

Note that $f_n \geqslant \alpha \mathbf{1}_{B_n} h$ and thus

$$\int f_n \, d\mu \geqslant \alpha \int \mathbf{1}_{B_n} h \, d\mu = \alpha \sum_{i=1}^m a_i \mu(A_i \cap B_n).$$

Since (f_n) is nondecreasing and $\alpha < 1$, then for any $x \in S$ there exists n such that $\alpha h(x) \leqslant f_n(x)$. Therefore, we have that $S = \bigcup_n B_n$ and that the B_n's are nondecreasing. Therefore, for every $i \leqslant m$, we have $(A_i \cap B_n) \uparrow A_i$ and by the monotone convergence lemma $\mu(A_i \cap B_n) \uparrow \mu(A_i)$. Thus we obtain

$$\lim_n \int f_n \, d\mu \geqslant \alpha \sum_{i=1}^m a_i \mu(A_i) = \alpha \int h \, d\mu.$$

This being true for any $\alpha < 1$, we let α go to 1 to get

$$\lim_n \int f_n \, d\mu \geqslant \int h \, d\mu.$$

This being true for any step function $h \leqslant f$, then we take the supremum on the right hand side to obtain

$$\lim_n \int f_n \, d\mu \geqslant \sup_{\{h \text{ nonnegative step function}, h \leqslant f\}} \int h \, d\mu = \int f \, d\mu.$$

□

Remark 2.15

We have already seen a particular case of the previous theorem, under the name of the monotone convergence lemma. Indeed, if A_n is a nondecreasing sequence of events then $(\mathbf{1}_{A_n})$ is a nondecreasing sequence of nonnegative measurable functions. We can therefore apply the monotone convergence theorem to $(\mathbf{1}_{A_n})$ to recover the monotone convergence lemma.

This result gain importance when combined with the preceding proposition. For every nonnegative function, the proposition ensures that it could be written as the nondecreasing limit of nonnegative step functions, and the monotone convergence theorem guarantees the commutation between limit and integral to obtain the integral of any nonnegative function as the limit of the integrals of nonnegative step functions.

With this in mind, we can deduce the following properties from their analogous ones for step functions.

Proposition 2.16

(i) Let f and g be two measurable functions taking values in $[0,\infty]$ and let $\alpha,\beta \geqslant 0$. We have

$$\int (\alpha f + \beta g)\, d\mu = \alpha \int f\, d\mu + \beta \int g\, d\mu.$$

(ii) If $(f_n)_{n\in\mathbb{N}}$ is a sequence of measurable functions taking values in $[0,\infty]$ then

$$\int \sum_n f_n\, d\mu = \sum_n \int f_n\, d\mu.$$

(iii) Let f be a measurable function taking values in $[0,\infty]$. Then $\int f\, d\mu = 0$ if and only if $f = 0$ μ-almost surely.

(iv) Let f, g be two measurable functions taking values in $[0,\infty]$. If $f = g$ μ-almost surely then $\int f\, d\mu = \int g\, d\mu$.

Proof

(i) Let $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be two nondecreasing sequences of nonnegative step functions whose limits are f and g respectively. Clearly $\alpha f_n + \beta g_n$ is a nondecreasing sequence of nonnegative step functions whose limit is $\alpha f + \beta g$. By the monotone convergence theorem, we obtain that

$$\int (\alpha f + \beta g)\, d\mu = \lim_n \int (\alpha f_n + \beta g_n)\, d\mu.$$

Moreover, we have seen that the integral of step functions is linear, thus

$$\int (\alpha f_n + \beta g_n)\, d\mu = \alpha \int f_n\, d\mu + \beta \int g_n\, d\mu.$$

Taking the limit and putting together the previous equations, we obtain the result.

(ii) Notice first that what we just proved implies that

$$\int \sum_{k=1}^{n} f_k\, d\mu = \sum_{k=1}^{n} \int f_k\, d\mu.$$

We therefore have

$$\sum_{n\in\mathbb{N}} \int f_n\, d\mu = \lim_{n\to\infty} \sum_{k=1}^{n} \int f_k\, d\mu = \lim_{n\to\infty} \int \sum_{k=1}^{n} f_k\, d\mu.$$

But $\sum_{k=1}^{n} f_k$ is nondecreasing since the f_k's are nonnegative, therefore by the monotone convergence theorem, we have

$$\sum_{n\in\mathbb{N}} \int f_n\, d\mu = \lim_{n\to\infty} \int \sum_{k=1}^{n} f_k\, d\mu = \int \lim_{n\to\infty} \sum_{k=1}^{\infty} f_k\, d\mu = \int \sum_{n\in\mathbb{N}} f_n\, d\mu$$

(iii) Exercise.

(iv) Exercise.

□

Corollary 2.17

Let f be a nonnegative measurable function. For every $A \in \mathcal{A}$, define $\nu(A) = \int f\mathbf{1}_A\, d\mu$ (which we also denote $\int_A f\, d\mu$). Then ν is a measure.

Proof Exercise. □

2.3 Integrals in the general case

Every function f can be decomposed into a nonnegative and a negative part by taking $f^+ = \sup(0, f)$ and $f^- = \sup(0, -f)$ so that $f = f^+ - f^-$. Using this, we can now extend the definition of the integral of nonnegative functions to the general setting.

Definition 2.18 (*Integrable functions*)

Let $(S, \mathcal{A}, \mu)$ be a measure space and $f : S \to \mathbb{R}$ be a measurable function. We say that f is μ-integrable if

$$\int |f|\, d\mu < \infty.$$

In this case, we set

$$\int f\, d\mu = \int f^+\, d\mu - \int f^-\, d\mu.$$

We denote $L^1(S, \mathcal{A}, \mu)$ the set of all functions which are μ-integrable.

We have the following properties.

Proposition 2.19

(i) Let $f \in L^1(S, \mathcal{A}, \mu)$. We have $|\int f\, d\mu| \leqslant \int |f|\, d\mu$.

(ii) If $f, g \in L^1(S, \mathcal{A}, \mu)$ and $\alpha, \beta \in \mathbb{R}$ then $\int(\alpha f + \beta g)\, d\mu = \alpha \int f\, d\mu + \beta \int g\, d\mu$.

(iii) If $f, g \in L^1(S, \mathcal{A}, \mu)$ and $f \leqslant g$ then $\int f\, d\mu \leqslant \int g\, d\mu$.

(iv) If $f, g \in L^1(S, \mathcal{A}, \mu)$ and $f = g$ μ-almost surely then $\int f\, d\mu = \int g\, d\mu$.

Proof Exercise. □

Therefore, we extended the properties which we had for the integrals of step functions to the general case. However, it remains to check the property established in the monotone convergence theorem which allowed us to interchange the order between limit and integral. In the general case, we will be able to do so under certain constraints as we will see in the dominated convergence theorem. We first need the following lemma.

Lemme 2.20 (*Fatou Lemma*)

Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative measurable functions. Then we have

$$\int \liminf f_n\, d\mu \leqslant \liminf \int f_n\, d\mu.$$

Proof We write $\liminf f_n = \lim_{m\to\infty} \inf_{n\geqslant m} f_n = \lim_{m\to\infty} g_m$. But $(g_m)_{m\in\mathbb{N}}$ is a nondecreasing sequence of measurable nonnegative functions, so by the monotone convergence theorem, we have

$$\int \lim_{m\to\infty} g_m \, d\mu = \lim_{m\to\infty} \int g_m \, d\mu.$$

But $g_m \leqslant f_n$ for every $n \geqslant m$, so $\int g_m \, d\mu \leqslant \int f_n \, d\mu$ for every $n \geqslant m$. We deduce that $\int g_m \, d\mu \leqslant \inf_{n\geqslant m} \int f_n \, d\mu$. Putting together the above, we get

$$\int \lim_{m\to\infty} g_m \, d\mu = \lim_{m\to\infty} \int g_m \, d\mu \leqslant \lim_{m\to\infty} \inf_{n\geqslant m} \int f_n \, d\mu.$$

□

Remark 2.21

We have already proved a particular case of this lemma in the previous Chapter. Indeed if A_n are events and $\mu = \mathbb{P}$ is a probability measure, we set $f_n = \mathbf{1}_{A_n}$ and note that $\int f_n \, d\mu = \int \mathbf{1}_{A_n} \, d\mathbb{P} = \mathbb{P}(A_n)$. On the other hand, we have seen that $\liminf \mathbf{1}_{A_n} = \mathbf{1}_{\liminf A_n}$ and thus

$$\int \liminf f_n \, d\mu = \int \mathbf{1}_{\liminf A_n} \, d\mathbb{P} = \mathbb{P}(\liminf A_n).$$

Therefore, this lemma implies that $\mathbb{P}(\liminf A_n) \leqslant \liminf \mathbb{P}(A_n)$.

Theorem 2.22 (*Dominated convergence theorem*)

Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions in $L^1(S, \mathcal{A}, \mu)$ such that $\lim_n f_n(x) = f(x)$ μ-almost surely and there exists a nonnegative, measurable, integrable function g with $\forall n \in \mathbb{N}$, $|f_n| \leqslant g$ μ-almost surely. Then, we have $f \in L^1(S, \mathcal{A}, \mu)$ and

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu, \quad \text{and} \quad \lim_{n\to\infty} \int |f_n - f| \, d\mu = 0.$$

Proof Let $A = \{x \in S : \lim_n f_n(x) = f(x) \text{ and } |f_n(x)| \leqslant g(x) \ \forall n \in \mathbb{N}\}$. By hypothesis, we have $\mu(A^c) = 0$. Set

$$\widetilde{f}_n(x) = f_n(x)\mathbf{1}_A(x) \quad \text{and} \quad \widetilde{f}(x) = f(x)\mathbf{1}_A(x).$$

We clearly have that $\widetilde{f}_n = f_n$ and $\widetilde{f} = f$ μ-almost surely. Thus, we have $\int \widetilde{f}_n \, d\mu = \int f_n \, d\mu$, $\int \widetilde{f} \, d\mu = \int f \, d\mu$ and $\int |\widetilde{f}_n - \widetilde{f}| \, d\mu = \int |f_n - f| \, d\mu$. We can therefore work with $\widetilde{f}_n$ and $\widetilde{f}$ in place of f_n and f.

Since $|\widetilde{f}| \leqslant g$ and $\int g \, d\mu < \infty$ then $\int |f| \, d\mu = \int |\widetilde{f}| \, d\mu < \infty$ and $f \in L^1(S, \mathcal{A}, \mu)$.

We have $2g - |\widetilde{f} - \widetilde{f}_n| \geqslant 0$ then by Fatou lemma, we have

$$\liminf \int (2g - |\widetilde{f} - \widetilde{f}_n|) \, d\mu \geqslant \int \liminf (2g - |\widetilde{f} - \widetilde{f}_n|) \, d\mu = 2 \int g \, d\mu.$$

We deduce that $\limsup \int |\widetilde{f} - \widetilde{f}_n| \, d\mu = 0$ so in particular $\lim \int |f - f_n| \, d\mu = 0$. Finally, by the linearity of integrals, we have

$$\Big| \int f \, d\mu - \int f_n \, d\mu \Big| = \Big| \int (f - f_n) \, d\mu \Big| \leqslant \int |f - f_n| \, d\mu \underset{n\to\infty}{\to} 0.$$

□

2.4 Product measure

We are given two measurable spaces $(S_1, \mathcal{A}_1)$ and $(S_2, \mathcal{A}_2)$, and we would like to define the product space.

Definition 2.23 (*Product σ-algebra*)

Let $(S_1, \mathcal{A}_1)$ and $(S_2, \mathcal{A}_2)$ be two measurable spaces. The product σ-algebra on $S_1 \times S_2$ is given by $\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma(A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2)$.

Remark 2.24

1. *Denote ρ_1 and ρ_2 the canonical projections i.e. $\rho_1 : S_1 \times S_2 \to S_1$ is defined by $\rho_1(s_1, s_2) = s_1$, and $\rho_2 : S_1 \times S_2 \to S_2$ is defined by $\rho_2(s_1, s_2) = s_2$. Then we can verify that $\mathcal{A}_1 \otimes \mathcal{A}_2$ is the smallest σ-algebra which make ρ_1 and ρ_2 measurable. More precisely, we have that $\mathcal{A}_1 \otimes \mathcal{A}_2$ is the σ-algebra generated by sets of the form $\rho_1^{-1}(A_1)$, $A_1 \in \mathcal{A}_1$, and $\rho_2^{-1}(A_2)$, $A_2 \in \mathcal{A}_2$. Therefore, we have $\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma(A_1 \times S_2, S_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2)$.*

2. *We can of course extend this definition to a finite product of measurable spaces.*

The next step is to characterize the measurable maps on product spaces.

Proposition 2.25

Let $f : (S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \to \mathbb{R}$ be a $\mathcal{A}_1 \otimes \mathcal{A}_2$-measurable map. Then for any $s_1 \in S_1$, the map $s_2 \to f(s_1, s_2)$ is $\mathcal{A}_2$-measurable. Similarly, for any $s_2 \in S_2$, the map $s_1 \to f(s_1, s_2)$ is $\mathcal{A}_1$-measurable.

Proof We will prove that every $\mathcal{A}_1 \otimes \mathcal{A}_2$-measurable step function verifies the conclusion of the proposition. Since every measurable function is the limit of step functions, we would obtain the result. It is enough to prove that for any $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$, $\mathbf{1}_A$ verifies the conclusion of (i). Let $\mathcal{M} = \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 : \mathbf{1}_A \text{ verifies the conclusion of (i)}\}$. We easily check that $\mathcal{I} \subset \mathcal{M}$ and that $\mathcal{M}$ is a monotone class. Therefore, by the monotone classes lemma, we obtain that $\mathcal{M} = \sigma(\mathcal{I}) = \mathcal{A}_1 \otimes \mathcal{A}_2$. □

Our goal now is to define a measure on $\mathcal{A}_1 \otimes \mathcal{A}_2$. We would like, as in the simple case, to say that the measure of a set is nothing but the integral with respect to this measure of the set indicator function. The next proposition, which also serves as a definition for the product measure, allows us to assert that the order of integration with respect to μ_1 and μ_2 does not matter.

Proposition 2.26 (*Product measure*)

Let $(S_1, \mathcal{A}_1, \mu_1)$ and $(S_2, \mathcal{A}_2, \mu_2)$ be two measure space with μ_1 and μ_2 σ-finite. Then, we have that

$$\mu(A) = \int_{S_1} \Big(\int_{S_2} \mathbf{1}_A(s_1, s_2)\, d\mu_2 \Big)\, d\mu_1 = \int_{S_2} \Big(\int_{S_1} \mathbf{1}_A(s_1, s_2)\, d\mu_1 \Big)\, d\mu_2,$$

is a measure on $\mathcal{A}_1 \otimes \mathcal{A}_2$, called the product measure of μ_1 and μ_2. We will denote μ by $\mu_1 \otimes \mu_2$.

Proof Suppose that μ_1 and μ_2 are finite. We should verify that

$$\int_{S_1} \Big(\int_{S_2} \mathbf{1}_A(s_1, s_2)\, d\mu_2 \Big)\, d\mu_1 = \int_{S_2} \Big(\int_{S_1} \mathbf{1}_A(s_1, s_2)\, d\mu_1 \Big)\, d\mu_2$$

and that μ such as defined is a measure. For the first point, set

$$\mathcal{M} = \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 : \ \mathbf{1}_A \text{ verifies the equality}\}.$$

Clearly if $A = A_1 \times A_2$ then $\mathbf{1}_A(s_1, s_2) = \mathbf{1}_{A_1}(s_1)\mathbf{1}_{A_2}(s_2)$ and thus $\mathbf{1}_A$ verifies the equality. Therefore $\mathcal{I} \subseteq \mathcal{M}$ and it is enough to show that $\mathcal{M}$ is a monotone class to deduce that $\mathcal{M} = \mathcal{A}_1 \otimes \mathcal{A}_{\in}$. Let $A, B \in \mathcal{M}$ with $A \subset B$ and let us show that $B \backslash A \in \mathcal{M}$. But $\mathbf{1}_{B\backslash A} = \mathbf{1}_B - \mathbf{1}_A$. Since μ_1 and μ_2 are finite, we can assert that $\mathbf{1}_B - \mathbf{1}_A$ is integrable and we can use linearity to deduce that $B\backslash A \in \mathcal{M}$. Finally, we use the monotone convergence theorem to show that $\mathcal{M}$ is stable by nondecreasing union. To extend to the case of σ-finite measures, we take nondecreasing sequence $(C_n)_{n\in\mathbb{N}} \in \mathcal{A}_1$ and $(D_n)_{n\in\mathbb{N}} \in \mathcal{A}_2$ such that $\mu_1(C_n)\infty$, $\mu_2(D_n) < \infty$, and $S_1 = \bigcup_{n\in\mathbb{N}} C_n$ and $S_2 = \bigcup_{n\in\mathbb{N}} D_n$. We repeat the same procedure with $\mu_{1,n}(\cdot) = \mu_1(\cdot \cap C_n)$ and $\mu_{2,n}(\cdot) = \mu_2(\cdot \cap D_n)$ which are finite measures and which converge to μ_1 and μ_2 respectively. Using the monotone convergence theorem, we finish the proof.

It remains to verify that this indeed defines a measure. To this aim, we use that if A and A' are disjoint then $\mathbf{1}_{A\cup A'} = \mathbf{1}_A + \mathbf{1}_{A'}$ and we use the linearity of integrals. This allows us to get that $\mu(A \cup A') = \mu(A) + \mu(A')$. To extend this for a countable union, we make use of the monotone convergence theorem. □

Remark 2.27

1. *Note that the finiteness hypothesis is crucial. Indeed, if* $(S_1, \mathcal{A}_1) = (S_2, \mathcal{A}_2) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *and* $\mu_1 = \lambda$ *and* $\mu_2 = \nu$ *is the counting measure. Then if* $A = \{(x, x) : x \in \mathbb{R}\}$, *we clearly have*

$$\int_{\mathbb{R}} \mathbf{1}_A(s_1, s_2)\, d\lambda = \lambda(\{s_2\}) = 0$$

and

$$\int_{\mathbb{R}} \mathbf{1}_A(s_1, s_2)\, d\nu = \nu(\{s_1\}) = 1.$$

Therefore

$$0 = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbf{1}_A(s_1, s_2)\, d\lambda \Big)\, d\nu \neq \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbf{1}_A(s_1, s_2)\, d\nu \Big)\, d\lambda = \infty.$$

2. *To identify the variable on which we integrate we may denote* $\int_{S_2} \mathbf{1}_A(s_1, s_2)\, d\mu_2 = \int_{S_2} \mathbf{1}_A(s_1, s_2)\, \mu_2(ds_2)$. *We will omit this when the variable is clear from the context.*

Theorem 2.28 (*Fubini*)

Let $(S_1, \mathcal{A}_1, \mu_1)$ *and* $(S_2, \mathcal{A}_2, \mu_2)$ *be two measure spaces and* μ_1, μ_2 σ*-finite.*

(i) The product measure $\mu_1 \otimes \mu_2$ *is the unique measure on* $(S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ *which satisfies*

$$\mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2),$$

for all $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$.

(ii) Let $f : (S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \to [0, \infty]$ *nonnegative measurable. Then*

$$\int f\, d\mu_1 \otimes \mu_2 = \int_{S_1} \Big(\int_{S_2} f(s_1, s_2)\, d\mu_2 \Big)\, d\mu_1 = \int_{S_2} \Big(\int_{S_1} f(s_1, s_2)\, d\mu_1 \Big)\, d\mu_2.$$

(iii) If $f \in L^1(S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu_1 \otimes \mu_2)$, then

$$\int f \, d\mu_1 \otimes \mu_2 = \int_{S_1} \Big(\int_{S_2} f(s_1, s_2) \, d\mu_2 \Big) \, d\mu_1 = \int_{S_2} \Big(\int_{S_1} f(s_1, s_2) \, d\mu_1 \Big) \, d\mu_2.$$

Proof

(i) By definition, we have that for all $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$

$$\mu_1 \otimes \mu_2(A_1 \times A_2) = \int_{S_1} \Big(\int_{S_2} \mathbf{1}_{A_1 \times A_2}(s_1, s_2) \, d\mu_2 \Big) \, d\mu_1.$$

But $\mathbf{1}_{A_1 \times A_2}(s_1, s_2) = \mathbf{1}_{A_1}(s_1)\mathbf{1}_{A_2}(s_2)$, thus

$$\mu_1 \otimes \mu_2(A_1 \times A_2) = \int_{S_1} \Big(\int_{S_2} \mathbf{1}_{A_1}(s_1)\mathbf{1}_{A_1}(s_2) \, d\mu_2 \Big) \, d\mu_1 = \mu_1(A_1)\mu_2(A_2).$$

The unicity follows from Corollary 1.17. Indeed, if there exists another measure μ such that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ then $\mu_1 \otimes \mu_2$ and μ coincide on the measurable rectangles which form a generating π-system. Therefore, by Corollary 1.17 we deduce that $\mu = \mu_1 \otimes \mu_2$.

(ii) The previous proposition/definition shows that this is true for indicator functions. By linearity of the integral, this shows its validity for any nonnegative step function. Since every nonnegative measurable function is a nondecreasing limit of nonnegative step functions, then the monotone convergence theorem implies the result.

(iii) It is enough to decompose f into a nonnegative and negative part which are both integrable and verify the corresponding equality obtained in (ii). The linearity finishes the proof.

□

Remark 2.29

1. *The fact that $f \in L^1(S_1 \times S_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu_1 \otimes \mu_2)$ is crucial. Indeed if $f(x, y) = 2e^{-2xy} - e^{-xy}$ is defined on $(0, \infty) \times (0, 1]$. Then*

$$\int_{(0,\infty} f(x,y)dx = 0 \quad \text{and} \quad \int_{(0,1]} f(x,y)dy = \frac{e^{-x} - e^{-2x}}{x}.$$

Therefore, we would get

$$0 = \int_{(0,1]} \Big(\int_{(0,\infty)} f(x,y)dx \Big) dy \neq \int_{(0,\infty)} \Big(\int_{(0,1]} f(x,y) \, dy \Big) \, dx > 0.$$

2. *A particular case which we will use in the sequel is the following. Let μ_1 be the Lebesgue measure (which is σ-finite) and μ_2 be a probability measure. If $f(x_1, x_2)$ is nonnegative and $x_1 \to f(x_1, x_2)$ is integrable with respect to μ_1 and $x_2 \to \int f(x_1, x_2) \, \mu_1(dx_1)$ is bounded, then $f \in L^1$.*

Indeed, since f is nonnegative we can apply Fubini to get

$$\int f \, d(\mu_1 \otimes \mu_2) = \int \Big(\int f(x_1, x_2) \, \mu_1(dx_1) \Big) \, \mu_2(dx_2)$$

Since $x_2 \to \int f(x_1, x_2)\, \mu_1(dx_1)$ *is bounded and* μ_2 *is a probability, then the above quantity is finite and* $f \in L^1$.

Therefore, if we have an arbitrary function g *such that* $|g| \leqslant f$ *and* f *verifies the above conditions, we can assert that* $g \in L^1$ *and use Fubini to interchange the integrals for* g *as well.*

Chapter 3

Random Variables

3.1 Random variable

When we switch to the probabilistic language, measurable maps are referred to as random variables.

Definition 3.1 (*Random variable*)

Let Ω be a set and $\mathcal{A}$ be a family of events on Ω. A real random variable X is a map $X : \Omega \to \mathbb{R}$ which is $\mathcal{A}$-measurable.

Remark 3.2

1. *One should not forget that a random variable is a map itself. This is something to check, even if it is often clear, before verifying measurability.*
2. *Similarly, we can talk about random variables taking values in other measurable spaces, for instance $\mathbb{R}^n$. The measurability is then with respect to $\big(\mathcal{A}, \mathcal{B}(\mathbb{R}^n)\big)$.*
3. *Intuitively, a real random variable is a machine which produces numbers. This machine accepts say different types of coins (the samples) and for each one of them produces a number (say the number written on the coin). What we just described is the "mapping" aspect of the random variable. It turns out that this machine has an encrypted code to associate numbers in a manner which is not completely arbitrary. For simplicity, imagine that the machine produces a finite number (or countable) of outputs. The crypt rule of the machine stipulates that for each output, the set of coins which have this number written on it belongs to the kit $\mathcal{A}$. Therefore, $\mathcal{A}$ plays the role of a crypt which we add to the machine. However, this crypt is not very constraining since we can construct several machines which satisfy it which means that there are still several choices of outputs and the input of $\mathcal{A}$ is not enough to identify them, thus the word "random".*
4. *Sometimes the starting set is ambiguous, and as we will see, we can often define it as the set of outcomes of an experiment.*
5. *We define a random variable to keep track of some quantity which is of interest to us during an experiment.*

Example 3.3

1. *During this year, Al Jazira and Al Wahda played* 6 *times. There are* 3^6 *possible results (either one of them wins, or there is a draw). We are interested in the number of games won by Al Jazira and thus define* X *to be the random variable equal to the number of games won by Al Jazira. Let* $X : \Omega \to \{0, \dots, 6\}$, *we can choose* Ω *to be the set of all possible results of the experiment; for instance* $\{AlJazira, AlJazira, AlJazira, AlJazira, AlJazira, draw\}$ *is an element of* Ω, *it corresponds to a sequence indicating the name of the winner to each of the* 6 *games and "draw" indicates a draw game. When we do not specify the family of events, it means that we work with the set of all subsets of* Ω.

2. *We roll a dice twice and we are interested in the sum of the numbers obtained. We denote* Ω *the set of all possible outcomes of the experiment i.e.* $\Omega = \big\{(i,j) : i,j \in \{1,\dots,6\}\big\}$. *Therefore* $X : \Omega \to \{2, \dots, 12\}$ *is defined by* $X(i,j) = i + j$. *If we equip* Ω *of the* σ*-algebra* $\mathcal{A} := \big\{\emptyset, \Omega, \{(1,1)\}, \Omega\backslash\{(1,1)\}\big\}$ *then* X *is not a random variable since* $X^{-1}(\{3\}) = \{(1,2),(2,1)\} \notin \mathcal{A}$. *However, once we equip it with the* σ*-algebra* $\mathcal{P}(\Omega)$, *it becomes a properly defined random variable.*

3. *We are interested in the number of heads obtained when we toss a coin* n *times. Let* $\Omega = \{H, T\}^n$ *and let* $\mathcal{A} = \mathcal{P}(\Omega)$. *For every* $i \in \{1, \dots, n\}$, *we define* $X_i : \Omega \to \{0, 1\}$ *by* $X_i(\omega) = 1$ *if* $\omega_i = H$ *and* 0 *otherwise, for every* $\omega = (\omega_1, \dots, \omega_n) \in \Omega$. *Thus* X_i *tells us whether the* i*-th coin toss is head or not. Clearly* X_i *is a random variable.*

   To count the number of heads in n *tosses, we define* $X = X_1 + X_2 + \ldots + X_n$. *Since each* X_i *is a random variable and* X *is the sum then* X *is also a random variable.*

There exists a relatively simple class of random variables which take a finite number of values. These will be very useful for us.

Definition 3.4 (*Staircase random variables*)

A random variable $X : (\Omega, \mathcal{A}) \to \mathbb{R}$ *is said to be staircase if there exist numbers* $a_1, \dots, a_n \in \mathbb{R}$ *and events* $A_1, \dots, A_n \in \mathcal{A}$ *such that* $X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$.

It turns out that any random variable is the limit of a sequence of staircase random variables. This is what we have seen and proved in the previous chapter.

Lemme 3.5

Let $X : (\Omega, \mathcal{A})$ *be a random variable.*

(i) If X *is nonnegative, then* X *is a limit of a nondecreasing sequence of staircase random variables i.e. there exists a sequence of nondecreasing sequence of staircase random variables* $X_1 \leqslant X_2 \ldots$ *such that for any* $\omega \in \Omega$, *we have* $X_n(\omega) \underset{n\to\infty}{\to} X(\omega)$.

(ii) If X *is arbitrary, then* X *is a limit of a sequence of staircase random variables.*

Proof

(i) Already done.

(ii) Just write $X = X^+ - X^-$ where X^+ and X^- are defined by $X^+(\omega) = \max\big(0, X(\omega)\big)$ and $X^-(\omega) = -\min\big(0, X(\omega)\big)$. It remains to apply (i) for X^+ and X^- to get the result.

□

As we have noted in Example 3.3, taking the set of all subsets of Ω as σ-algebra guaranteed the measurability of the random variable. However, we do not need this and we can sometimes choose a much smaller σ-algebra. In a similar manner to what is done for continuous functions, where we equip the space with the topology which makes the function continuous. We will do the same here.

Definition 3.6 (*σ-algebra generated by a map*)

Let $X : \Omega \to \mathbb{R}$ be a map. Then we define $\sigma(X)$ as the smallest σ-algebra on which X is a random variable.

Remark 3.7

1. *If $X : (\Omega, \mathcal{A}) \to \mathbb{R}$ is a random variable, then $\sigma(X) \subseteq \mathcal{A}$.*
2. *The same definition stays valid for random variables taking values in another measurable space.*
3. *We have $\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}$. Indeed, it is clear that X is $\{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}$-measurable which implies that $\sigma(X) \subseteq \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}$. On the other hand, since X is $\sigma(X)$-measurable then for every $A \in \mathcal{B}(\mathbb{R})$ we have $X^{-1}(A) \in \sigma(X)$ which implies that $\{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\} \subseteq \sigma(X)$.*

Example 3.8

1. *Let $c \in \mathbb{R}$ and let $X : \mathbb{R} \to \mathbb{R}$ defined by $X(a) = c$ for every $a \in \mathbb{R}$. Then $\sigma(X) = \{\varnothing, \mathbb{R}\}$.*
2. *Let Ω be a set and $A \subset \Omega$ be a subset of Ω. Let $X = \mathbf{1}_A$ be the indicator of A. Then $\sigma(X) = \{\varnothing, A, A^c, \Omega\}$.*
3. *In the three example of Example 3.3, we had $\sigma(X) = \mathcal{P}(\Omega)$.*

Proposition 3.9

Let $X, Y : (\Omega, \mathcal{A}) \to \mathbb{R}$ be two random variables. Then the following two assertions are equivalent:

(i) Y is $\sigma(X)$-measurable.

