# Discrete-time Stochastic Processes

M1 Applied Mathematics and Statistics
Institut Polytechnique de Paris

Cyril Marzouk[1]

10th January 2024

[1]CMAP, École polytechnique. cyril.marzouk@polytechnique.edu

**What is this?** This document serves as a guide for the probability course in the first term of this master program. It contains mostly the mathematical details, and it should not be thought of as complete lecture notes with many examples and applications. Students are advised to use well-written books such as those cited below to either cover some missing prerequisites, see a different presentation from this one, or also to go further. Some sections will also not be covered in class and are included for the curious reader, they are denoted by a ($\star$).

This document is still improving, all remarks, including typos, are *very* welcome.

**Presentation & Prerequisites.** The goal of this course is to introduce the theory of stochastic processes in discrete time, namely sequences of random variables which are usually not independent, but rather in which the law at a given time depends on the past. The two main theories which constitute central objects in modern probability and statistics, both from a theoretical perspective as well as in applications, are *Markov chains* and *martingales*. They form the main content of this course.

This course is meant to be an advanced course in probability. Familiarity with basic measure theory and probability such as random variables, their law and expectation, independence, $L^p$ spaces, the different notion of convergences, Law of Large Numbers & Central Limit Theorem will be assumed. These notions are recalled in Chapter 1 and 2 and will not be covered in class: **this course starts with Chapter 3**. Some knowledge on Markov chains on a finite state-space is useful but not at all mandatory.

**Lectures session by session (prevision)**

*Part I: Markov Chains*

1) Introduction to Markov chains, transition matrices (Sections 3.1 & 3.2)
2) Random recursion and the strong Markov property (Sections 3.3 & 3.4)
3) Recurrence & transience (Sections 4.1 & 4.2)
4) Stationary measures (Sections 4.2 & 4.3)
5) Ergodic Theorem, aperiodicity (Sections 5.1 & 5.2)
6) Convergence to equilibrium (Section 5.2)
7) Monte–Carlo simulation (Section 5.3)
8) *Midterm exam* (up to session 6 included)

*Part II: Martingales*

9) Abstract conditional expectation, properties (Chapter 6)
10) Generalities on stochastic processes, filtrations, stopping times, martingales, stopping thm (Chapter 7 & Section 8.1 & 8.2)
11) Almost sure convergence, closed martingales, $L^1$ convergence (Section 9.1 & 9.2)
12) Maximal inequalities, $L^p$ convergence, the $L^2$ case (Sections 9.4 & 9.6)
13) Central Limit Theorems (Section 9.7)
14) Applications: Optimal Stopping (Section 8.5), Robins–Morro (Section 9.8)
15) *Final exam* (up to session 13 included)

**References.** Here are some books that can be useful in relation with this course, some of them inspired these notes. This is a personal list of references that I used as a student (especially the books by Durrett, by Williams, and Baldi–Mazliak–Priouret for the exercises) or to prepare this course. Feel free to look outside this list, the important point is to find one or more that you enjoy reading and find complementary to the lectures.

- The following references cover the basics of measure theory and probability, as well as the conditional expectation and martingale part of this course. They offer a complete course with also many exercises and examples.
  - Rick Durrett. *Probability: Theory and Examples.* Available at `https://services.math.duke.edu/~rtd/PTE/pte.html`.
    Comprehensive book that covers the prerequisites, the material of this course, and much more; many examples and exercises.
  - Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes.*
    Excellent as well, starts from the basics of probability, complemented by an exercise book listed below.
  - Allan Gut. *Probability A Graduate Course.*
    More focus on the background of probability, the last chapter covers martingale theory.
  - Jean Jacod and Philip Protter. *Probability Essentials.*
    Same remark, shorter book.
  - David Williams. *Probability with Martingales.*
    Short and straight to the point, excellent reference but no Markov chains.
- The following references only contain a short recap of the definitions and main results (mostly without proof), but are a great source of *solved* exercises.
  - Paolo Baldi, Laurent Mazliak, and Pierre Priouret. *Martingales and Markov Chains - Solved Exercises and Elements of Theory.* (also available in French)
    Content adapted to this course, no more, no less.
  - Geoffrey Grimmett and David Stirzaker. *One Thousand Exercises in Probability.*
    Complementary book to that in the previous list, covers many topics.
- Finally, the last references are entirely dedicated to Markov chains and extend way beyond the scope of this course.
  - James Norris. *Markov Chains.*
    Easy to read, very nice introduction to the topic; Chapter 2 and 3 relate to the course in the second term.
  - Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains.*
    More advanced than this course; for the interested students who wish to go further.
  - David Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times.*
    Rather disjoint from this course; for the interested students who wish to go further in another direction.

# Contents

# Part I

# Foundations of Probability

# Chapter 1

# Basics of Measure Theory & Integration (⋆)

The content of this chapter will not be discussed in class and is only here to recall some technical details on measure theory and integration with respect to a measure, which constitutes the foundation of probability.

**Contents**



We first introduce the basic definitions of $\sigma$-algebras and measures in Section 1.1, we present a technical result that we shall use a few times in the next chapters, see Theorem 1.1.13. In Section 1.2 we define the notion of measurable functions, then we construct the Lebesgue integral of nonnegative functions in Section 1.3, and of general functions in Section 1.4, where we derive key results that we use all the time (monotone convergence, dominated convergence, Fatou's lemma). Finally we discuss product measures in Section 1.5, which are the foundation of independence in probability.

## 1.1 Measured spaces

Throughout this section we let $S$ be a set.

**Definition 1.1.1.** A *$\sigma$-algebra* on $S$ is a collection $\Sigma$ of subsets of $S$ such that

(i) $S \in \Sigma$,

(ii) $A \in \Sigma \implies A^c \in \Sigma$,

(iii) $A_n \in \Sigma$ for all $n \geq 1 \implies \bigcup_{n\geq 1} A_n \in \Sigma$.

The pair $(S, \Sigma)$ is called a *measurable space*; elements of $\Sigma$ are said to be *measurable*.

**Remark 1.1.2.** We also have $\emptyset = S^c \in \Sigma$ and $A_n \in \Sigma$ for all $n \geq 1 \implies \bigcap_{n\geq 1} A_n = (\bigcup_{n\geq 1} A_n^c)^c \in \Sigma$.

**Exercise 1.1.3.** The intersection of any collection of $\sigma$-algebras on $S$ is a $\sigma$-algebra.

**Definition 1.1.4.** Let $\mathscr{C}$ be a collection of subsets of $S$, then we let

$$\sigma(\mathscr{C}) = \bigcap \{\Sigma : \Sigma \text{ is a } \sigma\text{-algebra and } \Sigma \supset \mathscr{C}\}$$

denote the smallest $\sigma$-algebra that contains $\mathscr{C}$. It is called the $\sigma$-algebra *generated by* $\mathscr{C}$.

**Remark 1.1.5.** In general, if $(\mathcal{F}_n)_{n\geq 1}$ are $\sigma$-algebras, then $\bigcup_n \mathcal{F}_n$ is not, so we instead consider

$$\sigma(\mathcal{F}_n, n \geq 1) = \sigma\Big(\bigcup_n \mathcal{F}_n\Big).$$

**Exercise 1.1.6.** The $\sigma$-algebra generated by the singletons is

$$\sigma(\{\{\omega\} : \omega \in S\}) = \{A \subset S : A \text{ or } A^c \text{ is at most countable}\}.$$

**Example 1.1.7.**

- If $S$ is at most countable, we usually consider the $\sigma$-algebra of all subsets of $S$, which is generated by the singletons.
- In a topological space $(E, \mathcal{O})$, we usually consider the *Borel* $\sigma$-algebra $\mathcal{B}(E) = \sigma(\mathcal{O})$ generated by the open sets. In $\mathbb{R}^d$, we have

$$\mathcal{B}(\mathbb{R}^d) = \sigma\Big(\Big\{\prod_{i=1}^d (a_i, b_i) : a_i < b_i \text{ for all } 1 \leq i \leq d\Big\}\Big).$$

**Exercise 1.1.8.** In $\mathbb{R}$, we have

$$\mathcal{B}(\mathbb{R}) = \sigma(\{\dagger a, b\dagger : a < b\}) = \sigma(\{\dagger a, \infty) : a \in \mathbb{R}\}) = \sigma(\{(-\infty, b\dagger : b \in \mathbb{R}\}),$$

for any choice $\dagger = ($ or $\dagger = [$ on the left and $\dagger =)$ or $\dagger =]$ on the right. A similar property holds in $\mathbb{R}^d$ with products of intervals.

**Exercise 1.1.9.** Let $(\mathcal{F}_n)_{n\geq 1}$ be $\sigma$-algebras, then $\sigma(\mathcal{F}_n, n \geq 1)$ is also generated by the intersections of elements in each $\mathcal{F}_n$, namely:

$$\sigma(\mathcal{F}_n, n \geq 1) = \sigma\Big(\bigcup_{I\subset\mathbb{N} \text{ finite}} \Big\{\bigcap_{i\in I} A_i; A_i \in \mathcal{F}_i\Big\}\Big).$$

**Definition 1.1.10.** A *measure* $\mu$ on $(S, \Sigma)$ is a function $\mu : \Sigma \longrightarrow [0, \infty]$ such that

(i) $\mu(\emptyset) = 0$,

(ii) If $A_n \in \Sigma$ for all $n \geq 1$ are disjoint, then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.

The triple $(S, \Sigma, \mu)$ is called a *measured space*. The measure is said to be:

- *$\sigma$-finite* if there exists a countable collection $(A_n)_n$ of elements of $\Sigma$ such that $S = \bigcup_n A_n$ and $\mu(A_n) < \infty$ for all $n$,
- *finite* if $\mu(S) < \infty$,
- a *probability* if $\mu(S) = 1$.

**Remark 1.1.11.** We will not encounter non-$\sigma$-finite measures, but they do exist (take e.g. the cardinal of a set in $\mathbb{R}$). Also in other contexts, one sometimes considers signed measures, taking values in $\mathbb{R}$, or complex measures, in $\mathbb{C}$.

**Notation.** For $A \in \Sigma$, we say that $A$ holds $\mu$-a.e. (for 'almost everywhere') when $\mu(A^c) = 0$. If $\mu$ is a probability, this means $\mu(A) = 1$ and we say instead that $A$ holds $\mu$-a.s. (for 'almost surely').

The following simple properties of measures are used all the time.

**Lemma 1.1.12** (Key properties)**.** *Let $(S, \Sigma, \mu)$ be a $\sigma$-finite measured space and let $A_n \in \Sigma$ for all $n \geq 1$.*

*(i) If $(A_n)_n$ is increasing in that $A_n \subset A_{n+1}$, then $\mu(A_n) \uparrow \mu(\bigcup_n A_n)$.*

*(ii)* *If $(A_n)_n$ is decreasing in that $A_n \supset A_{n+1}$, and if there exists $k \geq 1$ such that $\mu(A_k) < \infty$, then $\mu(A_n) \downarrow \mu(\bigcap_n A_n)$.*

*(iii)* *In any case, $\mu(\bigcup_n A_n) \leq \sum_n \mu(A_n)$.*

A consequence of the last item is that

$$\mu(A_n) = 0 \text{ for any fixed } n \implies \mu\Big(\bigcup_n A_n\Big) = 0.$$

In particular if $\mu$ is a probability, then taking the complement we get

$$\mu(A_n) = 1 \text{ for any fixed } n \implies \mu\Big(\bigcap_n A_n\Big) = 1.$$

This means that, in a *countable* collection, if each event occurs with probability one (we say that it occurs *almost surely*), then they all occur almost surely simultaneously! Using arguments such as the density of the rational numbers in $\mathbb{R}$ and monotonicity or continuity arguments, this can sometimes (depending on the context) be extended to uncountable collections of events, which makes it a very powerful tool.

### 1.1.1 The $\pi - \lambda$ lemma

A $\sigma$-algebra can be complicated and very often, we aim at considering only simpler subsets. For example, in Exercise 1.1.8, one prefers to work only with the intervals $(a, b)$ or the intervals $(-\infty, x]$ instead of all the Borel sets of $\mathbb{R}$. The question we need to answer is: Given a $\sigma$-algebra $\Sigma$ on a set $S$ and a collection $\mathscr{C}$ of subsets all in $\Sigma$, when is it to sufficient to prove that a property holds for any element of $\mathscr{C}$ to ensure that it holds for any element of $\Sigma$?

**Theorem 1.1.13.** *Let $\mathscr{C}$ be a collection of subsets and let $\mu$ and $\nu$ be two measures on $(S, \sigma(\mathscr{C}))$ Suppose that $\mu(A) = \nu(A)$ for all $A \in \mathscr{C}$ and that for all $A, B \in \mathscr{C}$, we have $A \cap B \in \mathscr{C}$. Assume that:*

*(i)* *either $\mu(S) = \nu(S) < \infty$,*

*(ii)* *or there exists an increasing sequence $(A_n)_n$ of subsets of $\mathscr{C}$ such that $\bigcup_n A_n = S$ and $\mu(A_n) = \nu(A_n) < \infty$ for all $n$.*

*Then $\mu(A) = \nu(A)$ for all $A \in \sigma(\mathscr{C})$.*

In particular two probability measures that agree on such a collection $\mathscr{C}$ agree more generally on the $\sigma$-algebra that it generates. As an immediate application, in probabilistic words, we deduce that two random variables with the same distribution function have the same law.

**Corollary 1.1.14.** *Two probability measures on $\mathbb{R}$ that agree on any interval $(a, b)$, or on any interval $(-\infty, x]$ must agree on $\mathscr{B}(\mathbb{R})$.*

Let us prove Theorem 1.1.13.

**Definition 1.1.15.** A collection $\mathscr{C}$ of subsets of $S$ is called a *$\lambda$-system* when:

(i) $S \in \mathscr{C}$,

(ii) If $A, B \in \mathscr{C}$ and $A \subset B$, then $B \setminus A \in \mathscr{C}$,

(iii) If $A_n \in \mathscr{C}$ and $A_n \subset A_{n+1}$ for all $n$, then $\bigcup_n A_n \in \mathscr{C}$.

A collection $\mathscr{C}$ of subsets of $S$ is called a *$\pi$-system* when it is stable under finite intersections, namely for all $A, B \in \mathscr{C}$, we have $A \cap B \in \mathscr{C}$.

**Remark 1.1.16.** A $\lambda$-*system* is also sometimes called *d-system* instead; one also finds the name of *monotone* class, although the latter might also refer to something else depending on the author.

**Lemma 1.1.17.** *A collection $\mathscr{C}$ of subsets of $S$ is a $\sigma$-algebra if and only if it is both a $\lambda$-system and a $\pi$-system.*

*Proof.* Suppose that $\mathscr{C}$ is a $\sigma$-algebra. Then indeed it is a $\pi$-system, it is even stable under countable intersections. It also clearly satisfies the first and last property of a $\lambda$-system; for the second one, if $A, B \in \mathscr{C}$ with $A \subset B$, then $B \setminus A = B \cap A^c \in \mathscr{C}$,

Suppose that $\mathscr{C}$ is both a $\pi$-system and a $\lambda$-system. First for every $A \in \mathscr{C}$, we have $A^c = S \setminus A \in \mathscr{C}$. Next let $A_n \in \mathscr{C}$ for all $n$, then for any $n \geq k \geq 1$, we have $A_k^c \in \mathscr{C}$ and so $\bigcap_{k\leq n} A_k^c \in \mathscr{C}$ since it is a $\pi$-system. Hence $B_n = \bigcup_{k\leq n} A_k = (\bigcap_{k\leq n} A_k^c)^c \in \mathscr{C}$. Now $B_n \in \mathscr{C}$ and $B_n \subset B_{n+1}$ for all $n$, so $\bigcup_n A_n = \bigcup_n B_n \in \mathscr{C}$ since it is a $\lambda$-system. □

As for $\sigma$-algebras the intersection of $\lambda$-systems is always a $\lambda$-system so one can define for any collection $\mathscr{C}$ of subsets of $S$

$$\Lambda(\mathscr{C}) = \bigcap \{\Lambda : \Lambda \text{ is a } \lambda\text{-system and } \Lambda \supset \mathscr{C}\}$$

the $\lambda$-system *generated by* $\mathscr{C}$.

**Lemma 1.1.18** ($\pi - \lambda$ Lemma)**.** *If $\mathscr{C}$ is a $\pi$-system, then so is $\Lambda(\mathscr{C})$. The latter is therefore a $\sigma$-algebra, and actually $\Lambda(\mathscr{C}) = \sigma(\mathscr{C})$.*

*Proof.* Let us prove that $\Lambda(\mathscr{C})$ is a $\pi$-system. Fix $A \in \mathscr{C}$ (beware, not in $\Lambda(\mathscr{C})$) and define

$$\mathscr{L} = \{B \in \Lambda(\mathscr{C}) : A \cap B \in \Lambda(\mathscr{C})\}.$$

Since $A \in \mathscr{C}$ which is a $\pi$-system, then $\mathscr{C} \subset \mathscr{L}$. Further,

(i) $A \cap S = A \in \Lambda(\mathscr{C})$, so $S \in \mathscr{L}$,

(ii) If $B, C \in \mathscr{L}$ and $B \subset C$, then $A \cap (C \setminus B) = (A \cap C) \setminus (A \cap B) \in \Lambda(\mathscr{C})$,

(iii) If $B_n \in \mathscr{L}$ and $B_n \subset B_{n+1}$ for all $n$, then $A \cap (\bigcup_n B_n) = \bigcup_n (A \cap B_n) \in \Lambda(\mathscr{C})$.

Thus $\mathscr{L}$ is a $\lambda$-system, and since it contains $\mathscr{C}$, then it contains $\Lambda(\mathscr{C})$, i.e. for every $A \in \mathscr{C}$ and every $B \in \Lambda(\mathscr{C})$ we have

$$A \cap B \in \Lambda(\mathscr{C}).$$

Then the exact same reasoning with now $A \in \Lambda(\mathscr{C})$ instead shows that $\{B \in \Lambda(\mathscr{C}) : A \cap B \in \Lambda(\mathscr{C})\}$ is also a $\lambda$-system which contains $\mathscr{C}$ and so $\Lambda(\mathscr{C})$, i.e. for every $A \in \Lambda(\mathscr{C})$ and every $B \in \Lambda(\mathscr{C})$ we have

$$A \cap B \in \Lambda(\mathscr{C}),$$

that is, $\Lambda(\mathscr{C})$ is a $\pi$-system.

Combined with the previous lemma, we infer that $\Lambda(\mathscr{C})$ is a $\sigma$-algebra. Since it contains $\mathscr{C}$, then it also contains the smallest such $\sigma$-algebra, namely $\Lambda(\mathscr{C}) \subset \sigma(\mathscr{C})$. Similarly, $\sigma(\mathscr{C})$ is a $\lambda$-system which contains $\mathscr{C}$, so it also contains the smallest such $\lambda$-system, namely $\sigma(\mathscr{C}) \subset \Lambda(\mathscr{C})$ and this concludes the proof. □

The proof of Theorem 1.1.13 follows easily.

*Proof of Theorem 1.1.13.* (i) Let $\Lambda = \{A \in \sigma(\mathscr{C}) : \mu(A) = \nu(A)\}$. One can check that it is a $\lambda$-system that contains $\mathscr{C}$ so it contains $\Lambda(\mathscr{C}) = \sigma(\mathscr{C})$ by Lemma 1.1.18.

(ii) For every $n \geq 1$ and every $B \in \sigma(\mathscr{C})$, let $\mu_n(B) = \mu(B \cap A_n)$ and $\nu_n(B) = \nu(B \cap A_n)$. According to the first item for every $B \in \sigma(\mathscr{C})$, we have $\mu_n(B) = \nu_n(B)$ for every $n \geq 1$ and so by monotonicity,

$$\mu(B) = \uparrow \lim_{n\to\infty} \mu_n(B) = \uparrow \lim_{n\to\infty} \nu_n(B) = \nu(B),$$

which completes the proof. □

## 1.2 Measurable functions

**Definition 1.2.1.** Given to measurable spaces $(S, \Sigma)$ and $(E, \mathscr{E})$, a function $f : S \longrightarrow E$ is *measurable* if for every $B \in \mathscr{E}$ the set $f^{-1}(B) = \{\omega \in S : f(\omega) \in B\}$ belongs to $\Sigma$.

We shall also denote by $\{f \in B\}$ the subset $f^{-1}(B)$.

**Remark 1.2.2.** In our typical use, the space $(E, \mathscr{E})$ will be fixed, typically $(E, \mathscr{E}) = (\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, or a countable set, as well as the set $S = \Omega$ (unspecified), and we will consider several $\sigma$-algebras on $S$ so we shall insist on which one and write $f \textcircled{m} \Sigma$ to mean that $f$ is measurable for $\Sigma$.

**Exercise 1.2.3.** For any function $f$ from a set $S$ to another one $E$ and for any collection of subsets $(A_i)_{i\in I}$ of $E$, we have

$$f^{-1}(A_i^c) = \left(f^{-1}(A_i^c)\right)^c \qquad \text{and} \qquad f^{-1}\Big(\bigcup_{i\in I} A_i\Big) = \bigcup_{i\in I} f^{-1}(A_i).$$

The next lemma shows that it suffices to check the measurability on a collection of subsets that generates the $\sigma$-algebra $\mathscr{E}$.

**Lemma 1.2.4.** *Let $\mathscr{C}$ be a collection of subsets of $E$ and let $f : S \longrightarrow E$ be such that $f^{-1}(B) \in \Sigma$ for any $B \in \mathscr{C}$. Then $f^{-1}(B) \in \Sigma$ for any $B \in \sigma(\mathscr{C})$.*

*Proof.* Let $\mathscr{E} = \{B \in \sigma(\mathscr{C}) : f^{-1}(B) \in \Sigma\}$. Then it contains $\mathscr{C}$ by assumption and the exercise shows that it is a $\sigma$-algebra, thus it must contain $\sigma(\mathscr{C})$. □

**Example 1.2.5.** An example of application of this lemma is when $(E, \mathcal{O})$ is a topological space and $\mathscr{E} = \mathscr{B}(E) = \sigma(\mathcal{O})$ is the Borel $\sigma$-algebra. Then it suffices to check that $f^{-1}(O) \in \Sigma$ for any open set $O$. In the case $E = \mathbb{R}^d$, recall that

$$\mathscr{B}(\mathbb{R}^d) = \sigma\Big(\Big\{\prod_{i=1}^d (a_i, b_i) : a_i < b_i \text{ for all } 1 \le i \le d\Big\}\Big) = \sigma\Big(\Big\{\prod_{i=1}^d (-\infty, x_i] : x_i \in \mathbb{R}\Big\}\Big),$$

so it suffices to look at one of these two types of sets.

As an important example, if $E$ and $F$ are both topological spaces equipped with their Borel $\sigma$-algebra $\mathscr{B}(E)$ and $\mathscr{B}(F)$, then a continuous function from $E \longrightarrow F$ is automatically measurable.

**Exercise 1.2.6.** Take $(E, \mathscr{E}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, then measurability of functions is preserved by summation, products, multiplication by constants, composition, limit (lim sup and lim inf), supremum and infimum, also $\{s \in S : (f_n(s))_n \text{ converges}\} \in \Sigma$ and $\{s \in S : f \text{ is continuous at } s\} \in \Sigma$ if $\Sigma$ is the Borel $\sigma$-algebra on $S$.

Recall that we usually work with a fixed image space $(E, \mathscr{E})$ and different $\sigma$-algebras on the starting space $S$. A function may be measurable for some $\sigma$-algebras on $S$ and not for other ones.

**Definition 1.2.7.** Let $(f_i)_{i\in I}$ be a collection of functions from $S \longrightarrow E$. Define

$$\sigma(f_i : i \in I) = \sigma(\{f_i^{-1}(B) : i \in I, B \in \mathscr{E}\}),$$

the smallest $\sigma$-algebra on $S$ such that all functions $f_i$ are measurable.