(ii) There exists a Borel function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$.

Proof We have already seen that the composition of measurable maps produces a measurable map. This means that if $Y = f(X)$ then Y is $\sigma(X)$-measurable. It remains to prove that (i) implies (ii). Suppose first that Y is staircase and write $Y = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ where $A_i \in \sigma(X)$. For every $i \leqslant n$, take $B_i \in \mathcal{B}(\mathbb{R})$ such that $A_i = X^{-1}(B_i)$. Therefore

$$Y = \sum_{i=1}^n a_i \mathbf{1}_{X^{-1}(B_i)} = \sum_{i=1}^n a_i \mathbf{1}_{B_i} \circ X = f \circ X,$$

with $f = \sum_{i=1}^{n} a_i \mathbf{1}_{B_i}$ which is Borelian.
In the general case, we have seen that Y is a limit of staircase random variables which are of the form $Y_n = f_n(X)$ with f_n Borelian. We set $f(x) = \lim_{n\to\infty} f_n(x)$ if this limit exists and 0 otherwise. f is then measurable, and $f(X(\omega)) = \lim_n f_n(X(\omega)) = Y(\omega)$. □

3.2 Distribution of a random variable

We now move to the next step by adding a probability to our measurable space.

Definition 3.10

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X be a random variable on it. We define the distribution (or law) of X as the function $\mathcal{L}_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ defined by $\mathcal{L}_X(A) = \mathbb{P}\big(X^{-1}(A)\big)$ for every $A \in \mathcal{B}(\mathbb{R})$.

Remark 3.11

1. *The distribution of X is a probability measure on $\mathcal{B}(\mathbb{R})$. (Exercise)*
2. *The idea is that we transport, using X, the probability measure $\mathbb{P}$ we have on $\mathcal{A}$ to get another one on $\mathcal{B}(\mathbb{R})$. This probability measure $\mathcal{L}_X$ is referred to as the pushforward measure of $\mathbb{P}$ by the measurable map X.*
3. *In the sequel, given a random variable, we are interested in the mapping aspect of the random variable, its measurability, as well as the probability measure which is generated by its distribution.*

Example 3.12

We recall the three examples of Example 3.3 in the same order:

1. *We define a probability measure on Ω by stipulating that $\mathbb{P}(\{\omega\}) = 0$ if ω contains the word "draw". In other words, the probability that there is at least one draw game among the six is equal to zero. We stipulate that the other outcomes are all equi-probable. Therefore, $\mathcal{L}_X$ is defined by*

$$\mathcal{L}_X(\{k\}) = \mathbb{P}\big(X^{-1}(\{k\})\big) = \frac{\binom{6}{k}}{2^6},$$

   for every $k = 0, \dots, 6$. Note that since the sets $\{k\}$, $k = 0, \dots, 6$, form (when adding the empty set) a generatingπ-system of $\sigma(\{0, \dots, 6\})$, and by Theorem 1.18, it is enough to define the probability on the generating π-system since the extension is unique.

2. *We suppose that all outcomes have the same probability. Therefore, the distribution of X is given by*

$$\mathcal{L}_X(\{2\}) = \mathbb{P}(\{(1,1)\}) = \frac{1}{36},\ \mathcal{L}_X(\{3\}) = \mathbb{P}(\{(1,2),(2,1)\}) = \frac{1}{18},\ \mathcal{L}_X(\{4\}) = \frac{1}{12},$$

   and so on. As before, it is enough to define $\mathcal{L}_X$ on the generating π-system.

3. *We also suppose that all outcomes are equi-probable, therefore $\mathbb{P}(\{\omega\}) = 2^{-n}$ for every $\omega \in \{P, F\}^n$. To obtain $X = k$ for some $k \in \{0, \dots, n\}$, we should have an element ω containing k*

times H. We denote Ω_k the set of all ω's containing k times H. Clearly, we have $|\Omega_k| = \binom{n}{k}$. Therefore,

$$\mathcal{L}_X(\{k\}) = \mathbb{P}(X^{-1}(\{k\})) = \mathbb{P}(\Omega_k) = \sum_{\omega\in\Omega_k} \mathbb{P}(\{\omega\}) = 2^{-n}|\Omega_k| = 2^{-n}\binom{n}{k}.$$

Definition 3.13 (*Discrete random variable*)

Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a probability space and X be a random variable on this space. If the set of values taken by X is countable then we say X is a discrete random variable. The distribution of X is then given by $\mathcal{L}_X = \sum_{a\in\operatorname{Im} X} \mathbb{P}(X = a)\,\delta_a$ where δ_ω is the Dirac measure on ω which is given by $\delta_\omega(A) = \mathbf{1}_A(\omega)$ for every $A \in \mathcal{P}(\Omega)$.

Remark 3.14

1. *Note that the notation $\mathbb{P}(X = a)$ is an abbreviation for $\mathbb{P}(\{X = a\})$ which is the probability of the event $\{\omega \in \Omega : X(\omega) = a\} = X^{-1}(\{a\})$.*
2. *A staircase random variable is discrete.*
3. *Writing the distribution as $\sum_{\omega\in\Omega} \mathbb{P}(\{\omega\})\,\delta_\omega$ means that it is enough to calculate the probability of each value taken by X to derive the distribution. This is coherent with we have previously seen since the set of singletons (together with the empty set) is a π-system and it is enough to define the probability on it to obtain it in a unique way over the whole σ-algebra using Theorem 1.18.*
4. *We would like to verify manually the claim we just made. Indeed, we can write*

$$\mathcal{L}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}\Big(\bigcup_{a\in A}\{X = a\}\Big) = \sum_{a\in A} \mathbb{P}(X = a) = \sum_{a\in\operatorname{Im} X} \mathbb{P}(X = a)\delta_a(A).$$

5. *If we revisit the calculation of the distributions in Example 3.12, then for (1) we can write that $\mathcal{L}_X = \sum_{k=0}^{6} 2^{-6}\binom{6}{k}\delta_k$. For (2), we can write $\mathcal{L}_X = \frac{1}{36}\delta_2 + \frac{1}{18}\delta_3 + \frac{1}{12}\delta_4 + \dots$*

Definition 3.15 (*Random variable with density*)

Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a probability space and X be a random variable on this space. We say that X has a density if there exists a Borelian function $f : \mathbb{R} \to \mathbb{R}_+$ such that $\mathcal{L}_X(A) = \int_A f(x)dx$ for every $A \in \mathcal{B}(\mathbb{R})$.

The above definition as the one preceding it extends in the trivial way to random variables taking value in $\mathbb{R}^n$.

Example 3.16 (*Classical discrete distributions*)

1. *Uniform distribution: Let S be a finite set of cardinality n. We say that a random variable X follows a uniform distribution on S if $\mathbb{P}(X = a) = \frac{1}{n}$ for every $a \in S$.*
2. *Bernoulli distribution with parameter $p \in [0, 1]$. This is the distribution of the random variable X taking values in $\{0, 1\}$ such that $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.*

Note that this describes only the "distribution aspect" of the random variable and does not say anything about the "map aspect" apart from the range of the random variable. Note that every Bernoulli random variable can be represented as the indicator of some event with probability p. On the canonical space, an example of a Bernoulli random variable is given by $\mathbf{1}_{[0,p]}$.

3. *The binomial distribution $\mathcal{B}(n, p)$. This is the distribution of the random variable X taking values in $\{0, \dots, n\}$ such that $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for every $k \in \{0, \dots, n\}$.*

4. *The geometric distribution with parameter p. This is the distribution of the random variable X taking values in $\mathbb{N}^*$ such that $\mathbb{P}(X = k) = (1-p)^{k-1} p$ for every $k \in \mathbb{N}^*$.*

5. *The Poisson distribution with parameter $\alpha > 0$. This is the distribution of the random variable X taking values in $\mathbb{N}$ such that $\mathbb{P}(X = k) = \frac{\alpha^k}{k!} e^{-\alpha}$ for every $k \in \mathbb{N}$.*

Example 3.17 (*Classical distributions with density*)

1. *The uniform distribution on $[a, b]$. This is the distribution of the random variable X taking values in $\mathbb{R}$ with density function $f(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x)$.*

2. *The exponential distribution with parameter $\alpha > 0$. This is the distribution of the random variable X taking values in $\mathbb{R}$ with density function$f(x) = \alpha e^{-\alpha x} \mathbf{1}_{\mathbb{R}_+}(x)$.*

3. *The Gaussian (Normal) distribution $\mathcal{N}(m, \sigma^2)$. This is the distribution of the random variable X taking values in $\mathbb{R}$ with density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\big(- \frac{(x-m)^2}{2\sigma^2} \big)$.*

3.3 Cumulative distribution function

Definition 3.18 (*CDF of a random variable*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X be a real random variable. We define the Cumulative Distribution Function (CDF for short) of X by $F_X : \mathbb{R} \to [0, 1]$ with

$$F_X(t) = \mathcal{L}_X(-\infty, t] = \mathbb{P}(X \leqslant t),$$

for every $t \in \mathbb{R}$.

Remark 3.19

1. *The CDF determines completely the distribution of X. Indeed, the set $\{(-\infty, t] : t \in \mathbb{R}\}$ forms a π-system which generates $\mathcal{B}(\mathbb{R})$, and thus by Corollary 1.17, F_X characterizes uniquely $\mathcal{L}_X$. Concretely, this means that to find the distribution of a real random variable, it is enough to find its CDF. However, one should keep in mind that this does not determine the random variable itself but only the "distribution" aspect of it.*

2. *Clearly, F_X is nondecreasing since if $t \leqslant t'$ then $(-\infty, t] \subseteq (-\infty, t']$ and thus $\mathbb{P}((-\infty, t]) \leqslant \mathbb{P}((-\infty, t'])$.*

3. *We also have* $\lim_{t\to-\infty} F_X(t) = \lim_{t\to-\infty} \mathbb{P}(X \in (-\infty,t]) = \mathbb{P}(X \in \bigcap_{t\in\mathbb{R}}(-\infty,t]) = \mathbb{P}(X \in \varnothing) = 0$.

4. *We also have* $\lim_{t\to\infty} F_X(t) = \lim_{t\to\infty} \mathbb{P}(X \in (-\infty,t]) = \mathbb{P}(X \in \bigcup_{t\in\mathbb{R}}(-\infty,t]) = \mathbb{P}(X \in \mathbb{R}) = 1$.

5. *Finally,* F_X *is continuous from the right. Indeed, we have*
$$\lim_{n\to\infty} \mathbb{P}(X \leqslant t + n^{-1}) = \mathbb{P}(X \in \bigcap(-\infty, t+n^{-1}]) = \mathbb{P}(X \leqslant t).$$
This shows that F_X *is continuous from the right.*

In the previous remark, we regrouped all properties satisfied by the CDF of a random variable. It turns out that any function satisfying these properties is the CDF of some (or many) random variable.

Proposition 3.20

Let $F : \mathbb{R} \to [0,1]$ *be a function satisfying the following properties:*

1. F *is nondecreasing.*

2. $\lim_{t\to-\infty} F(t) = 0$ *and* $\lim_{t\to\infty} F(t) = 1$.

3. F *is continuous from the right.*

Then there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ *and a random variable* X *such that* $F = F_X$*. In the sequel, any function satisfying the above properties will be called a CDF.*

Proof We set $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$ and $\mathbb{P} = \lambda$, the Lebesgue measure. This space is usually called the canonical probability space. We set $X(\omega) = \sup\{y : F(y) < \omega\}$ for every $\omega \in [0,1]$. Bu definition, we have that if $X(\omega) < t$ then $F(t) \geqslant \omega$. This being true for every $t > X(\omega)$ and since F is continuous from the right then we deduce that $F(X(\omega)) \geqslant \omega$. Thus, for every $t \geqslant X(\omega)$ we have $\omega \leqslant F(X(\omega)) \leqslant F(t)$.
On the other hand, note that $X(\omega) = \inf\{z : F(z) \geqslant \omega\}$. Indeed, otherwise, there would exist $z < X(\omega)$ such that $F(z) \geqslant \omega$. By the definition of $X(\omega)$, this implies that there exists $y > z$ such that $F(y) < \omega$ contradicting the fact that F is nondecreasing. Thus we deduce that $X(\omega) = \inf\{z : F(z) \geqslant \omega\}$ which implies that if $F(t) \geqslant \omega$ then $X(\omega) \leqslant t$. We finally deduce that
$$X(\omega) \leqslant t \Leftrightarrow F(t) \geqslant \omega.$$
Therefore,
$$F_X(t) = \lambda\{\omega : X(\omega) \leqslant t\} = \lambda\{\omega : \omega \leqslant F(t)\} = \lambda\{[0, F(t)]\} = F(t).$$
□

Remark 3.21

1. *With the CDF in hand, one can easily calculate probabilities. Indeed, we have*
$$\mathbb{P}(]a,b]) = F(b) - F(a),\ \mathbb{P}([a,b]) = F(b) - F(a^-),\ \mathbb{P}([a,b[) = F(b^-) - F(a^-).$$

2. *If* X *has a density* f *then* $\mathbb{P}(]a,b]) = \int_a^b f(x)dx$*. More generally,* $F(t) = \int_{-\infty}^t f(x)dx$.

3. *Since $\mathbb{P}(\{a\}) = F(a) - F(a^-)$ then if X has a density, we have $\mathbb{P}(\{a\}) = 0$.*

3.4 Moments and expectation

In the previous chapter, we defined the integral with respect to an arbitrary measure of a measurable function. This of course applies in the same manner in the case of a probability measure, and will be called expectation.

Definition 3.22

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We define the expectation of X by

$$\mathbb{E}\, X = \int_\Omega X(\omega)\, \mathbb{P}(d\omega) = \int X\, d\mathbb{P}.$$

This expectation exists if $X \geqslant 0$ (in this case $\mathbb{E}\, X \in [0, \infty]$) or if $\mathbb{E}|X| < \infty$. When $\mathbb{E}\,|X| < \infty$, we say that $X \in L^1$.

Remark 3.23

1. *Indeed, to construct the expectation, we start by defining it for nonnegative staircase random variables, then we extend the definition to any nonnegative random variable using that any such variable is the limit of a nondecreasing sequence of nonnegative staircase random variables and use the monotone convergence theorem. Then by decomposing a random variable into a nonnegative and negative part, one gets the definition in the general case.*
2. *If $X = (X_1, \dots, X_n)$ is a random variable taking values in $\mathbb{R}^n$ then $\mathbb{E}\, X = (\mathbb{E}\, X_1, \dots, \mathbb{E}\, X_n) \in \mathbb{R}^n$.*
3. *A particular case to keep in mind is when $X = \mathbf{1}_A$ with $A \in \mathcal{A}$. In this case, we have $\mathbb{E}\, \mathbf{1}_A = \mathbb{P}(A)$.*

Let us translate the results of the previous chapter to a probabilistic language. Recall that a property holds almost surely if the set of ω which satisfies it has probability equal to one.

Proposition 3.24

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

(i) (Monotone convergence) If X_n is a nondecreasing sequence of nonnegative random variables and $X = \lim_{n\to\infty} X_n$. Then $\lim_{n\to\infty} \mathbb{E}\, X_n = \mathbb{E}\, X$.

(ii) (Fatou) If $X_n \geqslant 0$ then $\mathbb{E}\, \liminf_n X_n \leqslant \liminf_n \mathbb{E}\, X_n$.

(iii) (Dominated convergence) If $X_n \in L^1$ and $X_n \to X$ almost surely and $|X_n| \leqslant Y$ almost surely with $Y \in L^1$ then

$$\mathbb{E}\, X_n \underset{n\to\infty}{\to} X \quad \text{and} \quad \mathbb{E}\, |X_n - X| \underset{n\to\infty}{\to} 0.$$

It turns out that we can calculate the expectation of discrete random variables and that of random variables with a density in a very practical way.

Proposition 3.25

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

(i) If X is a discrete random variable and $X \in L^1$, then $\mathbb{E}\, X = \sum_{a \in \operatorname{Im} X} a \mathbb{P}(X = a)$.

(ii) If X is a random variable with a density function f and $h : \mathbb{R} \to \mathbb{R}$ a measurable function and $h(X) \in L^1$, then $\mathbb{E}\, h(X) = \int h(x) f(x)\, dx$. In particular, $\mathbb{E}\, X = \int x f(x)\, dx$.

Proof

(i) We know that any staircase random variable is discrete, and we have by definition that $\mathbb{E}\, X = \sum_{a \in \operatorname{Im} X} a \mathbb{P}(X = a)$ for any staircase random variable. In the general case, we can suppose that X is nonnegative, otherwise we decompose it into a nonnegative and negative part. Let $(a_i)_{i \in \mathbb{N}}$ be the set of values taken by X. For every $n \in \mathbb{N}$, we set $X_n = \sum_{i=1}^n a_i \mathbf{1}_{\{X = a_i\}}$ which is a staircase random variable and we have

$$\mathbb{E}\, X_n = \sum_{i=1}^n a_i \mathbb{P}(X = a_i) \underset{n \to \infty}{\to} \sum_{i=1}^\infty a_i \mathbb{P}(X = a_i).$$

On the other hand, $(X_n)_{n \in \mathbb{N}}$ is a nondecreasing sequence which converges to X and by the monotone convergence theorem $\mathbb{E}\, X_n \underset{n \to \infty}{\to} \mathbb{E}\, X$, which finishes the proof.

(ii) Start by taking $h = \mathbf{1}_A$. We have

$$\mathbb{E}\, \mathbf{1}_A \circ X = \mathbb{P}(X \in A).$$

Since X has density f, we deduce that

$$\mathbb{E}\, \mathbf{1}_A \circ X = \mathbb{P}(X \in A) = \int_A f(x)\, dx = \int \mathbf{1}_A(x) f(x)\, dx.$$

Therefore the conclusion is verified for $\mathbf{1}_A$. By linearity, we deduce it for any nonnegative staircase function and by the monotone convergence theorem for any nonnegative function. It is enough to decompose h into a nonnegative and negative part to obtain the general result.

□

An important point to notice is that the expectation of a random variable depends only on its distribution and not on the “map” aspect of it. The previous proposition illustrates already this, and more generally we have the following which can be proved in a similar manner.

Proposition 3.26

Let X be a random variable taking values in $(E, \mathcal{E})$ and $h : E \to [0, \infty]$ be a measurable function. We have

$$\mathbb{E}\,[h(X)] = \int_E h\, d\mathcal{L}_X (= \int_E h(x)\, \mathcal{L}_X(dx).)$$

Remark 3.27

The statement of the proposition means that knowing $\mathbb{E}\left[h(X)\right]$ for every nonnegative function is equivalent to knowing the distribution of X. Indeed, if we know $\mathcal{L}_X$, then we would know $\mathbb{E}\, h(X)$. Reversely, if we know $\mathbb{E}\left[h(X)\right]$ for every nonnegative measurable function h, then in particular for every $A \in \mathcal{E}$, we would know $\mathbb{E}\, \mathbf{1}_A(X) = \mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathcal{L}_X(A)$ which is the distribution of X. We can actually even restrict to bounded functions.

A very useful inequality is Markov inequality which allows to control the tails of the distribution using the expectation.

Proposition 3.28 (*Markov inequality*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, X be a random variable and h be a nonnegative nondecreasing measurable function. Then for any $t \in \mathbb{R}$, we have

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}\left[h(X)\right]}{h(t)}.$$

Proof Since h is nondecreasing then

$$\mathbb{P}(X \geqslant t) = \mathbb{P}\big(h(X) \geqslant h(t)\big) = \mathbb{E}\, \mathbf{1}_{\{h(X) \geqslant h(t)\}}.$$

On the other hand,

$$h(t)\mathbf{1}_{\{h(X) \geqslant h(t)\}} \leqslant h(X),$$

and since $h(t)$ is nonnegative then $\mathbf{1}_{\{h(X) \geqslant h(t)\}} \leqslant \frac{h(X)}{h(t)}$. By taking the expectation on both sides, we obtain the result. □

Definition 3.29

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X be a random variable. We say that $X \in L^p$, for some $p \geqslant 1$, if $\mathbb{E}\,|X|^p < \infty$. We define in this case $\|X\|_{L^p} = \left(\mathbb{E}\,|X|^p\right)^{\frac{1}{p}}$.

One can check that $\|\cdot\|_{L^p}$ defines a norm which is nondecreasing in p. Let us first state a very useful inequality.

Proposition 3.30 (*Jensen inequality*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X \in L^1$ be a random variable and $g : \mathbb{R} \to \mathbb{R}_+$ be a convex function. Then

$$g(\mathbb{E}\, X) \leqslant \mathbb{E}\left[g(X)\right].$$

The proof uses the fact that the claim holds when g is affine, then that any convex function is the supremum of affine functions. Using this inequality with $g(x) = x^{r/p}$ for $1 \leqslant p \leqslant r < \infty$ we get the following.

Corollary 3.31 (*Monotonicity of L^p-norms*)

Let $1 \leqslant p \leqslant r < \infty$ and $X \in L^r$. Then $X \in L^p$ and $\|X\|_{L^p} \leqslant \|X\|_{L^r}$.

An important particular case is when $p = 2$ since L^2 turns out to be a Hilbert space.

Proposition 3.32 (*Cauchy-Schwarz inequality*)

Let $X, Y \in L^2$. Then $XY \in L^1$ and $|\mathbb{E}[XY]| \leqslant \mathbb{E}|XY| \leqslant \|X\|_{L^2}\|Y\|_{L^2}$.

Proof We can suppose that both X and Y are nonnegative, otherwise we would work with $|X|$ and $|Y|$.
We set $X_n = \min(X, n)$ and $Y_n = \min(Y, n)$ which are bounded and thus X_nY_n is integrable. We have for any $a \in \mathbb{R}$ that

$$\mathbb{E}\left[(aX_n + Y_n)^2\right] \geqslant 0.$$

Expanding, this means that for any $a \in \mathbb{R}$, we have $a^2\mathbb{E}[X_n^2] + 2aE[X_nY_n] + \mathbb{E}[Y_n^2] \geqslant 0$. This means that the discriminant is nonpositive, which implies

$$\mathbb{E}[X_nY_n] \leqslant \|X_n\|_{L^2}\|Y_n\|_{L^2} \leqslant \|X\|_{L^2}\|Y\|_{L^2}$$

But X_nY_n is nondecreasing and converges to XY therefore by the monotone convergence theorem we get the result. □

Definition 3.33 (*Variance and covariance*)

Let $X \in L^2$. Then $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}[(X - \mathbb{E}X)^2]$. If in addition $Y \in L^2$, then $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - (\mathbb{E}X)(\mathbb{E}Y)$.

Remark 3.34

In particular we have $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. The previous proposition ensures that $\mathbb{E}[XY]$ is well defined since $XY \in L^1$ because $X, Y \in L^2$.

One way to calculate the moments of a random variable is through the use of the moment generating function, also known as the Laplace transform.

Definition 3.35 (*Moment generating function*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X be a random variable. The moment generating function of X is given by

$$M_X(\gamma) = \mathbb{E}\, e^{\gamma X}.$$

Remark 3.36

1. *When $\mathbb{E}\, e^{\gamma X}$ exists, it allows to calculate the moments. Indeed, $M_X'(\gamma) = \mathbb{E}[Xe^{\gamma X}]$ in such a way that $\mathbb{E}X = M_X'(0)$. Similarly, we have $\mathbb{E}X^p = M_X^{(p)}(\gamma)$.*
2. *If $X = (X_1, \ldots, X_n)$ takes values in $\mathbb{R}^n$, then we can define $M_X : \mathbb{R}^n \to \mathbb{R}$ by*

$$M_X(\gamma) = \mathbb{E}\left[e^{\langle\gamma, X\rangle}\right] = \mathbb{E}\prod_{i=1}^{n} e^{\gamma_i X_i},$$

for every $\gamma = (\gamma_1, \ldots, \gamma_n) \in \mathbb{R}^n$.

Example 3.37

1. *If $X \sim \mathcal{N}(0,1)$ then $M_X(\gamma) = e^{\gamma^2/2}$. Indeed*
$$\mathbb{E}\, e^{\gamma X} = \int_{\mathbb{R}} e^{\gamma x} e^{-\frac{x^2}{2}}\, dx = e^{\frac{\gamma^2}{2}} \int_{\mathbb{R}} e^{-\frac{(x-\gamma)^2}{2}}\, dx = e^{\frac{\gamma^2}{2}}.$$

2. *If $X \sim Rad(1/2)$ then*
$$M_X(\gamma) = \frac{e^{\gamma} + e^{-\gamma}}{2} = \cosh(\gamma) \leqslant e^{\frac{\gamma^2}{2}}.$$

3.5 Characteristic functions

We have already characterized the distribution of a random variable using the CDF. Another way to do so, is through the characteristic function whose main advantage is its tensorization property i.e. one can easily express the characteristic function of a sum of random variables in terms of each.

Definition 3.38 (*Characteristic function*)

Let X be a real random variable. The characteristic function $\phi_X : \mathbb{R} \to \mathbb{C}$ of X at $t \in \mathbb{R}$ is defined by
$$\phi_X(t) = \mathbb{E}\, e^{itX} = \mathbb{E}[\cos(tX)] + i\mathbb{E}\,[\sin(tX)].$$

Remark 3.39

1. *If $Z = X + iY$ with $X, Y \in L^1$. Then $|\mathbb{E}\, Z| \leqslant \mathbb{E}\, |Z|$ where $|Z|$ here refers to the modulus of Z. Indeed, we write*
$$|\mathbb{E}\, Z| = \mathbb{E}\, Z \cdot \frac{\overline{\mathbb{E}\, Z}}{|\mathbb{E}\, Z|} = \mathrm{Re}\big(\mathbb{E}\, Z \cdot \frac{\overline{\mathbb{E}\, Z}}{|\mathbb{E}\, Z|}\big) = \mathbb{E}\,\mathrm{Re}\big(Z \cdot \frac{\overline{\mathbb{E}\, Z}}{|\mathbb{E}\, Z|}\big),$$
but $\mathrm{Re}\big(Z \cdot \frac{\overline{\mathbb{E}\, Z}}{|\mathbb{E}\, Z|}\big) \leqslant |Z|$ which finishes the proof.

   Therefore, we always have $|\phi_X(t)| \leqslant 1$.

2. *We have $\phi_X(0) = 1$. We can verify that $\overline{\phi_X} = \phi_{-X}$ and that for any $a, b \in \mathbb{R}$, we have $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.*

3. *ϕ_X is uniformly continuous. Indeed, we have*
$$\sup_{t\in\mathbb{R}} |\phi_X(t+\varepsilon) - \phi_X(t)| \leqslant \mathbb{E}\, |e^{i\varepsilon X} - 1|.$$
Since $|e^{i\varepsilon X} - 1| \underset{\varepsilon\to 0}{\to} 0$ then by the dominated convergence theorem we obtain that $\mathbb{E}\, |e^{i\varepsilon X} - 1| \underset{\varepsilon\to 0}{\to} 0$.

Example 3.40

Let $X \sim \mathcal{N}(0, \sigma^2)$. Then $\phi_X(t) = e^{-\frac{\sigma^2 t^2}{2}}$ for every $t \in \mathbb{R}$.

Indeed, we have by definition that

$$\phi_X(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x^2}{2\sigma^2}} e^{itx}\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x^2}{2\sigma^2}}\cos(tx)\,dx + \frac{i}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x^2}{2\sigma^2}}\sin(tx)\,dx$$

Since $e^{-\frac{x^2}{2\sigma^2}}\sin(tx)$ is odd, then the second integral above is zero. Therefore, we have

$$\phi_X(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x^2}{2\sigma^2}}\cos(tx)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} g(t,x)\,dx$$

But, for every $t \in \mathbb{R}$, we have that $g(t,x)$ is integrable with respect to x (since it is bounded by $e^{-\frac{x^2}{2\sigma^2}}$ which is integrable). Moreover $g(t,x)$ is continuously differentiable with respect to t since $\frac{\partial g}{\partial t} = -xe^{-\frac{x^2}{2\sigma^2}}\sin(tx)$ is continuous. Finally, we have for every t that $\frac{\partial g}{\partial t} \leqslant |x|e^{-\frac{x^2}{2\sigma^2}}$ which is integrable. Therefore, we can differentiate inside the integral and write

$$\phi'_X(t) = -\frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} xe^{-\frac{x^2}{2\sigma^2}}\sin(tx)\,dx$$

Integrate by parts to get

$$\phi'_X(t) = -\frac{\sigma}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x^2}{2\sigma^2}} t\cos(tx)\,dx = -\sigma^2 t\phi_X(t).$$

Resolving the differential equation and using that $\phi_X(0) = 1$, we get $\phi_X(t) = e^{-\frac{\sigma^2 t^2}{2}}$ which is equal to the density of the Nromal $\mathcal{N}(0,\sigma^{-2})$ properly normalized.

Characteristic functions are useful because they also characterize the distribution. To see this, we will first give an inversion formula which allows to recover the measure from the characteristic function.