The following important result (extensively used in the sequel) characterises measurable functions with respect to the $\sigma$-algebra generated by another function.

**Lemma 1.2.8.** *Let $f : S \longrightarrow E$ be measurable. Then a function $g : S \longrightarrow \mathbb{R}$ is measurable for $\sigma(f)$ if and only if there exists a measurable function $h : E \longrightarrow \mathbb{R}$ such that $g = h \circ f$.*

*Proof.* First, if $g = h \circ f$ with $h$ measurable, then for any $B \in \mathscr{B}(\mathbb{R})$, we have $h^{-1}(B) \in \mathscr{E}$ and thus $g^{-1}(B) = f^{-1}(h^{-1}(B)) \in \sigma(f)$. Let us prove the direct implication. Let $g : S \longrightarrow \mathbb{R}$ be measurable for $\sigma(f)$ and suppose first that $g$ only takes finitely many values so it takes the form $g = \sum_{k=1}^{K} a_k \mathbb{1}_{A_k}$ where $K \geq 1$, $a_k \in \mathbb{R}$, and $A_k \in \sigma(f)$ for every $k$. For each $k$, let $B_k \in \mathscr{E}$ be such that $A_k = f^{-1}(B_k)$, then $\mathbb{1}_{A_k} = \mathbb{1}_{B_k} \circ f$. Define then

$$h := \sum_{k=1}^{K} a_k \mathbb{1}_{B_k}$$

which is indeed $\mathscr{E} \longrightarrow \mathscr{B}(\mathbb{R})$-measurable, and $g = h \circ f$. If $g \geq 0$ can take infinitely many values, then we can write it as the limit of functions $g_n$ which only take finitely many values, e.g. explicitly:

$$g_n = \sum_{i=0}^{n2^n - 1} \frac{i}{2^n} \mathbb{1}_{i \leq 2^n g < i+1} + n \mathbb{1}_{g \geq n}.$$

Since $g$ is $\sigma(f)$-measurable, then each set $\{i \leq 2^n g < i+1\} = g^{-1}([2^{-n}i, 2^{-n}(i+1)) \in \sigma(f)$ and $\{g \geq n\} = g^{-1}([n, \infty)) \in \sigma(f)$ as well, so in turn each $g_n$ is $\sigma(f)$-measurable. By the previous case, they take the form $g_n = h_n \circ f$ with $h_n : E \longrightarrow \mathbb{R}$ measurable. Define then for every $x \in E$

$$h(x) = \lim_{n \longrightarrow \infty} h_n(x) \text{ if this limit exists} \qquad \text{and} \qquad h(x) = 0 \text{ otherwise,}$$

which is measurable by the previous exercise. Since for every $s \in S$ we have $g(s) = \lim_n g_n(s) = \lim_n h_n(f(s))$, then the sequence $h_n$ converges at the point $f(s)$ so finally $g(s) = h(f(s))$ and the proof is complete. □

Note that the converse implication is clear by composition of measurable functions. The representation given by the direct implication is very useful. This results extends to the $\sigma$-algebra generated by finitely many functions $f_1, \dots, f_n$, in which case $g$ takes the form $h(f_1, \dots, f_n)$.

## 1.3 Integration of nonnegative functions

Let us fix throughout this section a $\sigma$-finite measured space $(S, \Sigma, \mu)$. All functions considered here on $S$ are real-valued and measurable. We make the convention that $0 \times \infty = 0$.

**Definition 1.3.1.** A nonnegative function $f$ is called *simple* if it takes only finitely many values, in which case it can be written as $f = \sum_{i=1}^{k} a_i \mathbb{1}_{A_i}$ for some $a_i \in \mathbb{R}_+$ and $A_i \in \Sigma$. We then define for such a function $f$:

$$\mu(f) = \sum_{i=1}^{k} a_i \mu(A_i) \in [0, \infty].$$

**Notation.** We shall also write $\mu(f)$ as $\int f \, \mathrm{d}\mu$ or $\int_S f(x) \mu(\mathrm{d}x)$. Also for $A \in \Sigma$, we write $\int_A f \, \mathrm{d}\mu$ or $\int_A f(x) \mu(\mathrm{d}x)$ for $\mu(f \mathbb{1}_A)$.

**Remark 1.3.2.** There are several choices of decompositions of a simple function in such a form but we can always fix one by requiring that $a_i < a_{i+1}$.

**Exercise 1.3.3.** Prove that this definition of $\mu(f)$ does not depend on the decomposition of $f$. Prove also that it has the usual properties of linearity and monotonicity, namely that for two simple functions $f$ and $g$ and two positive real numbers $a$ and $b$, we have

$$\mu(af + bg) = a\mu(f) + b\mu(g) \qquad \text{and} \qquad f \leq g \implies \mu(f) \leq \mu(g).$$

**Definition 1.3.4.** For any nonnegative (measurable) function, we set

$$\mu(f) = \sup\{\mu(g) : g \text{ simple and such that } g \leq f\} \in [0, \infty].$$

This preserves the monotonicity property. We also have the following result used all the time.

**Lemma 1.3.5.** *If $f \geq 0$ and $\mu(f) = 0$, then $\mu(f > 0) = \mu(\{x \in S : f(x) > 0\}) = 0$.*

*Proof.* For every $n \geq 1$, we have $f \geq n^{-1} \mathbb{1}_{f \geq n^{-1}}$ which is a simple function so

$$0 = \mu(f) \geq \mu(n^{-1} \mathbb{1}_{f \geq n^{-1}}) = n^{-1} \mu(f \geq n^{-1}).$$

Thus $0 = \mu(f \geq n^{-1}) \uparrow \mu(f > 0)$ since the sequence of sets is increasing. □

The building block of this integration theory is the following result.

**Theorem 1.3.6** (Monotone convergence)**.** *Let $(f_n)_{n \geq 1}$ be nonnegative measurable functions such that $f_n \leq f_{n+1}$ for all $n \geq 1$. Then*

$$\mu(\uparrow \lim_{n \to \infty} f_n) = \uparrow \lim_{n \to \infty} \mu(f_n) \in [0, \infty].$$

*Proof.* Let $f = \uparrow \lim_n f_n$. By monotonicity, $\mu(f) \geq \mu(f_n)$ for all $n$ and thus $\mu(f) \geq \uparrow \lim_n \mu(f_n)$. For the converse inequality, let $0 \leq g \leq f$ be a simple function and let $c \in [0, 1)$. Write $g = \sum_{i=1}^k a_i \mathbb{1}_{A_i}$, then by monotonicity,

$$\mu(f_n) \geq \mu(f_n \mathbb{1}_{f_n \geq cg}) \geq c\, \mu(g \mathbb{1}_{f_n \geq cg}) = c \sum_{i=1}^k a_i \mu(A_i \cap \{f_n \geq cg\}).$$

Now observe that the sets $\{f_n \geq cg\}$ are increasing and since $c < 1$, then $\bigcup_n \{f_n \geq cg\} = S$. Thus $A_i \cap \{f_n \geq cg\} \uparrow A_i$ and so

$$\uparrow \lim_{n \to \infty} \mu(f_n) \geq \uparrow \lim_{n \to \infty} c \sum_{i=1}^k a_i \mu(A_i \cap \{f_n \geq cg\}) = c \sum_{i=1}^k a_i \mu(A_i) = c\, \mu(g).$$

Now let $c \uparrow 1$ to get $\uparrow \lim_{n \to \infty} \mu(f_n) \geq \mu(g)$ for all simple function $g \leq f$ and thus finally $\uparrow \lim_{n \to \infty} \mu(f_n) \geq \mu(f)$. □

Using this theorem, we can construct explicitly for any given measurable nonnegative function $f$ a sequence of simple functions whose integral converges to that of $f$.

**Corollary 1.3.7.** *Let $f$ be measurable nonnegative and for every $n \geq 1$ define*

$$f_n = \sum_{i=0}^{n2^n - 1} \frac{i}{2^n} \mathbb{1}_{i \leq 2^n f < i+1} + n \mathbb{1}_{f \geq n}.$$

*Then $f_n \uparrow f$ so $\mu(f_n) \uparrow \mu(f)$.*

This allows to transfer properties of the integral of simple functions to general nonnegative functions, such as the linearity.

**Exercise 1.3.8.** Let $f$ and $g$ be two nonnegative measurable functions. Prove that

(i) For every $a, b > 0$, we have $\mu(af + bg) = a\mu(f) + b\mu(g)$.

(ii) $\mu(f) < \infty \implies \mu(f = \infty) = 0$.

(iii) $\mu(f) = 0 \iff \mu(f > 0) = 0$.

(iv) $\mu(f \neq g) = 0 \implies \mu(f) = \mu(g)$.

**Remark 1.3.9.** Thanks to the last point of the exercise, we can slightly relax the assumption of the monotone convergence theorem by requiring that the monotonicity assumption only holds $\mu$-a.e. in the sense that the set $A = \{s \in S : f_n(s) \leq f_{n+1}(s) \text{ for all } n\}$ has $\mu(A^c) = 0$. For definiteness, we then set $\uparrow \lim_n f_n(s) = 0$ for $s \in A^c$. Indeed, we can apply the theorem in its previous form to the functions $f_n \mathbb{1}_A$ to deduce that $\mu(\uparrow \lim_n f_n \mathbb{1}_A) = \uparrow \lim_n \mu(f_n \mathbb{1}_A)$ and note that for any $n$, we have $\mu(f_n) = \mu(f_n \mathbb{1}_A)$ and further $\mu(\uparrow \lim_n f_n \mathbb{1}_A) = \mu(\uparrow \lim_n f_n)$. All the next results can be extended in this way.

**Lemma 1.3.10.** *Let $f$ be a nonnegative measurable function and define for every $A \in \Sigma$:*

$$\nu(A) = \mu(f\,\mathbb{1}_A).$$

*Then $\nu$ is a measure and $f$ is called the* density *of $\nu$ with respect to $\mu$. The function $f$ is unique $\mu$-a.e.*

*Proof.* For $A = \emptyset$, we have $f\,\mathbb{1}_\emptyset = 0$ so $\nu(\emptyset) = 0$. Let $(A_n)_n$ be disjoint measurable sets, and write $f_n = f\,\mathbb{1}_{A_n}$, then linearity and monotone convergence combined justify the following identity:

$$\int \sum_{n\geq 1} f_n \,\mathrm{d}\mu = \int \uparrow \lim_{N\to\infty} \sum_{n\leq N} f_n \,\mathrm{d}\mu = \uparrow \lim_{N\to\infty} \int \sum_{n\leq N} f_n \,\mathrm{d}\mu = \uparrow \lim_{N\to\infty} \sum_{n\leq N} \int f_n \,\mathrm{d}\mu = \sum_{n\geq 1} \int f_n \,\mathrm{d}\mu.$$

Note that $\int f_n \,\mathrm{d}\mu = \nu(A_n)$ whereas $\int \sum_{n\geq 1} f_n \,\mathrm{d}\mu = \int f\,\mathbb{1}_{\bigcup_n A_n} \,\mathrm{d}\mu = \nu(A)$ so we have prove the $\sigma$-additivity and $\nu$ is indeed a measure.

If $g \geq 0$ is another function such that $\mu(f\,\mathbb{1}_A) = \mu(g\,\mathbb{1}_A)$ for every $A \in \Sigma$, then for $A = \{f > g\}$, we have by linearity $0 \leq \mu((f-g)\,\mathbb{1}_{f>g}) = \mu(f\,\mathbb{1}_{f>g}) - \mu(g\,\mathbb{1}_{f>g}) = 0$, hence $(f-g)\,\mathbb{1}_{f>g} = 0$ $\mu$-a.e. which means that $f \leq g$ $\mu$-a.e. By exchanging $f$ and $g$ we also have that $f \geq g$ $\mu$-a.e. so $f = g$ $\mu$-a.e. □

Note that if $\mu(A) = 0$, then $f\,\mathbb{1}_A = 0$ $\mu$-a.e. and so $\nu(A) = 0$; in such a case, we say that $\nu$ is *absolutely continuous* with respect to $\mu$. A more difficult theorem provides the converse implication.

**Theorem 1.3.11** (Radon–Nikodým)**.** *If $\mu$ and $\nu$ are $\sigma$-finite measures on $(S, \Sigma)$ such that for any $A \in \Sigma$ we have $\nu(A) = 0$ as soon as $\mu(A) = 0$, then there exists a nonnegative function $f$ such that $\nu(A) = \mu(f\,\mathbb{1}_A)$ for all $A \in \Sigma$. The function $f$ is unique $\mu$-a.e.*

Uniqueness of the density is provided by the previous lemma. The argument we use is due to Anton Schep.

*Proof.* Step 1: reduction to *finite* measures. If $\mu$ and $\nu$ are $\sigma$-finite measures, then there exits a countable collection $(A_n)_n$ of disjoint sets in $\Sigma$ such that $\bigcup_n A_n = S$ and $\mu(A_n) < \infty$ and $\nu(A_n) < \infty$ for every $n$. Define then finite measures by $\mu_n(B) = \mu(A_n \cap B)$ and $\nu_n(B) = \nu(A_n \cap B)$ for every $B \in \Sigma$. Notice $\mu_n(B) = 0$, i.e. $\mu(A_n \cap B) = 0$, implies $\nu(A_n \cap B) = 0$, i.e. $\mu_n(B) = 0$. If the theorem holds for finites measures, then for every $n$, there exists $f_n$ such that for every $B \in \Sigma$, we have:

$$\nu(A_n \cap B) = \mu(f_n\,\mathbb{1}_{A_n \cap B}).$$

Let $f = \sum_n f_n\,\mathbb{1}_{A_n}$; since the sets $A_n$ are disjoint and cover $S$, then for every $s \in S$ exactly one indicator is non zero in the definition of $f(s)$. Then summing over $n$ the previous display yields: $\nu(B) = \mu(f\,\mathbb{1}_B)$.

From now on, we assume that $\mu$ and $\nu$ are finite measures. Replacing $\mu$ by $\mu(S)^{-1}\mu(\cdot)$, let us assume further that $\mu(S) = 1$.

Step 2: a first bound. Consider the set $H$ of all the measurable functions $f : S \longrightarrow [0, \infty)$ which satisfy:

$$\mu(f\,\mathbb{1}_A) \leq \nu(A) \qquad \text{for every} \qquad A \in \Sigma.$$

Note that it contains the constant function 0 so $H$ is not empty. Let then $M = \sup\{\mu(f), f \in H\}$. Taking the set $A$ above to be $S$, we have $0 \leq \mu(f) \leq \nu(S)$ which we assume here is finite. Hence $0 \leq M < \infty$ and there exists a sequence of functions $(f_n)_n$ all in $H$ such that $\mu(f_n) \longrightarrow M$. We can take this sequence to be nondecreasing by replacing it by $f_n' = \sup_{k\leq n} f_k$. For this, notice that the maximum of two functions in $H$ remains in $H$. Indeed, if $g, h \in H$, then for every $A \in \Sigma$, we have:

$$\mu(\max(g,h)\,\mathbb{1}_A) = \mu(g\,\mathbb{1}_{A\cap\{g\geq h\}}) + \mu(h\,\mathbb{1}_{A\cap\{g<h\}}) \leq \nu(A\cap\{g\geq h\}) + \nu(A\cap\{g<h\}) = \nu(A).$$

This extends to finitely many functions by induction so each $f_n' \in H$. Denote by $f = \uparrow \lim_n f_n'$ their limit, which is measurable and nonnegative. By monotone convergence $\mu(f) = \uparrow \lim_n \mu(f_n') = M$. We claim that $f \in H$. Indeed, by monotone convergence again, for every $A \in \Sigma$, we have

$$\mu(f\,\mathbb{1}_A) = \uparrow \lim_{n\to\infty} \mu(f_n'\,\mathbb{1}_A) \leq \nu(A).$$

Thus $f \in H$. The formula $\tilde{\nu}(A) = \nu(A) - \mu(f\,\mathbb{1}_A)$ for every $A \in \Sigma$ then defines a finite nonnegative measure, and we want to prove that it is constant equal to 0.

Step 3: a contradiction. Suppose by contradiction that $\tilde{\nu}(S) > 0$. We claim that in this case there exists $A \in \Sigma$ such that

$$\mu(A) > 0 \qquad \text{and} \qquad \tilde{\nu}(A \cap B) \geq \tilde{\nu}(S)\mu(A \cap B) \quad \text{for every} \quad B \in \Sigma. \tag{1.1}$$

Let us conclude from here and prove this claim after. Let $g = f + \tilde{\nu}(S)\,\mathbb{1}_A$, we claim that it belongs to the set $H$. Indeed for every $B \in \Sigma$, one has by the previous display for the first inequality and $\mu(f\,\mathbb{1}_A) \leq \nu(A)$ for the second one:

$$\begin{aligned}
\mu(g\,\mathbb{1}_B) &= \mu(f\,\mathbb{1}_B) + \tilde{\nu}(S)\mu(A \cap B) \\
&\leq \mu(f\,\mathbb{1}_B) + \tilde{\nu}(A \cap B) \\
&= \mu(f\,\mathbb{1}_B) + \nu(A \cap B) - \mu(f\,\mathbb{1}_{A\cap B}) \\
&= \mu(f\,\mathbb{1}_{A^c\cap B}) + \nu(A \cap B) \\
&\leq \nu(A^c \cap B) + \nu(A \cap B) \\
&\leq \nu(B).
\end{aligned}$$

Hence $g \in H$, so in particular $\mu(g) \leq M$. On the other hand, since $\mu(A) > 0$, then $\mu(g) = \mu(f) + \tilde{\nu}(S)\mu(A) > \mu(f) = M$. This contradiction shows that $A$ cannot exist, and thus $\tilde{\nu}$ is the 0 measure, namely $\tilde{\nu}(A) = \nu(A) - \mu(f\,\mathbb{1}_A) = 0$ for all $A \in \Sigma$ as we wanted.

It remains to prove (1.1). Let $\pi(A) = \tilde{\nu}(S)\mu(A) - \tilde{\nu}(A)$ for every $A \in \Sigma$ to simplify notation. Recall that for any $A \in \Sigma$, if $\mu(A) = 0$, then $\nu(A) = 0$ and then further $\tilde{\nu}(A) = \nu(A) - \mu(f\,\mathbb{1}_A) = 0$, so $\pi(A) = 0$. If $A = S$ does not satisfy (1.1), this means that there exists $B \in \Sigma$ such that $\pi(B) > 0$. Let then $n_1 \geq 1$ be the smallest integer such that there exists $B \in \Sigma$ with $\pi(B) > 1/n_1$, let $B_1$ be any such set, and let $A_1 = B_1^c$ be its complement. Then again if $A_1$ does not satisfy (1.1) then there exists $B \in \Sigma$ such that $\pi(A_1 \cap B) > 0$ and we let $n_2 \geq 1$ be the smallest integer such that there exists $B \in \Sigma$ with $\pi(A_1 \cap B) > 1/n_2$, then we let $B_2$ be any such a set $B$ and finally $A_2 = A_1 \cap B_2^c = (B_1 \cup B_2)^c$. By induction, for every $k \geq 1$, if $A_k = (\bigcup_{i=1}^k B_i)^c$ does not satisfy (1.1) then there exists $B \in \Sigma$ such that $\pi(A_k \cap B) > 0$ and we let $n_{k+1}$ be the smallest integer such that there exists $B \in \Sigma$ with $\pi(A_k \cap B) > 1/n_{k+1}$, then we let $B_k$ be any such a set $B$, and finally we let $A_{k+1} = A_k \cap B_{k+1}^c = (\bigcup_{i=1}^{k+1} B_i)^c$. If no $A_k$ satisfies (1.1) then we claim that $A = \bigcap_k A_k = (\bigcup_{i\geq 1} B_i)^c$ does. Indeed, first we have (recall that $\mu$ is a probability):

$$0 = \tilde{\nu}(S)\mu(S) - \tilde{\nu}(S) = \tilde{\nu}(S)\mu(A) - \tilde{\nu}(A) + \tilde{\nu}(S)\mu(A^c) - \tilde{\nu}(A^c).$$

Notice that $A^c = \bigcup_{i\geq 1} B_i$ is a disjoint union, and by construction we have $\tilde{\nu}(S)\mu(B_i) - \tilde{\nu}(B_i) \geq 1/n_i$ for every $i$. Consequently

$$\tilde{\nu}(S)\mu(A^c) - \tilde{\nu}(A^c) = \sum_i \tilde{\nu}(S)\mu(B_i) - \tilde{\nu}(B_i) \geq \sum_i 1/n_i.$$

Combining the two displays, we first infer that

$$\sum_i 1/n_i \leq \tilde{\nu}(A) - \tilde{\nu}(S)\mu(A) \leq \tilde{\nu}(S) < \infty.$$

In particular $n_k \longrightarrow \infty$ as $k \longrightarrow \infty$. Next, for any $B \in \Sigma$, for every $k \geq 1$, we have $A \cap B \subset A_k \cap B$ so $\pi(A \cap B) \leq \pi(A_k \cap B) \leq 1/(n_{k+1} - 1)$ by definition of $n_{k+1}$. Since the right-hand side tends to 0, then we conclude that $\pi(A \cap B) \leq 0$, that is $A$ satisfies (1.1). It only remains to check that $\mu(A) > 0$. Recall that if $\mu(A) = 0$, then $\tilde{\nu}(A) = 0$ and so $\tilde{\nu}(S)\mu(A) - \tilde{\nu}(A) = 0$, which implies by the previous display that $\sum_i 1/n_i \leq 0$, which is a contradiction. This concludes the proof of (1.1). □

A last key result is Fatou's lemma. Given a sequence of functions $f_n$, define $\liminf_n f_n$ and $\limsup_n f_n$ as the pointwise limits in $[-\infty, \infty]$:

$$\liminf_{n\to\infty} f_n = \uparrow \lim_{n\to\infty} \inf_{p\geq n} f_p \qquad \text{and} \qquad \limsup_{n\to\infty} f_n = \downarrow \lim_{n\to\infty} \sup_{p\geq n} f_p.$$

**Theorem 1.3.12** (Fatou). *Let $(f_n)_{n\geq 1}$ be nonnegative measurable functions. Then in $[0,\infty]$,*

$$\mu(\liminf_{n\to\infty} f_n) \leq \liminf_{n\to\infty} \mu(f_n).$$

*If in addition there exists $h$ such that $f_n \leq h$ for all $n$ and $\mu(h) < \infty$, then*

$$\mu(\limsup_{n\to\infty} f_n) \geq \limsup_{n\to\infty} \mu(f_n).$$

*Proof.* Put $g_n = \inf_{p\geq n} f_p$, which is a nondecreasing sequence and note that $f_p \geq g_n$ for all $n \geq p$ so by monotone convergence,

$$\mu(\uparrow \lim_{n\to\infty} g_n) = \uparrow \lim_{n\to\infty} \mu(g_n) \leq \uparrow \lim_{n\to\infty} \inf_{p\geq n} \mu(f_p).$$

The left-hand side equals $\mu(\liminf_n f_n)$, while the right-hand side equals $\liminf_n \mu(f_n)$. The second claim follows by applying the first one to the nonnegative functions $h - f_n$. □

## 1.4 Integration of general functions

**Definition 1.4.1.** A measurable function $f : S \longrightarrow \mathbb{R}$ is said to be *integrable* when $\mu(|f|) < \infty$. Let us set

$$f^+ = \max(f, 0) \qquad \text{and} \qquad f^- = -\min(f, 0) = \max(-f, 0),$$

so that

$$f = f^+ - f^- \qquad \text{and} \qquad |f| = f^+ + f^-.$$

Then $f$ is integrable if and only if both $\mu(f^+), \mu(f^-) < \infty$ and we define

$$\mu(f) = \mu(f^+) - \mu(f^-) \in \mathbb{R},$$

which we also denote by $\int f \,\mathrm{d}\mu = \int_S f(x)\mu(\mathrm{d}x)$.