Theorem 3.41 (*Inversion formula*)

Let μ be a probability measure on $\mathbb{R}$ and $\phi(t) = \int e^{itx}\,\mu(dx)$. Then for any $a < b$, we have

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

Proof We start by writing

$$\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi(t)\,dt = \int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\Big(\int_{\mathbb{R}} e^{itx}\,\mu(dx)\Big)\,dt$$

and we note that we can use Fubini since for every $t \in \mathbb{R}$ we have

$$|\frac{e^{-ita}-e^{-itb}}{it}e^{itx}| = \Big|\int_a^b e^{-ity}\,dy\Big| \leqslant |b-a|$$

and we have two finite measures (Lebesgue on $[-T,T]$ and a probability measure). Therefore, we have

$$\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi(t)\,dt = \int\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}e^{itx}\,dt\,\mu(dx)$$

Using that $e^{iu} = \cos(u) + i\sin(u)$, we expand the above integral and use the cosine and sine part to get

$$\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\phi(t)\,dt = 2\int\int_0^T \frac{\sin\big(t(x-a)\big) - \sin\big(t(x-b)\big)}{t}\,dt\,\mu(dx).$$

But with a change of variables, we have that the equalities below hold when taking the limit in T.

$$\operatorname{sgn}(x-a)\int_0^T \frac{\sin\big(t(x-a)\big)}{t}\,dt = \int_0^T \frac{\sin(t)}{t}\,dt = \int_0^T\int_0^\infty e^{-ty}\sin(t)\,dy\,dt = \int_0^\infty\int_0^T e^{-ty}\sin(t)\,dt\,dy$$

An integration by parts allows to prove that $\int_0^T \frac{\sin\big(t(x-a)\big)}{t}\,dt \underset{T\to\infty}{\longrightarrow} \frac{\pi}{2}\operatorname{sgn}(x-a)$. Therefore,

$$\sup_{T,x}\Big|\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}e^{itx}\,dt\Big| < \infty,$$

and by the dominated convergence theorem we have

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\phi(t)\,dt &= \frac{1}{\pi}\int\lim_{T\to\infty}\int_0^T\frac{\sin\big(t(x-a)\big) - \sin\big(t(x-b)\big)}{t}\,dt\,\mu(dx)\\ &= \frac{1}{2}\int\big(\operatorname{sgn}(x-a) - \operatorname{sgn}(x-b)\big)\,\mu(dx)\\ &= \frac{\mu(\{a\}) + \mu(\{b\})}{2} + \mu(a,b).\end{aligned}$$

□

Theorem 3.42

The characteristic function characterizes the distribution. More precisely, if X and Y are two real random variables then $\phi_X = \phi_Y$ if and only if $\mathcal{L}_X = \mathcal{L}_Y$.

Proof This follows easily from the inversion formula. Let F_X and F_Y be the CDFs of X and Y and denote $\mathcal{C}_X = \{c \in \mathbb{R} : \mathcal{L}_X(\{c\}) \neq 0\}$ and $\mathcal{C}_Y = \{c \in \mathbb{R} : \mathcal{L}_Y(\{c\}) \neq 0\}$. These two sets are countable and form the set of discontinuity points of F_X and F_Y. Since these two functions are continuous from the right, it is enough to find the values at the continuity points. Let b be an arbitrary continuity point of F_X and F_Y, then by taking $a \to -\infty$ (while being a continuity point), we have

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

and we finish the proof. □

Chapter 4

Independence

4.1 Independent events and σ-algebra

We are given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 4.1 (*Independent events*)

We say that two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
We say that n events $A_1, \ldots, A_n$ are mutually independent if

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2})\ldots\mathbb{P}(A_{i_k})$$

for every subset $\{i_1, \ldots, i_k\}$ de $\{1, \ldots, n\}$.

Remark 4.2

1. *One shouldn't confuse the notion of mutual independence with that of pairwise independence. Indeed, if toss a coin twice and consider the events $A = \{Head\ at\ first\ toss\}$, $B = \{Head\ at\ second\ toss\}$, and $C = \{same\ result\ on\ both\ tosses\}$. We can check that these events are pairwise independent but not mutually independent.*

2. *If $A_1, \ldots, A_n$ are independent then $\mathbb{P}(B_1 \cap \ldots B_n) = \mathbb{P}(B_1)\ldots\mathbb{P}(B_n)$ for every $B_i \in \sigma(A_i) = \{\emptyset, A_i, A_i^c, \Omega\}$, $i = 1, \ldots, n$. This can be easily shown and it implies that if $A_1, \ldots, A_n$ are independent then their complements are as well (even more when replacing some of the events by their complements and others not).*

More generally, we will define the independence between σ-algebras.

Definition 4.3 (*Independent σ-algebras*)

Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be σ-algebras contained in $\mathcal{A}$. We say that they are independent if

$$\forall A_1 \in \mathcal{A}_1, \ldots, \forall A_n \in \mathcal{A}_n,\ \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\ldots\mathbb{P}(A_n).$$

With this definition, we have seen in the previous remarks that the events $A_1, \ldots, A_n$ are mutually independent if and only if the σ-algebras $\sigma(A_1), \ldots, \sigma(A_n)$ are independent. Once more, it is enough to verify the condition in the definition only on a generating π-system.

Proposition 4.4

Let $\mathcal{A}_1, \ldots, \mathcal{A}_n \subset \mathcal{A}$ be σ-algebras. For every $i \in \{1, \ldots, n\}$, let $\mathcal{I}_i$ be a generating π-system of $\mathcal{A}_i$. If for any $A_1 \in \mathcal{I}_1, \ldots, A_n \in \mathcal{I}_n$ we have

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

then $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent.

Proof We first fix $A_2 \in \mathcal{I}_1, \ldots, A_n \in \mathcal{I}_n$ and set

$$\mathcal{M}_1 = \{A_1 \in \mathcal{A}_1 : \mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \ldots \mathbb{P}(A_n)\}.$$

We check that $\mathcal{M}_1$ is a monotone class and by hypothesis $\mathcal{I}_1 \subset \mathcal{A}_1$. Therefore, by the monotone classes lemma $\mathcal{M}_1 = \sigma(\mathcal{I}_1) = \mathcal{A}_1$.
Now we fix $A_1 \in \mathcal{A}_1, A_3 \in \mathcal{A}_3, \ldots, A_n \in \mathcal{A}_n$ and set

$$\mathcal{M}_2 = \{A_2 \in \mathcal{A}_2 : \mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \ldots \mathbb{P}(A_n)\}.$$

We similarly show that $\mathcal{M}_2$ is a monotone class and use the monotone classes lemma to deduce that $\mathcal{M}_2 = \mathcal{A}_2$. We continue in a similar manner to finish the proof. □

One of the consequences of this proposition is that we can assert the independence of classes of σ-algebras if these classes are formed from independent σ-algebras.

Corollary 4.5 (*Grouping by classes*)

Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be independent σ-algebras. Let $n_0 < n_1 < n_2 \ldots < n_k = n$. Then the σ-algebras

$$\mathcal{D}_1 = \sigma(\mathcal{A}_1, \ldots, \mathcal{A}_{n_1}),\ \mathcal{D}_2 = \sigma(\mathcal{A}_{n_1+1}, \ldots, \mathcal{A}_{n_2}), \ldots,\ \mathcal{D}_k = \sigma(\mathcal{A}_{n_{k-1}+1}, \ldots, \mathcal{A}_{n_k})$$

are independent.

Proof For every $j \in \{1, \ldots, k\}$, we define $\mathcal{I}_j$ as the set of parts of the form

$$A_{n_{j-1}+1} \cap \ldots \cap A_{n_j},$$

with $A_i \in \mathcal{A}_i$ for every $i \in \{n_{j-1}+1, \ldots, n_j\}$. Clearly, the $\mathcal{I}_j$'s are generating π-systems of $\mathcal{D}_j$. By the previous proposition, the result follows. □

4.2 Independent random variables

Definition 4.6 (*Independent random variables*)

We say that n random variables $X_1, \ldots, X_n$ are independent if the σ-algebras $\sigma(X_1), \ldots, \sigma(X_n)$ are independent. More precisely if for every $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$ we have

$$\mathbb{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \ldots \cap \{X_n \in A_n\}) = \mathbb{P}(\{X_1 \in A_1\})\mathbb{P}(\{X_2 \in A_2\}) \ldots \mathbb{P}(\{X_n \in A_n\}).$$

Remark 4.7

1. *The definition extends in a similar manner to random variables taking values in other mea-*

surable spaces i.e. if $X_1, \ldots, X_n$ are random variables taking values in $(\Omega_1, \mathcal{B}_1), \ldots, (\Omega_n, \mathcal{B}_n)$ respectively. Then we say that they are independent if for every $A_1 \in \mathcal{B}_1, \ldots, A_n \in \mathcal{B}_n$ we have

$$\mathbb{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \ldots \cap \{X_n \in A_n\}) = \mathbb{P}(\{X_1 \in A_1\})\mathbb{P}(\{X_2 \in A_2\}) \ldots \mathbb{P}(\{X_n \in A_n\}).$$

2. *Let $\mathcal{A}_1, \ldots, \mathcal{A}_n \subset \mathcal{A}$ be independent σ-algebras. If for every $i \in \{1, \ldots, n\}$, X_i is $\mathcal{A}_i$-measurable then $X_1, \ldots, X_n$ are independent. Indeed, for every $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$, we have $X_1^{-1}(A_1) \in \mathcal{A}_1, \ldots, X_n^{-1}(A_n) \in \mathcal{A}_n$ (since X_i is $\mathcal{A}_i$-measurable) and these events are thus independent; since this is true for any choice of $A_1, \ldots, A_n$ then the variables $X_1, \ldots, X_n$ are independent.*

3. *Similarly, we have that n real random variables $X_1, \ldots, X_n$ are independent if and only if for every $t_1, \ldots, t_n \in \mathbb{R}$ we have*

$$\mathbb{P}(X_1 \leqslant t_1, X_2 \leqslant t_2, \ldots, X_n \leqslant t_n) = \mathbb{P}(X_1 \leqslant t_1)\mathbb{P}(X_2 \leqslant t_2) \ldots \mathbb{P}(X_n \leqslant t_n).$$

Definition 4.8 (*Independence of an infinite sequence*)

All previous definitions concerning the independence can be extended to an infinite number of random variables/σ-algebras/events. For example, we say that an infinite sequence of events are independent if every finite subsequence of events form a family of independent events. One can define in a similar manner the independence for infinite families of σ-algebras and random variables.

As for σ-algebras, one can also assert independence by grouping in classes random variables.

Proposition 4.9

Let $X_1, \ldots, X_n$ be independent random variables and let $n_0 < n_1 < n_2 \ldots < n_k = n$. Then the random variables $Y_1 = (X_1, \ldots, X_{n_1}), \ldots, Y_k = (X_{n_{k-1}+1}, \ldots, X_{n_k})$ are independent.

In other words, if the random variables are functions of disjoint classes of independent random variables, then these random variables are independent. For example, if X_1, X_2, X_3, X_4 are independent then $Z_1 = X_1 X_3$ and $Z_2 = X_2^2 + X_4 X_2$ are independent.

Definition 4.10 (*Joint distribution/law*)

Let $X_1, \ldots, X_n$ are random variables taking values in $(S_1, \mathcal{A}_1), \ldots, (S_n, \mathcal{A}_n)$ respectively. The joint law of $X_1, \ldots, X_n$ is the distribution of the n-tuple $(X_1, \ldots, X_n)$ in $S_1 \times \ldots \times S_n$ equipped with the product σ-algebra $\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n$.

Example 4.11

If X is a standard Bernoulli random variable and $Y = 1 - X$. Then the joint law of X, Y is given by

$$\mathcal{L}_{(X,Y)}(\{(1,0)\}) = \mathcal{L}_{(X,Y)}(\{(0,1)\}) = \frac{1}{2}.$$

We say that $\mathcal{L}_{X_i}$ are the marginal distributions of $\mathcal{L}_{(X_1,\ldots,X_n)}$. We have the following characterization of the independence of random variables.

Proposition 4.12

Let $X_1, \dots, X_n$ be random variables taking values in $(S_1, \mathcal{A}_1), \dots, (S_n, \mathcal{A}_n)$ respectively. Then $X_1, \dots, X_n$ are independent if and only if the joint distribution is equal to the product of the marginal distributions i.e. $\mathcal{L}_{(X_1,\dots,X_n)} = \mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}$.

Proof Suppose that $X_1, \dots, X_n$ are independent and let us show that $\mathcal{L}_{(X_1,\dots,X_n)} = \mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}$. Let $A_1 \in \mathcal{A}_1, \dots, A_n \in \mathcal{A}_n$ and write

$$\mathcal{L}_{(X_1,\dots,X_n)}(A_1 \times \dots \times A_n) = \mathbb{P}\big((X_1, \dots, X_n) \in A_1 \times \dots A_n\big) = \mathbb{P}\big(\{X_1 \in A_1\} \cap \dots \cap \{X_n \in A_n\}\big)$$

and by independence of the X_i's, we obtain

$$\mathcal{L}_{(X_1,\dots,X_n)}(A_1 \times \dots \times A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

which by definition is nothing else but $\mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}(A_1 \times \dots \times A_n)$. This shows that $\mathcal{L}_{(X_1,\dots,X_n)}$ and $\mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}$ coincide on all measurable rectangles (which form a π-system) thus by Corollary 1.17, we have the equality.
For the reverse, if $\mathcal{L}_{(X_1,\dots,X_n)} = \mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}$, then in particular

$$\mathcal{L}_{(X_1,\dots,X_n)}(A_1 \times \dots \times A_n) = \mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}(A_1 \times \dots \times A_n)$$

for all $A_1 \in \mathcal{A}_1, \dots, A_n \in \mathcal{A}_n$. But

$$\mathcal{L}_{(X_1,\dots,X_n)}(A_1 \times \dots \times A_n) = \mathbb{P}\big((X_1, \dots, X_n) \in A_1 \times \dots A_n\big) = \mathbb{P}\big(\{X_1 \in A_1\} \cap \dots \cap \{X_n \in A_n\}\big)$$

and

$$\mathcal{L}_{X_1} \otimes \dots \otimes \mathcal{L}_{X_n}(A_1 \times \dots \times A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i),$$

which implies the independence of the X_i's. □

Remark 4.13

Combining this proposition with Fubini, we can deduce that if $X_1, \dots, X_n$ are real random variables with density f_i each then $(X_1, \dots, X_n)$ has a density in $\mathbb{R}^n$ given by the product of the densities. The reverse is also true in the sense that if $(X_1, \dots, X_n)$ has a density decomposable in a product of marginal densities then the X_i's are independent.

Corollary 4.14

If X and Y are real independent random variables then $\mathcal{L}_{X+Y} = \mathcal{L}_X * \mathcal{L}_Y$ where $*$ is the convolution product i.e. for every $A \in \mathcal{B}(\mathbb{R})$ we have

$$\mathcal{L}_{X+Y}(A) = \int_{\mathbb{R}^2} \mathbf{1}_A(x+y)\, \mathcal{L}_X(dx)\mathcal{L}_Y(dy).$$

Proof The proof is trivial since $\mathcal{L}_{X+Y}(A) = \mathbb{E}\, \mathbf{1}_A \circ h(X,Y)$ where $h(X,Y) = X + Y$. Thus

$$\mathcal{L}_{X+Y}(A) = \int \mathbf{1}_A(x+y)\, \mathcal{L}_{(X,Y)}(dx, dy).$$

Since $\mathcal{L}_{(X,Y)} = \mathcal{L}_X \otimes \mathcal{L}_Y$, we finish the proof. □

Another characterization of independence can be done using the expectation.

Proposition 4.15

Let $X_1, \dots, X_n$ be random variables taking values in $(S_1, \mathcal{A}_1), \dots, (S_n, \mathcal{A}_n)$ respectively. Then $X_1, \dots, X_n$ are independent if and only if for all nonnegative measurable functions $f_i : S_i \to [0, \infty]$, we have

$$\mathbb{E}\,[\prod_{i=1}^{n} f_i(X_i)] = \prod_{i=1}^{n} \mathbb{E}\,[f_i(X_i)].$$

Proof One of the implications is trivial by taking the f_i's to be indicators of events A_i's. For the other implication, we use as before that the joint law is equal to the product of the marginal distributions to write that

$$\mathbb{E}\,[\prod_{i=1}^{n} f_i(X_i)] = \int_{S_1\times\ldots\times S_n} \prod_{i=1}^{n} f_i(X_i)\, \mathcal{L}_{(X_1,\ldots,X_n)}(dx_1, \ldots, dx_n) = \int_{S_1\times\ldots\times S_n} \prod_{i=1}^{n} f_i(X_i)\, \mathcal{L}_{X_1}(dx_1) \ldots \mathcal{L}_{X_n}(dx_n).$$

By Fubini, we obtain that

$$\int_{S_1\times\ldots\times S_n} \prod_{i=1}^{n} f_i(X_i)\, \mathcal{L}_{X_1}(dx_1) \ldots \mathcal{L}_{X_n}(dx_n) = \prod_{i=1}^{n} \int_{S_i} f_i(X_i)\, \mathcal{L}_{X_i}(dx_i) = \prod_{i=1}^{n} \mathbb{E}\,[f_i(X_i)],$$

and finish the proof. □

Remark 4.16

1. *We use the fact that the f_i's are nonnegative measurable in order to apply Fubini. However, the proposition remains valid if we had integrable functions f_i i.e. if $\mathbb{E}\,|f_i(X_i)| < \infty$ for every $i = 1, \ldots, n$.*
2. *In particular if X and Y are two real independent random variables then $\mathrm{Cov}(X, Y) = 0$ since $\mathbb{E}\,XY = \mathbb{E}\,X\mathbb{E}\,Y$. The reverse is false as we can see by taking $X \sim \mathcal{N}(0, \sigma^2)$ and $Y = \xi X$ where $\xi \sim \mathrm{Ber}(1/2)$ is independent of X.*

For the real random variables, we have the following.

Proposition 4.17

Let $X_1, \ldots, X_n$ be real random variables. Then $X_1, \ldots, X_n$ are independent if and only if $\phi_{(X_1,\ldots,X_n)}(t_1, \ldots, t_n) = \prod_{i=1}^{n} \phi_{X_i}(t_i)$ for any $t_1, \ldots, t_n \in \mathbb{R}$.

Proof Suppose first that the X_i's are independent and note that

$$\phi_{(X_1,\ldots,X_n)}(t_1, \ldots, t_n) = \mathbb{E}\, e^{i\sum_{j=1}^{n} t_j X_j} = \mathbb{E} \prod_{j=1}^{n} e^{it_j X_j} = \prod_{j=1}^{n} \mathbb{E}\, e^{it_j X_j} = \prod_{j=1}^{n} \phi_{X_j}(t_j).$$

The other way follows from the multi-dimensional inversion formula which extends Theorem 3.41. □

4.3 Second Borel-Cantelli lemma

Lemme 4.18 (*Second Borel-Cantelli lemma*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(A_n)_{n\in\mathbb{N}}$ be a sequence of independent events. Then

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty \quad \Rightarrow \quad \mathbb{P}(\limsup A_n) = 1.$$

Proof Recall that $\limsup A_n = \bigcap_{m\in\mathbb{N}} \bigcup_{n\geqslant m} A_n$. By independence of the A_n's, we have that for every $m \in \mathbb{N}$,

$$\mathbb{P}\big(\bigcap_{n\geqslant m} A_n^c\big) = \prod_{n=m}^{\infty} \big(1 - \mathbb{P}(A_n)\big) \leqslant \exp\big(-\sum_{n\geqslant m} \mathbb{P}(A_n)\big),$$

where we used that $1-x \leqslant e^{-x}$ for every $x \geqslant 0$. Since $\sum_n \mathbb{P}(A_n)$ is a divergent sequence, we deduce that for every $m \in \mathbb{N}$ we have $\mathbb{P}\big(\bigcap_{n\geqslant m} A_n^c\big) = 0$.
This means that $\mathbb{P}\big(\bigcup_{n\geqslant m} A_n\big) = 1$ and that for every $m \in \mathbb{N}$ the event $\bigcup_{n\geqslant m} A_n$ holds almost-surely. Since the countable intersection of almost-sure events is almost-sure, we deduce that $\mathbb{P}(\bigcap_{m\in\mathbb{N}} \bigcup_{n\geqslant m} A_n) = 1$ and finish the proof. □

Let us look at some examples of application of the previous lemma.

Example 4.19

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables following an exponential distribution with parameter 1 i.e. $\mathbb{P}(X_n > t) = e^{-t}$ for every $t \geqslant 0$. Then

$$\limsup \frac{X_n}{\ln n} = 1 \quad a.s.$$

Indeed, we have for every $\alpha > 0$

$$\mathbb{P}(X_n > \alpha \ln n) = n^{-\alpha},$$

which forms a convergent sequence if and only if $\alpha > 1$. Therefore, by the two Borel-Cantelli lemmas, we have

$$\mathbb{P}(X_n > \alpha \ln n \textit{ for an infinity of } n) = 0 \textit{ if } \alpha > 1 \textit{ and } 1 \textit{ if } \alpha \leqslant 1.$$

Clearly, if $X_n > \ln n$ for an infinity of n then $\limsup \frac{X_n}{\ln n} \geqslant 1$. Therefore,

$$\mathbb{P}(\limsup \frac{X_n}{\ln n} \geqslant 1) \geqslant \mathbb{P}(X_n > \ln n \textit{ for an infinity of } n) = 1.$$

Thus we have $\limsup \frac{X_n}{\ln n} \geqslant 1$ a.s. and it remains to prove that $\mathbb{P}(\limsup \frac{X_n}{\ln n} > 1) = 0$. We write

$$\big\{\limsup \frac{X_n}{\ln n} > 1\big\} = \bigcup_{k\in\mathbb{N}} \big\{\limsup \frac{X_n}{\ln n} > 1 + \frac{2}{k}\big\}.$$

On the other hand, if $\limsup \frac{X_n}{\ln n} > 1 + \frac{2}{k}$ then $\frac{X_n}{\ln n} > 1 + \frac{1}{k}$ for an infinity of n. Thus

$$\mathbb{P}(\limsup \frac{X_n}{\ln n} > 1 + \frac{2}{k}) \leqslant \mathbb{P}(X_n > (1 + \frac{1}{k}) \ln n \textit{ for an infinity of } n) = 0.$$

Since $\big\{\limsup \frac{X_n}{\ln n} > 1\big\}$ is the countable union of the previous events, we deduce that $\mathbb{P}(\limsup \frac{X_n}{\ln n} > 1) = 0$.

Example 4.20

For every $n \in \mathbb{N}$, we denote A_n the set of multiples of n. There is no probability $\mathbb{P}$ on $\mathbb{N}$ such that $\mathbb{P}(A_n) = \frac{1}{n}$ for every $n \in \mathbb{N}$.
Indeed, suppose there exists one. Let $\mathcal{P}$ be the set of prime numbers. The $(A_p)_{p \in \mathcal{P}}$'s are independent since for every $p_1, \ldots, p_k \in \mathcal{P}$ we have

$$\mathbb{P}(A_{p_1} \cap \ldots \cap A_{p_k}) = \mathbb{P}(A_{p_1 \ldots p_k}) = \frac{1}{p_1 \ldots p_k} = \prod_{i=1}^{k} \mathbb{P}(A_{p_i}).$$

On the other hand, we have $\sum_{p \in \mathcal{P}} \frac{1}{p} = \infty$. Therefore, by the second Borel-Cantelli lemma, we have that almost every prime number is multiple of an infinity of prime numbers, which is a contradiction.

4.4 Zero/One law

Theorem 4.21 (*Kolmogorov* 0/1 *law*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent random variables. For every $n \in \mathbb{N}$, we set $\mathcal{T}_n = \sigma(X_n, X_{n+1}, \ldots)$ and $\mathcal{T}_\infty = \bigcap_{n \in \mathbb{N}} \mathcal{T}_n$. Then we have

$$\forall A \in \mathcal{T}_\infty, \quad \mathbb{P}(A) \in \{0, 1\}.$$

Proof For every $n \in \mathbb{N}$, we set $\mathcal{B}_n = \sigma(X_1, \ldots, X_n)$. By the grouping by classes independence, we have that $\mathcal{B}_n$ and $\mathcal{T}_{n+1}$ are independent. Therefore, we have that $\mathcal{B}_n$ and $\mathcal{T}_\infty$ are independent. This being true for every $n \in \mathbb{N}$, we can write that

$$\forall A \in \mathcal{T}_\infty,\ \forall B \in \bigcup_{n=1}^{\infty} \mathcal{B}_n,\ \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

But $\bigcup_{n=1}^{\infty} \mathcal{B}_n$ is stable by finite intersection and thus is a π-système. We have seen that this implies that $\mathcal{T}_\infty$ and $\sigma(\bigcup_{n=1}^{\infty} \mathcal{B}_n)$ are independent. But $\sigma(\bigcup_{n=1}^{\infty} \mathcal{B}_n) = \sigma(X_n : n \geqslant 1)$. Thus $\mathcal{T}_\infty$ is independent from itself which implies that for any $A \in \mathcal{T}_\infty$ we have $\mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$. This means that $\mathbb{P}(A) \in \{0, 1\}$ and finishes the proof. □

Remark 4.22

We have already seen a result similar in spirit. Indeed, the two Borel-Cantelli lemmas stipulate that if $A_1, A_2, \ldots$ are independent events then $\mathbb{P}(\limsup A_n) \in \{0, 1\}$. If we set $X_n = \mathbf{1}_{A_n}$ for every $n \in \mathbb{N}$ then $\sigma(X_n) = \sigma(A_n)$ and the 0/1 law implies in particular that $\mathbb{P}(\limsup A_n) \in \{0, 1\}$ since $\limsup A_n = \bigcap_{n \in \mathbb{N}} \bigcup_{m \geqslant n} A_m \in \bigcap_{n \in \mathbb{N}} \sigma(X_m : m \geqslant n)$.

The following corollary gives an intuition on the previous result. To summarize, if a random variable is measurable with respect to an infinity of independent random variables then it is almost surely constant.

Corollary 4.23

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(X_n)_{n \in \mathbb{N}}$ is a sequence of independent random variables and $\mathcal{T}_\infty = \bigcap_{n \in \mathbb{N}} \sigma(X_m : m \geqslant n)$. Then any $\mathcal{T}_\infty$-measurable random variable X is almost surely constant *i.e.* $\mathbb{P}(X = c) = 1$ *for some* $c \in \mathbb{R}$.

Proof For every $t \in \mathbb{R}$, we have $\{X \leqslant t\} \in \mathcal{T}_\infty$ thus by the zero/one law we have $\mathbb{P}(T \leqslant t) \in \{0, 1\}$. This shows that the CDF of X only takes the values 0 and 1. If we define $c := \inf\{t \in \mathbb{R} : \mathbb{P}(X \leqslant t) = 1\}$ then we have $\mathbb{P}(X \leqslant c) = 1$ and $\mathbb{P}(X < c) = 0$ which implies that $\mathbb{P}(X = c) = 1$. □

Example 4.24

Let $X_1, X_2, \ldots$ *be real independent random variables. Then* $\mathbb{P}(\sum_{n=1}^{\infty} X_n$ *converge*$) \in \{0, 1\}$. *In other words, a series whose terms are independent random variables either converge almost surely or diverge almost surely.*

Indeed, we have for every $m \in \mathbb{N}$ *that* $\{\sum_{n=1}^{\infty} X_n$ *converge*$\} = \{\sum_{n=1}^{\infty} X_{n+m}$ *converge*$\} \in \sigma(X_{m+1}, \ldots)$. *Thus* $\{\sum_{n=1}^{\infty} X_n$ *converge*$\} \in \bigcap_{n \in \mathbb{N}} \sigma(X_m : m \geqslant n)$ *and by the* 0/1 *law we get the result.*

Example 4.25

A drunk man is walking on the real line starting from the origin. At each step, he walks one step to the right with probability p *and a step to the left with probability* $1 - p$. *We suppose that the steps made at each step are independent. Therefore the event that the man escapes to infinity is either almost sure or almost null.*

We model the problem in the following way. We set $X_0 = 0$ *and for every* $n \geqslant 1$, *we define independent random variables* $(X_n)_{n \geqslant 1}$ *by* $\mathbb{P}(X_n = -1) = 1 - p$ *and* $\mathbb{P}(X_n = 1) = p$. *The position of the man after* n *steps is given by* $W_n = \sum_{k=1}^{n} X_k$. *The question is to prove*

$$\mathbb{P}(\limsup_n W_n = \infty) \in \{0, 1\}.$$

As before, this follows from the 0/1 *law since the event* $\{\limsup_n W_n = \infty\} \in \sigma(X_m : m \geqslant n)$ *for every* $n \in \mathbb{N}$.

Chapter 5

Convergence of random variables

5.1 Different notions of convergence

The goal of this chapter is to introduce the different notions of convergence of random variables. Recall that a random variable has two aspects, one being the "mapping aspect". We could then simply define convergence of random variables X_n by looking at the convergence of X_n as functions i.e. look at the convergence of the sequence $X_n(\omega)$ for every ω. However, ignoring completely the "random aspect" doesn't seem interesting from a probabilistic point of vue since it doesn't allow to capture phenomena which we look to understand. Take the example of tossing a fair coin repeatedly, denote Ω the set of possible outcomes and set $(X_n)_{n\in\mathbb{N}}$ be the sequence of independent random variables where $X_n = 1$ if we obtain head at the n-th trial and 0 otherwise. Since the coin is fair, we have $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = 0) = \frac{1}{2}$.

Intuitively, if we repeat the experiment an infinite number of times, we should end up getting the same number of heads and tails. This means that we would like to define a notion of convergence which would allow us to assert that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\to} \frac{1}{2},$$

which is nothing but saying that the average number of heads obtained is half, which would be compatible with the fact the the probability of getting a head is 1/2. If we define our notion of convergence as being the regular one for functions, then we wouldn't capture the above phenomenon since if $\omega \in \Omega$ contains a finite number of heads, then $\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\to} 0$. It turns out here that this event is almost null, and we will define the notion of convergence as being the regular function convergence but outside bad events which are of zero probability.