The following results are easily derived from the definition.

**Exercise 1.4.2.** Let $f$ and $g$ be two integrable functions. Prove the following properties:

(i) $|\int f \,\mathrm{d}\mu| \leq \int |f| \,\mathrm{d}\mu$.

(ii) For every $a, b \in \mathbb{R}$, the function $af + bg$ is integrable and $\int (af + bg) \,\mathrm{d}\mu = a \int f \,\mathrm{d}\mu + b\mu(g)$.

(iii) If $f \leq g$ then $\int f \,\mathrm{d}\mu \leq \mu(g)$.

(iv) If $\mu(\{f \neq g\}) = 0$ then $\int f \,\mathrm{d}\mu = \mu(g)$.

**Remark 1.4.3.** We shall also need to consider vector-valued functions, in $\mathbb{R}^d$ for $d \geq 1$. Such a function $f = (f_1, \dots, f_d)$ is said to be integrable when its norm $|f|$ (any equivalent norm in $\mathbb{R}^d$) is integrable, or equivalently when each coordinate is an integrable real-valued function and then we define

$$\int f \,\mathrm{d}\mu = \left( \int f_1 \,\mathrm{d}\mu, \dots, \int f_d \,\mathrm{d}\mu \right) \in \mathbb{R}^d.$$

The above properties extend, except the third one. In the particular case of complex-valued functions, we have

$$\int f \,\mathrm{d}\mu = \int \mathrm{Re}\, f \,\mathrm{d}\mu + i \int \mathrm{Im}\, f \,\mathrm{d}\mu \in \mathbb{C}.$$

An important tool is the dominated convergence theorem.

**Theorem 1.4.4** (Dominated convergence). *Let $(f_n)_{n\geq 1}$ be measurable functions which converge pointwise to a function $f$. Suppose that there exists a measurable function $h$ such that $|f_n| \leq h$ for all $n$ and $\mu(h) < \infty$. Then*

$$\mu(|f_n - f|) \underset{n\to\infty}{\longrightarrow} 0$$

*and consequently $\mu(f_n) \longrightarrow \mu(f)$.*

*Proof.* First we can note that the function $f$ is integrable since $|f| \leq h$. Moreover $|f_n - f| \leq 2h$ so by Fatou's lemma,

$$\limsup_{n\to\infty} \mu(|f_n - f|) \leq \mu(\limsup_{n\to\infty} |f_n - f|) = \mu(0) = 0.$$

Moreover, we have

$$|\mu(f_n) - \mu(f)| = |\mu(f_n - f)| \leq \mu(|f_n - f|),$$

and the second claim follows. □

**Corollary 1.4.5** ($L^p$ dominated convergence). *Let $p \geq 1$. Let $(f_n)_{n\geq 1}$ be measurable functions which converge pointwise to a function $f$. Suppose that there exists a measurable function $h$ such that $|f_n| \leq h$ for all $n$ and $\mu(h^p) < \infty$. Then*

$$\mu(|f_n - f|^p) \underset{n\to\infty}{\longrightarrow} 0.$$

*Proof.* The sequence of functions $g_n = |f_n - f|^p$ converges pointwise to 0 and satisfies $|f_n - f|^p \leq (|f_n| + |f|)^p \leq (2h)^p$, which is integrable, so the result follows from the previous theorem. □

**Lemma 1.4.6** (Scheffé). *Let $(f_n)_{n\geq 1}$ be integrable functions which converge pointwise to an integrable function $f$. Then*

$$\mu(|f_n - f|) \underset{n\to\infty}{\longrightarrow} 0 \qquad \textit{if and only if} \qquad \mu(|f_n|) \underset{n\to\infty}{\longrightarrow} \mu(|f|).$$

*Proof.* For the direct implication, note that $||f_n| - |f|| \leq |f_n - f|$ so

$$|\mu(|f_n|) - \mu(|f|)| \leq \mu(||f_n| - |f||) \leq \mu(|f_n - f|).$$

For the converse one, assume first that all the functions are nonnegative, then $\min(f_n, f) \longrightarrow f$ and $0 \leq \min(f_n, f) \leq f$. Since $f$ is integrable, then we infer from dominated convergence that $\mu(\min(f_n, f)) \longrightarrow \mu(f)$. Now observe that

$$f_n + f = \max(f_n, f) + \min(f_n, f),$$

so after integration,

$$\mu(\max(f_n, f)) = \mu(f_n) + \mu(f) - \mu(\min(f_n, f)) \underset{n\to\infty}{\longrightarrow} \mu(f).$$

Hence

$$\mu(|f_n - f|) = \mu(\max(f_n, f)) - \mu(\min(f_n, f)) \underset{n\to\infty}{\longrightarrow} 0.$$

For general functions, write $f_n = f_n^+ - f_n^-$ and $f = f^+ - f^-$ and note that $f_n^\pm \longrightarrow f^\pm$ pointwise. Then the assumption for the converse implication reads

$$\mu(f_n^+) + \mu(f_n^-) \underset{n\to\infty}{\longrightarrow} \mu(f^+) + \mu(f^-).$$

Fatou's lemma implies that both $\liminf_n \mu(f_n^+) \geq \mu(f^+)$ and $\liminf_n \mu(f_n^-) \geq \mu(f^-)$, so we infer that

$$\mu(f_n^+) \underset{n\to\infty}{\longrightarrow} \mu(f^+) \qquad \text{and} \qquad \mu(f_n^-) \underset{n\to\infty}{\longrightarrow} \mu(f^-).$$

By the nonnegative case, this further implies that

$$\mu(|f_n - f|) \leq \mu(|f_n^+ - f^+|) + \mu(|f_n^- - f^-|) \underset{n\to\infty}{\longrightarrow} 0,$$

ans the proof is complete. □

In order to prove that a certain property holds for any integrable function, we often rely on the following reasoning:

- We first prove that it holds for indicator functions.
- We extend the property by linearity to simple functions.
- We extend it by monotone convergence to nonnegative functions.
- We finally extend it by monotone convergence to integrable functions by linearity after splitting the positive and negative part.

Let us illustrate this with the image measure by a function (also called *push-forward*), which allows to transfer measures from a measurable space to another.

**Definition 1.4.7** (Image measure)**.** Let $f : (S, \Sigma) \longrightarrow (E, \mathscr{E})$ be a measurable function and let $\mu$ be a measure on $(S, \Sigma)$. Then the function defined for all $B \in \mathscr{E}$ by

$$\mu_f(B) = \mu(f^{-1}(B)) = \mu(f \in B)$$

is a measure on $(E, \mathscr{E})$ called the *image measure* of $f$.

**Lemma 1.4.8** (Transfer)**.** *Let $f : (S, \Sigma) \longrightarrow (E, \mathscr{E})$ be a measurable function, $\mu$ be a measure on $(S, \Sigma)$, and $g : (E, \mathscr{E}) \longrightarrow (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be measurable. Then $g$ is $\mu_f$-integrable if and only if $g \circ f$ is $\mu$-integrable and in this case it holds*

$$\mu_f(g) = \int_E g \, \mathrm{d}\mu_f = \int_S g \circ f \, \mathrm{d}\mu = \mu(g \circ f).$$

*Proof.* When $g$ is the indicator of a set $B \in \mathscr{B}(\mathbb{R})$, then the identity is the definition of $\mu_f$. By linearity, the identity extends to nonnegative simple functions and then to any nonnegative functions by monotone convergence. Hence if $g$ is any measurable function, then $\mu_f(|g|) = \mu(|g(f)|)$ so the left-hand side is finite if and only if the right-hand side is and then the identity extends by linearity again after the splitting $g = g^+ + g^-$. □

Recall from Lemma 1.3.10 that given a measurable and nonnegative function $h$, we can define a measure $\nu$ by $\nu(A) = \mu(h \, \mathbb{1}_A)$. The measure $\nu$ is said to have a density $h$ with respect to $\mu$. Then the same proof as above shows that another measurable function $g$ is integrable for $\nu$ if and only if $gh$ is integrable for $\mu$ and in this case

$$\nu(g) = \int g \, \mathrm{d}\nu = \int gh \, \mathrm{d}\mu = \mu(gh).$$

Combined with Lemma 1.4.8, we obtain the following very useful criterion.

**Proposition 1.4.9.** *Let $\mu$ be a measure on $(S, \Sigma)$ and $f : (S, \Sigma) \longrightarrow (R, \mathscr{B}(\mathbb{R}))$ be a measurable function. Let $\lambda$ be a measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and let $h : \mathbb{R} \longrightarrow \mathbb{R}_+$ be measurable as well. Then the measure $\mu_f$ has density $h$ with respect to $\lambda$ if and only if for any nonnegative and measurable function $g : \mathbb{R} \longrightarrow \mathbb{R}_+$, we have*

$$\mu_f(g) = \mu(g \circ f) = \lambda(gh).$$

*In this case, for any measurable function $g$, we have that $g$ is $\mu_f$-integrable if and only if $gh$ is $\lambda$-integrable and then the above identity extends.*

*Proof.* Suppose the identity holds for any nonnegative and measurable function $g : \mathbb{R} \longrightarrow \mathbb{R}_+$, then taking $g = \mathbb{1}_B$ for any $B \in \mathscr{B}(\mathbb{R})$, we have $\mu_f(B) = \lambda(h \, \mathbb{1}_b)$, which is the definition of the fact that $\mu_f$ has density $h$ with respect to $\lambda$. For the converse implication, we can repeat the proof of Lemma 1.4.8. By definition, if $\mu_f$ has density $h$ with respect to $\lambda$, the identity holds for indicator functions, so it extends to nonnegative simple functions by linearity and then to any nonnegative functions by monotone convergence. □

Let us mention that sometimes (as in the next subsection), we cannot apply this simple reasoning because controlling the indicator function of any measurable set is already too complicated. Following the discussion of Section 1.1.1, it is possible to extend this reasoning by only controlling the indicator functions of a collection of subsets. The following result is based on Lemma 1.1.18, which explains the similarity in the name.

**Theorem 1.4.10** (Monotone class)**.** *Let $\mathscr{H}$ be a set of bounded functions from $S$ to $\mathbb{R}$ satisfying the following conditions:*

*(i) It is a real vector space in that if $f, g \in \mathscr{H}$ and $a, b \in \mathbb{R}$, then $af + bg \in \mathscr{H}$.*

*(ii) It contains the constant function equal to 1.*

*(iii) If $f_n \geq 0$ belongs to $\mathscr{H}$ for all $n$ and $f_n \uparrow f$ where $f$ is a bounded function, then $f \in \mathscr{H}$.*

*Suppose there exists a $\pi$-system $\mathscr{P}$ such that for any $A \in \mathscr{P}$, we have $\mathbb{1}_A \in \mathscr{H}$. Then every bounded and $\sigma(\mathscr{P})$-measurable function belongs to $\mathscr{H}$.*

*Proof.* Let $\mathscr{M} = \{A \subset S : \ \mathbb{1}_A \in \mathscr{H}\}$. By the three properties of $\mathscr{H}$ we see that $\mathscr{M}$ is $\lambda$-system. Since we also assume that it contains the $\pi$-system $\mathscr{P}$, then by Lemma 1.1.18, we have $\sigma(\mathscr{P}) \in \mathscr{M}$.

Let $K > 0$ and $f$ be a $\sigma(\mathscr{P})$-measurable function with $0 \leq f \leq K$ and for any $n \geq 1$, define

$$f_n = \sum_{i=0}^{\lfloor K2^n \rfloor} \frac{i}{2^n} \mathbb{1}_{i \leq 2^n f < i+1}.$$

Then $f_n$ is a simple function and $f_n \uparrow f$. Note that each set $A_{n,i} = \{i \leq 2^n f < i+1\} \in \sigma(\mathscr{P})$ since $f$ is $\sigma(\mathscr{P})$-measurable, so by the properties of $\mathscr{H}$, we have $f_n \in \mathscr{H}$ and then $f \in \mathscr{H}$.

Given any bounded and $\sigma(\mathscr{P})$-measurable function $f$, we infer from this that both $f^+, f^- \in \mathscr{H}$ and then $f \in \mathscr{H}$. □

## 1.5 Product measures

**Definition 1.5.1.** Let $(E_i, \mathscr{E}_i)_{i \leq n}$ be measurable spaces, then we can define a $\sigma$-algebra on $\prod_{i=1}^n E_i$ by

$$\bigotimes_{i=1}^n \mathscr{E}_i = \sigma\Big(\prod_{i=1}^n A_i, A_i \in \mathscr{E}_i \text{ for all } i \leq n\Big).$$

**Notation.** For any two sets $E$ and $F$, any pair $(x, y) \in E \times F$, and any subset $C \subset E \times F$, we let

$$C_x = \{y \in F : \ (x, y) \in C\} \qquad \text{and} \qquad C^y = \{x \in E : \ (x, y) \in C\}.$$

For a function $f$ from $(E, F)$ we also set

$$f_x : \ y \in F \mapsto f(x, y) \qquad \text{and} \qquad f^y : \ x \in E \mapsto f(x, y).$$

**Lemma 1.5.2.** *Fix three measurable spaces $(E, \mathscr{E})$, $(F, \mathscr{F})$, and $(G, \mathscr{G})$. The following holds.*

*(i) For every $A \in \mathscr{E} \otimes \mathscr{F}$ we have*

$$A_x \in \mathscr{F} \ \text{ for every } x \in E \qquad \text{and} \qquad A^y \in \mathscr{E} \ \text{ for every } y \in F.$$

*(ii) For any measurable function $f : (E \times F, \mathscr{E} \otimes \mathscr{F}) \longrightarrow (G, \mathscr{G})$, we have*

$$f_x \ \textcircled{m} \ \mathscr{F} \text{ for every } x \in E \qquad \text{and} \qquad f^y \ \textcircled{m} \ \mathscr{E} \text{ for every } y \in F.$$

*Proof.* (i) Fix $x \in E$ and for any $A \in \mathscr{E} \otimes \mathscr{F}$, let $A_x = \{y \in F : (x, y) \in A\}$ and $\mathscr{A}_x = \{A \in \mathscr{E} \otimes \mathscr{F} : A_x \in \mathscr{F}\}$. One easily checks that $\mathscr{A}_x$ is a sub-$\sigma$-algebra of $\mathscr{E} \otimes \mathscr{F}$. Moreover, for any pair $(B, C) \in \mathscr{E} \otimes \mathscr{F}$, if $A = B \times C$, then either $x \in B$ and then $A_x = C$, or $x \notin B$ and then $A_x = \emptyset$. In any case $B \times C \in \mathscr{A}_x$ for any $(B, C) \in \mathscr{E} \otimes \mathscr{F}$ so $\mathscr{A}_x = \mathscr{E} \otimes \mathscr{F}$.

(ii) Fix again $x \in \mathscr{E}$, then for any $D \in \mathscr{G}$, we have $f^{-1}(D) \in \mathscr{E} \otimes \mathscr{F}$ and thus

$$f_x^{-1}(D) = \{y \in F : f(x, y) \in D\} = \{y \in F : (x, y) \in f^{-1}(D)\} = (f^{-1}(D))_x \in \mathscr{F}$$

so $f_x$ is indeed $\mathscr{F}$-measurable. □

**Theorem 1.5.3.** *Let $\mu$, $\nu$ be two $\sigma$-finite measures on $(E, \mathscr{E})$ and $(F, \mathscr{F})$ respectively. The following holds.*

*(i) There exists a unique measure, which we denote by $\mu \otimes \nu$ on $(E \times F, \mathscr{E} \otimes \mathscr{F})$ such that for any $A \in \mathscr{E}$ and $B \in \mathscr{F}$, it holds*

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B).$$

*Moreover $\mu \otimes \nu$ is $\sigma$-finite. If both $\mu$ and $\nu$ are probability, then so is $\mu \otimes \nu$.*

*(ii) For every $C \in \mathscr{E} \otimes \mathscr{F}$ the functions*

$$x \mapsto \nu(C_x) \qquad \text{and} \qquad y \mapsto \mu(C^y)$$

*are measurable, with respect to $\mathscr{E}$ and to $\mathscr{F}$ respectively.*

*(iii) For every $C \in \mathscr{E} \otimes \mathscr{F}$ it holds*

$$\mu \otimes \nu(C) = \int_E \nu(C_x)\mu(\mathrm{d}x) = \int_F \mu(C^y)\nu(\mathrm{d}y).$$

The proof is based on Theorem 1.4.10 and will be omitted.

**Remark 1.5.4.** The assumptions that both measures are $\sigma$-finite is important. For example take $\mu$ to be the Lebesgue measure on $\mathbb{R}$ and $\nu$ the counting measure on $\mathbb{R}$, then for $C = \{(x, x), x \in \mathbb{R}\}$, we have $\int_E \nu(C_x)\mu(\mathrm{d}x) = \infty$ but $\int_F \mu(C^y)\nu(\mathrm{d}y) = 0$.

**Theorem 1.5.5** (Fubini–Tonelli)**.** *Let $\mu$, $\nu$ be two $\sigma$-finite measures on $(E, \mathscr{E})$ and $(F, \mathscr{F})$ respectively and let $f : E \times F \longrightarrow [0, \infty]$ be measurable. The following holds.*

*(i) The functions $x \mapsto \int_F f(x, y)\nu(\mathrm{d}y)$ and $y \mapsto \int_E f(x, y)\mu(\mathrm{d}x)$ are measurable with respect to $\mathscr{E}$ and to $\mathscr{F}$ respectively.*

*(ii) We have*

$$\int_{E\times F} f(x, y)\mu \otimes \nu(\mathrm{d}x\,\mathrm{d}y) = \int_E \left( \int_F f(x, y)\nu(\mathrm{d}y) \right) \mu(\mathrm{d}x) = \int_F \left( \int_E f(x, y)\mu(\mathrm{d}x) \right) \nu(\mathrm{d}y).$$

*Proof.* (i) For any $A \in \mathscr{E} \otimes \mathscr{F}$, if $f = \mathbb{1}_A$ then $x \mapsto \int_F f(x, y)\nu(\mathrm{d}y) = \nu(A_x)$ is measurable by the previous theorem. Measurability is preserved by linear combination and limits so we can extend it to simple functions and then nonnegative functions by monotone convergence.

(ii) Again, the identity for indicator functions reduces to the previous theorem, and we conclude by linearity and then monotone convergence. □

**Theorem 1.5.6** (Fubini–Lebesgue)**.** *Let $\mu$, $\nu$ be two $\sigma$-finite measures on $(E, \mathscr{E})$ and $(F, \mathscr{F})$ respectively and let $f : E \times F \longrightarrow \mathbb{R}$ be integrable (for $\mu \otimes \nu$). The following holds.*

*(i) For $\mu$-a.e. $x \in E$, the function $f_x$ is $\nu$-integrable and for $\nu$-a.e. $y \in F$, the function $f^y$ is $\mu$-integrable.*

*(ii) The functions* $x \mapsto \int_F f(x,y)\nu(\mathrm{d}y)$ *and* $y \mapsto \int_E f(x,y)\mu(\mathrm{d}x)$ *are well-defined and integrable.*

*(iii) We have*

$$\int_{E\times F} f(x,y)\mu\otimes\nu(\mathrm{d}x\,\mathrm{d}y) = \int_E\left(\int_F f(x,y)\nu(\mathrm{d}y)\right)\mu(\mathrm{d}x) = \int_F\left(\int_E f(x,y)\mu(\mathrm{d}x)\right)\nu(\mathrm{d}y).$$

*Proof.* (i) By the previous theorem, we have

$$\int_E\left(\int_F |f(x,y)|\nu(\mathrm{d}y)\right)\mu(\mathrm{d}x) = \int_{E\times F}|f(x,y)|\mu\otimes\nu(\mathrm{d}x\,\mathrm{d}y) < \infty.$$

Consequently $\int_F |f_x(y)|\nu(\mathrm{d}y) < \infty$ for $\mu$-a.e. $x \in E$.

(ii) It follows that for $\mu$-a.e. $x \in E$, the integral $\int_F f(x,y)\nu(\mathrm{d}y)$ is well-defined and moreover we have $\int_E(|\int_F f(x,y)\nu(\mathrm{d}y)|)\mu(\mathrm{d}x) \le \int_E(\int_F |f(x,y)|\nu(\mathrm{d}y))\mu(\mathrm{d}x) < \infty$.

(iii) We use the previous theorem and linearity, decomposing positive and negative parts. □

# Chapter 2

# Independent Random Variables (⋆)

The content of this chapter should for a large part be already known from a bachelor course in probability and will not be covered in class. Some developments are often excluded in a first course and are included here for interesting readers such as uniform integrability in Section 2.3.2, some generalities on weak convergence in Section 2.5.1, the Skorokhod representation theorem in Section 2.5.2, and Lindeberg's version of the Central Limit Theorem in Section 2.7.1.

**Contents**



We first translate in Section 2.1 the vocabulary from measure theory to probability, then we focus on the notion of independence of $\sigma$-algebras and of random variables and some key related results such as the Borel–Cantelli lemma. In Section 2.2 we recall some very useful inequalities such as the Markov inequality and Hölder's inequality and we discuss $L^p$ spaces, with some emphasis on $L^2$ which will be used to develop the theory of conditional expectation in Chapter 6. In Section 2.3 we discuss the notions of convergence in probability, in $L^p$, and almost surely and their relations; this is pushed to the limit with the theory of uniform integrability. Then in Section 2.4 we recall the first fondamental result in probability: the Law of Large Numbers. In Section 2.5 we focus on the convergence in distribution with some general useful results and an interesting development with the Skorokhod representation theorem. Then in Section 2.6 we present the characteristic function of a random vector, how it characterises its law and the convergence in distribution. Finally we focus in Section 2.7 on the Central Limit Theorem with first a version in dimension 1 for independent random variables but not necessarily with the same law, and then a version for i.i.d. vectors in higher dimensions, relying the notion of Gaussian vectors.

## 2.1 Probability & Independence

Probability theory is developed using measure theory whose basics are recalled in Chapter 1. From now on we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a $\sigma$-algebra on a set $\Omega$ and $\mathbb{P}$ is a probability measure. Elements of $\mathcal{F}$ are called *events*.

### 2.1.1 Random variables and distribution functions

**Definition 2.1.1.** Let us translate some vocabulary from measure theory.

- A measurable function $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ is called a *random variable* (abbreviated r.v.). When $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ for $d \geq 2$, we speak of a *random vector*, and for $d = 1$, of a real random variable, abbreviated r.r.v.
- The image measure $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$ for every $B \in \mathcal{E}$ as in Definition 1.4.7 is called the *law of* $X$. A random vector is said to have a *density* $f$ when its law $\mathbb{P}_X$ has a density $f$ with respect to the Lebesgue measure in $\mathbb{R}^d$ in the sense of Lemma 1.3.10.
- Finally, when $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we let $\mathbb{E}[X] = \mathbb{P}(X)$ denote the integral of $X$, which we call its *expectation*, provided it exists (either when $X \in [0, \infty)^d$ or when $\mathbb{E}[|X|] < \infty$).

Let $X$ be a r.v. in a general space $E$. Lemma 1.4.8 reads: for any measurable function $g : E \to \mathbb{R}_+$, we have:

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega))\, \mathbb{P}(\mathrm{d}\omega) = \int_E g(x)\, \mathbb{P}_X(\mathrm{d}x).$$

Also Proposition 1.4.9 yields the following criterion: let $X$ be a random vector in $\mathbb{R}^d$, then it has a density $f$ if and only if for any nonnegative and measurable function $g : \mathbb{R} \to \mathbb{R}_+$, we have:

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x) f(x)\, \mathrm{d}x.$$

In each case, the identity extends to integrable functions $g$.

A random vector may not have a density, but it always has a *distribution function*.

**Definition 2.1.2.** For any random vector $X = (X_1, \dots, X_d)$ in $\mathbb{R}^d$, we define its distribution function by:

$$F_X(x) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$$

for any $x = (x_1, \dots, x_d) \in \mathbb{R}^d$.

We usually consider distribution functions mainly in dimension $d = 1$, but the next result could be generalised to any finite dimension.

**Theorem 2.1.3.** *For any r.r.v. $X$, its distribution function $F_X$ satisfies:*

*(i) It is nondecreasing: $x \leq y \implies F_X(x) \leq F_X(y)$.*

*(ii) It is right-continuous: For any $x \in \mathbb{R}$ and any sequence $x_n \downarrow x$, we have $F_X(x_n) \downarrow F_X(x)$.*

*(iii) It has the limits $F_X(x) \to 1$ as $x \to \infty$ and $F_X(x) \to 0$ as $x \to -\infty$.*

*Conversely, for any such function $F$, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a r.r.v. $X$ such that $F = F_X$. Finally, if $Y$ is a r.r.v. with distribution function $F_X$, then $X$ and $Y$ have the same law in the sense that for any $B \in \mathcal{B}(\mathbb{R})$ and any nonnegative and measurable function $g$, we have*

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \qquad \text{and} \qquad \mathbb{E}[g(X)] = \mathbb{E}[g(Y)].$$

*Proof.* The three properties are easily checked from the monotonicity of measures. The second part is treated in the exercise sheet, where it is proved that if one defines for any $u \in (0, 1)$,

$$G(u) = \inf\{x \in \mathbb{R} : F(x) > u\} = \sup\{x \in \mathbb{R} : F(x) \leq u\},$$

and if $U$ has the uniform distribution on $(0, 1)$, then $X = G(U)$ has precisely distribution function $F$. The last point comes from Theorem 1.1.13 since the probabilities $\mathbb{P}_X$ and $\mathbb{P}_Y$ agree on the $\pi$-system $\{(-\infty, x], x \in \mathbb{R}\}$ that generates $\mathcal{B}(\mathbb{R})$. The identity for expectations follows as in the end of Section 1.4. □

The second part of the theorem has important applications in numerical simulations for it allows to generate from a uniform law (coded in any good language) any law for which $G$ is explicit. We shall also use it in some proofs.

### 2.1.2 Independence

We say that $\mathcal{G}$ is a sub-$\sigma$-algebra of $\mathcal{F}$ if it is a $\sigma$-algebra and $\mathcal{G} \subset \mathcal{F}$. Recall from Section 1.2 that if $X$ is a random variable with values in $(E, \mathcal{E})$, then

$$\sigma(X) = \{\omega \in \Omega : \ X(\omega) \in B, B \in \mathcal{E}\}$$

is a sub–$\sigma$-algebra of $\mathcal{F}$, which is the smallest one that makes $X$ measurable.

**Exercise 2.1.4.** Prove that for every set $A$, we have $\sigma(\mathbb{1}_A) = \sigma(A)$.

**Definition 2.1.5.** Let $(\mathcal{F}_n)_{n\geq 1}$ be sub-$\sigma$-algebras of $\mathcal{F}$. They are said to be *independent* when for every *finite* subset of indices $I \subset \mathbb{N}$ and every $A_i \in \mathcal{F}_i$ for $i \in I$, we have:

$$\mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = \prod_{i\in I} \mathbb{P}(A_i).$$

Random variables $(X_n)_{n\geq 1}$ are said to be independent when $(\sigma(X_n))_{n\geq 1}$ are and events $(A_n)_{n\geq 1}$ are said to be independent when $(\sigma(A_n))_{n\geq 1}$ are.

**Remark 2.1.6.**

- For any event $A$, we have $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$, so one can relate this definition with the more familiar one of independence of events.
- In the definition, one can always take $I$ of the form $\{1, \dots, N\}$ for $N \geq 1$. Indeed, for other subsets of indices, just take $A_i = \Omega$ for the indices $i$ that you do not want to appear.
- If $(X_n)_{n\geq 1}$ are independent, then so are $(f_n(X_n))_{n\geq 1}$ for any measurable functions $(f_n)_{n\geq 1}$ since each $f_n(X_n)$ is $\sigma(X_n)$-measurable.
- If $I \subset \mathbb{N}$ is infinite, then letting $I_n = I \cap \{1, \dots, n\}$, we have by monotonicity:

$$\mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = \mathbb{P}\Big(\bigcap_{n\geq 1}\bigcap_{i\in I_n} A_i\Big) = \downarrow \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{i\in I_n} A_i\Big) = \downarrow \lim_{n\to\infty} \prod_{i\in I_n} \mathbb{P}(A_i) = \prod_{i\in I} \mathbb{P}(A_i).$$

The following reformulation of independence of r.v.'s is very useful. Recall the product measure from Theorem 1.5.3.

**Theorem 2.1.7.** *For every $n \geq 1$, let $X_n$ be a r.v. with value in a measurable space $(E_n, \mathcal{E}_n)$. Then the r.v.'s $(X_n)_{n\geq 1}$ are independent if and only if for every $n \geq 1$, the law of $(X_1, \dots, X_n)$ on $E_1 \times \cdots \times E_n$ is the product law:*

$$\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n},$$

*which is equivalent to having*

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

*for all measurable functions $f_i : \ E_i \longrightarrow \mathbb{R}_+$, $i \leq n$.*

*Proof.* For each $i \leq n$, let $A_i \in \mathcal{E}_i$. On the one hand,

$$\mathbb{P}_{(X_1,\dots,X_n)}\Big(\prod_{i=1}^n A_i\Big) = \mathbb{P}\Big((X_1, \dots, X_n) \in \prod_{i=1}^n A_i\Big) = \mathbb{P}\Big(\bigcap_{i=1}^n \{X_i \in A_i\}\Big).$$

On the other hand,

$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}\Big(\prod_{i=1}^n A_i\Big) = \prod_{i=1}^n \mathbb{P}_{X_i}(A_i) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i).$$

Therefore $(X_n)_{n\geq 1}$ are independent if and only if for every $n \geq 1$, the laws $\mathbb{P}_{(X_1,\dots,X_n)}$ and $\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ coincide on the sets of the form $\prod_{i=1}^n A_i$, and thus on $\bigotimes_{i=1}^n \mathcal{E}_i$ which is generated by this $\pi$-system. The second assertion follows from the first one by Fubini's Theorem:

$$\begin{aligned}\mathbb{E}\Big[\prod_{i=1}^n f_i(X_i)\Big] &= \int_{E_1\times\cdots\times E_n} \prod_{i=1}^n f_i(x_i) \prod_{i=1}^n \mathbb{P}_{X_i}(\mathrm{d}x_i)\\ &= \prod_{i=1}^n \int_{E_i} f_i(x_i)\, \mathbb{P}_{X_i}(\mathrm{d}x_i)\\ &= \prod_{i=1}^n \mathbb{E}[f_i(X_i)].\end{aligned}$$

Conversely, the second assertion immediately implies the first one by taking $f_i = \mathbb{1}_{A_i}$. □

**Remark 2.1.8.** If the r.v.'s $(X_n)_{n\geq 1}$ are independent, then by Fubini's Theorem the identity

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

holds as soon as $\mathbb{E}[|f_i(X_i)|] < \infty$ for each $i \leq n$. Note that the left-hand side is well-defined since in this case, we have $\mathbb{E}[|\prod_{i=1}^n f_i(X_i)|] = \prod_{i=1}^n \mathbb{E}[|f_i(X_i)|] < \infty$ by the previous theorem.

As often, independence needs not to be checked for all possible sets, but sufficiently many. Recall the notion of a $\pi$-system from Section 1.1.1 and especially Theorem 1.1.13.

**Lemma 2.1.9.** *Let $(\pi_n)_{n\geq 1}$ be $\pi$-systems each included in $\mathcal{F}$ and containing $\Omega$ and such that for every finite subset of indices $I \subset \mathbb{N}$ and every $A_i \in \pi_i$, we have:*

$$\mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = \prod_{i\in I} \mathbb{P}(A_i).$$

*Then the sub-$\sigma$-algebras $(\sigma(\pi_n))_{n\geq 1}$ are independent.*

*Proof.* Consider two $\pi$-systems. Fix an event $A \in \pi_1$ and define two measures $\mu$ and $\nu$ on $\Omega$ by:

$$\mu(B) = \mathbb{P}(A \cap B) \qquad \text{and} \qquad \nu(B) = \mathbb{P}(A)\,\mathbb{P}(B),$$

for every $B \in \mathcal{F}$. Then they have same finite total mass $\mathbb{P}(A)$ and they agree on $\pi_2$ by assumption so they agree on $\sigma(\pi_2)$ by Theorem 1.1.13. We can therefore use the same reasoning with $B \in \sigma(\pi_2)$ fixed instead of $A$ and obtain that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

for any $A \in \sigma(\pi_1)$ and $B \in \sigma(\pi_2)$, i.e. that $\sigma(\pi_1)$ and $\sigma(\pi_2)$ are independent. The general case then follows by induction: suppose that we have proved that $(\sigma(\pi_n))_{n\leq N}$ are independent for some $N \geq 2$. Then fix events $A_1 \in \pi_1, \dots, A_N \in \pi_N$ and define two measures $\mu$ and $\nu$ by

$$\mu(B) = \mathbb{P}\Big(B \cap \bigcap_{i=1}^N A_i\Big) \qquad \text{and} \qquad \nu(B) = \mathbb{P}(B) \prod_{i=1}^N \mathbb{P}(A_i).$$

They have the same total mass $\mathbb{P}(\bigcap_{i=1}^N A_i) = \prod_{i=1}^N \mathbb{P}(A_i)$ and agree on $\pi_{N+1}$ so they agree on $\sigma(\pi_{N+1})$. Then we can fix $B \in \sigma(\pi_{N+1})$ an iteratively replace each $A_i \in \pi_i$ by $A_i \in \sigma(\pi_i)$. □

**Corollary 2.1.10.** *Let $(X_n)_{n\geq 1}$ be r.r.v.'s such that for any $N \geq 1$ and any $x_1, \dots, x_N \in \mathbb{R}^N$ we have*

$$\mathbb{P}(X_1 \leq x_1, \dots, X_N, \leq x_N) = \prod_{n=1}^N \mathbb{P}(X_n \leq x_n).$$

*Then they are independent.*

*Proof.* Apply the previous lemma with the $\pi$-system $\{(-\infty, x], x \in \mathbb{R}\}$ that generates $\mathcal{B}(\mathbb{R})$. □

Another consequence of Lemma 2.1.9 is that it allows to group $\sigma$-algebras that are independent.

**Corollary 2.1.11** (Grouping property)**.** *Let $(\mathcal{F}_n)_{n\geq 1}$ be independent $\sigma$-algebras. Let $(I_n)_{n\geq 1}$ be a partition of $\mathbb{N}$ and for every $n \geq 1$, define a $\sigma$-algebra by*

$$\mathcal{G}_n = \sigma(\mathcal{F}_k, k \in I_n) = \sigma\Big(\bigcup_{k\in I_n} \mathcal{F}_k\Big).$$

*Then $(\mathcal{G}_n)_{n\geq 1}$ are independent.*

*Proof.* Recall from Exercise 1.1.9 that each $\sigma$-algebra $\mathcal{G}_n$ is also generated by the $\pi$-system:

$$\pi_n = \bigcup_{I\subset I_n \text{ finite}} \Big\{\bigcap_{i\in I} B_i; B_i \in \mathcal{F}_i\Big\}\Big).$$

Fix indices $n_1 < \cdots < n_k$ and events $A_{n_1} \in \pi_{n_1}, \dots, A_{n_k} \in \pi_{n_k}$, namely each $A_{n_j}$ take the form $A_{n_j} = \bigcap_{i\in I^j} B_i$ for some finite subset $I^j \subset I_{n_j}$ and $B_i \in \mathcal{F}_i$. Using twice the independence of the $\mathcal{F}_k$'s, we have:

$$\mathbb{P}\Big(\bigcap_{j=1}^k A_{n_j}\Big) = \mathbb{P}\Big(\bigcap_{j=1}^k \bigcap_{i\in I^j} B_i\Big) = \prod_{j=1}^k \prod_{i\in I^j} \mathbb{P}(B_i) = \prod_{j=1}^k \mathbb{P}\Big(\bigcap_{i\in I^j} B_i\Big) = \prod_{j=1}^k \mathbb{P}(A_{n_j}).$$

We conclude by applying Lemma 2.1.9 that the $\sigma$-algebras $\mathcal{G}_n = \sigma(\pi_n)$ are indeed independent. □

**Example 2.1.12.** Let $(X_n)_{n\geq 1}$ be independent r.v.'s then

$$\sigma(X_{2n}, n \geq 1) \qquad \text{and} \qquad \sigma(X_{2n-1}, n \geq 1) \qquad \text{are independent.}$$

Also, for every $n \geq 1$,

$$\sigma(X_k, k \leq n) \qquad \text{and} \qquad \sigma(X_k, k \geq n+1) \qquad \text{are independent.}$$

### 2.1.3 $0-1$ laws

**Definition 2.1.13.** Let us define for events $(A_n)_{n\geq 1}$:

$$\limsup_{n\to\infty} A_n = \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n = \{\omega \in \Omega : \ \omega \in A_n \text{ for infinitely many indices } n\},$$

and

$$\liminf_{n\to\infty} A_n = \bigcup_{N\geq 1} \bigcap_{n\geq N} A_n = \{\omega \in \Omega : \ \omega \in A_n \text{ for all but finitely many indices } n\}.$$

Note that both $\limsup_n A_n \in \mathcal{F}$ and $\liminf_n A_n \in \mathcal{F}$ and that

$$\Big(\limsup_{n\to\infty} A_n\Big)^c = \liminf_{n\to\infty} A_n^c.$$

We have a simple Fatou's lemma for events.

**Lemma 2.1.14** (Fatou)**.** *We have*

$$\mathbb{P}\Big(\liminf_{n\to\infty} A_n\Big) \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \qquad \textit{and} \qquad \mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) \geq \limsup_{n\to\infty} \mathbb{P}(A_n).$$

*Proof.* For every $N \geq 1$ we have $\mathbb{P}(\bigcap_{n\geq N} A_n) \leq \inf_{n\geq N} \mathbb{P}(A_n)$; the right-hand side increases to $\liminf_n \mathbb{P}(A_n)$ as $N \to \infty$ while the left-hand side increases to $\mathbb{P}(\liminf_n A_n)$ by Lemma 1.1.12 since the sequence $(\bigcap_{n\geq N} A_n)_N$ is increasing. The second property follows similarly. □

The following simple result is also a powerful tool to prove that events occur with probability 0 (or 1 by taking the complement).

**Theorem 2.1.15** (Borel–Cantelli). *Let $(A_n)_{n\geq 1}$ be a sequence of events. Then*

*(i) If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_n A_n) = 0$.*

*(ii) If $\sum_n \mathbb{P}(A_n) = \infty$ and the $(A_n)_{n\geq 1}$ are independent, then $\mathbb{P}(\limsup_n A_n) = 1$.*

*Proof.* (i) By Lemma 1.1.12, for every $N \geq 1$ we have $\mathbb{P}(\bigcup_{n\geq N} A_n) \leq \sum_{n\geq N} \mathbb{P}(A_n)$. The right-hand side tends to 0 as $N \longrightarrow \infty$ by our assumption while the left-hand side tends to $\mathbb{P}(\limsup_n A_n) \geq \limsup_n \mathbb{P}(A_n)$ by Lemma 2.1.14.

(ii) By independence and the easy bound $e^x \geq 1 - x$ for $x \geq 0$, we infer that:

$$\mathbb{P}\Big(\bigcap_{n=N}^{\infty} A_n^c\Big) = \prod_{n=N}^{\infty}(1 - \mathbb{P}(A_n)) \leq \exp\Big(-\sum_{n=N}^{\infty} \mathbb{P}(A_n)\Big).$$

The left-hand side converges to $\mathbb{P}(\liminf_n A_n^c)$ as $N \longrightarrow \infty$ and the right-hand side to 0. □

Therefore, as soon as $\sum_n \mathbb{P}(A_n) < \infty$, we have $\mathbb{P}(\liminf_n A_n^c) = 1$, i.e. almost surely, $A_n$ occurs for only finitely many indices $n$. On the other hand, if $\sum_n \mathbb{P}(A_n) = \infty$, then for independent events, almost surely, $A_n$ occurs for infinitely many indices $n$.

We shall provide a second proof of the following result later using martingale theory.

**Theorem 2.1.16** (Kolmogorov's 0-1 law). *Let $(X_n)_{n\geq 1}$ be independent r.v.'s and consider the $\sigma$-algebras*

$$\mathcal{T}_n = \sigma(X_k, k \geq n+1) \qquad \text{and} \qquad \mathcal{T} = \bigcap_n \mathcal{T}_n.$$

*Then $\mathcal{T}$ is trivial in the sense that $\mathbb{P}(A) \in \{0,1\}$ for all events $A \in \mathcal{T}$ and that any $\mathcal{T}$-measurable r.v. is constant a.s.*

The $\sigma$-algebra $\mathcal{T}$ is called the *tail* $\sigma$-algebra. It contains all events that do not depend on any finite subset of r.v.'s such as

$$\{(X_n)_n \text{ converges}\}, \qquad \Big\{\sum_n X_n \text{ converges}\Big\}, \qquad \Big\{\frac{X_1 + \cdots + X_n}{n} \text{ converges}\Big\}.$$

Therefore all these events have probability either 0 or 1; of course the theorem does not tell us which case occurs! Also, r.v.'s such as

$$\liminf_n X_n \qquad \text{and} \qquad \liminf_n \frac{X_1 + \cdots + X_n}{n}$$

and the limsup, are measurable with respect to $\mathcal{T}$ so they are a.s. constant (possibly infinite).

*Proof.* Let $\mathcal{F}_n = \sigma(X_k, k \leq n)$. We observed already that by the grouping property, $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent for any $n \geq 1$. Thus $\mathcal{T}$ is independent from $\mathcal{F}_n$ for any $n \geq 1$. We infer that for any events $A \in \mathcal{T}$ and $B \in \bigcup_n \mathcal{F}_n$, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Note that $\bigcup_n \mathcal{F}_n$ is a $\pi$-system and $\mathcal{F} = \sigma(\bigcup_n \mathcal{F}_n) = \sigma(X_k, k \geq 1)$ so by Lemma 2.1.9, $\mathcal{T}$ and $\mathcal{F}$ are independent: for any $A \in \mathcal{T}$ and $B \in \mathcal{F}$, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Finally, $\mathcal{T} \subset \mathcal{F}$ so for any $A \in \mathcal{T}$, we have

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2,$$

thus $\mathbb{P}(A) \in \{0,1\}$. □

### 2.1.4 Independent random variables exist!

Theorem 2.1.7 combined with Theorem 1.5.3 shows that for any finitely many laws $\mathbb{P}_1, \dots, \mathbb{P}_n$, there exist independent r.v.'s $X_1, \dots, X_n$ such that each $X_i$ has law $\mathbb{P}_i$. The question is then to extend this to (countably) infinitely many laws. For laws on general measurable spaces, this extension may fail, but on $\mathbb{R}$ (and more generally on complete separable metric spaces), they do exist. The general result is called *Kolmogorov's extension Theorem*. We will content ourself with the result on $\mathbb{R}$ (it could as well be $\mathbb{R}^d$) for which we provide a constructive proof.

**Theorem 2.1.17.** *Given distribution functions $(F_n)_{n\geq 1}$, there exists a sequence of independent r.r.v.'s $(X_n)_{n\geq 1}$ such that $F_{X_n} = F_n$ for all $n \geq 1$.*

The proof goes in three steps. Starting from a single r.v. with the uniform distribution on $[0,1]$, we first construct a sequence of independent r.v.'s taking values 0 or 1 with probability 1/2. Then we use it to to construct a sequence of independent r.v.'s all having the uniform distribution on $[0,1]$. Finally, we prove the general form using these uniform r.v.'s.

*Proof in the case of coin tossing.* Take $\Omega = [0,1)$, $\mathscr{F} = \mathscr{B}(\Omega)$ the Borel $\sigma$-algebra, and $\mathbb{P} = \mathrm{Leb}$ the Lebesgue measure. Let us write any element $\omega \in \Omega$ using its binary expansion:

$$\omega = \sum_{n\geq 1} \varepsilon_n(\omega) 2^{-n},$$

where each $\varepsilon_n(\omega)$ is either 0 or 1, and can be defined explicitly by $\varepsilon_n(\omega) = \lfloor 2^n \omega \rfloor - 2\lfloor 2^{n-1}\omega \rfloor$. Fix $p \geq 1$ and $i_1, \dots, i_p \in \{0,1\}$ and note that we have $\varepsilon_1(\omega) = i_1, \dots, \varepsilon_p(\omega) = i_p$ if and only if $\omega \in [\sum_{n=1}^p i_n 2^{-n}, \sum_{n=1}^p i_n 2^{-n} + 2^{-p}]$. Therefore, we have

$$\mathbb{P}(\varepsilon_1 = i_1, \dots, \varepsilon_p = i_p) = \mathrm{Leb}\Big( \Big[ \sum_{n=1}^p i_n 2^{-n}, \sum_{n=1}^p i_n 2^{-n} + 2^{-p} \Big] \Big) = 2^{-p}.$$

This proves that the r.v.'s $(\varepsilon_n)_{n\geq 1}$ are independent and Bernoulli distributed with parameter 1/2. Indeed, for any $p \geq 1$, and $i_p \in \{0,1\}$, we have

$$\mathbb{P}(\varepsilon_p = i_p) = \sum_{i_1,\dots,i_{p-1} \in \{0,1\}} \mathbb{P}(\varepsilon_1 = i_1, \dots, \varepsilon_p = i_p) = 2^{p-1} \cdot 2^{-p} = 1/2,$$

and independence follows from the product form above. □

**Remark 2.1.18.** By grouping the variables, the random vectors $(\varepsilon_{np+1}, \dots, \varepsilon_{n(p+1)})_{n\geq 0}$ are independent for any given $p \geq 1$. Since each sequence $i_1, \dots, i_p \in \{0,1\}$ has a fixed probability $2^{-p} > 0$ to appear for each such vector, then by the Borel–Cantelli lemma, with probability one, any finite sequence of 0 and 1 appears infinitely many times in the binary expansion of a uniform random number! Of course, the same would hold for any other numerical basis.

*Proof in the case of the uniform distribution.* Let us continue with $\Omega = [0,1)$, $\mathscr{F} = \mathscr{B}(\Omega)$, and $\mathbb{P}$ the Lebesgue measure, and the previous sequence $(\varepsilon_n)_{n\geq 1}$ of independent r.v.'s with the law $\mathbb{P}(\varepsilon_n = 1) = \mathbb{P}(\varepsilon_n = 0) = 1/2$. Let $\varphi : \mathbb{N}^2 \longrightarrow \mathbb{N}$ be an injective map and let $\delta_{p,q} = \varepsilon_{\varphi(p,q)}$ for all $(p,q) \in \mathbb{N}^2$. Then the r.v.'s $(\delta_{p,q})_{(p,q)\in\mathbb{N}^2}$ are independent and $\mathbb{P}(\delta_{p,q} = 1) = \mathbb{P}(\delta_{p,q} = 0) = 1/2$. By grouping them, the sequences $((\delta_{p,q})_{q\geq 1})_{p\geq 1}$ are independent and thus if we set for each $p \geq 1$,

$$U_p = \sum_{q\geq 1} \delta_{p,q} 2^{-q} \in [0,1),$$

then the r.v.'s $(U_p)_{p\geq 1}$ are independent and they all have the same law $\mathbb{P}$, i.e. the uniform distribution on $[0,1)$. □

*Proof in the general case.* Let $(F_n)_{n\geq 1}$ be distribution functions and let us denote by $G_n$ their pseudo-inverse as in the proof of Theorem 2.1.3. Let $(U_n)_{n\geq 1}$ be independent r.v.'s with the uniform distribution on $[0,1)$. Then the r.v.'s $(G_n(U_n))_{n\geq 1}$ are independent and they have distribution function $F_n$ respectively. □

## 2.2 $L^p$ spaces in probability

### 2.2.1 Important inequalities

**Theorem 2.2.1** (Markov's inequality). *Let $X$ be a nonnegative r.r.v. then for every $a > 0$, we have*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

*Proof.* Simply note that $\mathbb{1}_{X \geq a} \leq X/a$ and take the expectation. □

This very simple inequality can become very powerful when applied to a transformation of $X$. For example, for any r.r.v. $X$, we have for any $a \in \mathbb{R}$ and $t > 0$:

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq e^{-ta}\,\mathbb{E}[e^{tX}],$$
$$\mathbb{P}(X \leq a) = \mathbb{P}(e^{-tX} \geq e^{-ta}) \leq e^{ta}\,\mathbb{E}[e^{-tX}].$$

Therefore

$$\mathbb{P}(X \geq a) \leq \inf_{t>0} e^{-ta}\,\mathbb{E}[e^{tX}] \qquad \text{and} \qquad \mathbb{P}(X \leq a) \leq \inf_{t>0} e^{ta}\,\mathbb{E}[e^{-tX}].$$

**Exercise 2.2.2.** If $X$ has the binomial distribution with parameters $n \geq 1$ and $p \in (0,1)$, find the optimal $t > 0$ in the above inequalities and thus the tightness bounds using this method.