Definition 5.1 (*Almost sure convergence*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that X_n converge almost surely to a random variable X if

$$\mathbb{P}(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

We denote $X_n \xrightarrow[n\to\infty]{a.s.} X$ or $\lim_{n\to\infty} X_n = X$ a.s.

Remark 5.2

The starting space will sometimes be omitted. We take a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and the *sequence $(X_n)_{n\in\mathbb{N}}$ would be constituted of random variables defined on the is space. As often, we will omit sometimes detailing the event and write for example $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$ instead of $\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\})$.*

Example 5.3

We consider the canonical space $([0,1], \mathcal{B}([0,1]), \mathbb{P})$ with $\mathbb{P} = \lambda$ the Lebesgue measure. We define $X_n = \mathbf{1}_{[0,\frac{n+1}{2n})}$ for every n. Then, we have $X_n \xrightarrow[n\to\infty]{a.s.} \mathbf{1}_{[0,\frac{1}{2})}$.
Indeed, let $A = \{\omega \in [0,1] : \lim_{n\to\infty} X_n(\omega) = \mathbf{1}_{[0,\frac{1}{2})}\}$ and let us show that $\mathbb{P}(A) = 1$. Note first that if $\omega < \frac{1}{2}$, then $X_n(\omega) = \mathbf{1}_{[0,\frac{1}{2})}(\omega) = 1$ and $\omega \in A$. Therefore $[0,\frac{1}{2}) \subset A$. On the other hand, if $\omega > \frac{1}{2}$ then $X_n(\omega) = \mathbf{1}_{[0,\frac{1}{2})}(\omega) = 0$ for every $n > \frac{1}{2\omega-1}$. Therefore we also have that $\lim_{n\to\infty} X_n(\omega) = \mathbf{1}_{[0,\frac{1}{2})}(\omega)$ for every $\omega \in (\frac{1}{2}, 1]$. Thus we deduce that

$$[0,1]\backslash\{\frac{1}{2}\} \subset A$$

and therefore that $\mathbb{P}(A) \geqslant \mathbb{P}([0,1]\backslash\{\frac{1}{2}\}) = 1$. Thus $\mathbb{P}(A) = 1$ and $X_n \xrightarrow[n\to\infty]{a.s.} \mathbf{1}_{[0,\frac{1}{2})}$.
Note here that $X_n(\frac{1}{2}) = 1$ and thus that $\mathbf{1}_{[0,\frac{1}{2})}(\frac{1}{2}) = 0$, therefore we have $\lim_{n\to\infty} X_n(\frac{1}{2}) \neq \mathbf{1}_{[0,\frac{1}{2})}(\frac{1}{2})$.

In a similar manner where almost sure convergence allowed us to capture phenomena that the usual functional convergence was able to capture, there exists other notions of convergence which will allow us to study more precise phenomena which can't always be captured by the almost sure convergence.
One of the forms of this convergence is the one given by the L_p norms.

Definition 5.4 (*Convergence in L_p norms*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that X_n converges to X in L_p, and we denote $X_n \xrightarrow[n\to\infty]{L_p} X$, if $X_n, X \in L_p$ and

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0.$$

Remark 5.5

1. *If $X_n \xrightarrow[n\to\infty]{L_1} X$ then $\mathbb{E}\,X_n \xrightarrow[n\to\infty]{} \mathbb{E}\,X$ and $\mathbb{E}\,|X_n| \xrightarrow[n\to\infty]{} \mathbb{E}\,|X|$.*
   Indeed, we can write
   $$|\mathbb{E}\,X_n - \mathbb{E}\,X| \leqslant \mathbb{E}\,|X_n - X| \quad \text{and} \quad \big|\mathbb{E}\,|X_n| - \mathbb{E}\,|X|\big| \leqslant \mathbb{E}\,|X_n - X|,$$
   then take the limit to deduce the assertion.
2. *We have already seen the monotonicity of the L_p norms. Therefore if $1 \leqslant p < q < \infty$ and $X_n \xrightarrow[n\to\infty]{L_q} X$ then $X_n \xrightarrow[n\to\infty]{L_p} X$.*

Example 5.6

Let X_n following a uniform distribution on $(0, \frac{1}{n})$. Then for any $p \geqslant 1$, we have $X_n \xrightarrow[n\to\infty]{L_p} 0$. Indeed, we have that the density of X_n is given by $f_{X_n}(x) = n\mathbf{1}_{[0,\frac{1}{n}]}(x)$, and therefore

$$\mathbb{E}\,|X_n - 0|^p = \int_0^{\frac{1}{n}} x^p n\,dx = \frac{1}{(p+1)n^p} \xrightarrow[n\to\infty]{} 0.$$

Definition 5.7 (*Convergence in probability*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that X_n converges to X in probability, and we denote $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, if for every $\varepsilon > 0$ we have

$$\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

Equivalently, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, if for every $\varepsilon > 0$ and every $\delta > 0$, there exists $N = N(\delta)$ such that

$$\mathbb{P}(|X_n - X| > \varepsilon) < \delta,$$

for every $n \geqslant N$.

Example 5.8

Let $X_n \sim \mathcal{E}(n)$ be an exponential random variable with parameter n. Then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. Indeed, for every $\varepsilon > 0$ we have

$$\mathbb{P}(|X_n - 0| > \varepsilon) = e^{-n\varepsilon}$$

which converges to 0 when n goes to infinity.

One way which would help in verifying the convergence in probability is the following.

Theorem 5.9

Let $f : \mathbb{R} \to \mathbb{R}_+$ be an even function bounded by M, increasing on $[0,1]$, continuous, and with $f(0) = 0$. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. Then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ if and only if $\lim_{n\to\infty} \mathbb{E}\left[f(X_n - X)\right] = 0$.

Proof Without loss of generality, we can suppose that $X = 0$.

Suppose first that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$ and let us show that $\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = 0$. Then we have that for every $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n| > \varepsilon) = 0$. We write then

$$f(X_n) = f(X_n)\mathbf{1}_{\{|X_n|>\varepsilon\}} + f(X_n)\mathbf{1}_{\{|X_n|\leqslant\varepsilon\}} \leqslant M\mathbf{1}_{\{|X_n|>\varepsilon\}} + f(\varepsilon),$$

where we used that f is even and increasing to say that $f(X_n) \leqslant f(\varepsilon)$ on $\{|X_n| \leqslant \varepsilon\}$ and that f is bounded. Taking the expectation on both sides, we get that

$$\mathbb{E}\left[f(X_n)\right] \leqslant M\mathbb{P}(|X_n| > \varepsilon) + f(\varepsilon),$$

for every $\varepsilon > 0$. Taking the limit in n, we have

$$\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] \leqslant f(\varepsilon),$$

for every $\varepsilon > 0$. Letting ε go to 0 and using the continuity of f with the fact that $f(0) = 0$, we get that $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = 0$.

Reversely, suppose that $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = 0$ and let us show that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. Since f is increasing, then for every $\varepsilon > 0$ we have

$$f(-\varepsilon)\mathbf{1}_{\{|X_n|>\varepsilon\}} \leqslant f(X_n)\mathbf{1}_{\{|X_n|>\varepsilon\}} \leqslant f(X_n),$$

which gives by taking the expectation that

$$f(\varepsilon)\mathbb{P}(|X_n| > \varepsilon) \leqslant \mathbb{E}[f(X_n)].$$

Since f is increasing on $[0,1]$ then $f(\varepsilon) > 0$ and taking the limit in the above inequality, we get that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. □

Remark 5.10

Therefore, taking for example $f(x) = \frac{|x|}{|x|+1}$, *we have* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *if and only if* $\lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n-X|}{1+|X_n-X|}\right] = 0$.
Similarly, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *if and only if* $\lim_{n\to\infty} \mathbb{E} \min(|X_n - X|, 1) = 0$.

5.2 Comparison of the notions of convergence

It turns out that the notion of convergence in probability is the weakest of the three introduced ones, in the sense that if we have almost sure convergence, or convergence in L_p, then we have convergence in probability. Then this convergence is designed to observe certain phenomena which are not captured by the other stronger notions of convergence (almost sure and L_p convergenece).

Theorem 5.11 (*Conv a.s.*$\Rightarrow$ *Conv* $\mathbb{P}$, *Conv* L_p $\Rightarrow$ *Conv* $\mathbb{P}$)

Let $(X_n)_{n\in\mathbb{N}}$ *be a sequence of random variables. Then we have the following:*

(i) If $X_n \xrightarrow[n\to\infty]{L_p} X$, *then* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

(ii) If $X_n \xrightarrow[n\to\infty]{a.s.} X$, *then* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Proof

(i) Let $\varepsilon > 0$, we write

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) \leqslant \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p},$$

where we used Markov inequality. It is enough to take the limit on both sides to deduce the assertion.

(ii) By the previous theorem, it is enough to prove that

$$\lim_{n\to\infty} \frac{|X_n - X|}{1 + |X_n - X|} = 0.$$

Since for every $n \in \mathbb{N}$ on a $\frac{|X_n - X|}{1+|X_n - X|} \leqslant 1$ and 1 is integrable with respect to a probability measure, then by the dominated convergence theorem, we have

$$\lim_{n\to\infty} \mathbb{E}\, \frac{|X_n - X|}{1+|X_n - X|} = \mathbb{E}\, \lim_{n\to\infty} \frac{|X_n - X|}{1+|X_n - X|} = 0,$$

since $|X_n - X| \xrightarrow[n\to\infty]{a.s.} 0$.

□

Remark 5.12

1. *The convergence in probability does not necessarily imply the almost sure convergence. Indeed, we take $(0,1]$ equipped with the Lebesgue measure. Consider the set $\Gamma = \{(n,j) : n \in \mathbb{N}, j = 1,\dots,n\}$ which we order with the lexicographical order. More precisely $(n,j) < (n',j')$ if $n < n'$ or if ($n = n'$ and $j < j'$). Graphically, if we place these points in the plane, $(n,j) < (n',j')$ if to go from (n,j) to (n',j') one has to use only "right" and "up" movements. We enumerate Γ and for every $m = (n,j) \in \Gamma$, we set $X_m = \mathbf{1}_{(\frac{j-1}{n},\frac{j}{n}]} = Y_{n,j}$.*

   Therefore $X_1 = Y_{1,1}$, $X_2 = Y_{2,1}$, $X_3 = Y_{2,2}$, ... Since for every n, the intervals $(\frac{j-1}{n},\frac{j}{n}]$ form a partition of $(0,1]$ then for every ω and every n, there exists $j, j' \leqslant n$ such that $Y_{n,j} = 1$ and $Y_{n,j'} = 0$. Therefore, we clearly have that $\limsup X_m = 1$ and $\liminf X_m = 0$ and thus the sequence X_n does not converge almost surely. However, since the X_m's are indicators of intervals of size $1/n$ then $\mathbb{P}(|X_m| > \varepsilon) \leqslant \frac{1}{n}$ which goes to zero, which shows that X_m converges to 0 in probability.

2. *The convergence in probability does not imply the convergence in L_p. Indeed, we take $(0,1]$ equipped with the Lebesgue measure. We set $Y_{n,j} = n^{\frac{1}{p}}\mathbf{1}_{(\frac{j-1}{n},\frac{j}{n}]}$ for every $n \geqslant 1$ and $j \in \{1,\dots,n\}$. We order them by lexicographical order to form a sequence which, as before, converges to 0 in probability. However $\mathbb{E}\left[|Y_{n,j}|^p\right] = 1$.*

3. *The almost sure convergence does not imply the convergence in L_p. The same previous example is an illustration of this.*

As is shown in the previous remark, the reverses to the theorem are false. However, we give below sufficient conditions allowing to obtain the reverse implications.

Theorem 5.13 (*Conv $\mathbb{P}$ ⇒ Conv a.s. for a subsequence*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then there exists a subsequence n_k such that $X_n \xrightarrow[k\to\infty]{a.s.} X$.

Proof Since $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then $\lim_{n\to\infty} \mathbb{E}\, \min(|X_n - X|, 1) = 0$ by our characterization of the convergence in probability. Therefore, there exists a subsequence n_k such that

$$\mathbb{E}\, \min(|X_{n_k} - X|, 1) \leqslant \frac{1}{2^k},$$

for every $k \in \mathbb{N}$. Thus, we have

$$\sum_{k=1}^{\infty} \mathbb{E}\, \min(|X_{n_k} - X|, 1) < \infty,$$

and by the monotone convergence theorem, we have that

$$\mathbb{E}\sum_{k=1}^{\infty}\min(|X_{n_k}-X|,1)<\infty.$$

Therefore (by the remark below), we deduce that $\sum_{k=1}^{\infty}\min(|X_{n_k}-X|,1)<\infty$ almost surely. But a convergent series has its general term going to zero, therefore we have that $|X_n-X|\xrightarrow[n\to\infty]{a.s.}0$ and we finish the proof. □

Remark 5.14

In the above proof, we have used that if $\mathbb{E}\,Y<\infty$ then $Y<\infty$ a.s. Indeed, we haven't insisted on this when defining the integral of a function, but if we allow a function to take the value ∞ and if we work on $\bar{\mathbb{R}}$, then when the value ∞ is taken over a set of non zero measure, we automatically have that the integral is equal to ∞. This implies that if the integral is finite, then the set of points where the function is infinite is actually of measure zero, and thus the function is almost surely finite.

Theorem 5.15 (*Conv $\mathbb{P}$ + bounded $\Rightarrow$ Conv L_p*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables with $|X_n|\leqslant Y$, $n\in\mathbb{N}$, and $Y\in L_p$. If $X_n\xrightarrow[n\to\infty]{\mathbb{P}}X$, then $X\in L_p$ and $X_n\xrightarrow[n\to\infty]{L_p}X$.

Proof We clearly have that $X_n\in L_p$ for every $n\in\mathbb{N}$ since $\mathbb{E}\left[|X_n|^p\right]\leqslant\mathbb{E}\left[Y^p\right]<\infty$. Let $\varepsilon>0$. Using that $|X_n|\leqslant Y$, we have that for every $n\in\mathbb{N}$

$$\mathbb{P}(|X|>Y+\varepsilon)\leqslant\mathbb{P}(|X|-|X_n|>\varepsilon)\leqslant\mathbb{P}(|X-X_n|>\varepsilon),$$

where we used that $|X|-|X_n|\leqslant|X-X_n|$ for the last inequality. Taking the limit in n and using that $X_n\xrightarrow[n\to\infty]{\mathbb{P}}X$, we deduce that for every $\varepsilon>0$

$$\mathbb{P}(|X|>Y+\varepsilon)=0$$

Therefore, taking the limit as ε goes to 0 and using the monotone convergence lemma, we get that $\mathbb{P}(|X|>Y)=0$ and thus that $|X\leqslant Y$ a.s. which implies that $X\in L_p$.
It remains to show that X_n converges to X in L_p. Suppose that it is not the case, then there would exist $\varepsilon>0$ and a subsequence (n_k) such that $\mathbb{E}\left[|X_n-X|^p\right]\geqslant\varepsilon$ for every $k\in\mathbb{N}$. But the sequence X_{n_k} converges to X in probability, then by the previous theorem it admits a subsequence $X_{n_{k_j}}$ which converges almost surely to X. Therefore, when $j\to\infty$, we have $|X_{n_{k_j}}-X|$ goes to 0 a.s. and $|X_{n_{k_j}}-X|\leqslant 2Y$, thus by the dominated convergence theorem, we get that $\mathbb{E}\left[|X_{n_{k_j}}-X|^p\right]\xrightarrow[j\to\infty]{}0$. This contradicts our assumption that $\mathbb{E}\left[|X_n-X|^p\right]\geqslant\varepsilon$ for every $k\in\mathbb{N}$. The result follows. □

Theorem 5.16

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables and f be a continuous function. Then we have

(i) If $X_n\xrightarrow[n\to\infty]{a.s.}X$, then $f(X_n)\xrightarrow[n\to\infty]{a.s.}f(X)$.

(ii) If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then $f(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} f(X)$.

Proof

(i) Let $A = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$. By hypothesis, we have $\mathbb{P}(A) = 1$. Let $\omega \in A$, and notice that by continuity of f, we can write

$$\lim_{n\to\infty} f\big(X_n(\omega)\big) = f\big(\lim_{n\to\infty} X_n(\omega)\big) = f\big(X(\omega)\big).$$

Therefore $A \subset \{\omega : \lim_{n\to\infty} f\big(X_n(\omega)\big) = f\big(X(\omega)\big)\}$. Since $\mathbb{P}(A) = 1$ then $\mathbb{P}(\{\omega : \lim_{n\to\infty} f\big(X_n(\omega)\big) = f\big(X(\omega)\big)\})$ and we deduce that $f(X_n) \xrightarrow[n\to\infty]{a.s.} f(X)$.

(ii) We know that a continuous functions is uniformly continuous over a bounded interval. Let $[-k, k]$, $k \in \mathbb{N}$, be such an interval and $\varepsilon > 0$. Therefore, there exists $\delta > 0$ such that for any $x, y \in [-k, k]$ with $|x - y| \leqslant \delta$, we have $|f(x) - f(y)| \leqslant \varepsilon$.

Using this, we have

$$\{|X_n - X| \leqslant \delta, |X| \leqslant k\} \subseteq \{|f(X_n) - f(X)| \leqslant \varepsilon, |X| \leqslant k\}$$

and by taking the complements, we get

$$\{|f(X_n) - f(X)| > \varepsilon, |X| \leqslant k\} \subseteq \{|X_n - X| > \delta, |X| \leqslant k\} \subseteq \{|X_n - X| > \delta\}.$$

Therefore

$$\{|f(X_n) - f(X)| > \varepsilon\} \subseteq \{|X_n - X| > \delta\} \bigcup \{|X| > k\}.$$

By taking the probabilities, we get

$$\mathbb{P}(|f(X_n) - f(X)| > \varepsilon) \leqslant \mathbb{P}(|X_n - X| > \delta) + \mathbb{P}(|X| > k),$$

for every $k \in \mathbb{N}$. Since $\mathbb{P}(|X| > k) \xrightarrow[k\to\infty]{} 0$, then for every $\gamma > 0$ there exists k_0, such that $\forall k \geqslant k_0$ we have

$$\mathbb{P}(|f(X_n) - f(X)| > \varepsilon) \leqslant \mathbb{P}(|X_n - X| > \delta) + \gamma,$$

for some fixed δ depending on k_0. We let n go to infinity and use that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ to deduce that

$$\lim_{n\to\infty} \mathbb{P}(|f(X_n) - f(X)| > \varepsilon) \leqslant \gamma.$$

Since this is true for every $\gamma > 0$, we have $\lim_{n\to\infty} \mathbb{P}(|f(X_n) - f(X)| > \varepsilon) = 0$ and we finish the proof.

□

5.3 Law of large numbers

In statistics, an important problem consists of trying to predict the law behind a certain phenomenon by looking at experiments done on the population. We would like for example to have an idea of the expectation of the distribution, its variance, etc. These parameters are a priori unknown because it is very costly to test all elements of a large population. Let us take the concrete example where we aim at studying the number of accidents on a certain route. Let X be the random variable defined by $X(\omega) = 1$ if the person ω had an accident and 0 otherwise. We would like to understand $\mathbb{E}\, X$ which is unknown to us. Instead of investigating all the population, we choose first a finite sample and check if they had an accident. To this aim, let $X_1, \ldots, X_n$ be independent copies of X (these are independent random variables with the same distribution than X). If we test on n people, then the average number of accidents would be given by $\frac{1}{n}\sum_{j=1}^{n} X_j$ and and what we would like to claim is that this average is in fact close to the actual mean if n is sufficiently large. This is exactly the content of the law of large numbers. We will treat several versions of this statement depending on the notion of convergence that we use.

Theorem 5.17 (*Weak law of large numbers*)

Let $(X_n)_{n\geqslant 1}$ be a sequence of independent random variables and with the same distribution as X.

(i) If $\mathbb{E}\,|X| < \infty$, then $\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\, X$.

(ii) If in addition $\mathbb{E}\,[X^2] < \infty$, then $\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow[n\to\infty]{L_2} \mathbb{E}\, X$.

Proof

(ii) We start with the second assertion which is easier to prove. This is simply a calculation which needs verification since

$$\mathbb{E}\Big[\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big|^2\Big] = \frac{\mathbb{E}\Big[\Big|\sum_{j=1}^{n}(X_j - \mathbb{E}\, X_j)\Big|^2\Big]}{n^2},$$

where we used that$\mathbb{E}\, X = \mathbb{E}\, X_j$ since they all have the same distribution. Now we expand the square to get

$$\Big|\sum_{j=1}^{n}(X_j - \mathbb{E}\, X_j)\Big|^2 = \sum_{j=1}^{n}(X_j - \mathbb{E}\, X_j)^2 + \sum_{j\neq k}(X_j - \mathbb{E}\, X_j)(X_k - \mathbb{E}\, X_k).$$

Taking the expectation and using linearity, we get

$$\mathbb{E}\Big[\Big|\sum_{j=1}^{n}(X_j - \mathbb{E}\, X_j)\Big|^2\Big] = \sum_{j=1}^{n}\mathrm{Var}(X_j) + \sum_{j\neq k}\mathbb{E}\,[(X_j - \mathbb{E}\, X_j)(X_k - \mathbb{E}\, X_k)].$$

Since the X_j's all have the same distribution then $\mathrm{Var}(X_j) = \mathrm{Var}(X)$ for every j. Moreover, using the independence of X_j and X_k for $j \neq k$, we have

$$\mathbb{E}\Big[\Big|\sum_{j=1}^{n}(X_j - \mathbb{E}\, X_j)\Big|^2\Big] = n\mathrm{Var}(X) + \sum_{j\neq k}\mathbb{E}\,[(X_j - \mathbb{E}\, X_j)]\,\mathbb{E}\,[(X_k - \mathbb{E}\, X_k)].$$

We notice now that by linearity $\mathbb{E}\left[(X_j - \mathbb{E}\, X_j)\right] = \mathbb{E}\, X_j - \mathbb{E}\left[\mathbb{E}\, X_j\right] = 0$ since the expectation of a constant is itself. Therefore we deduce that

$$\mathbb{E}\Big[\Big|\sum_{j=1}^{n}(X_j - \mathbb{E}\, X_j)\Big|^2\Big] = n\mathrm{Var}(X),$$

which implies that

$$\mathbb{E}\Big[\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big|^2\Big] \leqslant \frac{\mathrm{Var}(X)}{n},$$

which goes to 0 as n tend to infinity since $\mathrm{Var}(X)$ is finite.

(i) Let $\varepsilon > 0$. We would like to bound $\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big| > \varepsilon\Big) = \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big|^2 > \varepsilon^2\Big)$.

We would like at this point to use Markov inequality in order to bound this probability using the moments. However, this assumes the existence of moment of second order. To avoid this shortcoming, we impose the finiteness of the second moment artificially through truncation of our random variables. More precisely, for every $N \in \mathbb{N}$ and every $j \leqslant n$, we set $X_{j,N} = X_j \mathbf{1}_{\{|X_j| \leqslant N\}}$ and $X_{j,N^c} = X_j \mathbf{1}_{\{X_j > N\}}$ in such a way that $X_j = X_{j,N} + X_{j,N^c}$. Similarly, we set $X_N = X\mathbf{1}_{\{|X| \leqslant N\}}$ and $X_{N^c} = X_j \mathbf{1}_{\{X > N\}}$. What we have gained is that for N fixed, our new random variables $X_{j,N}$ have finite second moment and we can use the previous calculation to write

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_{j,N} - \mathbb{E}\, X_N\Big| > \frac{\varepsilon}{2}\Big) \leqslant \frac{2\mathrm{Var}(X_N)}{n\varepsilon}$$

For the other parts, we use the fact that we have a finite first moment to write

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_{j,N^c} - \mathbb{E}\, X_{N^c}\Big| > \frac{\varepsilon}{2}\Big) \leqslant \frac{4\mathbb{E}\,|X_{N^c}|}{\varepsilon}.$$

But

$$\Big\{\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big| > \varepsilon\Big\} \subseteq \Big\{\Big|\frac{1}{n}\sum_{j=1}^{n} X_{j,N} - \mathbb{E}\, X_N\Big| > \frac{\varepsilon}{2}\Big\} \bigcup \Big\{\Big|\frac{1}{n}\sum_{j=1}^{n} X_{j,N^c} - \mathbb{E}\, X_{N^c}\Big| > \frac{\varepsilon}{2}\Big\}.$$

Therefore, using that $\mathbb{P}(A \cup B) \leqslant \mathbb{P}(A) + \mathbb{P}(B)$, we obtain

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big| > \varepsilon\Big) &\leqslant \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_{j,N} - \mathbb{E}\, X_N\Big| > \frac{\varepsilon}{2}\Big) + \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_{j,N^c} - \mathbb{E}\, X_{N^c}\Big| > \frac{\varepsilon}{2}\Big)\\ &\leqslant \frac{2\mathrm{Var}(X_N)}{n\varepsilon} + \frac{4\mathbb{E}\,|X_{N^c}|}{\varepsilon}.\end{aligned}$$

Taking the limit as $n \to \infty$ and using that $\mathrm{Var}(X_N)$ is finite, we get

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\, X\Big| > \varepsilon\Big) \leqslant \frac{4\mathbb{E}\,|X_{N^c}|}{\varepsilon},$$

for every $N \in \mathbb{N}$. It remains to note that $|X_{N^c}| \leqslant |X|$ which is integrable, therefore by the dominated convergence theorem we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}\,X\Big| > \varepsilon\Big) \leqslant \frac{4\lim_{N\to\infty}\mathbb{E}\,|X_{N^c}|}{\varepsilon} = \frac{4\mathbb{E}\,[\lim_{N\to\infty}|X_{N^c}|]}{\varepsilon} = 0,$$

and we finish the proof.

□

Theorem 5.18 (*Strong law of large numbers*)

Let $(X_n)_{n\geqslant 1}$ be a sequence of independent random variables, with the same distribution as X, and suppose that $X \in L^1$. Then $\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow[n\to\infty]{a.s.} \mathbb{E}\,X$.

Proof Denote $S_n = \sum_{j=1}^{n} X_j$ and $S_0 = 0$. Our aim is to show that $\lim_n \frac{S_n}{n} = \mathbb{E}\,X$ a.s.

To this aim, it is enough to show that $\limsup \frac{S_n}{n} \leqslant \mathbb{E}\,X$ a.s. since replacing X by $-X$ we would get that $\limsup \frac{-S_n}{n} \leqslant -\mathbb{E}\,X$ which would imply that $\liminf \frac{S_n}{n} \geqslant \mathbb{E}\,X$ a.s.

Let $a > 0$ be such that $\mathbb{E}\,X < a$. We will show that $\limsup \frac{S_n}{n} \leqslant a$ a.s. and since this is true for any a and that the intersection of almost sure events is almost sure, we would get $\limsup \frac{S_n}{n} \leqslant \mathbb{E}\,X$ a.s.

Write

$$S_n \leqslant na + \sup_{n\in\mathbb{N}}(S_n - na)$$

and note that to show that $\limsup \frac{S_n}{n} \leqslant a$ a.s. it is enough to show that $M := \sup_{n\in\mathbb{N}}(S_n - na) < \infty$ a.s. We will make use of the 0/1 law, and to this aim we have to verify that the event $\{M < \infty\}$ is measurable with respect to the σ-algebras $\sigma(X_{k+1}, X_{k+2}, \ldots)$ for every k. Indeed,

$$\{M < \infty\} = \{\sup_{n\in\mathbb{N}}(S_n - na) < \infty\} = \{\sup_{n\geqslant k}(S_n - S_k - (n-k)a) < \infty\},$$

and $S_n - S_k = X_{k+1} + \ldots + X_n$. Therefore, by the 0/1 law, we have that $\mathbb{P}(M < \infty) \in \{0, 1\}$. It is enough then to eliminate the possibility of a zero probability in order to finish the proof. We proceed by contradiction and suppose that $\mathbb{P}(M < \infty) = 0$ which implies that $\mathbb{P}(M = \infty) = 1$ and thus

$$\mathbb{E}\,\inf(a - X, M) = \mathbb{E}\,\inf(a - X, \infty) = \mathbb{E}\,[a - X] > 0.$$

On the other hand, if we denote $M_k = \sup_{n\in\{0,\ldots,k\}}(S_n - na)$ then clearly $\inf(a - X, M_k) \xrightarrow[k\to\infty]{} \inf(a - X, M)$. Moreover, we have

$$|\inf(a - X, M_k)| \leqslant |a - X|,$$

which is integrable. Thus, we can apply the dominated convergence theorem to write

$$\mathbb{E}\,\inf(a - X, M) = \mathbb{E}\,[\lim_{k\to\infty}\inf(a - X, M_k)] = \lim_{k\to\infty}\mathbb{E}\,[\inf(a - X, M_k)].$$

We will verify that $\mathbb{E}\,[\inf(a - X, M_k)] \leqslant 0$ which would be a contradiction with the fact that $\mathbb{E}\,\inf(a - X, M) > 0$ and would finish the proof. Indeed,

$$\mathbb{E}\,[\inf(a - X, M_k)] = \mathbb{E}\,M_k + \mathbb{E}\,[\inf(a - X - M_k, 0)] = \mathbb{E}\,M_k - \mathbb{E}\,[\sup(0, M_k + X - a)].$$

We assert that $\sup(0, M_k + X - a)$ has the same distribution as M_{k+1} since

$$\sup(0, M_k + X - a) = \sup\big(0, X - a, X + X_1 - 2a, X + X_1 + X_2 - 3a, \ldots, X + X_1 + \ldots + X_k - (k+1)a\big)$$

and

$$M_{k+1} = \sup\big(0, X_1 - a, X_1 + X_2 - 2a, X_1 + X_2 + X_3 - 3a, \ldots, X_1 + X_2 + \ldots + X_{k+1} - (k+1)a\big)$$

and $(X, X_1, \ldots, X_k) \sim (X_1, \ldots, X_{k+1})$ by independence of the X_i's and the fact that they have the same distribution. Therefore we deduce that

$$\mathbb{E}\left[\inf(a - X, M_k)\right] = \mathbb{E}\, M_k - \mathbb{E}\, M_{k+1} \leqslant 0,$$

since $M_k \leqslant M_{k+1}$. □

Chapter 6

Convergence in distribution and CLT

6.1 Convergence in distribution

In the previous chapter, we investigated the different notions of convergence of random variables while keeping in mind the "mapping" aspect as well as the "probabilistic" aspect. In this chapter, we will completely disregard the "mapping" aspect and will only look at the distribution of the random variables in hand. To this aim, we will study the convergence of a sequence of probability measures. Our motivation, as previously, is to capture phenomena which couldn't be seen by other types of convergence.