**Theorem 2.2.3** (Jensen's inequality). *Let $\phi$ be a convex function from an open interval $I$ to $\mathbb{R}$ and let $X$ be a r.v. such that $X \in I$ a.s. and $\mathbb{E}[|X|] < \infty$. Then*

*(i)* $\mathbb{E}[\phi(X)^-] < \infty$ *so* $\mathbb{E}[\phi(X)] = \mathbb{E}[\phi(X)^+] - \mathbb{E}[\phi(X)^-]$ *makes sense in* $\mathbb{R} \cup \{\infty\}$,

*(ii) We have*

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X]).$$

*(iii) The inequality is an equality if and only if either $X$ is a.s. constant or $\phi$ is affine $\mathbb{P}_X$-a.s.*

*Proof.* As a convex function on an open interval, $\phi$ is continuous so $\phi(X)$ is indeed measurable. Moreover, it is known that for every $a \in I$, there exists $\lambda_a \in \mathbb{R}$ such that for all $x \in I$ such that $x - a \in I$,

$$\phi(x) \geq \phi(a) + \lambda_a(x - a).$$

It follows that $\phi(x)^- \leq (\phi(a) + \lambda_a(x-a))^-$, and when applied to $x = X$, the right-hand side has finite mean so we can indeed define $\mathbb{E}[\phi(X)] \in \mathbb{R} \cup \{\infty\}$. Moreover, for $a = \mathbb{E}[X]$, we get after taking the expectation:

$$\mathbb{E}[\phi(X)] \geq \mathbb{E}[\phi(\mathbb{E}[X]) + \lambda_{\mathbb{E}[X]}(X - \mathbb{E}[X])] = \phi(\mathbb{E}[X]).$$

Finally, since $\phi(X) \geq \phi(\mathbb{E}[X]) + \lambda_{\mathbb{E}[X]}(X - \mathbb{E}[X])$ a.s. then the equality of the expectations holds iff the r.v.'s are a.s. equal. □

**Notation.** For $p \geq 1$, set $\|X\|_p = \mathbb{E}[|X|^p]^{1/p} \in [0, \infty]$ and write $X \in L^p$ if $\|X\|_p < \infty$.

**Theorem 2.2.4** (Hölder's inequality). *Let $p, q > 1$ satisfy $1/p + 1/q = 1$, then*

$$\|XY\|_1 \leq \|X\|_p \, \|Y\|_q.$$

*Consequently, $XY \in L^1$ as soon as $X \in L^p$ and $Y \in L^q$. Moreover, we have $\|XY\|_1 = \|X\|_p \, \|Y\|_q$ iff either $X = 0$ a.s. or $Y = 0$ a.s. or there exists $c > 0$ such that $|X|^p = c\,|Y|^q$ a.s.*

*Proof.* Let us start with the Young inequality: for every $a, b > 0$, by strict convexity of exp, it holds

$$ab = \exp\big(\log(a^p)/p + \log(b^q)/q\big) \le a^p/p + b^q/q,$$

with equality iff $a^p = b^q$. Suppose $\|X\|_p > 0$ and $\|Y\|_q > 0$ as otherwise either $X$ or $Y$ (or both) equals 0 a.s. so $XY = 0$ a.s. and the theorem clearly holds. Assume also they are finite. Then the Young inequality applied to $a = |X|/\|X\|_p$ and $b = |Y|/\|Y\|_q$ yields

$$\frac{|XY|}{\|X\|_p \|Y\|_q} \le \frac{1}{p}\Big(\frac{|X|}{\|X\|_p}\Big)^p + \frac{1}{q}\Big(\frac{|Y|}{\|Y\|_q}\Big)^q = \frac{|X|^p}{p\,\|X\|_p^p} + \frac{|Y|^q}{q\,\|Y\|_q^q},$$

with equality iff $(|X|/\|X\|_p)^p = (|Y|/\|Y\|_q)^q$, iff there exists $c > 0$ such that $|X|^p = c\,|Y|^q$ a.s. Taking the expectation, we obtain

$$\frac{\|XY\|_1}{\|X\|_p \|Y\|_q} \le \frac{1}{p} + \frac{1}{q} = 1,$$

and the result follows. □

**Corollary 2.2.5** (Inclusion of $L^p$ spaces)**.** *Let $p \ge q \ge 1$, then $\|X\|_q \le \|X\|_p$. Consequently $L^p \subset L^q$.*

*Proof.* Just apply Hölder's inequality to $X$ and 1 and the exponents $r = p/q$ and $s = r/(r-1) = p/(p-q)$ so $1/r + 1/s = 1$ to get

$$\|X\|_q = \mathbb{E}[|X|^q]^{1/q} \le \mathbb{E}[|X|^{qr}]^{1/(qr)}\,\|1\|_s = \|X\|_p.$$

In particular, if $\|X\|_p < \infty$ then $\|X\|_q < \infty$. □

**Remark 2.2.6.** The inclusion $L^p \subset L^q$ for $p \ge q \ge 1$ is quite characteristic of finite measures in the sense that if $\mu$ is a $\sigma$-finite measure such that there exists a pair $p > q \ge 1$ with $L^p(\mu) \subset L^q(\mu)$, then $\mu$ is in fact finite. To see that, let $I : L^p(\mu) \longrightarrow L^q(\mu)$ be the identity operator, then by an argument similar to that used in the proof of Theorem 2.2.8 below, its graph its closed in the sense that if $X_n \longrightarrow X$ in $L^p$ and if $I(X_n) = X_n \longrightarrow Y$ in $L^q$, then for each of these sequences we can extract a subsequence that converges a.s. so $X = Y$ a.s. By the closed graph theorem (applied in the quotient spaces $L^p$ and $L^q$), we infer that $I$ is a continuous linear operator, so there exists $C < \infty$ such that for every $X \in L^p(\mu)$, we have $\|X\|_q \le C\,\|X\|_p$. In particular, if $\mu(A) < \infty$, then for $X = \mathbb{1}_A$, we read $\mu(A)^{1/p} \le C\mu(A)^{1/q}$ and so $\mu(A) \le C^{1/(1/p-1/q)} < \infty$. Hence, if $\mu$ is $\sigma$-finite, then there exists $(A_n)_{n\ge1}$ with $\mu(A_n) < \infty$ for all $n$ and $E = \bigcup_n A_n$ so $\mu(E) = \lim_n \mu(A_n) = C^{1/(1/p-1/q)} < \infty$.

### 2.2.2 $L^p$ spaces are *almost* Banach spaces

Let us start with another famous inequality.

**Corollary 2.2.7** (Minkowski's inequality)**.** *Let $p \ge 1$, then*

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_q.$$

*Proof.* The claim is clear for $p = 1$ so fix $p > 1$ and notice that for any $x, y \in \mathbb{R}$, we have

$$|x + y|^p \le |x|\,|x + y|^{p-1} + |y|\,|x + y|^{p-1}.$$

Let $q = p/(p-1)$ be such that $1/p + 1/q = 1$, then by Hölder's inequality,

$$\begin{aligned}\mathbb{E}[|X + Y|^p] &\le \mathbb{E}[|X|\,|X + Y|^{p-1}] + \mathbb{E}[|Y|\,|X + Y|^{p-1}] \\ &\le \|X\|_p\,\||X + Y|^{p-1}\|_q + \|Y\|_p\,\||X + Y|^{p-1}\|_q \\ &\le (\|X\|_p + \|Y\|_p)\,\mathbb{E}[|X + Y|^p]^{1-1/p}.\end{aligned}$$

The result follows after rearranging the terms. □

As a consequence, the space $(L^p, \|\cdot\|_p)$ is basically a normed vector space, except that $\|X\|_p = 0$ if and only if $X = 0$ a.s. One can get a true normed vector space by taking the quotient by the equivalence relation $X \sim Y$ when $X = Y$ a.s. and actually, what we denote by $L^p$ here is usually denoted by $\Lambda^p$, whereas $L^p$ should refer to the quotient space. We prefer to work without quotienting and speak of $L^p$ as a metric space by abuse of notation. The next result shows that it is a Banach space.

**Theorem 2.2.8** (Completeness)**.** *For any $p \geq 1$, any Cauchy sequence in $L^p$ converges.*

*Proof.* Let $(X_n)_{n\geq 1}$ be a Cauchy sequence in $L^p$ i.e. $\|X_n\|_p < \infty$ for all $n \geq 1$ and

$$\sup_{s,t\geq n} \|X_s - X_t\|_p \underset{n\to\infty}{\longrightarrow} 0.$$

Then we can build a sequence of integers $(n_k)_{k\geq 1}$ such that $\sup_{s,t\geq n_k} \|X_s - X_t\|_p \leq 2^{-k}$ for every $k \geq 1$, and in particular,

$$\mathbb{E}\left[\sum_{k\geq 1} |X_{n_{k+1}} - X_{n_k}|\right] = \sum_{k\geq 1} \|X_{n_{k+1}} - X_{n_k}\|_1 \leq \sum_{k\geq 1} \|X_{n_{k+1}} - X_{n_k}\|_p < \infty.$$

Thus a.s. the series $\sum_{k\geq 1}(X_{n_{k+1}} - X_{n_k})$ converges absolutely and so $X_{n_k}$ converges to some $X$. Fix $k \geq 1$ and $s \geq n_k$, then for every $\ell \geq k$ we have that

$$\mathbb{E}[|X_s - X_{n_\ell}|^p] \leq 2^{-pk}$$

so by Fatou's lemma, letting $\ell \longrightarrow \infty$, we get

$$\mathbb{E}[|X_s - X|^p] \leq \liminf_{\ell\to\infty} \mathbb{E}[|X_s - X_{n_\ell}|^p] \leq 2^{-pk}.$$

Since $X_s \in L^p$, then this shows that $X \in L^p$, and furthermore,

$$\limsup_{s\longrightarrow\infty} \mathbb{E}[|X_s - X|^p] \leq 2^{-pk}.$$

since $k \geq 1$ is arbitrary, we conclude that $X_s \longrightarrow X$ in $L^p$. □

### 2.2.3 The case of $L^2$

Let us end with some extra words on the case $p = 2$. Here $L^2$ not only is a Banach space, but a Hilbert space since the norm $\|\cdot\|_2$ comes from an inner (or scalar) product, namely for $X, Y \in L^2$,

$$X \cdot Y = \mathbb{E}[XY],$$

which is well-defined by Hölder's inequality, which in the case $p = q = 2$ is the Cauchy–Schwarz inequality for inner products.

**Definition 2.2.9** (Covariance)**.** For $X, Y \in L^2$, define their *covariance* by

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

as well as their *variance* by

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

and similarly for $Y$, and finally define their *correlation coefficient* by

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \in [-1, 1]$$

by the Cauchy–Schwarz inequality.

**Notation.** If $X = (X_1, \dots, X_n)$ is a random vector in $\mathbb{R}^n$, we denote by $C_X$ its covariance matrix, given for every $1 \le i, j \le n$ by

$$(C_X)_{i,j} = \mathrm{Cov}(X_i, X_j).$$

From a geometrical point of view, suppose $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ (otherwise subtract the mean), then $\mathrm{Var}(X)$ is the square-norm of the "vector" $X$ and $\rho(X, Y)$ is the cosine of the angle between $X$ and $Y$. The case $\rho(X, Y) = 0$, equivalently $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] = 0$, corresponds to the orthogonality of the vectors, which is the case as soon as $X$ and $Y$ are independent but is weaker than independence in general).

**Remark 2.2.10.** The covariance is bilinear, so the variance satisfies for $X_1, \dots, X_n \in L^2$:

$$\mathrm{Var}\left(\sum_{k=1}^{n} a_k X_k\right) = \sum_{k=1}^{n} a_k^2 \,\mathrm{Var}(X_k) + 2 \sum_{1 \le k < \ell \le n} a_k a_\ell \,\mathrm{Cov}(X_k, X_\ell).$$

Let $\langle \cdot, \cdot \rangle$ denote the scalar product in $\mathbb{R}^n$, let $(X_1, \dots, X_n) \in \mathbb{R}^n$ and let $C_X$ denote its covariance matrix, then this reads equivalently: for every $a = (a_1, \dots, a_n) \in \mathbb{R}^n$,

$$\mathrm{Var}(\langle a, X \rangle) = \langle a, C_X a \rangle = a^t C_X a.$$

**Exercise 2.2.11.** Let $X = (X_1, \dots, X_n)$ be a random vector in $\mathbb{R}^n$ with covariance matrix $C_X$. Prove that $C_X$ is noninvertible if and only if one $X_k$ is an affine combination of the other ones.

## 2.3 Convergence of random variables

In this section, all the random variables are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and we let $|\cdot|$ denote the Euclidean norm in $\mathbb{R}^d$. The $L^p$ spaces considered are always for $p \ge 1$.

### 2.3.1 Definitions and first properties

**Definition 2.3.1.** We say that $X_n$ converges to $X$:

(i) *almost surely (a.s.)* if $\mathbb{P}(X_n \longrightarrow X) = \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \longrightarrow X(\omega)\}) = 1$,

(ii) *in $L^p$* if $X_n, X \in L^p$ and $\mathbb{E}[|X_n - X|^p] \longrightarrow 0$,

(iii) *in probability* if for every $\varepsilon > 0$ fixed, we have $\mathbb{P}(|X_n - X| > \varepsilon) \longrightarrow 0$.

Let us observe that $\{X_n \longrightarrow X\} = \bigcap_{k \ge 1} \bigcup_{N \ge 1} \bigcap_{n \ge N} \{|X_n - X| \le 1/k\}$ is indeed measurable so the a.s. convergence is well-defined.

**Proposition 2.3.2.** *These notions satisfy the following relations:*

*(i) If $X_n \longrightarrow X$ in $L^p$ for a given $p \ge 1$ then $X_n \longrightarrow X$ in $L^q$ for all $q \in [1, p)$ and also in probability.*

*(ii) If $X_n \longrightarrow X$ a.s. then also in probability.*

*(iii) If $X_n \longrightarrow X$ and $X_n \longrightarrow Y$ in probability then $X = Y$ a.s. Thus the same conclusion holds if we replace one or both convergences by a stronger one (a.s. or in $L^p$).*

*(iv) $X_n \longrightarrow X$ in probability if and only if every subsequence has a further subsequence that tends to $X$ a.s.*

*Proof.* (i) By the Hölder or Jensen inequality, we have the inclusion $L^p \subset L^q$ which reads here $\mathbb{E}[|X_n - X|^q] \le \mathbb{E}[|X_n - X|^p]^{q/p} \longrightarrow 0$. Also, the Markov inequality implies that for every $\varepsilon > 0$, we have $\mathbb{P}(|X_n - X| > \varepsilon) \le \varepsilon^{-p}\,\mathbb{E}[|X_n - X|^p] \longrightarrow 0$.

(ii) Fix $\varepsilon > 0$ and note that if $X_n \longrightarrow X$ a.s. then $\mathbb{1}_{|X_n - X| > \varepsilon} \longrightarrow 0$ a.s. so $\mathbb{P}(|X_n - X| > \varepsilon) \longrightarrow 0$ by dominated convergence.

(iii) By the triangle inequality, for every $\varepsilon > 0$,

$$\mathbb{P}(|X - Y| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon/2) + \mathbb{P}(|X_n - Y| > \varepsilon/2) \underset{n\to\infty}{\longrightarrow} 0.$$

Thus

$$\mathbb{P}(X \neq Y) = \mathbb{P}\Big(\bigcup_{k\geq 1}\{|X - Y| > 1/k\}\Big) \leq \sum_{k\geq 1} \mathbb{P}(|X - Y| > 1/k) = 0.$$

(iv) Suppose $X_n \longrightarrow X$ in probability, then define $n_1 = 1$ and the iteratively for every $k \geq 1$,

$$n_{k+1} = \inf\{j > n_k : \ \mathbb{P}(|X_j - X| > 2^{-(k+1)}) \leq 2^{-(k+1)}\}.$$

Then $\sum_{k\geq 1} \mathbb{P}(|X_{n_k} - X| > 2^{-k}) < \infty$ so by the Borel–Cantelli lemma, with probability 1 only finitely many indices $n_k$ have that $|X_{n_k} - X| > 2^{-k}$ and so $X_{n_k} \longrightarrow X$ with probability 1.

On the other hand, if $X_n$ does not converge to $X$ in probability, then there exists $\varepsilon > 0$ and an increasing sequence of integers $(n_k)_{k\geq 1}$ such that $\mathbb{P}(|X_{n_k} - X| > \varepsilon) > \varepsilon$ for all $k \geq 1$ so this subsequence has no further subsequence that converges in probability and so a.s. □

Let $(X_n)_n$ be independent r.v.'s such that $\mathbb{P}(X_n = n^{1/p}) = 1/n = 1 - \mathbb{P}(X_n = 0)$. Then $X_n \longrightarrow 0$ in probability but not in $L^p$ since $\mathbb{E}[|X_n|^p] = 1$ for each $n$. It does not converge a.s. either since the Borel–Cantelli lemma shows that a.s. there exists infinitely many indices $n$ such that $X_n = n^{1/p}$.

**Remark 2.3.3.** In a metric space a sequence $(x_n)_n$ converges to some $x$ if and only if every subsequence has a further subsequence that tends to $x$. Thus there is no metric on r.v.'s that corresponds to a.s. convergence. On the other hand we saw that if one does not distinguish r.v.'s that are equal a.s. then the $L^p$ convergence corresponds to a metric (even to a norm); this is also the case of convergence in probability, with e.g. the distance

$$d(X, Y) = \mathbb{E}[\max\{|X - Y|, 1\}].$$

See also the exercise sheet.

**Lemma 2.3.4** (Continuous mapping)**.** *Suppose that $f : \mathbb{R}^d \longrightarrow \mathbb{R}^e$ is continuous $\mathbb{P}_X$-a.s. then*

*(i) If $X_n \longrightarrow X$ a.s. then $f(X_n) \longrightarrow f(X)$ a.s.*

*(ii) If $X_n \longrightarrow X$ in probability then $f(X_n) \longrightarrow f(X)$ in probability.*

*Proof.* (i) Let $C_f = \{x \in \mathbb{R}^d : \ f \text{ is continuous at } x\} \in \mathscr{B}(\mathbb{R}^d)$. Then we have

$$X^{-1}(C_f) = \{\omega \in \Omega : \ f \text{ is continuous at } X(\Omega)\} \in \mathscr{F}$$

and $1 = \mathbb{P}_X(C_f) = \mathbb{P}(X^{-1}(C_f))$ so if we let $A = \{\omega \in \Omega : \ X_n(\omega) \longrightarrow X(\omega)\}$, then $\mathbb{P}(A) = 1$ so $\mathbb{P}(A \cap X^{-1}(C_f)) = 1$. Finally for every $\omega \in A \cap X^{-1}(C_f)$ we have $f(X_n(\omega)) \longrightarrow f(X(\omega))$.

(ii) If $X_n \longrightarrow X$ in probability then every subsequence $(n_k)_k$ has a further subsequence $(n_{k_j})_j$ that converges to $X$ a.s. so by the first point $f(X_{n_{k_j}}) \longrightarrow f(X)$ a.s. By the last item in Proposition 2.3.2 this is equivalent to $f(X_n) \longrightarrow f(X)$ in probability. □

**Corollary 2.3.5.** *If $X_n \longrightarrow X$ and $Y_n \longrightarrow Y$ in probability, then $(X_n, Y_n) \longrightarrow (X, Y)$ in probability and so $X_n + Y_n \longrightarrow X + Y$, $X_n Y_n \longrightarrow XY$ in probability etc. The same holds if all convergences are a.s.*

**Remark 2.3.6.** Deducing $f(X_n) \longrightarrow f(X)$ in $L^p$ from $X_n \longrightarrow X$ in $L^p$ where $f$ is continuous $\mathbb{P}_X$-a.s. is not automatic! We will see in Theorem 2.3.14 that the sequence $(|f(X_n)|^p)_{n\geq 1}$ must be *uniformly integrable*. This is one reason why this notion of convergence is sometimes less interesting than convergence in probability.

### 2.3.2 Uniform Integrability (⋆)

Uniform integrability is the key assumption that allows to improve a convergence in probability to an almost sure convergence. Let us motivate the forthcoming definition with an observation.

**Lemma 2.3.7.** *A r.v. $X$ is integrable if and only if*

$$\lim_{K\to\infty} \mathbb{E}\left[|X|\,\mathbb{1}_{|X|>K}\right] = 0.$$

*Proof.* Notice that $\mathbb{E}[|X|\,\mathbb{1}_{|X|\le K}] \le K$, so in $[0,\infty]$ it makes sense to write

$$\mathbb{E}[|X|\,\mathbb{1}_{|X|>K}] = \mathbb{E}[|X|] - \mathbb{E}[|X|\,\mathbb{1}_{|X|\le K}].$$

Then either $\mathbb{E}[|X|] = \infty$ and then $\mathbb{E}[|X|\,\mathbb{1}_{|X|>K}] = \infty$ for all $K$, or $\mathbb{E}[|X|] < \infty$ and by monotone convergence, the right-hand side converges 0 as $K \to \infty$. □

**Definition 2.3.8** (UI r.v.'s)**.** A collection $(X_i)_{i\in I}$ of integrable r.v.'s is said to be *uniformly integrable* when

$$\lim_{K\to\infty} \sup_{i\in I} \mathbb{E}\left[|X_i|\,\mathbb{1}_{|X_i|>K}\right] = 0.$$

The next result is a useful reformulation of the UI property.

**Proposition 2.3.9.** *A collection $(X_i)_{i\in I}$ of integrable r.v.'s is UI if and only if $\sup_{i\in I}\mathbb{E}[|X_i|] < \infty$ (we say it is bounded in $L^1$) and for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every $A \in \mathcal{F}$,*

$$\text{if} \quad \mathbb{P}(A) \le \delta \quad \text{then} \quad \sup_{i\in I} \mathbb{E}[|X_i|\,\mathbb{1}_A] \le \varepsilon.$$

*Proof.* First suppose that $(X_i)_{i\in I}$ is UI, then for $K$ large enough, we have

$$\sup_{i\in I}\mathbb{E}[|X_i|] \le \sup_{i\in I}\mathbb{E}\left[|X_i|\,\mathbb{1}_{|X_i|\le K}\right] + \sup_{i\in I}\mathbb{E}\left[|X_i|\,\mathbb{1}_{|X_i|>K}\right] \le K + 1.$$

Fix now $\varepsilon > 0$ and let $K$ be large enough so $\sup_{i\in I}\mathbb{E}[|X_i|\,\mathbb{1}_{|X_i|>K}] \le \varepsilon/2$. Put $\delta = \varepsilon/(2K)$ and take any $A \in \mathcal{F}$ such that $\mathbb{P}(A) \le \delta$, then for any $i \in I$ it holds

$$\mathbb{E}[|X_i|\,\mathbb{1}_A] \le \mathbb{E}[|X_i|\,\mathbb{1}_{|X_i|>K}] + K\,\mathbb{P}(\{|X_i| \le K\} \cap A) \le \varepsilon.$$

Conversely, suppose that $\sup_{i\in I}\mathbb{E}[|X_i|] \le C < \infty$, fix $\varepsilon > 0$ arbitrary and $\delta > 0$ as in the second property. Let $K$ be large enough so $\mathbb{P}(|X_i| > K) \le \mathbb{E}[|X_i|]/K \le C/K \le \delta$ for all $i \in I$, then by the second property, $\mathbb{E}[|X_i|\,\mathbb{1}_{|X_i|>K}] \le \varepsilon$ for all $i \in I$. □

**Remark 2.3.10.** As an example of a family of r.v.'s bounded in $L^1$ but not UI, take $\mathbb{P}(X_n = n) = 1/n$ and $\mathbb{P}(X_n = 0) = 1 - 1/n$. Then for every $K \ge 0$,

$$\sup_{n\ge1}\mathbb{E}[|X_n|] = 1 \qquad \text{and} \qquad \sup_{n\ge1}\mathbb{E}\left[|X_n|\,\mathbb{1}_{|X_n|>K}\right] = 1.$$

One can also check that the second property of Proposition 2.3.9 fails here.

The next result gives an explicit way of checking wether collection of r.v.'s if UI. Of course the converse implication is the one useful in practice.

**Theorem 2.3.11** (de la Vallée Poussin)**.** *A collection $(X_i)_{i\in I}$ of integrable r.v.'s is UI if and only if there exists $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\varphi(x) \to \infty$ as $x \to \infty$ and*

$$\sup_{i\in I}\mathbb{E}[|X_i|\varphi(|X_i|)] < \infty.$$

*Proof.* Suppose that there exists such a function $\varphi$, then for every $K > 0$, for every $i \in I$, we have:

$$\mathbb{E}[|X_i|\,\mathbb{1}_{|X_i|>K}] = \mathbb{E}\left[|X_i|\frac{\varphi(|X_i|)}{\varphi(|X_i|)}\,\mathbb{1}_{|X_i|>K}\right] \le \mathbb{E}\left[|X_i|\frac{\varphi(|X_i|)}{\inf_{x>K}\varphi(x)}\,\mathbb{1}_{|X_i|>K}\right] \le \frac{\mathbb{E}[|X_i|\varphi(|X_i|)]}{\inf_{x>K}\varphi(x)}.$$

The numerator is bounded uniformly in $i \in I$ and the denominator tends to $\infty$ as $K \to \infty$.

Conversely, suppose that $(X_i)_{i\in I}$ is UI and build inductively an increasing sequence $K_n \to \infty$ such that

$$\sup_{i\in I}\mathbb{E}[|X_i|\,\mathbb{1}_{|X_i|>K_n}] \le 2^{-n}$$

for every $n \ge 1$. Put $\varphi(0) = 0$ and for $x > 0$,

$$\varphi(x) = \frac{1}{x}\sum_{n\ge 1}(x - K_n)^+ = \sum_{n\ge 1}\left(1 - \frac{K_n}{x}\right)\mathbb{1}_{x>K_n}.$$

Then for every $N \ge 1$, if $x \ge 2K_{2N}$, then the sum contains at least $2N$ terms and each of them is larger than $1/2$ so $\varphi(x) \to \infty$. Moreover, for every $x > 0$, we have

$$x\varphi(x) = \sum_{n\ge 1}(x - K_n)\,\mathbb{1}_{x>K_n} \le \sum_{n\ge 1} x\,\mathbb{1}_{x>K_n}.$$

Recall the construction of $(K_n)_n$, then we conclude that

$$\sup_{i\in I}\mathbb{E}[|X_i|\varphi(|X_i|)] \le \sup_{i\in I}\sum_{n\ge 1}\mathbb{E}[|X_i|\,\mathbb{1}_{|X_i|>K_n}] \le 1,$$

which ends the proof. □

**Remark 2.3.12.** One can show that the previous function $\psi : x \mapsto x\varphi(x)$ is convex, increasing, and with moderate growth in that $\psi(x) \le x^2$ for all $x \ge 0$ so in the theorem, one can restrict to those functions.

We now list some sufficient conditions that imply uniform integrability.

**Exercise 2.3.13** (Sufficient conditions)**.** Prove the following:

(i) If the $X_i$'s have the same law and are in $L^1$, then they are UI.

(ii) If $\mathcal{C}_1, \dots, \mathcal{C}_n$ are collections of UI r.v.'s, then so is $\bigcup_{i=1}^n \mathcal{C}_i$.

(iii) If $(X_i)_{i\in I}$ and $(Y_i)_{i\in I}$ are UI, then so is $(aX_i + bY_i)_{i\in I}$ for any constants $a$, $b$.

(iv) If there exists $Y \in L^1$ such that $|X_i| \le Y$ for all $i \in I$, then $(X_i)_{i\in I}$ is UI.

(v) If there exists $p > 1$ such that $\sup_{i\in I}\mathbb{E}[|X_i|^p] < \infty$, then $(X_i)_{i\in I}$ is UI.

The reason to consider UI r.v.'s is the next result which is the central result of this section. Indeed, combined with the previous exercise, this theorem extends the dominated convergence to the best possible. Of course again, the implication (i) $\Longrightarrow$ (ii) is the one useful in practice.

**Theorem 2.3.14.** *Fix $p \ge 1$, a sequence $(X_n)_{n\ge 1}$ of r.v.'s in $L^p$, and a r.v. $X$. Then the following assertions are equivalent:*

*(i) $X_n \to X$ in probability and $(|X_n|^p)_{n\ge 1}$ is UI,*

*(ii) $X \in L^p$ and $X_n \to X$ in $L^p$.*

*Proof.* Let us only prove the claim for $p = 1$. For $p > 1$, all the arguments still apply, one can simply replace the triangle inequality in $\mathbb{R}^d$ by the easy bound $|x + y|^p \le 2^p(|x|^p + |y|^p)$ for all $x, y \in \mathbb{R}^d$.

Assume first that $X_n \longrightarrow X$ in probability and $(X_n)_{n\geq 1}$ is UI. In particular $(X_n)_{n\geq 1}$ is bounded in $L^1$. Let us extract a subsequence $X_{n_k} \longrightarrow X$ a.s. Then by Fatou's lemma,

$$\mathbb{E}[|X|] = \mathbb{E}[\liminf_{k\longrightarrow\infty} |X_{n_k}|] \leq \liminf_{k\longrightarrow\infty} \mathbb{E}[|X_{n_k}|] \leq \sup_{n\geq 1} \mathbb{E}[|X_n|] < \infty.$$

Hence $X \in L^1$. Consequently $(Y_n)_n = (X_n - X)_n$ is UI (Exercise 2.3.13) and converges to 0 in probability, and we aim at showing that it converges in $L^1$. For any $\varepsilon, K > 0$, it holds

$$\begin{aligned}\mathbb{E}[|Y_n|] &\leq \mathbb{E}[|Y_n|\,\mathbb{1}_{|Y_n|\leq\varepsilon}] + \mathbb{E}[|Y_n|\,\mathbb{1}_{\varepsilon<|Y_n|\leq K}] + \mathbb{E}[|Y_n|\,\mathbb{1}_{|Y_n|>K}] \\ &\leq \varepsilon + K\,\mathbb{P}(|Y_n| > \varepsilon) + \sup_{n\geq 1} \mathbb{E}[|Y_n|\,\mathbb{1}_{|Y_n|>K}].\end{aligned}$$

Let us make $n \longrightarrow \infty$, then $K \longrightarrow \infty$, and finally $\varepsilon \longrightarrow 0$, then the probability at the last line tends to 0 since $Y_n \longrightarrow 0$ in probability and the expectation tends to 0 by the UI property.

Suppose next that $\mathbb{E}[|X_n - X|] \longrightarrow 0$. Recall that the Markov inequality implies that $X_n \longrightarrow X$ in probability. Also $\mathbb{E}[|X_n|] \leq \mathbb{E}[|X_n - X|] + \mathbb{E}[|X|] \longrightarrow \mathbb{E}[|X|]$ so $(X_n)_n$ is bounded in $L^1$. We focus on the UI property. Let us write for any $K, L > 0$ the simpler but clever inequalities:

$$\begin{aligned}|X_n|\,\mathbb{1}_{|X_n|>K} &\leq |X_n - X|\,\mathbb{1}_{|X_n|>K} + |X|\,\mathbb{1}_{|X_n|>K,|X|\leq L} + |X|\,\mathbb{1}_{|X_n|>K,|X|>L} \\ &\leq |X_n - X|\,\mathbb{1}_{|X_n|>K} + \frac{L}{K}\,|X_n|\,\mathbb{1}_{|X_n|>K,|X|\leq L} + |X|\,\mathbb{1}_{|X_n|>K,|X|>L} \\ &\leq |X_n - X| + \frac{L}{K}\,|X_n| + |X|\,\mathbb{1}_{|X|>L}.\end{aligned}$$

Consequently, with $L = \sqrt{K}$, we obtain

$$\mathbb{E}[|X_n|\,\mathbb{1}_{|X_n|>K}] \leq \mathbb{E}[|X_n - X|] + \frac{1}{\sqrt{K}} \sup_{n\geq 1} \mathbb{E}[|X_n|] + \mathbb{E}[|X|\,\mathbb{1}_{|X|>\sqrt{K}}].$$

Fix $\varepsilon > 0$ and let $N$ be large enough, so $\mathbb{E}[|X_n - X|] \leq \varepsilon$ for all $n \geq N$. Recall that $\sup_{n\geq 1} \mathbb{E}[|X_n|]$, that $X \in L^1$ is UI, and similarly that the finite collection $(X_n)_{n\leq N}$ is UI. Then for $K$ large enough, we have

$$\sup_{n\leq N} \mathbb{E}[|X_n|\,\mathbb{1}_{|X_n|>K}] \leq \varepsilon \qquad \text{and} \qquad \sup_{n>N} \mathbb{E}[|X_n|\,\mathbb{1}_{|X_n|>K}] \leq 3\varepsilon,$$

so indeed $(X_n)_n$ is UI. □

**Corollary 2.3.15** (Boundedness in $L^p$)**.** *Suppose that $X_n \longrightarrow X$ in probability and $\sup_n \mathbb{E}[|X_n|^p] < \infty$ for some $p > 1$. Then $X_n \longrightarrow X$ in $L^q$ for all $q \in [1, p)$.*

*Proof.* By Exercise 2.3.13, for any $q \in [1, p)$, the family $(|X_n|^q)_{n\geq 1}$ is bounded in $L^r$ for $r = p/q > 1$ so it is UI. Then Theorem 2.3.14 implies the convergence in $L^q$. □

**Remark 2.3.16.** Under the assumptions of Corollary 2.3.15, Fatou's lemma implies $\mathbb{E}[|X|^p] \leq \sup_n \mathbb{E}[|X_n|^p] < \infty$ but it may be the case that $X_n$ does not converge to $X$ in $L^p$. Just adapt a previous example and take $\mathbb{P}(X_n = n^{1/p}) = 1/n = 1 - \mathbb{P}(X_n = 0)$; in this case $X_n \longrightarrow 0$ in probability so if it converges in $L^p$, then the limit must be 0 a.s. by Lemma 2.3.2, but $\mathbb{E}[|X_n|^p] = 1$ for all $n$.

## 2.4 Law of Large Numbers

The Law of Large Numbers formally links the mathematical notion of expectation and probability of an event with the conceptual idea of asymptotic frequenct. We restrict ourselves to real-valued random variables, but it extends to vectors by applying it separately on each component.

Let us start with a weak version, where convergence holds in probability, we next prove the strong law, with an almost sure convergence. The $L^p$ part of the statement uses the notion of uniform integrability from Section 2.3.2.

**Theorem 2.4.1** (WLLN). *Let $(X_n)_{n\geq 1}$ be i.i.d. r.v.'s in $\mathbb{R}$ with finite mean $\mathbb{E}[X_1] = m \in \mathbb{R}$. Then*

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to\infty}{\longrightarrow} m$$

*in probability and in $L^1$. If $X_1 \in L^p$ for some $p > 1$, then the convergence also holds in $L^p$.*

*Proof.* The $L^2$ case is immediate: Suppose in addition that $\sigma^2 = \mathrm{Var}(X_1) < \infty$, then by independence,

$$\mathbb{E}\left[\left(\frac{X_1 + \cdots + X_n}{n} - m\right)^2\right] = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2} = \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0.$$

Thus $n^{-1}\sum_{k=1}^n X_k$ converges to $m$ in $L^2$ and so in probability as well.

Now suppose only that $X_1 \in L^1$. For every $K > 0$, for every $i \geq 1$, let us set

$$X_i^K = X_i\, \mathbb{1}_{|X_i|\leq K} \qquad \text{and} \qquad Y_i^K = X_i\, \mathbb{1}_{|X_i|>K}.$$

On the one hand, by Lemma 2.3.7,

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n Y_i^K\right|\right] \leq \mathbb{E}[|Y_1^K|] = \mathbb{E}[|X_1|\, \mathbb{1}_{|X_1|>K}] \underset{K\to\infty}{\longrightarrow} 0.$$

On the other hand,

$$|m - \mathbb{E}[X_1^K]| \leq \mathbb{E}[|X_1 - X_1\, \mathbb{1}_{|X_1|\leq K}|] = \mathbb{E}[|Y_1^K|] \underset{K\to\infty}{\longrightarrow} 0.$$

Furthermore, by the $L^2$ case, for every $K > 0$ fixed, we have

$$\frac{X_1^K + \cdots + X_n^K}{n} \underset{n\to\infty}{\longrightarrow} \mathbb{E}[X_1^K] \qquad \text{in } L^2 \text{ and thus in } L^1.$$

Hence

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}\left[\left|\frac{X_1 + \cdots + X_n}{n} - m\right|\right] &\leq \limsup_{n\to\infty} \mathbb{E}\left[\left|\frac{X_1^K + \cdots + X_n^K}{n} - \mathbb{E}[X_1^K]\right|\right] + |m - \mathbb{E}[X_1^K]| + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n Y_i^K\right|\right] \\ &\leq 2\,\mathbb{E}[|Y_1^K|] \underset{K\to\infty}{\longrightarrow} 0,\end{aligned}$$

so $(X_1 + \cdots + X_n)/n \longrightarrow m$ in $L^1$ and thus in probability.

Suppose now that $X_1 \in L^p$ for some $p > 1$. Since $n^{-1}\sum_{k=1}^n X_k$ converges to $m$ in probability, it suffices to prove that the sequence $(|n^{-1}\sum_{k=1}^n X_k|^p)_n$ is UI to deduce the $L^p$ convergence from Theorem 2.3.14. Let us use the characterisation from Proposition 2.3.9. Since the $X_n$'s have the same law, then by Exercise 2.3.13 the sequence $(|X_n|^p)_{n\geq 1}$ is UI. Then for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every $A \in \mathcal{F}$,

$$\text{if} \quad \mathbb{P}(A) \leq \delta \quad \text{then} \quad \sup_{k\geq 1} \mathbb{E}[|X_k|^p\, \mathbb{1}_A] \leq \varepsilon.$$

Using the Minkowski inequality, we infer that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^n X_k\right|^p \mathbb{1}_A\right]^{1/p} \leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}[|X_k|^p\, \mathbb{1}_A]^{1/p} \leq \varepsilon^{1/p},$$

and the sequence $(|n^{-1}\sum_{k=1}^n X_k|^p)_n$ is indeed UI. □

Let us next strengthen the convergence to an almost sure one.

**Theorem 2.4.2** (SLLN). *Let $(X_n)_{n\geq 1}$ be i.i.d. r.r.v.'s with $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = m \in \mathbb{R}$. Then*

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to\infty}{\longrightarrow} m$$

*a.s. and in $L^1$. If $X_1 \in L^p$ for some $p > 1$, then the convergence also holds in $L^p$.*

*Conversely, if $n^{-1}(X_1 + \cdots + X_n)$ converges a.s. to some limit $X$ which is a.s. finite, then $X$ is a.s. constant equal to some $m \in \mathbb{R}$ and the r.v.'s have finite mean $\mathbb{E}[X_1] = m$.*

Let us point out that the fact that the limit of $n^{-1}(X_1 + \cdots + X_n)$ has to be constant can be shown by Theorem 2.1.16.

*Proof.* The idea is to introduce a cut-off, namely write:

$$\frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\sum_{k=1}^{n}\big(X_k\,\mathbb{1}_{|X_k|\le k} - \mathbb{E}[X_k\,\mathbb{1}_{|X_k|\le k}]\big) + \frac{1}{n}\sum_{k=1}^{n} X_k\,\mathbb{1}_{|X_k|>k} + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k\,\mathbb{1}_{|X_k|\le k}]. \tag{2.1}$$

Let us prove that the first two terms tend to 0 a.s. while the last one tends to $m$. Indeed, first by monotone convergence applied to each expectation, we have:

$$\mathbb{E}[X_1\,\mathbb{1}_{|X_1|\le k}] = \mathbb{E}[X_1^+\,\mathbb{1}_{|X_1|\le k}] - \mathbb{E}[X_1^-\,\mathbb{1}_{|X_1|\le k}] \underset{k\to\infty}{\longrightarrow} \mathbb{E}[X_1^+] - \mathbb{E}[X_1^-] = m.$$

This implies further that

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_1\,\mathbb{1}_{|X_1|\le k}] \underset{n\to\infty}{\longrightarrow} m,$$

and since the $X_k$'s have the same law, then $\mathbb{E}[X_k\,\mathbb{1}_{|X_k|\le k}] = \mathbb{E}[X_1\,\mathbb{1}_{|X_1|\le k}]$ for each $k$, hence the convergence to $m$ of the third term in the right-hand side of (2.1).

Similarly:

$$\mathbb{E}\Big[\sum_{k\ge1}\mathbb{1}_{|X_k|>k}\Big] = \sum_{k\ge1}\mathbb{E}[\mathbb{1}_{|X_k|>k}] = \sum_{k\ge1}\mathbb{E}[\mathbb{1}_{|X_1|>k}] = \mathbb{E}\Big[\sum_{k\ge1}\mathbb{1}_{|X_1|>k}\Big] \le \mathbb{E}[|X_1|] < \infty.$$

Consequently $\sum_{k\ge1}\mathbb{1}_{|X_k|>k} < \infty$ a.s. so with probability one, only finitely many indices $k$ satisfy $|X_k| > k$ and in particular:

$$\frac{1}{n}\sum_{k=1}^{n} X_k\,\mathbb{1}_{|X_k|>k} \underset{n\to\infty}{\overset{a.s.}{\longrightarrow}} 0,$$

since the sum a.s. only contains finitely many nonzero terms.

It remains to take care of the first term in our decomposition (2.1). Let us put

$$Y_k = X_k\,\mathbb{1}_{|X_k|\le k} - \mathbb{E}[X_k\,\mathbb{1}_{|X_k|\le k}] \qquad \text{and} \qquad Z_n = \sum_{k=1}^{n}\frac{Y_k}{k}.$$

Let us prove that $n^{-1}\sum_{k=1}^{n} Y_k$ converges to 0 a.s. We shall rely on Kronecker's lemma, namely write $Y_n = n(Z_n - Z_{n-1})$, so

$$\frac{1}{n}\sum_{k=1}^{n} Y_k = \frac{1}{n}\sum_{k=1}^{n} k(Z_k - Z_{k-1}) = \frac{1}{n}\left(\sum_{k=1}^{n} kZ_k - \sum_{k=1}^{n}(k-1)Z_{k-1} - \sum_{k=1}^{n} Z_{k-1}\right) = Z_n - \frac{1}{n}\sum_{k=1}^{n} Z_{k-1}.$$

We shall prove that a.s. $Z_n$ converges to a finite limit, which implies that $n^{-1}\sum_{k=1}^{n} Z_{k-1}$ converges to the same limit, which finally implies that $n^{-1}\sum_{k=1}^{n} Y_k$ converges to 0. Note that the $Y_k$'s are independent and have $\mathbb{E}[Y_k] = 0$, and each $Y_k$ is bounded (by $2k$), so

$$\mathbb{E}\Big[\Big(\sum_{k\ge1}\frac{Y_k}{k}\Big)^2\Big] = \sum_{k\ge1}\frac{\mathrm{Var}(Y_k)}{k^2} \le \sum_{k\ge1}\frac{1}{k^2}\,\mathbb{E}[X_1^2\,\mathbb{1}_{|X_1|\le k}] = \mathbb{E}\Big[X_1^2\sum_{k\ge1}\frac{1}{k^2}\,\mathbb{1}_{|X_1|\le k}\Big].$$

Now for any $k \ge 1$ and $t \in [k, k+1]$ we have $t^2 \le (k+1)^2 \le (2k)^2 = 4k^2$, so for any $x \in \mathbb{R}$, it holds:

$$\sum_{k\ge1}\frac{1}{k^2}\,\mathbb{1}_{|x|\le k} \le \sum_{k\ge1}\int_k^{k+1}\frac{4}{t^2}\,\mathbb{1}_{|x|\le k}\,\mathrm{d}t = \int_{\max(|x|,1)}^{\infty}\frac{4}{t^2}\,\mathrm{d}t = \frac{4}{\max(|x|,1)}.$$

Consequently, since $X_1 \in L^1$, then, combining the last to displays, we have

$$\mathbb{E}\Big[\Big(\sum_{k\ge1}\frac{Y_k}{k}\Big)^2\Big] \le 4\,\mathbb{E}\Big[\frac{X_1^2}{\max(|X_1|,1)}\Big] \le 4\,\mathbb{E}[1+|X_1|] < \infty.$$

This implies that a.s. the series $\sum_{k\geq 1} k^{-1} Y_k$ is convergent, namely that $Z_n$ has a finite limit as we wanted.

Let us finish with the converse implication: assume $n^{-1}(X_1 + \cdots + X_n)$ converges a.s. to some limit $X$ which is a.s. finite and let us prove that the r.v.'s have finite mean and $X = \mathbb{E}[X_1]$ a.s. By our assumption, we have

$$\frac{X_n}{n} = \frac{X_1 + \cdots + X_n}{n} - \frac{X_1 + \cdots + X_{n-1}}{n} \xrightarrow[n\to\infty]{a.s.} X - X = 0.$$

In particular, a.s. for $n$ large enough we have $|X_n| \leq n$, that is $\mathbb{P}(\limsup_n \{|X_n| \geq n\}) = 0$. Since the $X_n$'s are independent, then this probability is either 0 or 1 by the Borel–Cantelli lemma, according as wether the series of the probabilities converges or not. We infer that

$$\mathbb{E}[|X_1|] \leq \sum_{n\geq 1} \mathbb{P}(|X_1| + 1 \geq n) = 1 + \sum_{n\geq 1} \mathbb{P}(|X_n| \geq n) < \infty.$$

Hence $X_1 \in L^1$ and thus $n^{-1}(X_1 + \cdots + X_n) \longrightarrow \mathbb{E}[X_1]$ a.s. by the first part of the proof. Since we asuume that $n^{-1}(X_1 + \cdots + X_n) \longrightarrow X$ a.s. then we conclude that $X = \mathbb{E}[X_1]$ a.s. □

The almost sure convergence remains true in the case of infinite (but well-defined!) mean.