Definition 6.1 (*Convergence in distribution*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We say that X_n converges to X in distribution (or in law), and we denote $X_n \underset{n\to\infty}{\overset{\text{(law)}}{\longrightarrow}} X$, if the sequence of probabilities $(\mathcal{L}_{X_n})_{n\in\mathbb{N}}$ converges weakly to $\mathcal{L}_X$. In other words, if for every bounded continuous function h, we have

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

Remark 6.2

1. *If X_n is a sequence of discrete random variables taking values in the same space E and $\lim_{n\to\infty} \mathbb{P}(X_n = a) = \mathbb{P}(X = a)$ for every $a \in E$, then X_n converges to X in distribution.*

2. *If X_n is a sequence of random variables with density f_n satisfying $f_n \leqslant g$ λ-almost surely and g is integrable. If $\lim_{n\to\infty} f_n = f$ λ-almost surely, then X_n converges converges in distribution to the random variable with density f. This is a consequence of the dominated convergence theorem.*

Example 6.3

1. *Let X_n with uniform distributions on $\{\frac{1}{n}, \frac{2}{n}, \dots, 1\}$. Then X_n converges in distribution to the uniform distribution on $[0,1]$.*

   Indeed, if h is a continuous bounded function, then $\mathbb{E}\left[h(X_n)\right] = \frac{1}{n}\sum_{i=1}^{n} h(\frac{i}{n}) \underset{n\to\infty}{\longrightarrow} \int_0^1 h(x)dx$ by the definition of the Riemann integral. But if X follows a uniform distribution on $[0,1]$

then its density is $\mathbf{1}_{[0,1]}$ and thus $\mathbb{E}[h(X)] = \int_0^1 h(x)dx$.

2. *Let X_n be normal distributions $\mathcal{N}(0,\sigma_n^2)$ with σ_n going to 0. Then X_n converges in distribution to the random variable equal to zero (whose distribution is the Dirac mass at 0).*

   Indeed, if h is continuous bounded by M, then by a change of variables we have

$$\mathbb{E}[h(X_n)] = \frac{1}{\sigma_n\sqrt{2\pi}}\int_{\mathbb{R}} h(x)e^{-\frac{x^2}{2\sigma_n^2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} h(x\sigma_n)e^{-\frac{x^2}{2}}\,dx$$

   Notice that $|h(x\sigma_n)e^{-\frac{x^2}{2}}| \leqslant Me^{-\frac{x^2}{2}}$ which is integrable, therefore by the dominated convergence theorem we have

$$\lim_{n\to\infty}\mathbb{E}[h(X_n)] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\lim_{n\to\infty} h(x\sigma_n)e^{-\frac{x^2}{2}}\,dx$$

   and since h is continuous, we get

$$\lim_{n\to\infty}\mathbb{E}[h(X_n)] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} h(0)e^{-\frac{x^2}{2}}\,dx = h(0).$$

   But if $X = 0$ a.s. we also have $\mathbb{E}[h(X)] = h(0)$.

Proposition 6.4 (*Conv $\mathbb{P} \Rightarrow$ Conv in distribution*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ then $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$.

Proof Let h be a continuous bounded function. Since $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then by Theorem 5.16 we have $h(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} h(X)$. But $h(X_n)$ is bounded, then by Theorem 5.15 we have $h(X_n) \xrightarrow[n\to\infty]{L_1} h(X)$. This implies that $\lim_{n\to\infty}\mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ and $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$. □

In the definition of convergence in distribution, we can replace the continuous bounded functions by continuous functions with bounded support. This is the content of the next proposition.

Proposition 6.5

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. Then $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$ if and only if for every compactly supported continuous function g we have $\lim_{n\to\infty}\mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$.

Proof Since a compactly supported continuous function is bounded, then the first implication is trivial. Suppose now that $\lim_{n\to\infty}\mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$ for every compactly supported continuous function and let us show that $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$.

To this aim, let h be a continuous function bounded by M. To make h compactly supported, we would like to multiply it by $\mathbf{1}_{[-L,L]}$ but this unfortunately destroys the continuity. However, let us define a function ψ_L such that $\psi_L(x) = 1$ if $x \in [-L,L]$ and $\psi_L(x) = 0$ if $x \notin [-(L+1), L+1]$. Therefore, we have that $h\psi_L$ is continuous, compactly supported and thus

$$\lim_{n\to\infty}\mathbb{E}[h(X_n)\psi_L(X_n)] = \mathbb{E}[h(X)\psi_L(X)].$$

On the other hand, by the monotone convergence lemma, there exists L_0 such that if $L \geqslant L_0$, we have

$$\mathbb{E}\left[1-\psi_L(X)\right] \leqslant \mathbb{P}(|X| \geqslant L) \leqslant \varepsilon.$$

Therefore, if $L \geqslant L_0$ and n is sufficiently large we have $\mathbb{E}\left[1-\psi_L(X_n)\right] \leqslant \varepsilon$ since $\mathbb{E}\left[1-\psi_L(X_n)\right]$ goes to $\mathbb{E}\left[1-\psi_L(X)\right]$. Now, we put together the above by writing for $L \geqslant L_0$ and n sufficiently large that

$$\begin{aligned}\left|\mathbb{E}\left[h(X_n)-h(X)\right]\right| &\leqslant \left|\mathbb{E}\left[h(X_n)(1-\psi_L(X_n))\right]\right| + \left|\mathbb{E}\left[h(X_n)\psi_L(X_n)-h(X)\psi_L(X)\right]\right| + \left|\mathbb{E}\left[h(X)(1-\psi_L(X))\right]\right| \\ &\leqslant 2\varepsilon M + \left|\mathbb{E}\left[h(X_n)\psi_L(X_n)-h(X)\psi_L(X)\right]\right|.\end{aligned}$$

Taking the limit as n goes to infinity, we get $\lim_{n\to\infty}\left|\mathbb{E}\left[h(X_n)-h(X)\right]\right| \leqslant 2\varepsilon M$. This being true for every $\varepsilon > 0$, the result follows. □

Remark 6.6

We can even push this further by replacing continuous complacty supported by C^∞ (infinitely continuously differentiable) and compactly supported.

Theorem 6.7 (*Conv in law ⇔ Conv of CDF*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. Then $X_n \xrightarrow[n\to\infty]{(law)} X$ if and only if the sequence of CDF $F_{X_n}(x)$ converges to $F_X(x)$ at every continuity point x of F_X.

Proof Suppose first that $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$. Note that $F_{X_n}(t) = \mathbb{P}(X_n \leqslant t) = \mathbb{E}\left[\mathbf{1}_{(-\infty,t]}(X_n)\right]$. If the function $\mathbf{1}_{(-\infty,t]}$ was continuous then by the convergence in distribution we would obtain what is needed. However, it is discontinuous at t and to regularize it, we introduce for every $\varepsilon > 0$, the function

$$H_\varepsilon(t,x) = \begin{cases} 1 & \text{if } x \leqslant t \\ 1-\frac{x-t}{\varepsilon} & \text{if } x \in (t,t+\varepsilon) \\ 0 & \text{if } x \geqslant t+\varepsilon \end{cases}$$

Clearly, for every $\varepsilon > 0$ and every $t \in \mathbb{R}$, the function $H_\varepsilon(t,\cdot)$ is continuous and bounded. Therefore, the convergence in distribution implies that $\lim_{n\to\infty}\mathbb{E}\left[H_\varepsilon(t,X_n)\right] = \mathbb{E}\left[H_\varepsilon(t,X)\right]$. On the other hand, we have

$$\mathbf{1}_{(-\infty,t]}(x) \leqslant H_\varepsilon(t,x) \leqslant \mathbf{1}_{(-\infty,t+\varepsilon]}(x).$$

Therefore, we can write

$$\begin{aligned}\limsup_n F_{X_n}(t) &= \limsup_n \mathbb{E}\left[\mathbf{1}_{(-\infty,t]}(X_n)\right] \\ &\leqslant \lim_{n\to\infty}\mathbb{E}\left[H_\varepsilon(t,X_n)\right] = \mathbb{E}\left[H_\varepsilon(t,X)\right] \\ &\leqslant \mathbb{E}\left[\mathbf{1}_{(-\infty,t+\varepsilon]}(X)\right] = F_X(t+\varepsilon).\end{aligned}$$

We can repeat the same procedure with $-\varepsilon$ to get

$$\liminf_n F_{X_n}(t) \geqslant F_X(t-\varepsilon).$$

Therefore, we deduce that for every $\varepsilon > 0$,

$$F_X(t-\varepsilon) \leqslant \liminf_n F_{X_n}(t) \leqslant \limsup_n F_{X_n}(t) \leqslant F_X(t+\varepsilon).$$

Moreover, if F_X is continuous at t, we would have that $\lim_{\varepsilon\to 0} F_X(t-\varepsilon) = \lim_{\varepsilon\to 0} F_X(t-\varepsilon) = F_X(t)$ and thus that $\liminf_n F_{X_n}(t) \leqslant \limsup_n F_{X_n}(t) = F_X(t)$, which proves the first implication.

Let us now prove the reverse implications and suppose that F_{X_n} converges to F_X at every continuity point of F_X. Let h be a C^∞ function with compact support and let us prove that $\lim_{n\to\infty} \mathbb{E}\left[h(X_n)\right] = \mathbb{E}\left[h(X)\right]$. Note that

$$\mathbb{E}\left[h(X)\right] = \int_{\mathbb{R}} h(x)\mathcal{L}_X(dx) = \int_{\mathbb{R}} \Big(- \int_x^\infty h'(y)dy\Big)\mathcal{L}_X(dx) = \int_{\mathbb{R}} \Big(- \int_{\mathbb{R}} h'(y)\mathbf{1}_{[x,\infty)}(y)dy\Big)\mathcal{L}_X(dx)$$

But $|h'(y)\mathbf{1}_{[x,\infty)}(y)| \leqslant |h'(y)|$ and

$$\int_{\mathbb{R}^2} |h'(y)|dy\mathcal{L}_X(dx) = \int_{\mathbb{R}} |h'(y)|dy < \infty,$$

since h' is C^∞ with compact support. We can therefore apply Fubini and reverse the order of integrals to get

$$\mathbb{E}\left[h(X)\right] = \int_{\mathbb{R}} \Big(- \int_{\mathbb{R}} h'(y)\mathbf{1}_{[x,\infty)}(y)\mathcal{L}_X(dx)\Big)dy = -\int_{\mathbb{R}} h'(y)F_X(y)\,dy$$

In a similar manner, we have $\mathbb{E}\left[h(X_n)\right] = -\int_{\mathbb{R}} h'(y)F_{X_n}(y)\,dy$. To finish, we will use the dominated convergence theorem. To this aim, note that $|h'(y)F_{X_n}(y)| \leqslant |h'(y)|$ which is integrable since continuous and has compact support. Therefore, the sequence of functions $h'F_{X_n}$ is bounded in L^1 and we have $\lim_n h'(y)F_{X_n}(y) = h'(y)F_X(y)$ at every continuity point of X. Since the set of discontinuity points of F_X is at most countable then it is of Lebesgue measure zero and thus $\lim_n h'F_{X_n} = h'F_X$ almost everywhere. By the dominated convergence theorem, we have

$$\lim_n \mathbb{E}\left[h(X_n)\right] = -\lim_n \int_{\mathbb{R}} h'(y)F_{X_n}(y)\,dy = -\int_{\mathbb{R}} \lim_n \big(h'(y)F_{X_n}(y)\big)\,dy = -\int_{\mathbb{R}} h'(y)F_X(y)\,dy = \mathbb{E}\left[h(X)\right].$$

□

6.2 Levy theorem

The goal of this section is to give another characterization of the convergence in distribution, this time in terms of the characteristic function. First we have an easy implication.

Proposition 6.8

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables. If $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$ then $\lim_n \phi_{X_n}(t) = \phi_X(t)$ for every $t \in \mathbb{R}$.

Proof This is trivial since $x \to e^{itx}$ is bounded and continuous, thus the convergence in distribution implies that $\lim_n \mathbb{E}\left[e^{itX_n}\right] = \mathbb{E}\left[e^{itX}\right]$. The result follows. □

The proposition remains valid if we replace real random variables with random vectors in $\mathbb{R}^d$. For the reverse, if we have real random variables and $\lim_{n\to\infty} \phi_{X_n} = \phi$, we need to ensure that ϕ is the characteristic function of some random variable X in order to assert that $X_n \xrightarrow[n\to\infty]{\text{(law)}} X$. To this aim, the fact that ϕ is continuous at 0 will be sufficient and this is exactly the content of the next theorem.

Theorem 6.9 (*Continuity theorem of Levy*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables. If $\phi_{X_n} \underset{n\to\infty}{\longrightarrow} \phi$ and ϕ continuous at 0, then there exists a random variable X such that $\phi_X = \phi$ and $X_n \underset{n\to\infty}{\overset{\text{(law)}}{\longrightarrow}} X$.

Remark 6.10

- *Concretely, if we calculate the limit of a sequence of characteristic functions and get a known characteristic function (of some random variable we know), then this theorem allows us to directly conclude that our sequence of random variables converges in distribution to the corresponding random variable. We can apply this theorem when the limit is not a known characteristic function, one however needs to verify that the limit function is continuous at 0 in order to claim that the sequence of random variables converges in distribution.*
- *Levy theorem extends also to $\mathbb{R}^d$ by adapting the proof below. We can then keep in mind that if $(X_n)_{n\in\mathbb{N}}$ is a sequence of random vectors in $\mathbb{R}^d$ and if $\phi_{X_n}(t)$ converges to $\phi(t)$ for every $t \in \mathbb{R}^d$ and ϕ is continuous at 0, then X_n converges in distribution to a random vector X whose characteristic function is given by ϕ.*

The proof of this theorem is quite long and we need several intermediate steps.

Definition 6.11 (*Tight family*)

We say that a sequence of real random variables $(X_n)_{n\in\mathbb{N}}$ is tight if $\forall\varepsilon > 0$, there exists a compact set K such that $\mathbb{P}(X_n \in K) \geqslant 1 - \varepsilon$ for every $n \in \mathbb{N}$.

Remark 6.12

- *If the sequence consists of only one random variable (or a finite number) then it is tight. Indeed, let $\varepsilon > 0$ and note that since F_X goes to 0 at $-\infty$ then there exists $M_1 < 0$ such that $F_X(M_1) \leqslant \varepsilon/2$. On the other hand, since F_X goes to 1 at ∞ then there exists $M_2 > 0$ such that $F_X(M_2) \geqslant 1-\varepsilon/2$. Then we have $\mathbb{P}(X \in [M_1, M_2]) = F_X(M_2) - F_X(M_1^-) \geqslant 1-\varepsilon$. Since $[M_1, M_2]$ is compact, this shows that the family formed by X alone is tight.*
- *Clearly, the union of two tight families forms a tight family.*
- *A subsequence of a tight family is tight.*
- *A sequence of random vectors $(X_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ in tight if all sequences of coordinates are tight.*

Lemme 6.13

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables. If $\phi_{X_n} \underset{n\to\infty}{\longrightarrow} \phi$ and ϕ continuous at 0, then the sequence $(X_n)_{n\in\mathbb{N}}$ is tight.

Proof Let $u > 0$ and note that by Fubini

$$\frac{1}{u}\int_{-u}^{u} (1 - \phi_{X_n}(t))\, dt = \mathbb{E}\Big[\frac{1}{u}\int_{-u}^{u} \big(1 - e^{itX_n}\big)\, dt\Big] = 2\mathbb{E}\Big[1 - \frac{\sin(uX_n)}{uX_n}\Big].$$

On the other hand, if $|x| \geqslant 2$ then $|\frac{\sin x}{x}| \leqslant 1/2$; therefore

$$\mathbb{P}\big(|X_n| \geqslant \frac{2}{u}\big) = 2(1 - \frac{1}{2})\,\mathbb{E}\left[\mathbf{1}_{\{|X_n| \geqslant 2/u\}}\right] \leqslant 2\mathbb{E}\Big[1 - \frac{\sin(uX_n)}{uX_n}\Big].$$

We deduce that for any $n \in \mathbb{N}$, we have

$$\mathbb{P}\big(|X_n| \geqslant \frac{2}{u}\big) \leqslant \frac{1}{u}\int_{-u}^{u} |1 - \phi_{X_n}(t)|\, dt$$

By Fatou lemma, we have

$$\limsup_n \mathbb{P}\big(|X_n| \geqslant \frac{2}{u}\big) \leqslant \limsup_n \frac{1}{u}\int_{-u}^{u} |1 - \phi_{X_n}(t)|\, dt \leqslant \frac{1}{u}\int_{-u}^{u} |1 - \phi_X(t)|\, dt$$

But ϕ is continuous at 0, then

$$\limsup_{u\to 0}\limsup_{n\to\infty} \mathbb{P}\big(|X_n| \geqslant \frac{2}{u}\big) \leqslant \limsup_{u\to 0} \frac{1}{u}\int_{-u}^{u} |1 - \phi_X(t)|\, dt = 0$$

Therefore, if $\varepsilon > 0$, we can find $M > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\big(|X_n| \geqslant M\big) < \varepsilon,$$

and deduce that the family is tight. □

Lemme 6.14

Let $(X_n)_{n\in\mathbb{N}}$ be a tight family of real random variables. Then there exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ and a random variable X such that $X_{n_k} \xrightarrow[k\to\infty]{\text{(law)}} X$.

Proof Let us look at the sequence of corresponding CDFs. For every rational q, the sequence $(F_{X_n}(q))_{n\in\mathbb{N}}$ is a real bounded sequence, and by Bolzano-Weierstrass theorem, it admits a convergent subsequence. By the diagonal procedure, we form a subsequence $(F_{n_k})_{k\in\mathbb{N}}$ which converges at every rational point. We denote G the limit function obtained which we extend to $\mathbb{R}$ by setting

$$F(x) = \inf\{G(q) : q \in \mathbb{Q} \cap (x, \infty)\}.$$

Clearly F is nondecreasing and continuous to the right. Let us show that F_{n_k} converges to F at every continuity point of F.
Let $x \in \mathbb{R}$ be a continuity point of F and $\varepsilon > 0$. Since F is continuous at x, then there exists $\eta > 0$ such that $|F(x) - F(y)| \leqslant \varepsilon$ for every $y \in [x - \eta, x + \eta]$. Let $x - \eta \leqslant r \leqslant x \leqslant s \leqslant x + \eta$ be two rationals, then

$$F_{n_k}(r) \leqslant F_{n_k}(x) \leqslant F_{n_k}(s)$$

and taking the limit in k, we get that

$$F(r) \leqslant \liminf_k F_{n_k}(x) \leqslant \limsup_k F_{n_k}(x) \leqslant F(s).$$

But $F(r) \geqslant F(x) - \varepsilon$ and $F(s) \leqslant F(x) + \varepsilon$. Therefore, we deduce that

$$F(x) - \varepsilon \leqslant \liminf_k F_{n_k}(x) \leqslant \limsup_k F_{n_k}(x) \leqslant F(x) + \varepsilon,$$

for every $\varepsilon > 0$, which proves the convergence of the subsequence $(F_{n_k})_{k\in\mathbb{N}}$ to F at every continuity point of F.
It remains to show that $\lim_{-\infty} F(x) = 0$ and $\lim_{\infty} F(x) = 1$ to deduce that F is the CDF of some random variable X. We will make use of the fact that the sequence is tight. Let $\varepsilon > 0$. There exists $M > 0$ such that for every $k \in \mathbb{N}$ we have $F_{X_{n_k}}(t) \leqslant \varepsilon$ for every $t \leqslant -M$ and $F_{X_{n_k}}(t) \geqslant 1-\varepsilon$ for every $t \geqslant M$. Taking the limit in $k \in \mathbb{N}$, we deduce that $F_X(t) \leqslant \varepsilon$ for every $t \leqslant -M$ and $F_{X_{n_k}}(t) \geqslant 1 - \varepsilon$ for every $t \geqslant M$ (with t a continuity point). Taking the limit along continuity points, we deduce that $\lim_{-\infty} F(x) \leqslant \varepsilon$ and $\lim_{\infty} F(x) \geqslant 1 - \varepsilon$. This being true for every $\varepsilon > 0$, we deduce what is needed.
Finally, the characterization of the convergence in distribution in terms of CDFs allows to finish the proof. □

We are now ready to prove Levy theorem.
Proof [Proof of Theorem 6.9] By Lemma 6.13, we have that the sequence $(X_n)_{n\in\mathbb{N}}$ is tight. Let $(X_{n_\ell})_{\ell\in\mathbb{N}}$ be an arbitrary subsequence of $(X_n)_{n\in\mathbb{N}}$. Since it is also tight, we can extract from it a subsequence $(X_{n_{\ell_k}})_{k\in\mathbb{N}}$ which converges in distribution to some random variable X. By Proposition 6.8 the characteristic function of $X_{n_{\ell_k}}$ converges to ϕ_X. By the unicity of the limit, we have that $\phi_X = \phi$. Finally, we have proved that every subsequence of $(X_n)_{n\in\mathbb{N}}$, has a subsequence which converges in distribution to X. Therefore, we deduce that X_n converges in distribution to X. □

6.3 Central limit theorem

Given a sequence of independent real random variables with the same distribution, in L^1, we have already seen by the strong law of large numbers that $\frac{1}{n}\sum_{j=1}^n X_j \xrightarrow[n\to\infty]{\text{(a.s.)}} \mathbb{E}[X_1]$.
Now we would like to further quantify this convergence and understand the corresponding speed of convergence. To this aim, we look at studying the difference $\sum_{j=1}^n (X_j - \mathbb{E}\,X_j)$. What do we expect?
Suppose that the variables are also in L^2 and write

$$\mathbb{E}\big[\big(\sum_{j=1}^n (X_j - \mathbb{E}\,X_j)\big)^2\big] = \sum_{j=1}^n \mathbb{E}\big[(X_j - \mathbb{E}\,X_j)^2\big],$$

where we used independence. Since the X_j's have the same distribution, they have in particular the same variance which we denote by σ^2. We deduce that

$$\mathbb{E}\big[\big(\sum_{j=1}^n (X_j - \mathbb{E}\,X_j)\big)^2\big] = n\sigma^2.$$

This suggest that $\sum_{j=1}^n (X_j - \mathbb{E}\,X_j)$ is of the same order than $\sigma\sqrt{n}$. To make this more precise, let us study

$$\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^n (X_j - \mathbb{E}\,X_1).$$

Let us take an example in order to understand what to expect. Suppose that the X_j's are independent Rademacher variables with parameter 1/2 i.e. $\mathbb{P}(X_j = -1) = \mathbb{P}(X_j = 1) = 1/2$.

Note that $\mathbb{E}[X_j] = 0$ and $\mathrm{Var}(X_j) = \sigma^2 = 1$ and let us calculate the characteristic function of $S_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n X_j$

$$\phi_{S_n}(t) = \prod_{j=1}^n \phi_{X_j}(t/\sqrt{n}) = \prod_{j=1}^n \frac{e^{it/\sqrt{n}} + e^{-it/\sqrt{n}}}{2} = \cos^n\big(\frac{t}{\sqrt{n}}\big) = e^{n \ln \cos\big(\frac{t}{\sqrt{n}}\big)}.$$

But $\cos\big(\frac{t}{\sqrt{n}}\big)$ is equivalent to $1 - \frac{t^2}{2n}$ when $n \to \infty$ and $\ln\big(1 - \frac{t^2}{2n}\big)$ is equivalent to $-\frac{t^2}{2n}$. Therefore, we deduce that

$$\phi_{S_n}(t) \underset{n\to\infty}{\longrightarrow} e^{-\frac{t^2}{2}},$$

which is nothing but the characteristic function of the normal distribution. The central limit theorem asserts that this is not a coincidence and that this is always the case.

Theorem 6.15 (*Central limit theorem*)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent real random variables with the same distributions, in L^2. We denote $M = \mathbb{E}\, X_1$ the mean and $\sigma^2 = \mathrm{Var}(X_1)$ the variance. Then we have

$$\frac{X_1 + \ldots + X_n - nM}{\sigma\sqrt{n}} \underset{n\to\infty}{\overset{(law)}{\longrightarrow}} \mathcal{N}(0,1).$$

Proof We can suppose without loss of generality that the X_j's are centered i.e. $\mathbb{E}\, X_j = 0$, otherwise we work with $X_j - \mathbb{E}\, X_j$. Similarly, we can suppose that $\sigma = 1$ otherwise we work with $\frac{1}{\sigma}X$ instead of X.
Let as before $S_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n X_j$ and note that by the independence of the X_j's we can write

$$\phi_{S_n}(t) = \prod_{j=1}^n \phi_{X_j}(t/\sqrt{n}) = \phi_{X_1}^n\big(\frac{t}{\sqrt{n}}\big).$$

But by the Taylor expansion of the exponential function at 0, we can write

$$\mathbb{E}\, e^{i\frac{t}{\sqrt{n}}X_1} = 1 + i\frac{t}{\sqrt{n}}\mathbb{E}\, X_1 - \frac{t^2}{2n}\mathbb{E}\, X_1^2 + o\big(\frac{1}{n}\big) = 1 - \frac{t^2}{2n} + o\big(\frac{1}{n}\big).$$

Using this, we have

$$\lim_{n\to\infty} \phi_{S_n}(t) = \lim_{n\to\infty} \big(1 - \frac{t^2}{2n} + o\big(\frac{1}{n}\big)\big)^n = e^{-\frac{t^2}{2}},$$

which is the characteristic function of the standard normal distribution. By Levy theorem, we can conclude that S_n converges in distribution to $\mathcal{N}(0,1)$. □

Remark 6.16

- *Sometimes we state the theorem without normalization by writing*

$$\frac{X_1 + \ldots + X_n - nM}{\sqrt{n}} \underset{n\to\infty}{\overset{(law)}{\longrightarrow}} \mathcal{N}(0,\sigma^2).$$

- *We can translate the conclusion of the central limit theorem in terms of the different characterizations of the convergence in distribution. Therefore, we can state that if $X_1, \ldots, X_n$ is a* *sequence of independent random variables with the same distribution in L^2 and of variance σ^2, then for any $a < b$ we have*

$$\mathbb{P}\Big(\sum_{j=1}^{n} X_j \in [n\mathbb{E}\, X_1 + a\sqrt{n}, n\mathbb{E}\, X_1 + b\sqrt{n}]\Big) \underset{n\to\infty}{\longrightarrow} \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2\sigma^2}}\, dx.$$

- *We can also use the central limit theorem to calculate certain limits which are a priori difficult. For example, we can assert that for every $a < b$*

$$\frac{1}{2^n} \sum_{\frac{n}{2}+a\sqrt{n}\leqslant k\leqslant \frac{n}{2}+b\sqrt{n}} \binom{n}{k} \underset{n\to\infty}{\longrightarrow} \sqrt{\frac{2}{\pi}} \int_a^b e^{-2x^2}\, dx.$$

Indeed, if $(X_j)_{j\in\mathbb{N}}$ are independent Bernoulli variables with parameter $1/2$, then $\sum_{j=1}^{n} X_j$ follows a binomial distribution and we know that

$$\mathbb{P}\Big(\sum_{j=1}^{n} X_j = k) = \frac{1}{2^n}\binom{n}{k}.$$

Therefore since $\mathbb{E}\, X_1 = 1/2$ then

$$\mathbb{P}\Big(\sum_{j=1}^{n} X_j \in [n\mathbb{E}\, X_1 + a\sqrt{n}, n\mathbb{E}\, X_1 + b\sqrt{n}]\Big) = \frac{1}{2^n} \sum_{\frac{n}{2}+a\sqrt{n}\leqslant k\leqslant \frac{n}{2}+b\sqrt{n}} \binom{n}{k}.$$

Since in addition $\sigma^2 = 1/4$, then our previous remark implies the result.

Chapter 7

Gaussian vectors

7.1 Properties of Gaussian random variables

Recall that a Gaussian (or normal) random variable $g \sim \mathcal{N}(\mu, \sigma^2)$ is a random variable with density on $\mathbb{R}$ given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{\sigma^2}}.$$

Therefore we have $\mathbb{E}\, g = \mu$ and $\mathrm{Var}(g) = \sigma^2$. By convention, the zero random variable is Gaussian. We have already calculated the characteristic function of a normal random variable $g \sim \mathcal{N}(\mu, \sigma^2)$

$$\phi_g(t) := \mathbb{E}\, e^{itg} = e^{it\mu - t^2\sigma^2/2}.$$

The latter characterizes the distribution. In other words, there is no other random variable which shares the same characteristic function as the normal random variable. Another way to characterize the normal distribution is through Stein identity which is an integration by parts formula special for Gaussians.