**Corollary 2.4.3.** *Let $(X_n)_{n\geq 1}$ be i.i.d. r.r.v.'s with $\mathbb{E}[X_1^-] < \infty$ and $\mathbb{E}[X_1^+] = \infty$ so we can make sense of $\mathbb{E}[X_1] = \infty$. Then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \infty.$$

*Proof.* Fix $K > 0$, then the r.v.'s $(\max\{X_i, K\})_i$ are i.i.d. with finite mean so by the previous strong law,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \frac{1}{n}\sum_{i=1}^{n} (\max\{X_i, K\}) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[\max\{X_1, K\}].$$

By monotone convergence, the right-hand side further converges to $\mathbb{E}[X_1] = \infty$ as $K \longrightarrow \infty$. □

Finally, if the mean is not defined, then three possible cases may occur.

**Proposition 2.4.4.** *Let $(X_n)_{n\geq 1}$ be i.i.d. r.r.v.'s with both $\mathbb{E}[X_1^-] = \infty$ and $\mathbb{E}[X_1^+] = \infty$, then a.s.*

$$\textit{either} \qquad \liminf_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = -\infty \qquad \textit{or} \qquad \limsup_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \infty.$$

*Proof.* Indeed fix any integer $K \geq 1$. Then

$$\sum_{n\geq 1} \mathbb{P}(K^{-1} X_n^+ \geq n) = \sum_{n\geq 1} \mathbb{P}(K^{-1} X_1^+ \geq n) \geq \mathbb{E}[K^{-1} X_1^+] = \infty.$$

Since the r.v.'s $X_n^+$ are independent, then the Borel–Cantelli lemma shows that a.s. the events $\{X_n^+ \geq Kn\}$ occur for infinitely many indices $n$. Then obviously so do the events $\{X_n \geq Kn\}$. Now for any such index $n$, we have either $X_1 + \cdots + X_{n-1} \leq -Kn/2$ or $X_1 + \cdots + X_{n-1} \geq -Kn/2$ and in this second case $X_1 + \cdots + X_n \geq Kn/2$. We infer that a.s.

$$\text{either} \qquad \liminf_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \leq -\frac{K}{2} \qquad \text{or} \qquad \limsup_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \geq \frac{K}{2}.$$

Then finally this holds a.s. simultaneously for all integers $K$ since there are countably many of them. □

Let us mention that that actually, depending on the law of $X$, either $n^{-1}(X_1 + \cdots + X_n) \longrightarrow \infty$ a.s. or $n^{-1}(X_1 + \cdots + X_n) \longrightarrow -\infty$ a.s. or we have both $\limsup n^{-1}(X_1 + \cdots + X_n) = \infty$ and $\liminf n^{-1}(X_1 + \cdots + X_n) = -\infty$ a.s. and we have necessary and sufficient conditions to tell which case occurs. The interested reader can look at the original article by Erickson "The strong law of large numbers when the mean is undefined" at `https://www.ams.org/journals/tran/1973-185-00/S0002-9947-1973-0336806-5/home.html`.

## 2.5 Convergence in distribution

In this section, the random variables all take values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (but let us mention that this generalises to separable and complete metric spaces) but may be defined on different probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $C_b = C_b(\mathbb{R}^d, \mathbb{R})$ the set of continuous and bounded functions from $\mathbb{R}^d$ to $\mathbb{R}$.

### 2.5.1 Definitions and first properties

**Definition 2.5.1** (Weak convergence of measures)**.** Let $\mu$ and $(\mu_n)_{n\geq 1}$ be probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The sequence $(\mu_n)_{n\geq 1}$ is said to converge *weakly* (or *narrowly*) to $\mu$ when

$$\mu_n(f) = \int f \, \mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int f \, \mathrm{d}\mu = \mu(f) \qquad \text{for all } f \in C_b.$$

A sequence $(X_n)_{n\geq 1}$ of r.v.'s in $\mathbb{R}^d$ is said to converge *in distribution* to a r.v. $X$ when their laws converge weakly i.e. when

$$\mathbb{E}[f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f(X)] \qquad \text{for all } f \in C_b.$$

**Remark 2.5.2.** Note that speaking of convergence of r.v.'s is an abuse of language since only their laws converge. For example, if $X_n$ has the same law as $-X_n$, then they both converge to the same limit so one cannot simply take sums and products through this notion of convergence.

Convergence in distribution is the weakest notion we have seen, as shown in the next result.

**Proposition 2.5.3.** *Suppose that the r.v.'s are defined on the same probability space. Then*

*(i) If $X_n \to X$ in probability, then $X_n \to X$ in distribution.*

*(ii) If $X_n \to X$ in distribution and $X$ is a.s. constant, then $X_n \to X$ in probability.*

*Proof.* (i) Suppose that $X_n \to X$ in probability and let $f$ be continuous and bounded, then $f(X_n) \to f(X)$ in probability, and then by dominated convergence, $\mathbb{E}[f(X_{n_k})] \to \mathbb{E}[f(X)]$.

(ii) Suppose that $X_n \to X$ in distribution and $\mathbb{P}(X = c) = 1$ for some $c \in \mathbb{R}^d$. Fix $\varepsilon > 0$ and let $f_\varepsilon$ be a continuous and bounded function that satisfies $f_\varepsilon(x) = 1$ when $|x - c| > \varepsilon$ and $f_\varepsilon(x) = 0$ when $|x - c| \leq \varepsilon/2$. Then

$$\mathbb{P}(|X_n - c| > \varepsilon) \leq \mathbb{E}[f_\varepsilon(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[f_\varepsilon(c)] = 0.$$

□

Viewing a probability measure on $\mathbb{R}^d$ as a function $\mathcal{B}(\mathbb{R}^d) \to [0,1]$, a natural notion of convergence would be a pointwise convergence, i.e. $\mu_n \to \mu$ when $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$. This is actually quite strong because of boundary effects, for example, if $U_n$ has the uniform distribution on $\{k/n : 1 \leq k \leq n\}$ and $U$ the uniform distribution on $[0,1]$, then for $f$ continuous and bounded,

$$\mathbb{E}[f(U_n)] = \frac{1}{n}\sum_{k=1}^{n} f(k) \underset{n\to\infty}{\longrightarrow} \int_0^1 f(x)\,\mathrm{d}x = \mathbb{E}[f(U)]$$

so $U_n \to U$ in distribution, however

$$\mathbb{P}(U_n \in \mathbb{Q}) = 1 \qquad \text{whereas} \qquad \mathbb{P}(U \in \mathbb{Q}) = 0.$$

The next result compares this notion with the weak convergence.

**Theorem 2.5.4** (Portmanteau)**.** *Let $\mu$ and $(\mu_n)_{n\geq 1}$ be probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then the following are equivalent:*

*(i) $\mu_n \to \mu$ weakly,*

*(ii) For every open set $O \subset \mathbb{R}^d$ it holds* $\liminf_n \mu_n(O) \geq \mu(O)$,

*(iii) For every closed set $C \subset \mathbb{R}^d$ it holds* $\limsup_n \mu_n(C) \leq \mu(C)$,

*(iv) For every Borel set $B \subset \mathbb{R}^d$* such that $\mu(\partial B) = 0$ *it holds* $\lim \mu_n(B) = \mu(B)$.

*Proof.* Suppose that $\mu_n \longrightarrow \mu$ weakly and fix an open set $O$. For every $k \geq 1$ define the function

$$f_k(x) = \min\{kd(x, O^c), 1\}$$

which is continuous and bounded on $\mathbb{R}^d$. Moreover as $k \longrightarrow \infty$ it increases to the function $\mathbb{1}_O$. Thus

$$\liminf_{n\to\infty} \mu_n(O) \geq \liminf_{n\to\infty} \mu_n(f_k) = \mu(f_k) \underset{k\to\infty}{\longrightarrow} \mu(O),$$

by monotone convergence.

The 2nd and 3rd items are clearly equivalent by taking the complement. Suppose that they both hold and fix $B$ with $\mu(\partial B) = 0$. Let $O \subset B$ denote is interior and $C \supset B$ its closure, so $\partial B = C \setminus O$. Then we know that

$$\liminf_{n\to\infty} \mu_n(B) \geq \liminf_{n\to\infty} \mu_n(O) \geq \mu(O) \qquad \text{and} \qquad \limsup_{n\to\infty} \mu_n(B) \leq \limsup_{n\to\infty} \mu_n(C) \leq \mu(C),$$

and furthermore $\mu(C) = \mu(O) + \mu(C \setminus O) = \mu(O)$ so it also equals $\mu(B)$ and thus

$$\lim_{n\to\infty} \mu_n(B) = \mu(B).$$

Finally, suppose $\lim \mu_n(B) = \mu(B)$ whenever $\mu(\partial B) = 0$ and fix $f \in C_b$. Replacing $f$ by $f - \inf f$ if necessary, let us assume that $f \in [0, K]$ for some $K > 0$. Observe that for any $x \in \mathbb{R}^d$ we have

$$f(x) = \int_0^K \mathbb{1}_{t\leq f(x)} \, \mathrm{d}t$$

so by Fubini's theorem, if we let $A_t^f = \{x \in \mathbb{R}^d : f(x) \geq t\}$, then

$$\mu(f) = \int_{\mathbb{R}^d} f(x)\mu(\mathrm{d}x) = \int_0^K \left( \int_{\mathbb{R}^d} \mathbb{1}_{t\leq f(x)} \mu(\mathrm{d}x) \right) \mathrm{d}t = \int_0^K \mu(A_t^f) \, \mathrm{d}t.$$

Note that $\partial A_t^f = \{x \in \mathbb{R}^d : f(x) = t\}$ which are disjoint sets for different values of $t$. Therefore, for any $k \geq 1$, since $\mu(\mathbb{R}^d) = 1$, then there can only be at most $k$ values of $t$ for which $\mu(\partial A_t^f) \geq 1/k$. Thus the set $D = \{t \in [0, K] : \mu(\partial A_t^f) \neq 0\}$ is at most countable and in particular has zero Lebesgue measure. Now for any $t \in D^c$ we have $\mu(\partial A_t^f) = 0$ so $\mu(A_t^f) = \lim_n \mu_n(A_t^f)$. By dominated convergence,

$$\mu_n(f) = \int_0^K \mu_n(A_t^f) \, \mathbb{1}_{t\in D^c} \, \mathrm{d}t \underset{n\to\infty}{\longrightarrow} \int_0^K \mu(A_t^f) \, \mathbb{1}_{t\in D^c} \, \mathrm{d}t = \mu(f),$$

thus $\mu_n \longrightarrow \mu$ weakly. □

In practice, it may be useful to check the convergence of integrals of even more restrictive functions than all continuous and bounded ones. The next result shows many possibilities. It is very important here that both $\mu_n$ and $\mu$ are probability measures, even if their are finite, it fails if they do not have the same total mass.

**Theorem 2.5.5** (Restriction of test functions)**.** *Let $\mu_n, \mu$ be probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then the following are equivalent:*

*(i) $\mu_n(f) \longrightarrow \mu(f)$ for all $f$ continuous and bounded,*

*(ii) $\mu_n(f) \longrightarrow \mu(f)$ for all $f$ continuous and with compact support,*

*(iii)* $\mu_n(f) \longrightarrow \mu(f)$ *for all* $f \in H \subset C_b$ *such that the closure of* $H$ *for the sup norm contains all continuous functions with compact support,*

*(iv)* $\mu_n(f) \longrightarrow \mu(f)$ *for all* $f$ *uniformly continuous and bounded,*

*(v)* $\mu_n(f) \longrightarrow \mu(f)$ *for all* $f$ *Lipschitz and bounded,*

*(vi)* $\mu_n(f) \longrightarrow \mu(f)$ *for all* $f$ *continuous* $\mu$*-a.s. and bounded.*

*Proof.* Clearly, (i) $\implies$ (iv) $\implies$ (v). Suppose (v) holds, and recall the sequence of functions

$$f_k(x) = \min\{kd(x, O^c), 1\}$$

where $k \geq 1$ and $O$ is a fixed open set. Then this function is actually $k$-Lipschitz and the argument from the proof of Theorem 2.5.4 shows that the convergence $\mu_n(k) \longrightarrow \mu(k)$ for all $k$ implies $\liminf_n \mu_n(O) \geq \mu(O)$, which implies (i).

On the other hand, clearly, (i) $\implies$ (ii) $\implies$ (iii). Suppose (iii) holds and let $f$ be a continuous function with compact support. Then $f$ is the limit for the sup norm of a sequence of functions in $H$ so we can build a sequence $(f_k)_k$ in $H$ such that $\|f - f_k\|_\infty \leq 1/k$ for all $k \geq 1$. We infer that

$$\begin{aligned}
\limsup_{n\to\infty} |\mu_n(f) - \mu(f)| &\leq \limsup_{n\to\infty} |\mu_n(f) - \mu_n(f_k)| + |\mu_n(f_k) - \mu(f_k)| + |\mu(f_k) - \mu(f)| \\
&\leq 2\,\|f - f_k\|_\infty + \limsup_{n\to\infty} |\mu_n(f_k) - \mu(f_k)| \\
&\leq 2/k \underset{k\to\infty}{\longrightarrow} 0,
\end{aligned}$$

hence (iii) $\implies$ (ii).

Suppose next that (ii) holds and fix $f$ continuous and bounded. Let $(g_k)_k$ be continuous functions with compact support which satisfy $0 \leq g_k \leq 1$ and $g_k \uparrow 1$. Then $fg_k$ is continuous with compact support and $fg_k \uparrow f$ so

$$\begin{aligned}
\limsup_{n\to\infty} |\mu_n(f) - \mu(f)| &\leq \limsup_{n\to\infty} |\mu_n(f) - \mu_n(fg_k)| + |\mu_n(fg_k) - \mu(fg_k)| + |\mu(fg_k) - \mu(f)| \\
&\leq 2\,\|f\|_\infty \limsup_{n\to\infty} (1 - \mu_n(g_k)) \\
&\leq 2\,\|f\|_\infty (1 - \mu(g_k)),
\end{aligned}$$

which further converges to 0 as $k \longrightarrow \infty$ by monotone convergence.

Finally, (vi) $\implies$ (i) so it remains to prove the converse implication. Suppose that (i) holds, fix $f$ continuous $\mu$-a.s. and bounded and fix $\varepsilon > 0$. Let $K > 0$ be such that $|f| \leq K$ and note that as in the previous proof, for any $k \geq 1$ there can only be at most $k$ different values of $t \in \mathbb{R}$ such that $\mu(\{f = t\}) = \int \mathbb{1}_{f(x)=t}\, \mu(\mathrm{d}x) \geq 1/k$ so the set $D = \{t \in \mathbb{R} : \mu(\{f = t\}) > 0\} \subset [-K, K]$ is at most countable. Then there exists $k \geq 1$ and values $a_0 < \cdots < a_k$ such that: $a_0 < -K$, $a_k > K$, and $a_i - a_{i-1} \leq \varepsilon$ and $a_i \in D^c$ for all $i \leq k$. Define then $A_i = \{x \in \mathbb{R}^d : a_{i-1} < f(x) \leq a_i\}$ so $\partial A_i \subset \{x \in \mathbb{R}^d : f(x) \in \{a_{i-1}, a_i\}\} \cup C_f^c$ where we recall the notation $C_f = \{x \in \mathbb{R}^d : f \text{ is continuous at } x\}$. By our assumption $\mu(C_f^c) = 0$ and by our construction of $A_i$ we get $\mu(\partial A_i) = 0$ so $\mu_n(A_i) \longrightarrow \mu(A_i)$ for each $i \leq k$ and so

$$\sum_{i=1}^k a_i \mu_n(A_i) \underset{n\to\infty}{\longrightarrow} \sum_{i=1}^k a_i \mu(A_i).$$

Finally, from the construction we have $f \leq \sum_{i=1}^k a_i\, \mathbb{1}_{A_i} \leq f + \varepsilon$ so we infer that

$$\limsup_{n\to\infty} \mu_n(f) \leq \sum_{i=1}^k a_i \mu(A_i) \leq \mu(f) + \varepsilon \qquad \text{and} \qquad \liminf_{n\to\infty} \mu_n(f) \geq \sum_{i=1}^k a_i \mu(A_i) \geq \mu(f) - \varepsilon,$$

so $\mu_n(f) \longrightarrow \mu(f)$ since $\varepsilon > 0$ is arbitrary. □

Thanks to the last item, we can generalise easily Lemma 2.3.4 to convergences in distribution.

**Lemma 2.5.6** (Continuous mapping)**.** *Suppose that $X_n \longrightarrow X$ in distribution and that $f : \mathbb{R}^d \longrightarrow \mathbb{R}^e$ is continuous $\mathbb{P}_X$-a.s. then $f(X_n) \longrightarrow f(X)$ in distribution.*

*Proof.* Let $g$ be a continuous and bounded function, then $g \circ f$ is bounded and continuous $\mathbb{P}_X$-a.s. so by Theorem 2.5.5 we have

$$\mathbb{E}[g(f(X_n))] = \mathbb{E}[g \circ f(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[g \circ f(X)] = \mathbb{E}[g(f(X))],$$

i.e. $f(X_n) \longrightarrow f(X)$ in distribution. □

### 2.5.2 Distribution functions (⋆)

Although the discussion in this subsection can be made in $\mathbb{R}^d$, let us restrict to $\mathbb{R}$ for the sake of clarity. Recall that a function $F : \mathbb{R} \longrightarrow [0,1]$ is a *distribution function* when it is nondecreasing, right-continuous, and such that $F(x) \longrightarrow 0$ as $x \longrightarrow -\infty$ and $F(x) \longrightarrow 1$ as $x \longrightarrow \infty$. Recall that for any function $F$ we let $C_F = \{x \in \mathbb{R} : F \text{ is continuous at } x\}$ denote its continuity set. Recall finally that every r.r.v. $X$ has a distribution function $F_X : x \mapsto \mathbb{P}(X \leq x)$ and that for each distribution function $F$, there exists a r.r.v. $X$ such that $F = F_X$.

**Definition 2.5.7** (Weak convergence of distribution functions)**.** A sequence $(F_n)_{n\geq 1}$ of distribution function is said to converge *weakly* to a distribution function $F$ when

$$F_n(x) \underset{n\to\infty}{\longrightarrow} F(x) \qquad \text{for all } x \in C_F.$$

As for discontinuity points, basically one has to decide wether we take the left-limit or the right-limit.

**Example 2.5.8.** If $F_n(x) = (1 - \mathrm{e}^{-\lambda_n x})\,\mathbb{1}_{x>0}$ with $\lambda_n \longrightarrow \infty$, then $F_n(x) \longrightarrow \mathbb{1}_{x>0}$ for all $x \in \mathbb{R}$ and the limit is left-continuous. However $F_n$ converges weakly to $F : x \mapsto \mathbb{1}_{x\leq 0}$ which is a distribution function.

**Proposition 2.5.9.** *We have $X_n \longrightarrow X$ in distribution if and only if $F_{X_n} \longrightarrow F_X$ weakly.*

*Proof.* Note that for every $x \in \mathbb{R}$ we have $\partial(-\infty, x] = \{x\}$ and $\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x) = F(x) - F(x-)$ so the direct implication is a particular case of Theorem 2.5.4. For the converse implication, suppose that $F_{X_n} \longrightarrow F_X$ weakly and let $a < b$. Then

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}_{X_n}((a,b)) &= \liminf_{n\to\infty} \big(F_{X_n}(b-) - F_{X_n}(a)\big)\\ &\geq \liminf_{n\to\infty} F_{X_n}(b-) - \limsup_{n\to\infty} F_{X_n}(a+)\\ &\geq F_X(b-) - F_X(a) = \mathbb{P}_X((a,b)).\end{aligned}$$

Recall that any open set of $\mathbb{R}$ is a countable union of disjoint open interval, say $O = \bigcup_k (a_k, b_k)$, then

$$\liminf_{n\to\infty} \mathbb{P}_{X_n}(O) = \liminf_{n\to\infty} \sum_k \mathbb{P}_{X_n}((a_k,b_k)) \geq \sum_k \liminf_{n\to\infty} \mathbb{P}_{X_n}((a_k,b_k)) \geq \sum_k \mathbb{P}_X((a_k,b_k)) = \mathbb{P}_X(O).$$

By Theorem 2.5.4 again, this is equivalent to $X_n \longrightarrow X$ in distribution. □

The next theorem is quite useful in practice, but one should not misunderstand its statement.

**Theorem 2.5.10** (Skorokhod's representation)**.** *Suppose that $F_n, F$ are distribution functions such that $F_n \longrightarrow F$ weakly. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and r.v.'s $X_n, X$ all defined on it, which have distribution function $F_n, F$ respectively and such that $X_n \longrightarrow X$ a.s.*

*Proof.* Recall from Theorem 2.1.3 that given the distribution function $F$ we can construct $X$ on the space $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}(0,1), \text{Leb})$ by setting for any $u \in (0,1)$

$$X(u) = G(u) = \inf\{t \in \mathbb{R} : \ F(t) > u\} = \sup\{t \in \mathbb{R} : \ F(t) \leq u\}.$$

Define also

$$H(u) = \inf\{t \in \mathbb{R} : \ F(t) \geq u\} = \sup\{t \in \mathbb{R} : \ F(t) < u\} \geq G(u).$$

Define similarly $G_n$, $H_n$, and $X_n$ from $F_n$.

Fix $u \in (0,1)$. For $t < G(u)$ such that $F$ is continuous at $t$ we have $u > F(t) = \lim_n F_n(t)$ so for every $n$ large enough, $F_n(t) < u$ and thus $t \leq G_n(u)$, i.e. $\liminf_n G_n(u) \geq t$. Recall that $F$ has at most countably many discontinuity points so there exists a sequence of such $t$'s converging to $G(u)$, and passing to the limit, we get $\liminf_{n\to\infty} G_n(u) \geq G(u)$. The same reasoning with $t > H(u)$ leads to $\limsup_n H_n(u) \leq H(u)$. Thus for every $u \in (0,1)$,

$$G(u) \leq \liminf_{n\to\infty} G_n(u) \leq \limsup_{n\to\infty} G_n(u) \leq \limsup_{n\to\infty} H_n(u) \leq H(u).$$

Notice that $(G(u), H(u))$ is the largest open interval $(a,b)$ such that $F(t) = F(u)$ for all $t \in (a,b)$ so these intervals are either disjoint or equal for different values of $u$. In particular there can only be at most countably many non empty ones (since each one contains a rational number) so the set $\Omega_0 = \{u \in (0,1) : \ G(u) < F(u)\}$ is countable and in particular has Lebesgue measure 0. We conclude that for every $u \in \Omega \setminus \Omega_0$, which has probability 1, we have $X_n(u) = G_n(u) \longrightarrow G(u) = X(u)$. □

If one starts with r.v.'s $X_n, X$ in the first place with $X_n \longrightarrow X$ in distribution, then the theorem states that there exists another probability space with new r.v.'s $X'_n, X'$ defined on it, with the same law as $X_n, X$ and such that $X'_n \longrightarrow X'$ a.s. This does *not* mean that $X_n \longrightarrow X$ a.s!

**Example 2.5.11.** To see how this reasoning works, let us give another proof of Lemma 2.5.6. Suppose that $X_n \longrightarrow X$ in distribution and that $f : \mathbb{R}^d \longrightarrow \mathbb{R}^e$ is continuous $\mathbb{P}_X$-a.s. Then there exist $X'_n, X'$ with the same law as $X_n, X$ and such that $X'_n \longrightarrow X'$ a.s. Then $f(X'_n) \longrightarrow f(X')$ a.s. by Lemma 2.3.4 and so $f(X'_n) \longrightarrow f(X')$ in distribution. Since $f(X'_n), f(X')$ have the same law as $f(X_n), f(X)$ respectively, we conclude that $f(X_n) \longrightarrow f(X)$ in distribution.

Similarly, any theorem which assumes that $X_n \longrightarrow X$ a.s. and conclude about the behaviour of quantities of the form $\mathbb{E}[f(X_n)]$ also generalises to assuming only $X_n \longrightarrow X$ in distribution.