Theorem 7.1 (***Stein identity***)

Let X be a random variable. Then the following two assertions are equivalent

1. $X \sim \mathcal{N}(0, 1)$.
2. *For every differentiable function h such that $\mathbb{E}\,|h'(X)| < \infty$ and $\mathbb{E}\,|X\, h(X)| < \infty$, we have* $\mathbb{E}\, h'(X) = \mathbb{E}[X\, h(X)]$.

Proof

1⇒ 2. We denote $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ the density of the normal distribution. By integration by parts, we have

$$\mathbb{E}\, h'(X) = \int_{\mathbb{R}} h'(x) f(x)\, dx = \Big[h(x)f(x)\Big]_{-\infty}^{\infty} + \int_{\mathbb{R}} xh(x)f(x)\, dx.$$

It is enough to note that the first term above is surely zero since h is integrable (since $xh(x)$ is).

2⇒ 1. Taking $h(x) = e^{itx}$ we have $\phi'_X(t) = -t\phi_X(t)$ and $\phi_X(0) = 1$. We deduce that $\phi_X(t) = e^{-t^2/2}$ and that $X \sim \mathcal{N}(0, 1)$.

□

Therefore, it is only the standard normal distribution which satisfies the equation $\mathbb{E}\left[Xh(X)\right] = \mathbb{E}\, h'(X)$. This allows us to easily calculate the moments of the standard normal distribution.

Proposition 7.2 (*Moments*)

Let $g \sim \mathcal{N}(0,1)$. Then $\mathbb{E}[g^k] = 0$ if k is odd and

$$\mathbb{E}[g^k] = \frac{\Gamma(k+1)}{2^{k/2}\Gamma(k/2+1)},$$

if k is even.

Proof The first property follows by symmetry of g. For the other, we use Stein identity to write

$$\mathbb{E}[g^{k+2}] = (k+1)\mathbb{E}[g^k].$$

Therefore by induction, we prove that $\mathbb{E}[g^{2k}] = \frac{(2k)!}{2^k\, k!}$. □

7.2 Random vectors and covariance matrices

We now discuss higher dimensions and work in $\mathbb{R}^n$. Recall that if $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a random vector then $\mathbb{E}\, X$ is nothing but the vector $(\mathbb{E}\, X_1, \ldots, \mathbb{E}\, X_n)$. Now the covariance is recorded in a matrix by looking at covariances between all coordinates.

Definition 7.3 (*Covariance matrix*)

If $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a random vector. The covariance matrix of X is defined by

$$\Sigma = \Big(\mathrm{Cov}(X_i, X_j)\Big)_{i,j \in \{1,\ldots,n\}}.$$

The diagonal terms are givem by $\mathrm{Var}(X_i)$.

Remark 7.4

As the covariance of two independent random variables is zero, then if X is a random vector whose coordinates are independent then its covariance matrix is diagonal.

Proposition 7.5

If $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a random vector. Then its covariance matrix is given by $\Sigma = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^t]$ where $(X - \mathbb{E}X)^t$ denotes the transpose of $(X - \mathbb{E}X)$. Moreover Σ is positive semidefinite.

Proof The matrix product gives that the (i,j)-th entry of the matrix $(X - \mathbb{E}X)(X - \mathbb{E}X)^t$ is nothing but $(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)$. Therefore, the expectation of the (i,j)-th entry of the matrix $(X - \mathbb{E}X)(X - \mathbb{E}X)^t$ is nothing but $\mathrm{Cov}(X_i, X_j)$ which proves that $\Sigma = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^t]$. To check that Σ is positive semidefinite, note that $(X - \mathbb{E}X)(X - \mathbb{E}X)^t$ is positive semidefinite since for every $x \in \mathbb{R}^n$

$$\Big\langle (X - \mathbb{E}X)(X - \mathbb{E}X)^t x, x \Big\rangle = \langle (X - \mathbb{E}X), x\rangle^2 \geqslant 0.$$

□

Remark 7.6

- *It is interesting to have an operator point of vue of the covariance matrix. The point of vue of the definition is of statistical flavor, while from the above proposition we have*

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

We notice that $(X - \mathbb{E}X)(X - \mathbb{E}X)^t$ is a projection, it is the operator which projects in the direction of $X - \mathbb{E}X$. Therefore, the covariance matrix is the "average projection" in the directions of $X - \mathbb{E}X$.

- *We also have $\mathbb{E}\left[\|X - \mathbb{E}\,X\|_2^2\right] = \mathrm{Tr}(\Sigma)$. Indeed,*

$$\begin{aligned}\|X - \mathbb{E}\,X\|_2^2 &= \langle X - \mathbb{E}\,X, X - \mathbb{E}\,X\rangle = (X - \mathbb{E}\,X)^t(X - \mathbb{E}\,X) \\ &= \mathrm{Tr}\big((X - \mathbb{E}\,X)^t(X - \mathbb{E}\,X)\big) = \mathrm{Tr}\big((X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^t\big).\end{aligned}$$

Since the trace is linear, then we have

$$\mathbb{E}\left[\|X - \mathbb{E}\,X\|_2^2\right] = \mathrm{Tr}\big(\mathbb{E}\left[(X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^t\right]\big) = \mathrm{Tr}(\Sigma).$$

Recall the multi-dimensional Levy theorem.

Theorem 7.7 (*Multi-dimensional Levy theorem*)

Let $(Y_N)_{N\in\mathbb{N}}$ be a sequence of random vectors in $\mathbb{R}^n$. Then we have the followings:

(i) If Y_N converges in distribution to Y, then ϕ_{Y_N} converges to ϕ_Y.

(ii) If ϕ_{Y_N} converges to ϕ and ϕ is continuous at 0, then Y_N converges in distribution to a random vector Y whose characteristic function is given by ϕ.

A nice consequence of this is that the convergence in distribution of random vectors can be reduced the convergence in distribution of random variables. Indeed, we have the following:

Theorem 7.8

Let $(Y_N)_{N\in\mathbb{N}}$ be a sequence of random vectors in $\mathbb{R}^n$. Then Y_N converges in distribution if and only if for every $\theta \in \mathbb{R}^d$, the sequence $(\langle Y_N, \theta\rangle)_{N\in\mathbb{N}}$ converges in distribution. Moreover, if Y_N converges in distribution to Y then $\langle Y_N, \theta\rangle$ converges in distribution to $\langle Y, \theta\rangle$.

Proof The first implication is trivial, since if Y_N converges in distribution to Y, then since $\langle Y_N, \theta\rangle$ is a function of Y_N then $\langle Y_N, \theta\rangle$ converges in distribution to $\langle Y, \theta\rangle$.
For the reverse, note that if $\langle Y_N, \theta\rangle$ converges in distribution then $\phi_{\langle Y_N, \theta\rangle}$ converges to a characteristic function which we denote by h_θ for every $\theta \in \mathbb{R}^d$. Now note that by definition,

$$\phi_{Y_N}(\theta) = \phi_{\langle Y_N, \theta\rangle}(1)$$

which converges to $h_\theta(1)$. Therefore ϕ_{Y_N} converges to a function ϕ defined by $\phi(\theta) = h_\theta(1)$. If ϕ is continuous at 0, then Levy theorem allows us to conclude that Y_N converges in distribution. It is enough to notice that we can suppose that Y_N belongs to a compact; indeed, the sequence of coordinates of Y_N are tight since the corresponding characteristic functions converge to a function continuous at 0. Therefore, by definition, this implies that Y_N is tight and that for every $\varepsilon > 0$, we can find a ball $B(0,r)$ such that $\mathbb{P}\big(Y_N \in B(0,r)\big) \geqslant 1-\varepsilon$ for every N. Let $\delta > 0$ be such that $\forall x \in (-\delta r, \delta r)$ we have $|1-e^{ix}| \leqslant \varepsilon$. We then write for every $t \in B(0,\delta)$

$$\begin{aligned}|\phi(\theta+t)-\phi(\theta)| &= \lim_N \big|\mathbb{E}\,[e^{i\langle Y_N,\theta+t\rangle} - e^{i\langle Y_N,\theta\rangle}]\big| \\ &\leqslant 2\varepsilon + \lim_N \big|\mathbb{E}\,[(e^{i\langle Y_N,\theta+t\rangle} - e^{i\langle Y_N,\theta\rangle})\mathbf{1}_{\{Y_N\in B(0,r)\}}]\big| \\ &\leqslant 2\varepsilon + \lim_N \mathbb{E}\,\big|(1-e^{i\langle Y_N,t\rangle})\mathbf{1}_{\{Y_N\in B(0,r)\}}\big|.\end{aligned}$$

But if $Y_N \in B(0,r)$ then $\langle Y_N, t\rangle \in (-\delta r, \delta r)$, therefore $\big|(1-e^{i\langle Y_N,t\rangle})\mathbf{1}_{\{Y_N\in B(0,r)\}}\big| \leqslant \varepsilon$ and we deduce that we have found δ such that for every $t \in B(0,\delta)$ we have $|\phi(\theta+t)-\phi(\theta)| \leqslant 3\varepsilon$. This shows that ϕ is continuous at 0 and Levy theorem finishes the proof. □

7.3 Gaussian vectors

We denote S^{n-1} the unit Euclidean sphere of $\mathbb{R}^n$ i.e. $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = \sqrt{x_1^2+\ldots+x_n^2} = 1\}$.

Definition 7.9 (*Gaussian random vector*)

A random vector $X \in \mathbb{R}^n$ is called Gaussian if $\forall \theta \in S^{n-1}$, $\langle X,\theta\rangle$ is a Gaussian random variable. We will denote $X \sim \mathcal{N}(\mu,\Sigma)$ to mean that X is a Gaussian vector with mean μ and with covariance matrix Σ.

Remark 7.10

- *In other words, a vector is Gaussian if all of its projections in all directions are Gaussian random variables.*
- *Since we made the convention that 0 is a Gaussian random variable, then the zero vector is Gaussian. Moreover, if one (or many) of the projections of the vector is zero (and the others are Gaussians), it means that the vector is Gaussian. For example, if ξ_1 and ξ_2 are two independent Gaussian variables, then $(\xi_1,\xi_2,\xi_1+\xi_2)$ is a Gaussian vector. The same holds for (ξ_1,ξ_1). More generally, if all coordinates of a vector $X \in \mathbb{R}^n$ belong to $\mathrm{vect}(\xi_1,\ldots,\xi_s)$ with $\xi_1,\ldots,\xi_s$ independent Gaussian variables (and $s \leqslant n$), then X is Gaussian.*
- *We also have that $\forall x \in \mathbb{R}^n$, the variable $\langle X,x\rangle$ is Gaussian. It is enough to write $\langle X,x\rangle = \|x\|_2\langle X, \frac{x}{\|x\|_2}\rangle$ to see that.*
- *If X is a centered Gaussian vector, whose covariance matrix is the identity matrix, we will say that X is a **standard Gaussian** vector.*
- *If $X \sim \mathcal{N}(\mu,\Sigma)$ then for every $\theta \in \mathbb{R}^n$, we have $\langle X,\theta\rangle \sim \mathcal{N}(\langle\mu,\theta\rangle, \langle\Sigma\theta,\theta\rangle)$. Indeed, we know that $\langle X,\theta\rangle$ is a Gaussian random variable, it is enough to just calculate its expectation and variance. But the scalar product is linear, therefore $\mathbb{E}\langle X,\theta\rangle = \langle \mathbb{E}\,X,\theta\rangle = \langle\mu,\theta\rangle$. For the variance, we note that*

$$\mathbb{E}\,\langle X-\mu,\theta\rangle^2 = \mathbb{E}\,[\theta^t(X-\mu)(X-\mu)^t\theta] = \theta^t\mathbb{E}\,[(X-\mu)(X-\mu)^t]\theta,$$

by linearity. Thus we deduce that the variance is given by $\mathbb{E}\,\langle X-\mu,\theta\rangle^2 = \theta^t\Sigma\theta = \langle\Sigma\theta,\theta\rangle$.

Proposition 7.11

If $g_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \dots, g_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$ are independent Gaussian random variables, then $X = (g_1, \dots, g_n)$ is a Gaussian random vector and we have $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (\mu_1, \dots, \mu_n)$ and Σ is the diagonal matrix with σ_i^2 on its diagonal.

Proof Let $\theta \in S^{n-1}$, we need to check that $\langle X, \theta \rangle$ is a Gaussian random variable. To this aim, we calculate the characteristic function

$$\phi_{\langle X,\theta\rangle}(t) = \mathbb{E}\, e^{it\sum_{j=1}^n \theta_j g_j} = \prod_{j=1}^n \phi_{g_j}(t\theta_j),$$

where we used the independence of the X_j's. A simple calculation finishes the proof. □

Proposition 7.12

If $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1), \dots, X_n \sim \mathcal{N}(\mu_n, \Sigma_n)$ are independent Gaussian vectors, then for every $\theta \in \mathbb{R}^n$, we have that $\sum_{j=1}^n \theta_j X_j \sim \mathcal{N}(\sum_{j=1}^n \theta_j \mu_j, \sum_{j=1}^n \theta_j^2 \Sigma_j)$.

Proof Exercise. □

In a similar manner, we have:

Proposition 7.13

Let $X \sim N(\mu, \Sigma)$, then for every $N \times n$ matrix A we have $AX \sim \mathcal{N}(A\mu, A\Sigma A^t)$.

Remark 7.14

An important observation is that a standard Gaussian vector is invariant by rotation. Indeed, if X is standard Gaussian and U is a rotation, then $UX \sim \mathcal{N}(0, UU^t)$ and since $UU^t = I$ we deduce that $UX \sim X$.

As for real random variables, we define the characteristic function $\widehat{\Phi}_X : \mathbb{R}^n \to \mathbb{C}$ of a random vector X by

$$\widehat{\Phi}_X(t) = \mathbb{E}\, e^{i\langle X,t\rangle}.$$

Let us calculate the characteristic function of a Gaussian vector.

Proposition 7.15

If X is a Gaussien vector with mean μ and covariance matrix Σ. Then

$$\phi_X(t) = e^{i\langle t,\mu\rangle - \frac{\langle \Sigma t, t\rangle}{2}}.$$

Proof We note that $\langle X, t\rangle \sim \mathcal{N}\big(\langle \mu, t\rangle, \langle \Sigma t, t\rangle\big)$. Therefore

$$\phi_X(t) = \phi_{\langle X,t\rangle}(1) = e^{i\langle t,\mu\rangle - \langle \Sigma t, t\rangle/2}.$$

□

We have already seen that if two random variables are independent then their covariance is zero and that the reverse is false. However, the reverse becomes trues if these variables form a Gaussian vector. Therefore, for Gaussian random vectors, there is equivalence between the fact that the coordinates are independent and that the covariance matrix is diagonal.

Proposition 7.16

Let $X \sim \mathcal{N}(\mu, \Sigma)$. Then X has independent coordinates if and only if Σ is diagonal.

Proof It is enough to prove that if Σ is diagonal then the coordinates of X are independent since the reverse is always true. By the characterization of independence using characteristic functions, it is enough to prove that for every $u = (u_j)_{j=1,\dots,n} \in \mathbb{R}^n$ we have

$$\mathbb{E}e^{iu_1 g_1} \dots e^{iu_n g_n} = \mathbb{E}e^{iu_1 g_1}\mathbb{E}e^{iu_2 g_2} \dots \mathbb{E}e^{iu_n g_n},$$

where $g_1, \dots, g_n$ are the coordinates of X. Therefore, the task is to prove that

$$\phi_X(u) = \prod_{j=1}^{n} \phi_{g_j}(u_j).$$

Using previous calculations, the task is to verify that

$$e^{i\langle u,\mu\rangle - \frac{\langle \Sigma u, u\rangle}{2}} = \prod_{j=1}^{n} e^{iu_j\mu_j - \frac{u_j^2\sigma_{jj}^2}{2}}.$$

It remains to note that since Σ is diagonal then $\langle \Sigma u, u\rangle = \sum_{j=1}^{n} u_j^2 \sigma_{jj}^2$. □

To finish with the properties of Gaussian vectors, let us give the formula for the density of such vector when it exists.

Proposition 7.17

Let $X \sim \mathcal{N}(\mu, \Sigma)$. X has a density on $\mathbb{R}^n$ if and only if Σ is positive semidefinite, and in such case it is given by

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\Big(-\frac{\langle \Sigma^{-1}(x-\mu), x-\mu\rangle}{2}\Big).$$

In particular, the density of a standard Gaussian vector is given by

$$d\gamma_n = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{\|x\|_2^2}{2}\Big).$$

Proof We start with the case where X is a standard Gaussian vector. Since its covariance matrix is diagonal, then its coordinates are independent and thus is distribution is equal to the product of the marginal distributions and the result follows.

Now if $X \sim \mathcal{N}(\mu, \Sigma)$, we define $Y = \Sigma^{-1/2}(X - \mu)$ which is a standard normal vector and thus has the density

$$\frac{1}{(2\pi)^{n/2}} e^{-\|y\|_2^2/2}.$$

It remains to operate the change of variables $y = \Sigma^{-1/2}(x - \mu)$ to deduce the result.

For the reverse, suppose that Σ is not positive definite, then there exists $x \in \mathbb{R}^n$ such that $\Sigma x = 0$. Then $\mathbb{E}\langle (X-\mu), x\rangle^2 = \Sigma x, x\rangle = 0$ and thus $\langle (X-\mu), x\rangle = 0$ a.s. Therefore X belongs to the hyperplane $\mu + x^{\perp}$ almost surely. This implies that X has no density with respect to the Lebesgue measure since otherwise the probability of a hyperplane would be zero. □

7.4 Multi-dimensional central limit theorem

We have everything in place to state the central limit theorem in higher dimensions. The proof is essentially the same.

Theorem 7.18 (*Multi-dimensional CLT*)

Let $(X_N)_{N\in\mathbb{N}}$ be a sequence of independent random vectors with the same distribution in $\mathbb{R}^n$, and ans such that all coordinates are in L^2, and denote by Σ the covariance matrix of X_1. Then,

$$\frac{X_1+\ldots+X_N-N\mathbb{E}\,X_1}{\sqrt{N}} \xrightarrow[n\to\infty]{(law)} \mathcal{N}(0,\Sigma)$$

Proof Suppose for simplicity that $\mathbb{E}\,X_1=0$. We write

$$\phi_{\frac{1}{\sqrt{N}}\sum_{j=1}^N X_j}(t)=\mathbb{E}\,e^{i\langle\frac{1}{\sqrt{N}}\sum_{j=1}^N X_j,t\rangle}=\prod_{j=1}^N \mathbb{E}\,e^{i\frac{\langle t,X_j\rangle}{\sqrt{N}}},$$

where we used the independence of the X_j's. As before, we have

$$\mathbb{E}\,e^{i\frac{\langle t,X_j\rangle}{\sqrt{N}}}=1-\frac{1}{2N}\langle\Sigma t,t\rangle+o\big(\frac{1}{N}\big).$$

Therefore, we obtain

$$\lim_{N\to\infty}\phi_{\frac{1}{\sqrt{N}}\sum_{j=1}^N X_j}(t)=\lim_{N\to\infty}\big(1-\frac{1}{2N}\langle\Sigma t,t\rangle+o\big(\frac{1}{N}\big)\big)^N=e^{-\frac{\langle\Sigma t,t\rangle}{2}},$$

which is nothing but the characteristic function of $\mathcal{N}(0,\Sigma)$. The result now follows thanks to the multi-dimensional Levy theorem. □

Chapter 8

Conditional expectation

8.1 Conditioning with respect to an event

We work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Our goal as a first step is to define the expectation of a random variable conditioned on the realization of some event. We can take inspiration from the definition of the conditional probability.

Definition 8.1 (*Conditional probability*)

Let $B \in \mathcal{A}$ be an event such that $\mathbb{P}(B) > 0$. We define the conditional probability knowing B, which we denote $\mathbb{P}(\cdot \mid B)$, by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark 8.2

- *It is an easy exercise to verify that $\mathbb{P}(\cdot \mid B)$ defines a probability measure.*
- *The idea is that we restrict to B in such a way that the universe Ω is replaced by B. For any event $A \subset \Omega$, we replace it by $A \cap B$ which is its part present in B. Finally, the normalization by $\mathbb{P}(B)$ is to impose that the total probability is one.*

With this definition in mind, we already have an indication on how to define the conditional expectation for staircase random variables. Indeed, we can define

$$\mathbb{E}[\mathbf{1}_A \mid B] = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{E}[\mathbf{1}_A \mathbf{1}_B]}{\mathbb{P}(B)}.$$

Therefore, for any nonnegative staircase function, we would have the definition $\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)}$. This is exactly the definition we will use.

Definition 8.3 (*Conditional expectation knowing an event*)

Let X be a random variable and $B \in \mathcal{A}$ be an event. If $X \geqslant 0$ or $X \in L^1$, we define the conditional expectation of X knowing B by

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)}.$$

Remark 8.4

1. *This definition is well in accordance with what we expect since*

$$\mathbb{E}[X \mid B] = \mathbb{E}_{\mathbb{P}(\cdot|B)}[X].$$

This is automatic for $X = \mathbf{1}_A$*, thus by linearity for every nonnegative staircase random variable, and then by monotone convergence for any nonnegative random variable. The case of random variables in* L^1 *follows by decomposing into nonnegative and negative parts. This interpretation allows us to transfer all properties which we had establish for the expectation, to the case of the conditional expectation knowing an event.*

2. *As for the total probability formula, we have that if* B *and* B^c *have positive probabilities, then*

$$\mathbb{E}[X] = \mathbb{P}(B)\mathbb{E}[X \mid B] + \mathbb{P}(B^c)\mathbb{E}[X \mid B^c].$$

Indeed, this follows from the definition of the conditional expectation, the linearity of expectation and the fact that $\mathbf{1}_B + \mathbf{1}_{B^c} = 1$.

3. *Clearly, we also have* $\mathbb{E}[X] = \mathbb{E}[X \mid \Omega]$.

Example 8.5

Let X *be a random variable uniform on* $\{-2, -1, 0, 1, 2\}$ *and let* B *be the event that* X *is nonnegative. Then*

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)} = \frac{0.\frac{1}{5} + 1.\frac{1}{5} + 2.\frac{1}{5}}{\frac{3}{5}} = 1.$$

Indeed, X *follows a uniform distribution on* $\{-2, -1, 0, 1, 2\}$*, thus conditioned on being in* B*, it follows a uniform distribution on* B*, and the expected value on* B *is* 1.

8.2 Conditioning with respect to a discrete random variable

We proceed as we did in the whole course. We define the objects for the indicators, we extend to nonnegative staircase variables and then to any nonnegative random variable and finally to any integrable random variable.

Our goal here is to define the conditional expectation of a random variable X knowing another random variable Y. Let us start by looking at the simplest setting when $Y = \mathbf{1}_B$ for some event B. We know that $B = \{\mathbf{1}_B = 1\}$ and thus $\mathbb{E}[X \mid B] = \mathbb{E}[X \mid \{\mathbf{1}_B = 1\}]$. But $\mathbf{1}_B$ generates another event $\{\mathbf{1}_B = 0\} = B^c$. We would then like to define $\mathbb{E}[X \mid \mathbf{1}_B]$ in such a way as to have access to both $\mathbb{E}[X \mid \{\mathbf{1}_B = 1\}]$ and $\mathbb{E}[X \mid \{\mathbf{1}_B = 0\}]$. To this aim, we will define $\mathbb{E}[X \mid \mathbf{1}_B]$ as being a random variable which gives us access to both these quantities. Therefore, for every $\omega \in \Omega$, if $\omega \in B$ we set

$$\mathbb{E}[X \mid \mathbf{1}_B](\omega) = \mathbb{E}[X \mid \{\mathbf{1}_B = 1\}]$$

and if $\omega \notin B$ we set

$$\mathbb{E}[X \mid \mathbf{1}_B](\omega) = \mathbb{E}[X \mid \{\mathbf{1}_B = 0\}].$$

In other words, we have set for every $\omega \in \Omega$

$$\mathbb{E}[X \mid \mathbf{1}_B](\omega) = \mathbb{E}[X \mid \{\mathbf{1}_B = \mathbf{1}_B(\omega)\}],$$

where we recall that $\{\mathbf{1}_B = \mathbf{1}_B(\omega)\}$ is an abbreviation for $\{\omega' \in \Omega : \mathbf{1}_B(\omega') = \mathbf{1}_B(\omega)\}$.

Definition 8.6 (*Conditional expectation knowing a discrete random variable*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X, Y be random variables defined on this space and $X \in L^1$. Moreover, suppose that Y takes values in a countable space E and denote $E' = \{y \in E : \mathbb{P}(Y = y) > 0\}$. The conditional expectation of X knowing Y is defined by

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid \{Y = y\}],$$

if $Y(\omega) = y \in E'$ and 0 otherwise.

Remark 8.7

1. *One should keep in mind that $\mathbb{E}[X \mid Y]$ is a random variable and not a number.*
2. *One can also define the conditional expectation differently by changing the value 0 when $Y(\omega) \in E \backslash E'$. However, this would result in a random variable which is almost surely equal to the one we defined above since $\mathbb{P}(Y \in E \backslash E') = 0$.*

Example 8.8

1. *Let Y be a constant random variable i.e. $Y(\omega) = c$ for every $\omega \in \Omega$. Then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ i.e. $\mathbb{E}[X \mid Y]$ is a constant random variable equal to $\mathbb{E}[X]$.*

   Indeed, for every $\omega \in \Omega$, we have $Y(\omega) = c$ and thus $\{Y = c\} = \Omega$. It follows that $\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid \Omega] = \mathbb{E}[X]$.

2. *Let Y be a random variable taking values in a countable space. Then $\mathbb{E}[Y \mid Y] = Y$ a.s.*

   Indeed, let $\omega \in \Omega$, set $y = Y(\omega)$ and suppose that $\mathbb{P}(Y = y) > 0$. We write

$$\mathbb{E}[Y \mid Y](\omega) = \mathbb{E}[Y \mid \{Y = y\}] = \frac{\mathbb{E}[Y\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)} = \frac{\mathbb{E}[y\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)} = y = Y(\omega).$$

3. *Consider rolling a fair dice. We set $\Omega = \{1, \ldots, 6\}$ and define $X(\omega) = \omega$ for every $\omega \in \Omega$ and $Y(\omega) = 1$ if ω is even and 0 otherwise. Let us calculate $\mathbb{E}[X \mid Y]$.*

   Let $\omega \in \Omega$. If ω is even, then $Y(\omega) = 1$ and

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid \{Y = 1\}] = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=1\}}]}{\mathbb{P}(Y = 1)} = \frac{2.\frac{1}{6} + 4.\frac{1}{6} + 6\frac{1}{6}}{\frac{1}{2}} = 4.$$

   If ω is odd, then $Y(\omega) = 0$ and

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid \{Y = 0\}] = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=0\}}]}{\mathbb{P}(Y = 0)} = \frac{1.\frac{1}{6} + 3.\frac{1}{6} + 5\frac{1}{6}}{\frac{1}{2}} = 3.$$

Proposition 8.9

With the previous definition, $\mathbb{E}[X \mid Y]$ is a random variable $\sigma(Y)$-measurable.

Proof Indeed, we can write $\mathbb{E}[X \mid Y] = \phi(Y)$ with

$$\phi(y) = \mathbb{E}[X \mid \{Y = y\}] = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)}$$

if $y \in E'$ and 0 otherwise. Since ϕ is measurable and $\mathbb{E}[X \mid Y] = \phi(Y)$ then $\mathbb{E}[X \mid Y]$ is $\sigma(Y)$-measurable. □

Proposition 8.10

Let X, Y be two random variables such that $X \in L^1$ and Y takes values in a countable set. We have the following:

(i) We have $\mathbb{E}\left[|\mathbb{E}[X \mid Y]|\right] \leqslant \mathbb{E}|X|$ and thus $\mathbb{E}[X \mid Y] \in L^1$.

(ii) If Z is $\sigma(Y)$-measurable and bounded, then $\mathbb{E}[ZX] = \mathbb{E}\left[Z\mathbb{E}[X \mid Y]\right]$.

(iii) IfY' is a discrete random variable such that $\sigma(Y') = \sigma(Y)$, then $\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid Y']$ a.s.

Proof

(i) We use the definition of the expectation of a discrete random variable and we note that $\mathbb{E}[X \mid Y]$ takes the value$\frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y=y)}$ exactly when $Y = y$. Therefore

$$\mathbb{E}\left[|\mathbb{E}[X \mid Y]|\right] = \sum_{y\in E'} \mathbb{P}(Y = y)\frac{|\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]|}{\mathbb{P}(Y = y)} = \sum_{y\in E'} |\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]| \leqslant \sum_{y\in E'} \mathbb{E}\left[|X|\mathbf{1}_{\{Y=y\}}\right].$$

It is enough then to use the linearity of expectation together with the fact that $\sum_{y\in E'} \mathbf{1}_{\{Y=y\}} \leqslant 1$ to deduce that $\mathbb{E}\left[|\mathbb{E}[X \mid Y]|\right] \leqslant \mathbb{E}|X|$.

(ii) We have already seen that if Z est $\sigma(Y)$-measurable then $Z = h(Y)$ for a measurable function h. Moreover h is bounded since Z is. We start by writing the definition of the expectation as before

$$\mathbb{E}\left[Z\mathbb{E}[X \mid Y]\right] = \sum_{y\in E'} \mathbb{P}(Y = y)h(y)\frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)} = \sum_{y\in E'} h(y)\mathbb{E}[X\mathbf{1}_{\{Y=y\}}] = \sum_{y\in E'} \mathbb{E}[h(y)X\mathbf{1}_{\{Y=y\}}].$$

But $\mathbb{E}[h(y)X\mathbf{1}_{\{Y=y\}}] = 0$ for every $y \in E\backslash E'$ (we use here that h is bounded), thus

$$\mathbb{E}\left[Z\mathbb{E}[X \mid Y]\right] = \sum_{y\in E} \mathbb{E}[h(y)X\mathbf{1}_{\{Y=y\}}] = \sum_{y\in E} \mathbb{E}[h(Y)X\mathbf{1}_{\{Y=y\}}] = \mathbb{E}[h(Y)X],$$

where we used the linearity of expectation and the fact that $\sum_{y\in E} \mathbf{1}_{\{Y=y\}} = 1$.

(iii) We set $Z = \mathbf{1}_{\{\mathbb{E}[X|Y]>\mathbb{E}[X|Y']\}}$ which is $\sigma(Y)$ (and $\sigma(Y')$) measurable. Therefore by (ii), we have

$$\mathbb{E}[ZX] = \mathbb{E}\left[Z\mathbb{E}[X \mid Y]\right] = \mathbb{E}\left[Z\mathbb{E}[X \mid Y']\right].$$

Therefore, by linearity, we deduce that

$$\mathbb{E}\left[\mathbf{1}_{\{\mathbb{E}[X|Y]>\mathbb{E}[X|Y']\}}\big(\mathbb{E}[X \mid Y]-\mathbb{E}[X \mid Y']\big)\right]=\mathbb{E}\left[Z\big(\mathbb{E}[X \mid Y]-\mathbb{E}[X \mid Y']\big)\right]=0$$

This shows that necessarily $\mathbb{E}[X \mid Y] \leqslant \mathbb{E}[X \mid Y']$ a.s. We repeat the same in the other way to obtain that $\mathbb{E}[X \mid Y'] \leqslant \mathbb{E}[X \mid Y]$ a.s. and finally deduce that $\mathbb{E}[X \mid Y]=\mathbb{E}[X \mid Y']$ a.s.

□

Remark 8.11

The third item in the above proposition suggests that the conditional expectation knowing a random variable (discrete for the moment) depends actually on the σ-algebra generated by this variable; in the sense that it stays invariant even when we change the random variable as long as we do not change the corresponding σ-algebra. This observation will lead us in the sequel to define the conditional expectation knowing a σ-algebra.

8.3 Construction of the conditional expectation

Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space. We haven't really defined L^p spaces in this course; to do so, we will identify two random variables X and Y if they are equal almost surely. The space $L^p(\Omega,\mathcal{A},\mathbb{P})$ will consist of equivalence classes whose represent has a finite p-moment. It turns out that $L^2(\Omega,\mathcal{A},\mathbb{P})$ is a Hilbert space equipped with the scalar product $\langle X,Y\rangle=\mathbb{E}[XY]$. If $\mathcal{B}\subseteq\mathcal{A}$ is a σ-algebra, then $L^2(\Omega,\mathcal{B},\mathbb{P})$ is a closed subspace of $L^2(\Omega,\mathcal{A},\mathbb{P})$ which consists of all random variables of $L^2(\Omega,\mathcal{A},\mathbb{P})$ which are $\mathcal{B}$-measurable. We will start by defining the conditional expectation for random variables in L^2. To this aim, let us try to interpret the expectation differently. Given a random variable X, its expectation $\mathbb{E}\,X$ is exactly the quantity which minimizes the quadratic distance to X. More precisely, if X is a machine which produces random numbers and we would like to predict an outcome as close as possible to the number chosen by X, then we would look at a number a which minimizes for example $\mathbb{E}[(X-a)^2]=a^2-2a\mathbb{E}[X]+\mathbb{E}[X^2]$. The latter is a second degree polynomial which is minimized by $a=\mathbb{E}[X]$.

Now imagine that we have access to additional information concerning the machine X. We would then like to calculate the expectation knowing these information which are events encrypted into a σ-algebra. Previously, the answer was a number since we had no information and thus for any $\omega\in\Omega$, $\mathbb{E}[X \mid \mathcal{F}](\omega)=\mathbb{E}[X]$ with $\mathcal{F}=\{\varnothing,\Omega\}$. Now if the informations we know are encrypted in a σ-algebra $\mathcal{B}$, then then the answer would be a random variable which takes into account this encryption, thus a random variable which is $\mathcal{B}$-measurable. This motivates us to define $\mathbb{E}[X \mid \mathcal{B}]$ as the $\mathcal{B}$-measurable random variable which minimizes $\mathbb{E}[(X-a)^2]$ among all random variables a which are $\mathcal{B}$-measurable. This makes perfect sense in L^2 since this is exactly the definition of the orthogonal projection.

Definition 8.12 (*Conditional expectation in L^2*)

Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, $\mathcal{B}\subseteq\mathcal{A}$ be a σ-algebra and $X\in L^2(\Omega,\mathcal{A},\mathbb{P})$. Then we define the conditional expectation of X knowing $\mathcal{B}$, which we denote $\mathbb{E}[X \mid \mathcal{B}]$, as the orthogonal projection of X on $L^2(\Omega,\mathcal{B},\mathbb{P})$. More precisely, we have

$$\mathbb{E}[X \mid \mathcal{B}]=\arg\min\{\|X-Z\|_{L^2}:\ Z\in L^2(\Omega,\mathcal{B},\mathbb{P})\},$$

or also that $X-\mathbb{E}[X \mid \mathcal{B}]\perp L^2(\Omega,\mathcal{B},\mathbb{P})$ i.e. $\forall Y\in L^2(\Omega,\mathcal{B},\mathbb{P})$, $\mathbb{E}[(X-\mathbb{E}[X \mid \mathcal{B}])Y]=0$.

Remark 8.13

1. *Since the L^2 spaces are defined as equivalence classes, then every equality involving the conditional expectation is in fact an almost sure equality.*
2. *By definition, if $X \in L^2$, then $\mathbb{E}[X \mid \mathcal{B}] \in L^2$.*
3. *By definition, $\mathbb{E}[X \mid \mathcal{B}]$ is the unique $\mathcal{B}$-measurable random variable which satisfies that for every $Y \in L^2(\Omega, \mathcal{B}, \mathbb{P})$, we have $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[X \mid \mathcal{B}]]$.*
4. *If $\mathcal{B} = \{\varnothing, \Omega\}$ then $\mathbb{E}[X \mid \mathcal{B}] = \mathbb{E}[X]$. Indeed, the constant random variables are the only $\mathcal{B}$-mesurables variables, and since $\mathbb{E}[(X - \mathbb{E}[X])c] = 0$ for every constant c then $\mathbb{E}[X \mid \mathcal{B}] = \mathbb{E}[X]$.*
5. *We have $\mathbb{E}[X \mid \mathcal{A}] = X$ since $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$.*
6. *If $\mathcal{C} \subseteq \mathcal{B} \subseteq \mathcal{A}$, then $\mathbb{E}[X \mid \mathcal{C}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{B}] \mid \mathcal{C}]$ since $L^2(\Omega, \mathcal{C}, \mathbb{P}) \subset L^2(\Omega, \mathcal{B}, \mathbb{P})$.*
7. *The conditional expectation as defined is of course linear (Exercise).*
8. *If $X \geqslant 0$ then $\mathbb{E}[X \mid \mathcal{B}] \geqslant 0$. Indeed since $\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}}$ is $\mathcal{B}$-measurable, then we have*
$$\mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{B}])\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}}] = 0$$
and since $(X - \mathbb{E}[X \mid \mathcal{B}])\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}} \geqslant 0$, this implies that $\mathbb{E}[X \mid \mathcal{B}] \geqslant 0$ a.s.
9. *Similarly, we show that if X, Y are bounded and that $X \leqslant Y$ a.s. then $\mathbb{E}[X \mid \mathcal{B}] \leqslant \mathbb{E}[Y \mid \mathcal{B}]$.*

Let us check that all this is coherent with everything we previously saw.

Proposition 8.14

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ and Y be a random variable taking values in a countable space. Then $\mathbb{E}[X \mid \sigma(Y)] = \mathbb{E}[X \mid Y]$ a.s. where $\mathbb{E}[X \mid Y]$ was defined in the previous section.

Proof It is enough to show that for every $\sigma(Y)$-measurable random variable Z, we have
$$\mathbb{E}[(X - \mathbb{E}[X \mid Y])Z] = 0,$$
and by unicity of the orthogonal projection we would have $\mathbb{E}[X \mid \sigma(Y)] = \mathbb{E}[X \mid Y]$. But we already verified that if Z is $\sigma(Y)$-measurable, then we have $\mathbb{E}[XZ] = \mathbb{E}[X\mathbb{E}[X \mid Y]]$ which finishes the proof. □

We can now generalize this definition to integrable random variables.

Definition 8.15 (*Conditional expectation*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{B} \subseteq \mathcal{A}$ be a σ-algebra and $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. We define the conditional expectation of X knowing $\mathcal{B}$, which we denote $\mathbb{E}[X \mid \mathcal{B}]$, as the unique random variable $Z \in L^1(\Omega, \mathcal{B}, \mathbb{P})$ which satisfies
$$\mathbb{E}[XY] = \mathbb{E}[ZY],$$
for every bounded $\mathcal{B}$-measurable random variable Y.

This definition is in accordance with the case of random variables in L^2. However, in the L^2 case, we defined the conditional expectation explicitly as being the orthogonal projection. This is no longer the case here and we need to verify that this definition makes sense i.e. that such random variable in fact exists, and that it is uniquely defined.

Proposition 8.16 (*Unicity*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{B} \subseteq \mathcal{A}$ be a σ-algebra and $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Let Z and W be two random variables in $L^1(\Omega, \mathcal{B}, \mathbb{P})$ satisfying $\mathbb{E}[ZY] = \mathbb{E}[WY]$ for every bounded $\mathcal{B}$-measurable random variable Y. then $Z = W$ a.s.

Proof Since Z and W are $\mathcal{B}$-measurable then $\mathbf{1}_{\{Z>W\}}$ is $\mathcal{B}$-measurable (and bounded). Therefore, we have

$$\mathbb{E}[(Z-W)\mathbf{1}_{\{Z>W\}}] = 0,$$

which implies that $Z \leqslant W$ a.s.
Doing the same with $\mathbf{1}_{\{Z<W\}}$, we finish the proof. □

It remains now to show the existence promised by the definition.

Proposition 8.17 (*Existence*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ ube a probability space, $\mathcal{B} \subseteq \mathcal{A}$ be a σ-algebra and $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Then there exists $Z \in L^1(\Omega, \mathcal{B}, \mathbb{P})$ satisfying $\mathbb{E}[XY] = \mathbb{E}[ZY]$ for any bounded $\mathcal{B}$-measurable random variable Y.

Proof Suppose first that $X \geqslant 0$. We set $X_n = \min(X, n)$ and we note that $X = \lim_n X_n$. The idea is that the X_n's are bounded, and thus in L^2. We can therefore define the conditional expectation of X_n knowing $\mathcal{B}$ as being the orthogonal projection of X_n on $L^2(\Omega, \mathcal{B}, \mathbb{P})$. Then we have that $\mathbb{E}[X_n \mid \mathcal{B}]$ is a nonnegative nondecreasing sequence which is $\mathcal{B}$-measurable. We set $Z := \lim_n \mathbb{E}[X_n \mid \mathcal{B}]$, which is $\mathcal{B}$-measurable. Note that by the monotone convergence theorem, we have

$$\mathbb{E}[Z] = \mathbb{E}[\lim_n \mathbb{E}[X_n \mid \mathcal{B}]] = \lim_n \mathbb{E}[\mathbb{E}[X_n \mid \mathcal{B}]] = \lim_n \mathbb{E}[X_n] = \mathbb{E}[X].$$

Therefore, if X is integrable, then Z is as well. Let Y be bounded, nonnegative and $\mathcal{B}$-measurable. By the monotone convergence theorem, we have

$$\mathbb{E}[XY] = \lim_n \mathbb{E}[X_n Y].$$

But $\mathbb{E}[X_n Y] = \mathbb{E}[\mathbb{E}[X_n \mid \mathcal{B}]Y]$, thus we have

$$\mathbb{E}[XY] = \lim_n \mathbb{E}[\mathbb{E}[X_n \mid \mathcal{B}]Y] = \mathbb{E}[ZY],$$

where we used again the monotone convergence theorem. Now if Y is arbitrary, we write $Y = Y^+ - Y^-$ and we repeat the same procedure to deduce that $\mathbb{E}[XY] = \mathbb{E}[ZY]$. Therefore, when $X \geqslant 0$, we have shown the existence of a random variable Z satisfying the conclusion. If X is arbitrary, we write $X = X^+ - X^-$ to find $Z = Z^+ - Z^-$ and finish the proof. □

Remark 8.18

The proof above shows that if $X \geqslant 0$, then we can define the conditional expectation even if $\mathbb{E}[X]$ is infinite. However, for an arbitrary X, we need that $X \in L^1$. Therefore, all properties of conditional *expectations states in the sequel are valid for nonnegative random variables are valid even if these random variables are not integrable.*

Definition 8.19 (*Conditional expectation knowing a variable*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ and Y be a random variable. We define the conditional expectation of X knowing Y, which we denote by $\mathbb{E}[X \mid Y]$, as the conditional expectation of X knowing $\sigma(Y)$.

8.4 Properties of the conditional expectation

Let us summarize the properties of the conditional expectation.

Proposition 8.20 (*Properties of the conditional expectation*)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{B} \subseteq \mathcal{A}$ be a σ-algebra and $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. We have the followings:

(a) $\mathbb{E}[XY] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{B}]Y\big]$ *for every bounded $\mathcal{B}$-measurable random variable Y.*

(b) $\forall a, b \in \mathbb{R}$, *we have* $\mathbb{E}[(aX + bY) \mid \mathcal{B}] = a\mathbb{E}[X \mid \mathcal{B}] + b\mathbb{E}[Y \mid \mathcal{B}]$.

(c) If $X \geqslant 0$, then $\mathbb{E}[X \mid \mathcal{B}] \geqslant 0$.

(d) If a is a constant, then $\mathbb{E}[a \mid \mathcal{B}] = a$.

(e) If $\mathcal{F} = \{\varnothing, \Omega\}$, then $\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[X]$ and $\mathbb{E}[X \mid \mathcal{A}] = X$.

(f) If X is $\mathcal{B}$-measurable, then $\mathbb{E}[XZ \mid \mathcal{B}] = X\mathbb{E}[Z \mid \mathcal{B}]$ for every random variable $Z \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. In particular if X is $\mathcal{B}$-measurable, $\mathbb{E}[X \mid \mathcal{B}] = X$.

(g) If $\mathcal{C} \subseteq \mathcal{B}$, then $\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{B}] \mid \mathcal{C}\big] = \mathbb{E}[X \mid \mathcal{C}]$. In particular, $\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{B}]\big] = \mathbb{E}[X]$.

(h) If $\sigma(X)$ and $\mathcal{B}$ are independent, then $\mathbb{E}[X \mid \mathcal{B}] = \mathbb{E}[X]$. More generally, if Y is $\mathcal{B}$-measurable, then

$$\mathbb{E}[g(X, Y) \mid \mathcal{B}] = \int g(x, Y)\, \mathcal{L}_X(dx),$$

for every nonnegative measurable function g.

(i) (Jensen) If ϕ is convex and $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, then

$$\mathbb{E}[\phi(X) \mid \mathcal{B}] \geqslant \phi(\mathbb{E}[X \mid \mathcal{B}]).$$

Proof

(a) This follows from the definition.

(b) It is enough to show that for any bounded $\mathcal{B}$-measurable random variable Z, we have

$$\mathbb{E}[(a\mathbb{E}[X \mid \mathcal{B}] + b\mathbb{E}[Y \mid \mathcal{B}])Z] = \mathbb{E}[(aX + bY)Z],$$

and we would get by the unicity of the conditional expectation that $\mathbb{E}[(aX + bY) \mid \mathcal{B}] = a\mathbb{E}[X \mid \mathcal{B}] + b\mathbb{E}[Y \mid \mathcal{B}]$. The linearity of the expectation implies that

$$\mathbb{E}[(a\mathbb{E}[X \mid \mathcal{B}] + b\mathbb{E}[Y \mid \mathcal{B}])Z] = a\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{B}]Z\big] + b\mathbb{E}\big[\mathbb{E}[Y \mid \mathcal{B}]Z\big].$$

By (f), we have $\mathbb{E}[X \mid \mathcal{B}]Z = \mathbb{E}[XZ \mid \mathcal{B}]$ and by (g) we have $\mathbb{E}\big[\mathbb{E}[XZ \mid \mathcal{B}]\big] = \mathbb{E}[XZ]$. We do the same for Y, to obtain that

$$\mathbb{E}[(a\mathbb{E}[X \mid \mathcal{B}] + b\mathbb{E}[Y \mid \mathcal{B}])Z] = a\mathbb{E}[XZ] + b\mathbb{E}[YZ]$$

and the linearity of expectation finishes the proof.

(c) We know that $\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}}$ is $\mathcal{B}$-measurable and bounded. By the definition of the conditional expectation, we have

$$\mathbb{E}\big[X\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}}\big] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{B}]\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}}\big]$$

But $X\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}} \geqslant 0$ thus $\mathbb{E}\big[X\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}}\big] \geqslant 0$. But if $\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}} \neq 0$ then $\mathbb{E}[X \mid \mathcal{B}]\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}} < 0$, which would be contradictory. We deduce that $\mathbf{1}_{\{\mathbb{E}[X|\mathcal{B}]<0\}} = 0$ almost surely and thus that $\mathbb{E}[X \mid \mathcal{B}] \geqslant 0$ almost surely.

(d) Since a is $\mathcal{B}$-measurable (and that $\mathbb{E}[aY] = \mathbb{E}[aY]$ for every Y $\mathcal{B}$-measurable bounded), then by the unicity in the definition of the conditional expectation, we have that $\mathbb{E}[a \mid \mathcal{B}] = a$.

(e) The random variables which are $\mathcal{F}$-measurable are the constant random variables. Therefore, we have that $\mathbb{E}[X \mid \mathcal{F}]$ is a constant random variable and thus that $\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}]] = \mathbb{E}[X \mid \mathcal{F}]$. But by definition, $\mathbb{E}[X \mid \mathcal{F}]$ is the unique random variable which satisfies

$$\mathbb{E}\big[Y\mathbb{E}[X \mid \mathcal{F}]\big] = \mathbb{E}[YX],$$

for every random variable Y which is $\mathcal{F}$-measurable and bounded. Taking $Y = 1$ we get that $\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}]\big] = \mathbb{E}[X]$ and finish the proof.

For the second assertion, since X is $\mathcal{A}$-measurable then by (f) we have $\mathbb{E}[X \mid \mathcal{A}] = X$.

(f) Suppose that X is $\mathcal{B}$-measurable and let Y be $\mathcal{B}$-measurable and bounded. Set $X_N = X\mathbf{1}_{\{|X|\leqslant N\}}$ in such a way that $X_N Y$ is $\mathcal{B}$-measurable and bounded. By definition of the conditional expectation, we then have

$$\mathbb{E}\big[X_N ZY\big] = \mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{B}]X_N Y\big]$$

Taking the limit as $N \to \infty$ and using the dominated convergence theorem, we deduce that

$$\mathbb{E}\big[XZY\big] = \mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{B}]XY\big].$$

Since this is true for every Y $\mathcal{B}$-measurable and bounded, then it follows from the definition of the conditional expectation that $\mathbb{E}[XZ \mid \mathcal{B}] = X\mathbb{E}[Z \mid \mathcal{B}]$.

It remains to apply this with $Z = 1$ and use (d) to deduce the second part.

(g) Let Y $\mathcal{C}$-measurable and bounded. By definition, $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right] \mid \mathcal{C}\right]$ is the unique random variable satisfying

$$\mathbb{E}\left[Y\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right] \mid \mathcal{C}\right]\right] = \mathbb{E}\left[Y\mathbb{E}\left[X \mid \mathcal{B}\right]\right]$$

But Y is also $\mathcal{B}$-measurable and bounded, thus by definition of $\mathbb{E}\left[X \mid \mathcal{B}\right]$, we have

$$\mathbb{E}\left[Y\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right] \mid \mathcal{C}\right]\right] = \mathbb{E}\left[XY\right].$$

But by definition $\mathbb{E}\left[X \mid \mathcal{C}\right]$ is the unique random variable satisfying

$$\mathbb{E}\left[Y\mathbb{E}\left[X \mid \mathcal{C}\right]\right] = \mathbb{E}\left[XY\right].$$

Therefore, we deduce that $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right] \mid \mathcal{C}\right] = \mathbb{E}\left[X \mid \mathcal{C}\right]$.

For the second part, we take $\mathcal{C} = \mathcal{F} = \{\varnothing, \Omega\}$ and use (e) to write

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right] \mid \mathcal{C}\right].$$

The first part implies that $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{B}\right] \mid \mathcal{C}\right] = \mathbb{E}\left[X \mid \mathcal{C}\right]$ which is equl to $\mathbb{E}\left[X\right]$ by (e).

(h) It is of course enough to prove the second part then take $g(X, Y) = X$ to get the first. By definition, $\mathbb{E}\left[g(X, Y) \mid \mathcal{B}\right]$ is the unique random variable satisfying

$$\mathbb{E}\left[Z\mathbb{E}\left[g(X, Y) \mid \mathcal{B}\right]\right] = \mathbb{E}\left[Z\, g(X, Y)\right] = \int z\, g(x, y)\, \mathcal{L}_{(X,Y,Z)}(dx, dy, dz),$$

for every Z which is $\mathcal{B}$-measurable and bounded. But X and (Y, Z) are independent, thus $\mathcal{L}_{(X,Y,Z)} = \mathcal{L}_X\mathcal{L}_{(Y,Z)}$. Therefore,

$$\mathbb{E}\left[Z\mathbb{E}\left[g(X, Y) \mid \mathcal{B}\right]\right] = \int z\, g(x, y)\, \mathcal{L}_X(dx)\mathcal{L}_{(Y,Z)}(dy, dz) = \int z\Big(\int g(x, y)\, \mathcal{L}_X(dx)\Big)\, \mathcal{L}_{(Y,Z)}(dy, dz),$$

where we used Fubini. By unicity in the definition of the conditional expectation, we obtain the result.

(i) The proof is the same as for the classical Jensen inequality; we use that locally, ϕ is the supremum of affine functions.

□

Similarly, we can extend the convergence theorems to the setting of conditional expectations.

Proposition 8.21

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{B} \subseteq \mathcal{A}$ be a σ-algebra and $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.

(i) (MON) If $(X_n)_{n\in\mathbb{N}}$ is nonnegative nondecreasing and $X = \lim_n X_n$, then $\lim_n \mathbb{E}\left[X_n \mid \mathcal{B}\right] = \mathbb{E}\left[X \mid \mathcal{B}\right]$.

(ii) (Fatou) If $(X_n)_{n\in\mathbb{N}}$ is nonengative and $X = \lim_n X_n$, then

$$\mathbb{E}\left[\liminf X_n \mid \mathcal{B}\right] \leqslant \liminf \mathbb{E}\left[X_n \mid \mathcal{B}\right].$$

(iii) (Dominated convergence) If there exists an integrable Y such that $|X_n| \leqslant Y$ for every $n \in \mathbb{N}$,

then

$$\lim_n \mathbb{E}\left[|X_n - X| \mid \mathcal{B}\right] = 0 \quad \textit{and} \quad \lim_n \mathbb{E}\left[X_n \mid \mathcal{B}\right] = \mathbb{E}\left[X \mid \mathcal{B}\right],$$

where X is the almost sure limit of X_n.

Proof

(i) By definition, for every Y which is $\mathcal{B}$-measurable bounded and nonnegative, we have

$$\mathbb{E}\left[X_n Y\right] = \mathbb{E}\big[\mathbb{E}\left[X_n \mid \mathcal{B}\right] Y\big].$$

But $X_n Y$ and $\mathbb{E}\left[X_n \mid \mathcal{B}\right] Y$ form a nonnegative nondecreasing sequence, thus by the monotone convergence theorem we have

$$\mathbb{E}\left[XY\right] = \lim_n \mathbb{E}\left[X_n Y\right] = \lim_n \mathbb{E}\big[\mathbb{E}\left[X_n \mid \mathcal{B}\right] Y\big] = \mathbb{E}\big[\lim_n \mathbb{E}\left[X_n \mid \mathcal{B}\right] Y\big].$$

Decomposing into nonnegative and negative part, we obtain that for every Y which is $\mathcal{B}$-measurable bounded

$$\mathbb{E}\left[XY\right] = \mathbb{E}\big[\lim_n \mathbb{E}\left[X_n \mid \mathcal{B}\right] Y\big].$$

By definition of $\mathbb{E}\left[X \mid \mathcal{B}\right]$, this implies that $\lim_n \mathbb{E}\left[X_n \mid \mathcal{B}\right] = \mathbb{E}\left[X \mid \mathcal{B}\right]$.

(ii) We set $Y_m = \inf_{n \geqslant m} X_n$ and note that Y_m is a nonnegative nondecreasing sequence. By (i), we have

$$\mathbb{E}\left[\lim_m Y_m \mid \mathcal{B}\right] = \lim_m \mathbb{E}\left[Y_m \mid \mathcal{B}\right].$$

But $Y_m \leqslant X_n$ for every $n \geqslant m$, thus $\mathbb{E}\left[Y_m \mid \mathcal{B}\right] \leqslant \mathbb{E}\left[X_n \mid \mathcal{B}\right]$ for every $n \geqslant m$. The result follows.

(iii) By definition, for every Z which is $\mathcal{B}$-measurable bounded, we have

$$\mathbb{E}\left[|X_n - X| Z\right] = \mathbb{E}\big[\mathbb{E}\left[|X_n - X| \mid \mathcal{B}\right] Z\big].$$

But $|X_n - X| Z$ and $\mathbb{E}\left[|X_n - X| \mid \mathcal{B}\right] Z$ both form bounded sequences in L^1 (by $2Y|Z|$ for the first and $2\mathbb{E}\left[Y \mid \mathcal{B}\right]|Z|$ for the second), thus by the dominated convergence theorem we have

$$0 = \lim_n \mathbb{E}\left[|X_n - X| Z\right] = \lim_n \mathbb{E}\big[\mathbb{E}\left[|X_n - X| \mid \mathcal{B}\right] Z\big] = \mathbb{E}\big[\lim_n \mathbb{E}\left[|X_n - X| \mid \mathcal{B}\right] Z\big].$$

For the definition of the conditional expectation, this shows that $\lim_n \mathbb{E}\left[|X_n - X| \mid \mathcal{B}\right] = 0$.

□

8.5 Calculation of conditional expectations

Concretely, we have already see the case of discrete conditioning.

Proposition 8.22 (*Discrete case*)

Let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ and suppose that Y takes values in a countable space E. Then

$$\mathbb{E}\left[X \mid Y\right] = \phi(Y),$$

where ϕ is defined by $\phi(y) = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y=y)}$ for every $y \in E$ with $\mathbb{P}(Y=y) > 0$.

Proof We have already seen in Proposition 8.10, that for every Z $\sigma(Y)$-measurable and bounded, we have

$$\mathbb{E}[XZ] = \mathbb{E}[\phi(Y)Z],$$

which by definition of the conditional expectation, implies that $\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \sigma(Y)] = \phi(Y)$. □

Let us now look at the case of random variables with density.

Proposition 8.23 (*Density case*)

Let X and Y be two random variables such that the couple (X, Y) has a density function $f(x, y)$ in $\mathbb{R}^2$. Then $\mathbb{E}[h(X) \mid Y] = \phi(Y)$ with

$$\phi(y) = \frac{1}{\int_{\mathbb{R}} f(x,y)\,dx} \int_{\mathbb{R}} h(x) f(x,y)\,dx,$$

if $\int_{\mathbb{R}} f(x,y)\,dx > 0$ and $h(0)$ otherwise.

Proof We set $q(y) = \int_{\mathbb{R}} f(x,y)\,dx$. We take g bounded and Borelian, and we calcule

$$\mathbb{E}[h(X)g(Y)] = \int_{\mathbb{R}^2} h(x)g(y)f(x,y)\,dxdy = \int_{\mathbb{R}} \left(\frac{\int_{\mathbb{R}} h(x)f(x,y)\,dx}{q(y)}\right) g(y)q(y)\mathbf{1}_{\{q(y)>0\}}\,dy = \mathbb{E}[\phi(Y)g(Y)].$$

The result then follows by the definition of the conditional expectation. □

It turns out that the case of Gaussian vectors is significantly easier. In the general case, if we are given a random variable $X \in L^2$ and random variables $Y_1, \ldots, Y_n \in L^2$, then $\mathbb{E}[X \mid Y_1, \ldots, Y_n]$ is the orthogonal projection of X on $L^2(\sigma(Y_1, \ldots, Y_n))$. For Gaussian random vectors, this will translate in a geometric way since we will prove that it will correspond to the orthogonal projection on the (random) subspace generated by $Y_1, \ldots, Y_n$.

Theorem 8.24 (*Gaussian conditioning*)

Let $(Y_1, \ldots, Y_n, X)$ be a centered Gaussian vector. Then $\mathbb{E}[X \mid Y_1, \ldots, Y_n]$ is equal to the orthogonal projection of X on the vector space generated by $Y_1, \ldots, Y_n$, and thus we have

$$\mathbb{E}[X \mid Y_1, \ldots, Y_n] = \sum_{j=1}^{n} \gamma_j Y_j,$$

for some scalars $\gamma_1, \ldots, \gamma_n$.

Proof Let $Z = \sum_{j=1}^{n} \gamma_j Y_j$ be the orthogonal projection of X on the random subspace generated by $Y_1, \ldots, Y_n$. By definition, we have that $\mathbb{E}[(X - Z)Y_j] = 0$ or equivalently that $\mathrm{Cov}(X - Z, Y_j) = 0$ for every $j \in \{1, \ldots, n\}$. But $(Y_1, \ldots, Y_n, X - Z)$ is a centered Gaussian vector, then $X - Z$ is independent from the Y_j and thus we have

$$\mathbb{E}[(X - Z) \mid Y_1, \ldots, Y_n] = \mathbb{E}[X - Z] = 0.$$

Therefore, we deduce that

$$\mathbb{E}[X \mid Y_1, \ldots, Y_n] = \mathbb{E}[Z \mid Y_1, \ldots, Y_n].$$

But Z is $\sigma(Y_1,\dots,Y_n)$-measurable, thus we have $\mathbb{E}\left[Z \mid Y_1,\dots,Y_n\right] = Z$ which finishes the proof. □

Corollary 8.25

Let $(Y_1,\dots,Y_n,X)$ be a centered Gaussian vector. For every nonnegative Borelian function h, we have

$$\mathbb{E}\left[h(X) \mid Y_1,\dots,Y_n\right] = \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} h(x) \exp\Big(- \frac{\left(x - \mathbb{E}\left[X \mid Y_1,\dots,Y_n\right]\right)^2}{2\sigma^2} \Big)\, dx,$$

where $\sigma^2 = \mathbb{E}\left[(X - \mathbb{E}\left[X \mid Y_1,\dots,Y_n\right])^2\right]$.

Proof We set $W = X - Z$ with $Z = \mathbb{E}\left[X \mid Y_1,\dots,Y_n\right]$. By the previous theorem, W follows a normal distribution and W is independent of $Y_1,\dots,Y_n$. Therefore we have $W \sim \mathcal{N}(0,\sigma^2)$. We write

$$\mathbb{E}\left[h(X) \mid Y_1,\dots,Y_n\right] = \mathbb{E}\left[h(W+Z) \mid Y_1,\dots,Y_n\right].$$

By part (h) of Proposition 8.20, we have

$$\mathbb{E}\left[h(W+Z) \mid Y_1,\dots,Y_n\right] = \frac{1}{\sigma\sqrt{2\pi}} \int h(w+Z)\, e^{-\frac{w^2}{2\sigma^2}}\, dw$$

Thus, it is enough to operate a change of variables to finish the proof. □

Remark 8.26

The previous corollary suggests that the conditional distribution of X knowing $Y_1,\dots,Y_n$ is a normal distribution with mean $\mathbb{E}\left[X \mid Y_1,\dots,Y_n\right]$ and variance $\mathbb{E}\left[(X - \mathbb{E}\left[X \mid Y_1,\dots,Y_n\right])^2\right]$.

Chapter 9

Discrete Markov chains

9.1 Introduction and definitions

A random process models the evolution of a random system in time. We have already encountered such process when we considered a sequence of independent random variables; the evolution was made independently at each step. More generally, a discrete time random process is just a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ indexed by integer time, and one aims to study its dynamics i.e. how the system evolves in time. In this chapter, we will make the evolution depend on the current state but not on the past. This type of random process is referred to as Markov chain.

Definition 9.1 (*Markov chains*)

Let S be a discrete set (either finite or countably infinite). A discrete time stochastic process $(X_n)_{n\in\mathbb{N}}$ is called a Markov chain with state space S if its law of evolution is specified by the followings

- *An initial probability distribution ν on S.*
- *A one-step transition matrix $P = (p_{ij})_{i,j\in S}$ with $p_{ij} \in [0,1]$ satisfies $\sum_{j\in S} p_{ij} = 1$ for every $i \in S$.*

The law of evolution is given by

$$\mathbb{P}\big(X_0 = x_0,\, X_1 = x_1, \ldots,\, X_n = x_n\big) = \nu(x_0)p_{x_0x_1}\ldots p_{x_{n-1}x_n},$$

for all $x_0, \ldots, x_n \in S$.

Remark 9.2

- *Fix $i \in S$. All jumps from state i to any other state j happens with probability p_{ij}. Thus the random state where we land when we depart from i follows a probability distribution with probability mass function $(p_{ij})_{j\in S}$.*
- *Note that the future depends only on the present and not on the past. Indeed, one can easily check that*

  $$\mathbb{P}\big(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\big) = \mathbb{P}\big(X_n = x_n \mid X_{n-1} = x_{n-1}\big) = p_{x_{n-1}x_n},$$

  for all $x_0, \ldots, x_n \in S$ and $n \geqslant 1$. This is known as the Markov property.

- *Given $v \in S$, we abbreviate notation and write $\mathbb{P}_v(X_n = x)$ to mean $\mathbb{P}(X_n = x \mid X_0 = v)$.*

Example 9.3

1. *The simple random walk on Z. Let $S = \mathbb{Z}$, $\nu = \delta_a$ for some $a \in \mathbb{Z}$. And define $p_{i,i+1} = p$ and $p_{i,i-1} = 1 - p$ for some $p \in (0,1)$ and for every $i \in S$. That is, at each step toss a biased coin (head occurs with probability p), if you get head then move to the right, otherwise move to the left. Clearly, your next position depends only on your current one and not on the past.*

2. *Consider a graph on n vertices. Let S be the set of vertices and $\nu = \delta_v$ for some vertex v of the graph. Define for any two vertices u, w*

$$p_{u,w} = \frac{\mathbf{1}_{u\sim w}}{\deg(u)},$$

where $\deg(u)$ denotes the degree of u i.e. the number of vertices connected to it. In words, at a given step where the random walker is at vertex u, he picks uniformly at random one of the neighboring positions and moves to it.

When a system is evolving, we are led to natural questions such as the convergence to some equilibrium regime, or the time spent by the Markov chain at different states, the time it takes to move from a given state to another, etc. These are the type of questions we will consider in this chapter.

We have seen how to do a one step evolution, but the Markov property allows to iterate this. For instance considering a two-step transition, we have

$$\begin{aligned}\mathbb{P}\big(X_{n+2} = j \mid X_n = i\big) &= \sum_{k\in S} \mathbb{P}\big(X_{n+2} = j,\, X_{n+1} = k \mid X_n = i\big)\\ &= \sum_{k\in S} \mathbb{P}\big(X_{n+1} = k \mid X_n = i\big)\mathbb{P}\big(X_{n+2} = j \mid X_{n+1} = k,\, X_n = i\big)\\ &= \sum_{k\in S} p_{ik}p_{kj} = (P^2)_{ij}.\end{aligned}$$

More generally, we have the n-th transition given by

$$\mathbb{P}\big(X_{n+k} = j \mid X_k = i\big) = (P^n)_{ij},$$

and

$$\mathbb{P}(X_n = j) = (\nu P^n)_j,$$

where we identified the probability measure ν with the $1 \times |S|$ vector $\big(\nu(i)\big)_{i\in S}$.

Remark 9.4

An efficient way to calculate the n-th power of a matrix is through diagonalization. Indeed,

9.2 Strong Markov property

We have seen that the evolution of a Markov chain at a particular time depends only on that deterministic time and not on the past. What if we pick a time randomly in some way? As long as the time is properly measurable, we will able to say something.

Definition 9.5 (*Stopping time*)

Let $(X_n)_{n\in\mathbb{N}}$ be a stochastic process. A stopping time with respect to $(X_n)_{n\in\mathbb{N}}$ is a random variable τ such that for every $n \in \mathbb{N}$, the event $\{\tau = n\}$ belongs to the σ-algebra generated by $X_0, \ldots, X_n$.

Suppose that S is a discrete set and let $v \in S$. Then $\tau = \min\{n \geqslant 0 : X_n = v\}$ is a stopping time, called *hitting time* of state v.

Theorem 9.6 (*Strong Markov property*)

Let τ be a stopping time with values in $[0, \infty]$ adapted to the Markov chain $(X_n)_{n\in\mathbb{N}}$ with discrete state space S and transition matrix P. Then the Markov chain satisfies the strong Markov property which states that, conditioned on $\tau < \infty$ and $X_\tau = v_0$, the process $(X_{\tau+n})_{n\in\mathbb{N}}$ is a Markov chain started at v_0 with transition matrix P and independent of $(X_0, \ldots, X_T)$.

Proof Let $B \in \sigma(X_0, \ldots, X_\tau)$ i.e. $B \cap \{\tau = m\} \in \sigma(X_0, \ldots, X_m)$. By Markov property at time m we have

$$\begin{aligned} p_m &:= \mathbb{P}\big(B \cap \{\tau = m\} \cap \{X_\tau = v_0\} \cap \{X_\tau = v_0, \ldots, X_{\tau+n} = v_n\}\big) \\ &= \mathbb{P}_{v_0}\big(X_0 = v_0, \ldots, X_n = v_n\big)\mathbb{P}\big(B \cap \{\tau = m\} \cap \{X_\tau = v_0\}\big) \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathbb{P}\big(B \cap \{X_\tau = v_0, \ldots, X_{\tau+n} = v_n\} \mid T < \infty, X_\tau = v_0\big) &= \frac{\sum_{m\geqslant 0} p_m}{\sum_{m\geqslant 0} \mathbb{P}\big(\tau = m, X_\tau = v_0\big)} \\ &= \mathbb{P}_{v_0}\big(X_0 = v_0, \ldots, X_n = v_n\big)\mathbb{P}\big(B \mid \tau < \infty, X_\tau = v_0\big). \end{aligned}$$

□

The above property becomes increasingly useful when it is applied with a hitting time. Indeed, given $v \in S$, and taking $\tau_v = \min\{n \in \mathbb{N} : X_n = v\}$ then $\tau_v < \infty$ translates in $X_{\tau_v} = v$ and the formula above simplifies considerably. Let us record this in the following corollary.

Corollary 9.7

Let $v \in S$ and $\tau_v = \min\{n \in \mathbb{N} : X_n = v\}$. Conditioned on the event $\{\tau_v < \infty\}$, the process $(X_{\tau_v+n})_{n\geqslant 0}$ forms a Markov chain with initial state v and transition matrix P. Moreover, this chain is independent from the σ-algebra generated by τ_v.

9.3 Classification of states

Definition 9.8

We say that a state u is accessible from a state v if there exists $n \geqslant 0$ such that $(P^n)_{vu} > 0$ and we write $v \to u$.

Notice that we can represent these in graph language. Indeed, one can define a graph whose vertex set is S and where two vertices u, v are connected by a directed edge from u to v if $p_{uv} > 0$. Moreover u is accessible from v, if there exists some oriented path going from vertex v to u.

Proposition 9.9

The accessibility relation is reflexive and transitive.

Proof Exercise. □

Definition 9.10

We say that two states u and v communicate and we write $u \sim v$ if $u \to v$ and $v \to u$.

Proposition 9.11

The communication relation is an equivalence relation i.e. it is reflexive, symmetric and transitive.

The equivalence classes of this relation are the connected components of the graph. We often call them the irreducible classes or components. If C_1 and C_2 are two distinct classes, we could eventually transit from C_1 to C_2 but we could not make the return. However, all states inside the same class are connected.

When a class doesn't communicate with the outside, it is called *closed*. More precisely, if for every $u \in C$ with $u \to v$ we have $v \in C$ then C is closed. When there is only one communication class, that is, when all states are connected, we say that the chain is *irreducible*.

Example 9.12

1. *The simple random walk on $\mathbb{Z}$ is irreducible.*

2. *Consider the Markov chain on $S = \{0, 1, 2\}$ with transition matrix*

$$\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/4 & 1/4 \\ 0 & 1/3 & 2/3 \end{bmatrix}$$

   This chain is irreducible.

3. *Consider the Markov chain on $S = \{0, 1, 2, 3\}$ with transition matrix*

$$\begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

   This chain has three classes $\{0, 1\}, \{2\}, \{3\}$. We will see that the classes $\{0, 1\}$ and $\{3\}$ are recurrent while the class $\{2\}$ is transient.

Definition 9.13 (*Periodicity*)

Let $v \in S$. The period of v, denoted $d(v)$, is the greatest common divisor of all integers n such that $(P^n)_{vv} > 0$. If $d(v) = d \geqslant 2$, we say that v is periodic of period d. When $d(v) = 1$, we say that v is aperiodic and a chain is aperiodic if all its states are.

It turns out that states in the same communication class share the same period.

Proposition 9.14

Let $u \in S$ have finite period d and let $v \sim u$. Then v has period d.

Proof Since $u \to v$ and $v \to u$ then there exist two paths of length n and m respectively connecting u to v and v to u respectively. Let $\mathcal{C}$ be closed path starting and ending at v and with length ℓ. Then we can form a path of length $n + m + \ell$ which starts and ends at v. Moreover, we can form a path of length $n+m+2\ell$ as well. Since $d(v)$ divides both lengths by definition, then it divides the difference which is ℓ. Thus $d(v)$ divides $d(u)$ and doing the same procedure one shows that $d(u)$ divides $d(v)$. □

Example 9.15

1. *Consider the Markov chain with state space $S = \{0,1,2\}$, and with transition matrix*

$$\begin{bmatrix} 0 & 1 & 0 \\ p & 0 & 1-p \\ 1 & 0 & 0 \end{bmatrix}$$

We check that there is one class, and the chain is irreducible. All states communicate. Moreover, the two closed paths $0 \to 1 \to 0$ and $0 \to 1 \to 2 \to 0$ show that the 0 is aperiodic. Thus the chain is aperiodic.

2. *It is easy to check that the simple random walk on $\mathbb{Z}$ has period 2.*

3. *Consider the Markov chain on $S = \{0,1,2,3\}$ with transition matrix*

$$\begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

This chain is irreducible and the unique communication class is recurrent. There are exactly two paths $0 \to 2 \to 1 \to 0$ and $0 \to 3 \to 1 \to 0$ both of length 3. The class has period 3.

Proposition 9.16

If v is an aperiodic state, then there exists n_0 such that for any $n \geqslant n_0$ we have $P^n_{vv} > 0$.

Proof First note that if $P_{vv} > 0$ then there is nothing to prove. Now if $P_{vv} = 0$, then since v is aperiodic, there are two integers n_1, n_2 whose greatest common divisor is 1 and such that $P^{n_1}_{vv}, P^{n_2}_{vv} > 0$. Without loss of generality, suppose that $n_1 > n_2$. By Bezout theorem, there are $a, b \in \mathbb{N}$ such that $an_1 - bn_2 = 1$. Set $m_1 = an_1$, $m_2 = bn_2$ and take $n_0 = m_2^2$. For any $j \in \mathbb{N}$, write $j = km_2 + \ell$ and note that

$$P^{n_0+j}_{vv} = P^{m_2(m_2+k-\ell)+m_1\ell}_{vv} \geqslant P^{m_2(m_2+k-\ell)}_{vv} P^{m_1\ell}_{vv} \geqslant \left(P^{n_2}_{vv}\right)^{b(m_2+k-\ell)} \left(P^{n_1}_{vv}\right)^{a\ell} > 0.$$

□

Proposition 9.17

Suppose that the state space S is finite. If the chain is irreducible and aperiodic then there exists $n_0 \in \mathbb{N}$ such that for every $u, v \in S$ and every $n \geqslant n_0$ we have $P^n_{uv} > 0$.

Proof By the previous proposition, for every $i \in S$ there exists n_i such that for any $n \geqslant n_i$ we have $P^n_{ii} > 0$. Let $m = \max_{i \in S} n_i$ and note that for any $n \geqslant m$ and any $i \in S$ we have $P^n_{ii} > 0$. Now by irreducibility, we have that for every $i, j \in S$ there exists m_{ij} such that $P^{m_{ij}}_{ij} > 0$. Let

$M = \max_{ij}(m + m_{ij})$ and not that for every $n \geqslant M$ and every $i, j \in S$ we have

$$P_{ij}^{n} \geqslant P_{ii}^{n-m_{ij}} P_{ij}^{m_{ij}} > 0.$$

□

9.4 Recurrence and transience

Definition 9.18

We say that a state v is recurrent if, starting from v, the probability that the chain returns to v in a finite time is equal to one i.e. $\mathbb{P}_v(\tau_v < \infty) = 1$. Otherwise, we say that the state v is transient.

Remark 9.19

Recurrent states are further split into two categories: positive recurrent if $\mathbb{E}_v \tau_v < \infty$ and null recurrent if $\mathbb{E}_v \tau_v = \infty$. We won't deal with these notions in this course.

In other words, when v is recurrent then the chain will revisit it almost surely. However, when it is transient, there is a positive probability it won't return to it.

Proposition 9.20

If v is recurrent, then the number of returns to v (when the chain starts at v) is almost surely infinite (and in particular $\mathbb{E}_v N_v = \infty$). If v is transient, then the number of returns to v (when the chain starts at v) follows a geometric distribution and is therefore almost surely finite and has finite expectation.

Proof Denote by $N_v = \sum_{n \geqslant 0} \mathbf{1}_{\{X_n = v\}}$ be the number of visits to v. Let $k > 1$ and note that if $N_v = k$ then $\tau_v < \infty$. We write

$$\begin{aligned} \mathbb{P}_v\big(N_v = k\big) &= \mathbb{P}_v\big(N_v = k \mid \tau_v < \infty\big)\mathbb{P}_v(\tau_v < \infty) \\ &= \mathbb{P}_v\big(\sum_{n \geqslant 0} \mathbf{1}_{\{X_{\tau_v + n} = v\}} = k - 1 \mid \tau_v < \infty\big)\mathbb{P}(\tau_v < \infty) \end{aligned}$$

By the Strong Markov property, conditioned on $\tau_v < \infty$, $(X_{\tau_v + n})_{n \geqslant 0}$ has the same distribution as the markov chain started v. Thus, we have

$$\mathbb{P}_v\big(\sum_{n \geqslant 0} \mathbf{1}_{\{X_{\tau_v + n} = v\}} = k-1 \mid \tau_v < \infty\big) = \mathbb{P}_v\big(\sum_{n \geqslant 0} \mathbf{1}_{\{X_n = v\}} = k-1 \mid \tau_v < \infty\big) = \mathbb{P}_v\big(N_v = k-1 \mid \tau_v < \infty\big)$$

Therefore, we have

$$\mathbb{P}_v\big(N_v = k\big) = \mathbb{P}_v\big(N_v = k - 1\big)\mathbb{P}_v(\tau_v < \infty).$$

This shows that $\big(\mathbb{P}_v\big(N_v = k\big)\big)_{k \geqslant 1}$ form a geometric sequence.

Now if v is recurrent, then $\mathbb{P}(\tau_v < \infty) = 1$ and the sequence is constant. But $\sum_k \mathbb{P}_v\big(N_v = k\big) = \mathbb{P}(N_v < \infty)$, so it is a convergent series with constant terms. Thus, these terms should be zero and then we get that $\mathbb{P}_v\big(N_v = \infty\big) = 1$.

Now, if v is transient then $q = \mathbb{P}(\tau_v < \infty) < 1$, and the result follows. □

Remark 9.21

It follows from the above that if v is transient, then the expected number of returns to v is given by

$$\mathbb{E}_v N_v = \frac{1}{1 - \mathbb{P}_v(\tau_v < \infty)},$$

where we denoted by N_v the number of returns to v.

Corollary 9.22

A state v is recurrent if and only if $\sum_{n\geqslant 0} P_{v,v}^{(n)}$ is divergent. A state v is transient if and only if $\sum_{n\geqslant 0} P_{v,v}^{(n)}$ is convergent.

Proof It is enough to use monotone convergence to see that

$$\mathbb{E}N_v = \sum_{n\geqslant 0} P_{v,v}^{(n)}$$

and use the previous theorem to finish the proof. □

Proposition 9.23

Consider the simple random walk on $\mathbb{Z}^d$ started at 0. We have that 0 is recurrent if and only if $d \leqslant 2$.

Proof Let us first treat the case $d = 1$. We have

$$P_{0,0}{}^{(2n)} = \frac{\binom{2n}{n}}{2^{2n}} \sim \frac{1}{\sqrt{\pi n}}.$$

Since this series is divergent, then 0 is recurrent. More generally, in $\mathbb{Z}^d$, every "positive" move in any direction should be compensated by a corresponding negative move. Thus there should be n positive moves and n corresponding negative moves distributed across different directions. Denoting n_i the number of positive moves made in direction i, we have

$$\begin{aligned} P_{0,0}^{(2n)} &= \sum_{n_1+\ldots+n_d=n} \frac{\binom{2n}{n_1,n_1,n_2,n_2,\ldots,n_d,n_d}}{(2d)^{2n}} \\ &= \frac{\binom{2n}{n}}{2^{2n}} \sum_{n_1+\ldots+n_d=n} \frac{\binom{n}{n_1,n_2,\ldots,n_d}^2}{d^{2n}}. \end{aligned}$$

Note that if $d = 2$, we get $P_{0,0}^{(2n)} = \left(\frac{\binom{2n}{n}}{2^{2n}}\right)^2 \sim \frac{1}{\pi n}$. Since this series is divergent, then 0 is recurrent when $d = 2$. For $d \geqslant 3$, note that

$$P_{0,0}^{(2n)} \leqslant \frac{\binom{2n}{n}}{2^{2n}} \max_{n_1+\ldots+n_d=n} \frac{\binom{n}{n_1,n_2,\ldots,n_d}}{d^n} \sim \frac{1}{(\pi n)^{d/2}},$$

which is convergent. We deduce that 0 is transient for $d \geqslant 3$. □

Proposition 9.24

The recurrence property is a class property. That is, if a state u is recurrent and $u \sim v$ then v is *also recurrent.*

Proof As $u \sim v$ then $P_{u,v}^{(n_1)}, P_{v,u}^{(n_2)} > 0$ for some integers n_1, n_2. We deduce that

$$\sum_n P_{vv}^{(n_1+n_2+n)} \geqslant \sum_n P_{v,u}^{(n_2)} P_{uu}^{(n)} P_{uv}^{(n_1)} = \infty.$$

□

Therefore we can talk about recurrent and transient classes and decompose the state space accordingly.

Theorem 9.25 (*Decomposition of the state space*)

The state space can be partitioned into communication classes such that a non closed class is transient while a finite closed class is recurrent. In particular, when the state space is finite, the recurrent classes are the closed ones while the transient ones are the non closed. Moreover, a Markov chain on a finite state space has at least one recurrent class.

Proof Let C be a non closed class, then there exists $u \in C$ and $v \notin C$ such that $p_{uv} > 0$. On the other hand, since u and v are not in the same class then for any n we have $P_{vu}^{(n)} = 0$. Therefore,

$$\mathbb{P}_u(\tau_u = \infty) \geqslant p_{uv}\mathbb{P}_v(\tau_u = \infty) = p_{uv} > 0.$$

Thus u is transient and so is C.
We now consider a finite closed class C. Suppose that it is transient, and let $u \in C$. Since C is closed, then $\mathbb{P}_u(\sum_{v \in C} N_v = \infty) = 1$. Since C is finite, we deduce that $\mathbb{P}_u(\exists v \in C,\ N_v = \infty) = 1$. However, for any $v \in C$, since it is transient, we should have $\mathbb{P}_u(N_v = \infty) = 0$ which is a contradiction. □

Remark 9.26

When the state space is finite, if the chain starts at a recurrent state then the chain stays in the same class since it is closed. If the chain starts at a transient state then after an almost surely finite time it will exit the class and after some almost surely finite time it will reach a recurrent state and stays in it. Therefore, starting from a transient state, we reach in an almost surely finite time a recurrent state while starting from a recurrent state one cannot reach a transient state.

9.5 Stationary measures

Definition 9.27

Let ν be a probability measure on a countable state space S. We identify ν to a vector $(\nu_0, \nu_1, \ldots)$ with $\sum_i \nu_i = 1$. We say that ν is stationary or invariant with respect to a Markov chain with transition P if $\nu = \nu P$ i.e. for any $j \in S$ we have $\nu_j = \sum_{i \in S} \nu_i P_{ij}$.

Note that one can characterize a stationary measure by the fact that ν^t is an eigenvector of P^t associated with the eigenvalue 1. Since $\nu P^n = \nu$ then when started from the stationary measure, the corresponding Markov chain process is stationary i.e. the law of the n-th step is the same and is equal to ν.

Proposition 9.28

If ν is a stationary probability measure and i is a transient state, then $\nu(i) = 0$. In particular, if a *process only has transient states, then it doesn't admit a stationary measure.*

Proof Since ν is stationary then $\nu P^n = \nu$ for every n. Thus $\nu(i) = \sum_j \nu(j) P^n_{ji}$. Since i is transient, then P^n_{ji} goes to 0 as n goes to infinity. By dominated convergence, we get that $\nu(i) = 0$. The second part trivially follows. □

Let us first focus on the finite state space.

Theorem 9.29

Let $(X_n)_n$ be a Markov chain on a finite state space.

- *The chain admits at least one stationary measure.*
- *If the chain is irreducible, then there is a unique stationary measure ν given by $\nu(v) = \frac{1}{\mathbb{E}_v \tau_v}$.*
- *If in addition the chain is aperiodic, then P^n converges to the matrix whose rows are all equal to ν. In particular, for any initial distribution μ_0, X_n converges in distribution to ν.*

Proof Let us start by showing the existence of a stationary measure. We have seen that 1 is an eigenvalue of P and that the constant vector is the corresponding (right) eigenvector. Thus, 1 is also an eigenvalue of ${}^tP$ and our aim is to show the existence of an eigenvector whose coordinates are non-negative. Let x be an eigenvector of P associated with the eigenvalue 1. We will show that $y = (|x_i|)_i$ is also an eigenvector associated with the eigenvalue 1. Indeed, we have

$$\sum_j y_j P_{ji} - y_i = \sum_j |x_j| P_{ji} - |x_i| \geqslant |\sum_j x_j P_{ji}| - |x_i| = 0,$$

for every i. On the other hand, we have

$$\sum_i \big(\sum_j y_j P_{ji} - y_i \big) = 0.$$

Thus, we deduce that for every i we have $\sum_j y_j P_{ji} - y_i = 0$ and that y is indeed an eigenvector associated with the eigenvalue 1. It remains to normalize y in order to obtain the existence of a stationary probability measure.

We now suppose that P is irreducible. We will show that the eigenspace associated with the eigenvalue 1 is of dimension one which would imply the same for ${}^tP$. Let x be an eigenvector of P associated with the eigenvalue 1, and let $x_{i_0} = \max_i x_i$. Suppose that there exists j_0 such that $x_{j_0} < x_{i_0}$. Since P is irreducible then there exists n such that $P^n_{i_0 j_0} > 0$. Then since

$$x_{i_0} = P^n_{i_0 j_0} x_{j_0} + \sum_{j \neq j_0} P^n_{i_0 j} x_j,$$

we would get that

$$x_{i_0} < \frac{1}{1 - P^n_{i_0 j_0}} \sum_{j \neq j_0} P^n_{i_0 j} x_j \leqslant x_{i_0},$$

which is a contradiction. Therefore x is the constant vector.

Suppose now that the chain is irreducible and aperiodic. By Proposition 9.17, there exists n_0 such that P^{n_0} has all its entries positive. We will make use of the following consequence of Perron-Frobenius theorem

Theorem 9.30 (*Perron-Frobenius*)

Let A be a matrix whose entries are real positive numbers. Then there exists an eigenvalue of A of maximal modulus which is simple, and all other eigenvalues have modulus strictly less than the maximal one.

Using this, we deduce that P^{n_0} has a simple eigenvalue which has maximal modulus. It is easy to see that this eigenvalue is equal to one. Therefore, 1 is a simple eigenvalue of P as well and all other eigenvalues have modulus strictly less than one. By Jordan decomposition, we can write

$$P = Q \begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix} Q^{-1},$$

where A has all its eigenvalues of modulus strictly less than 1 and Q has its first column all equal to one since this is the eigenvector associated to the eigenvalue 1. Moreover, ν is the first row of Q^{-1} since it is the eigenvector of ${}^tP$ associated with the eigenvalue 1. Since A^n converge to 0 we deduce that P^n converge to

$$Q \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}$$

which is the matrix whose all rows are equal to ν. □

Remark 9.31

- *If P is symmetric then the eigenvectors of P and ${}^tP$ are the same, and we deduce that the uniform distribution is a stationary distribution.*
- *If the chain has only one recurrent class, then in a finite time it will reach the recurrent class and would act as an irreducible chain. Thus, the chain has a unique stationary measure supported on this recurrent class.*
- *If the chain has several recurrent classes, then it admits for each such class a unique stationary distribution supported on it. Therefore, any convex combination of these stationary measures is again a stationary measure. It turns out that these are the only stationary measures: the dimension of the eigenspace associated with the eigenvalue 1 is equal to the number of recurrent classes of the chain.*

Theorem 9.32

Let $(X_n)_n$ be an irreducible Markov chain on a finite state space S and we denote by ν its unique stationary measure. Then for any function f on S we have

$$\frac{1}{n}\sum_{i}^{n} f(X_i) \xrightarrow[n\to\infty]{a.s.} \int f\, d\nu.$$

Proof Fix $y \in S$ and let us first consider the function $f(x) = \mathbf{1}_{\{x=y\}}$. Therefore

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \frac{N_{y,n}}{n},$$

where $N_{y,n}$ is the number of visits to state y in an excursion of length n. First note that if y is transient, then since the number of visits to y is almost surely finite and that $\nu(y) = 0$ then there is nothing to prove. Therefore, we suppose that y is recurrent. Let τ_y^i, $i \geqslant 1$, be the (random) times of the i-th return of the chain to y. Note that by the strong Markov property, the random variables $(\tau_y^i - \tau_y^{i-1})_i$ are independent and identically distributed (if we suppose the chain starts at y, which we can since the chain is irreducible and y is recurrent). Therefore, by the strong law of large numbers we have

$$\frac{\tau_y^n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}\tau_y^1 = \frac{1}{\nu(y)}.$$

Using that $\{N_{y,n} \geqslant nt\} = \{\tau_y^{nt} \leqslant n\}$ for every $t < 1$, we can show that

$$\frac{N_{y,n}}{n} \xrightarrow[n\to\infty]{a.s.} \nu(y).$$

More generally, if f is arbitrary, note that

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \sum_{y\in S} \frac{N_{y,n}}{n} f(y),$$

and using the previous convergence we finish the proof. □