## 2.6 Characteristic functions

In this section, the random variables all take values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Recall that we denote by $\langle \cdot, \cdot \rangle$ the scalar product in $\mathbb{R}^d$, also $|\cdot|$ denotes the associated square norm (as well as the modulus in $\mathbb{C}$). We use $t$ to denote a real number and $u$ for a vector in $\mathbb{R}^d$. We let $(u_1, \dots, u_d)$ denote the coordinates of $u$ and $(u^n)_{n\geq 1}$ denote a sequence of vectors $(u_1^n, \dots, u_d^n)$. We stress that we use the line notation for vectors when writing in the text, but think of them as columns when it comes to matrix operations.

### 2.6.1 The characteristic function

**Definition 2.6.1.** The *characteristic function* of a r.v. $X$ in $\mathbb{R}^d$ is the function $\varphi_X : \ \mathbb{R}^d \longrightarrow \mathbb{C}$ defined by

$$\varphi_X(u) = \mathbb{E}[e^{i\langle u, X\rangle}] = \mathbb{E}[e^{i\sum_{k=1}^d u_k X_k}].$$

It is well-defined since $x \mapsto e^{i\langle u,x\rangle}$ is continuous and bounded for every given $u \in \mathbb{R}^d$.

We leave the proof of the basic properties as an exercise.

**Proposition 2.6.2.** *The characteristic function of $X$ satisfies the following properties:*

*(i)* $\varphi_X(0) = 1$,

*(ii)* $\varphi_X(-u) = \overline{\varphi_X(u)}$ *for every* $u \in \mathbb{R}^d$,

*(iii)* $|\varphi_X(u)| \leq 1$ *for every* $u \in \mathbb{R}^d$,

*(iv)* $|\varphi_X(u+h) - \varphi_X(u)| \leq \mathbb{E}[|e^{i\langle h,X\rangle} - 1|]$ *for every* $u, h \in \mathbb{R}^d$ *so* $\varphi_X$ *is uniformly continuous.*

*(v) If* $X$ *has dimension* $d_2$*, then for every* $d_1 \times d_2$ *matrix* $C$ *and vectors* $u, v \in \mathbb{R}^{d_1}$*, we have*

$$\varphi_{CX+v}(u) = \mathbb{E}[e^{i\langle u, CX+v\rangle}] = \mathbb{E}[e^{i\langle u,v\rangle + i\langle C^t u, X\rangle}] = e^{i\langle u,v\rangle}\, \varphi_X(C^t u).$$

*In particular,*

*(a) If* $d_1 = d_2 = d$ *and* $C = aI_d$ *with* $a \in \mathbb{R}$*, then* $\varphi_{aX+v}(u) = e^{i\langle u,v\rangle}\, \varphi_X(au)$.

*(b) Also, if* $d_1 = 1$*, then* $C = c \in \mathbb{R}^{d_2}$ *and* $CX = \langle c, X\rangle$*, and so for every* $s, t \in \mathbb{R}$*, we have* $\varphi_{\langle c,X\rangle + s}(t) = e^{ist}\, \varphi_X(tc)$.

*(vi) If* $X^1, \dots, X^n$ *are independent random vectors with dimension* $d_1, \dots, d_n$ *respectively, then for all* $u^1 \in \mathbb{R}^{d_1}, \dots, u^n \in \mathbb{R}^{d_n}$*, we have*

$$\varphi_{(X^1,\dots,X^n)}((u^1, \dots, u^n)) = \mathbb{E}\Big[\prod_{k=1}^{n} e^{i\langle u^k, X^k\rangle}\Big] = \prod_{k=1}^{n} \varphi_{X^k}(u^k).$$

*Combined with the previous item, if* $d_1 = \cdots = d_n = d$*, then for every* $u \in \mathbb{R}^d$*, we have*

$$\varphi_{X^1+\cdots+X^n}(u) = \prod_{k=1}^{n} \varphi_{X^k}(u).$$

*Indeed,* $X^1 + \cdots + X^n = CX$ *where* $C$ *is the* $d \times dn$ *matrix with all entries equal to* 1 *and* $X = (X^1, \dots, X^n)$ *is the concatenation of the* $X^k$*'s.*

The characteristic function corresponds to a Fourier transform of the law of the random vector. For explicit calculations, by the transfer lemma, if $X$ takes countably many values, say in $\mathbb{Z}^d$ for example, then

$$\varphi_X(u) = \sum_{x \in \mathbb{Z}^d} e^{i\langle u,x\rangle}\, \mathbb{P}(X = x),$$

whereas if $X$ admits a density $f_X$ with respect to the Lebesgue measure in $\mathbb{R}^d$, then

$$\varphi_X(u) = \int_{\mathbb{R}^d} e^{i\langle u,x\rangle} f_X(x)\, \mathrm{d}x.$$

The next exercise is treated in the exercise sheet.

**Example 2.6.3.** The characteristic function of a Gaussian r.v. $Z \sim \mathcal{N}(m, \sigma^2)$ is given for every $t \in \mathbb{R}$ by

$$\varphi_Z(t) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(itz - \frac{(z-m)^2}{2\sigma^2}\Big)\, \mathrm{d}z = \exp\Big(itm - \frac{t^2\sigma^2}{2}\Big).$$

More generally, if $Z_1, \dots, Z_d$ are i.i.d. with the law $\mathcal{N}(0, \sigma^2)$, by independence we get for every $u \in \mathbb{R}^d$:

$$\int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\Big(i\langle u, z\rangle - \frac{|z|^2}{2\sigma^2}\Big)\, \mathrm{d}z = \exp\Big(-\frac{|u|^2\sigma^2}{2}\Big). \tag{2.2}$$

We shall use this identity now to prove the main result of this section, which is that the characteristic function does indeed characterise the law.

**Theorem 2.6.4.** *If $X$ and $Y$ have the same characteristic function then they have the same law.*

*Proof.* Let $X$ and $Y$ both have characteristic function $\varphi$ and independently, let $Z = (Z_1, \dots, Z_d)$ be i.i.d. standard Gaussian r.r.v.'s. The key point is to prove that $X + Z/k$ has the same law as $Y + Z/k$ for any given $k \geq 1$. Indeed, for all measurable and nonnegative functions $g$ we have

$$\begin{aligned}\mathbb{E}[g(X + Z/k)] &= \int_{\mathbb{R}^d \times \mathbb{R}^d} g(x + z/k)\, \mathbb{P}_X(\mathrm{d}x) \otimes \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{|z|^2}{2}\Big)\, \mathrm{d}z \\ &\overset{\text{Fubini}}{=} \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}^d} g(x + z/k) \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{|z|^2}{2}\Big)\, \mathrm{d}z \right) \mathbb{P}_X(\mathrm{d}x) \\ &\overset{y = x + z/k}{=} \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}^d} g(y) \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{k^2\,|x - y|^2}{2}\Big) k^d\, \mathrm{d}y \right) \mathbb{P}_X(\mathrm{d}x).\end{aligned}$$

By (2.2) the exponential term can be rewritten as

$$\begin{aligned}\exp\Big(-\frac{k^2\,|x - y|^2}{2}\Big) &= \int_{\mathbb{R}^d} \frac{1}{(2\pi k^2)^{d/2}} \exp\Big(i\langle x - y, z\rangle - \frac{|z|^2}{2k^2}\Big)\, \mathrm{d}z \\ &= \int_{\mathbb{R}^d} \frac{1}{(2\pi k^2)^{d/2}} \exp\Big(i\langle x, z\rangle\Big) \exp\Big(-i\langle y, z\rangle - \frac{|z|^2}{2k^2}\Big)\, \mathrm{d}z.\end{aligned}$$

Using Fubini's theorem again, we arrive at

$$\begin{aligned}\mathbb{E}[g(X + Z/k)] &= \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}^d} g(y) \left( \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(i\langle x, z\rangle) \exp\Big(-i\langle y, z\rangle - \frac{|z|^2}{2k^2}\Big)\, \mathrm{d}z \right) \mathrm{d}y \right) \mathbb{P}_X(\mathrm{d}x) \\ &= \int_{\mathbb{R}^d} g(y) \left( \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \varphi(z) \exp\Big(-i\langle y, z\rangle - \frac{|z|^2}{2k^2}\Big)\, \mathrm{d}z \right) \mathrm{d}y.\end{aligned}$$

Therefore the term in parenthesis is a density for $X + Z/k$ and it only depends on $\varphi$ so $X + Z/k$ does have the same law as $Y + Z/k$.

Now observe that $X + Z/k \longrightarrow X$ a.s. and thus in distribution; similarly $Y + Z/k \longrightarrow Y$ in distribution and since $X + Z/k$ and $Y + Z/k$ have the same law, then the limit law is the same: $X$ has the same law as $Y$. □

**Remark 2.6.5** (Lévy's inversion formula)**.** One can recover more explicitly the law of a random variable with a given characteristic function $\varphi$. In dimension $d = 1$ to simplify, one always has

$$\frac{1}{2\pi} \int_{-K}^{K} \frac{\mathrm{e}^{-iat} - \mathrm{e}^{-ibt}}{it} \varphi(t)\, \mathrm{d}t \underset{K \to \infty}{\longrightarrow} \frac{F(b) + F(b-)}{2} - \frac{F(a) + F(a-)}{2},$$

where $F$ is the corresponding distribution function. Moreover, if $\int |\varphi| < \infty$, then the law admits a continuous density given by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-ixt}\, \varphi(t)\, \mathrm{d}t.$$

A consequence of the previous theorem is that the characteristic function of a vector characterises the independence in the following way, stronger than in Proposition 2.6.2.

**Corollary 2.6.6.** *Let $X^1, \dots, X^n$ be random vectors in $\mathbb{R}^d$. They are independent if and only if for every $u^1, \dots, u^n \in \mathbb{R}^d$ it holds*

$$\mathbb{E}\Big[\exp\Big(i\,(\langle u^1, X^1\rangle + \cdots + \langle u^n, X^n\rangle)\Big)\Big] = \mathbb{E}[\exp(i\langle u^1, X^1\rangle)] \times \cdots \times \mathbb{E}[\exp(i\langle u^n, X^n\rangle)].$$

*Proof.* The direct implication is clear since if the vectors $X^k$ are independent, then so are the $\exp(i\langle u^k, X^k\rangle)$'s and thus

$$\mathbb{E}\left[\prod_{k=1}^{n} \exp(i\langle u^k, X^k\rangle)\right] = \prod_{k=1}^{n} \mathbb{E}[\exp(i\langle u^k, X^k\rangle)].$$

Conversely, suppose that this identity holds for all $u^k$'s, then because the characteristic function characterises the law then the vector $(X^1, \dots, X^n)$ in $\mathbb{R}^{dn}$ has the same law as the concatenation of $n$ independent vectors so indeed these vectors are independent. □

Let us end with a useful computation: the moments of a r.v. can be obtained by differentiating the characteristic function at 0.

**Proposition 2.6.7** (Moments). *Suppose that $\mathbb{E}[|X|^n] < \infty$, then $\varphi_X \in C^n$ and every $u \in \mathbb{R}^d$, every $k \le n$, and for every $j_1, \dots, j_k \in \{1, \dots, d\}$, it holds*

$$\frac{\partial^k}{\partial u_{j_1} \cdots \partial u_{j_k}} \varphi_X(u) = i^k \, \mathbb{E}\left[X_{j_1} \cdots X_{j_k} \, \mathrm{e}^{i\langle u, X\rangle}\right].$$

*In particular, for $u = 0$, the right-hand side equals $i^k \, \mathbb{E}[X_{j_1} \cdots X_{j_k}]$.*

*Proof.* It is a matter of exchanging derivation and expectation; we use that $\mathbb{E}[|X|^k] < \infty$ to get the domination $|i^k \prod_{\ell=1}^k X_{j_\ell} \, \mathrm{e}^{i\langle u, X\rangle}| \le \prod_{\ell=1}^k |X_{j_\ell}| \in L^1$. □

### 2.6.2 Characteristic functions & Convergence in distribution

A second reason of the success of characteristic functions is that they characterise the convergence in distribution.

**Theorem 2.6.8.** *We have $X^n \longrightarrow X$ in distribution if and only if $\varphi_{X^n} \longrightarrow \varphi_X$ pointwise.*

*Proof.* The direct implication follows from the fact that the function $x \mapsto \mathrm{e}^{i\langle u, x\rangle}$ is continuous and bounded for every fixed $u \in \mathbb{R}^d$. For the converse one, we argue as in the proof of Theorem 2.6.4 and we consider $X^n + Z/k$ where $Z$ is an independent random vector which has the same law as $d$ i.i.d. standard Gaussian r.r.v.'s. Recall from that proof that $X^n + Z/k$ has a density given by

$$f_k^n(y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \varphi_{X^n}(z) \exp\left(-i\langle y, z\rangle - \frac{|z|^2}{2k^2}\right) \mathrm{d}z.$$

If $\varphi_{X^n} \longrightarrow \varphi_X$ pointwise, then by dominated convergence, $f_k^n$ converges pointwise to $f_k$, the density of $X + Z/k$. By a second application of the dominated convergence theorem we infer that for any continuous and bounded function $g$,

$$\mathbb{E}[g(X^n + Z/k)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[g(X + Z/k)],$$

i.e. that $X^n + Z/k \longrightarrow X + Z/k$ in distribution for any $k \ge 1$ fixed.

It remains to prove that $X^n \longrightarrow X$ in distribution. Fix $g$ bounded and $L$-Lipschitz, then for every $k \ge 1$,

$$\begin{aligned}
&|\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]| \\
&\le \mathbb{E}[|g(X_n + Z/k) - g(X_n)|] + |\mathbb{E}[g(X_n + Z/k)] - \mathbb{E}[g(X + Z/k)]| + \mathbb{E}[|g(X + Z/k) - g(X)|] \\
&\le \frac{2L}{k} \, \mathbb{E}[|Z|] + |\mathbb{E}[g(X_n + Z/k)] - \mathbb{E}[g(X + Z/k)]|.
\end{aligned}$$

The second term tends to 0 as $n \longrightarrow \infty$ and further the first one tends to 0 as $k \longrightarrow \infty$. Hence $\mathbb{E}[g(X_n)] \longrightarrow \mathbb{E}[g(X)]$ for every Lipschitz and bounded function, which is equivalent to the convergence in distribution by Theorem 2.5.5. □

This theorem shows that, given a sequence $(X_n)_n$ and a candidate $X$ for its limit in distribution, in order to prove the convergence one may rely on the characteristic functions, and we shall use this idea to prove the Central Limit Theorem in the next section. However sometimes one does not have a priori a candidate $X$. In this case, we have the following powerful extension.

**Theorem 2.6.9** (Lévy). *Let $X_n$ have characteristic function $\varphi_n$ for every $n \ge 1$ and suppose that there exists a function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{C}$ which is continuous at 0 and such that $\varphi_n \longrightarrow \varphi$ pointwise. Then there exists $X$ whose characteristic function is $\varphi$ and such that $X_n \longrightarrow X$ in distribution.*

We will omit the proof of this result. The key point is to prove that there exists a subsequence $(X_{n_k})_k$ which converges in distribution to some $X$ (key word is *tightness*). Then by Theorem 2.6.8 the characteristic functions converge along this subsequence, and by our assumption and Theorem 2.6.4 this means that $X$ has characteristic function $\varphi$. A second application of Theorem 2.6.8 allows us to conclude that $X_n \longrightarrow X$ in distribution.

Recall that the characteristic function of a random vector is always continuous at 0 so by Theorem 2.6.9 if $\varphi_n$ does not converge, or converges to a limit which is not continuous at 0, then $X_n$ does not converge in distribution.

## 2.7 Central Limit Theorems & Gaussian vectors

### 2.7.1 Central Limit Theorems in dimension 1 (⋆)

Let us start with the dimension $d = 1$. Recall that the Law of Large Number states that if $(X_n)_{n\geq 1}$ are i.i.d. r.v.'s with mean $\mathbb{E}[X_1] = m$, then

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to\infty}{\longrightarrow} m$$

in probability for the weak law and almost surely for the strong law. One then wonders at which speed does this convergence occur, or equivalently asks for the second order term. The Central Limit Theorem shows that under a finite variance assumption, these fluctuations are of order $\sqrt{n}$, and remain random in the limit.

**Theorem 2.7.1** (Standard CLT). *Let $(X_n)_{n\geq 1}$ be i.i.d. r.v.'s with $\mathbb{E}[X_1] = m \in \mathbb{R}$ and $\mathrm{Var}(X_1) = \sigma^2 \in (0,\infty)$. Then we have the convergence in distribution*

$$\sqrt{\frac{n}{\sigma^2}}\left(\frac{X_1 + \cdots + X_n}{n} - m\right) = \frac{X_1 + \cdots + X_n - nm}{\sqrt{n\sigma^2}} \underset{n\to\infty}{\overset{(d)}{\longrightarrow}} Z \sim \mathcal{N}(0,1).$$

We shall prove actually a more general version by removing the assumption that the r.v.'s have the same law. In the sequel we are given for every $n \geq 1$ a collection $(X_{n,k})_{k\leq n}$ of independent r.r.v.'s with $\mathbb{E}[X_{n,k}] = 0$ (otherwise subtract the mean) and $\mathbb{E}[X_{n,k}^2] = \mathrm{Var}(X_{n,k}) = \sigma_{n,k}^2 \in [0,\infty)$, and with at least one index such that $\sigma_{n,k}^2 \neq 0$. We let

$$s_n^2 = \sum_{k=1}^{n} \sigma_{n,k}^2 \in (0,\infty).$$

The following statement and first proof are due to Lindeberg, and Lévy then proposed a proof based on characteristic functions.

**Theorem 2.7.2** (Lindeberg's CLT). *Assume the so-called Lindeberg condition: for any $\varepsilon > 0$,*

$$\frac{1}{s_n^2}\sum_{k=1}^{n} \mathbb{E}\left[|X_{n,k}|^2\, \mathbb{1}_{|X_{n,k}|>\varepsilon s_n}\right] \underset{n\to\infty}{\longrightarrow} 0. \tag{2.3}$$

*Then we have the convergence in distribution*

$$\frac{X_{n,1} + \cdots + X_{n,n}}{s_n} \underset{n\to\infty}{\overset{(d)}{\longrightarrow}} Z \sim \mathcal{N}(0,1). \tag{2.4}$$

Let us defer the proof to the next section and immediately deduce the CLT for i.i.d. r.v.'s from this statement.

*Proof of Theorem 2.7.1.* In this case, the $X_{n,k}$ all have the same law so $s_n^2 = n\sigma^2$ and moreover, for any $\varepsilon > 0$, we have by dominated convergence

$$\frac{1}{s_n^2}\sum_{k=1}^{n} \mathbb{E}\left[|X_{n,k}|^2\, \mathbb{1}_{|X_{n,k}|>\varepsilon s_n}\right] = \frac{1}{\sigma^2}\,\mathbb{E}\left[|X_1 - m|^2\, \mathbb{1}_{|X_1-m|>\varepsilon\sqrt{n\sigma^2}}\right] \underset{n\to\infty}{\longrightarrow} 0.$$

Thus (2.3) is satisfied and we can apply Theorem 2.7.2. □

The Lindeberg condition (2.3) may be hard to check in practice and other stronger conditions, but simpler to verify, exist such as the Lyapunov condition.

**Theorem 2.7.3** (Lyapunov's CLT)**.** *Suppose that there exists $\delta > 0$ such that*

$$\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} \mathbb{E}\left[|X_{n,k}|^{2+\delta}\right] \xrightarrow[n\to\infty]{} 0. \tag{2.5}$$

*Then* (2.3) *holds so we have the convergence in distribution*

$$\frac{X_{n,1} + \cdots + X_{n,n}}{s_n} \xrightarrow[n\to\infty]{(d)} Z \sim \mathcal{N}(0,1).$$

*Proof.* Put $p = (2+\delta)/2$ and $q = (2+\delta)/\delta$ so $1/p + 1/q = 1$. Then the Hölder inequality and then the Markov inequality yield

$$\begin{aligned}
\mathbb{E}\left[|X_{n,k}|^2 \mathbb{1}_{|X_{n,k}|>\varepsilon s_n}\right] &\le \mathbb{E}\left[|X_{n,k}|^{2+\delta}\right]^{1/p} \mathbb{P}(|X_{n,k}| > \varepsilon s_n)^{1/q} \\
&\le \mathbb{E}\left[|X_{n,k}|^{2+\delta}\right]^{1/p} \left(\frac{\mathbb{E}[|X_{n,k}|^{2+\delta}]}{(\varepsilon s_n)^{2+\delta}}\right)^{1/q} \\
&= \frac{\mathbb{E}[|X_{n,k}|^{2+\delta}]}{(\varepsilon s_n)^{\delta}}.
\end{aligned}$$

Thus

$$\frac{1}{s_n^2} \sum_{k=1}^{n} \mathbb{E}\left[|X_{n,k}|^2 \mathbb{1}_{|X_{n,k}|>\varepsilon s_n}\right] \le \frac{1}{\varepsilon^{\delta} s_n^{2+\delta}} \sum_{k=1}^{n} \mathbb{E}[|X_{n,k}|^{2+\delta}],$$

which tends to 0 under (2.5). □

The Lyapunov condition (2.5) is often checked with $2+\delta = 3$ or 4 in practice (provided such a moment exists).

### 2.7.2 Proof of the Lindeberg CLT (⋆)

The proof of Theorem 2.7.2 will use the following two elementary results.

**Lemma 2.7.4.** *For every $n \ge 0$ and every $x \in \mathbb{R}$, it holds*

$$\left|\mathrm{e}^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}\right| \le \min\left(\frac{2\,|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!}\right).$$

*Proof.* Put $R_n(x) = \mathrm{e}^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}$ the rest of the Taylor expansion of $\mathrm{e}^{ix}$. Then for $n = 0$ we have

$$R_0(x) = \mathrm{e}^{ix} - 1 = \cos(x) - 1 + i\sin(x) = \int_0^x (-\sin(y) + i\cos(y))\,\mathrm{d}y = \int_0^x i\,\mathrm{e}^{iy}\,\mathrm{d}y$$

so indeed $|R_0(x)| \le \min(2, |x|)$. Then for $n \ge 1$, we have

$$R_n(x) = \int_0^x i\,R_{n-1}(y)\,\mathrm{d}y,$$

and the result follows by induction. □

**Remark 2.7.5.** We shall use this lemma with $n = 2$ and infer that, for any $\delta > 0$ and any $x \in \mathbb{R}$, we have:

$$\left|\mathrm{e}^{ix} - \left(1 + ix - \frac{x^2}{2}\right)\right| \le \min\left(x^2, \frac{|x|^3}{6}\right) \le x^2\,\mathbb{1}_{|x|>\delta} + \frac{|x|^3}{6}\,\mathbb{1}_{|x|\le\delta} \le x^2\,\mathbb{1}_{|x|>\delta} + \delta x^2.$$

**Lemma 2.7.6.** *Let $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n \in \mathbb{C}$ be such that $|\alpha_k|, |\beta_k| \le 1$ for all $k \le n$. Then*

$$\left| \prod_{k=1}^{n} \alpha_k - \prod_{k=1}^{n} \beta_k \right| \le \sum_{k=1}^{n} |\alpha_k - \beta_k|.$$

*Proof.* The claim is obvious for $n = 1$ and for $n \ge 2$, we have

$$\prod_{k=1}^{n} \alpha_k - \prod_{k=1}^{n} \beta_k = (\alpha_n - \beta_n) \prod_{k=1}^{n-1} \alpha_k + \beta_n \left( \prod_{k=1}^{n-1} \alpha_k - \prod_{k=1}^{n-1} \beta_k \right),$$

hence

$$\left| \prod_{k=1}^{n} \alpha_k - \prod_{k=1}^{n} \beta_k \right| \le |\alpha_n - \beta_n| + \left| \prod_{k=1}^{n-1} \alpha_k - \prod_{k=1}^{n-1} \beta_k \right|,$$

and the claim follows by induction. □

Let us make one last observation.

**Lemma 2.7.7.** *Under* (2.3) *we have*

$$\sup_{k \le n} \frac{\sigma_{n,k}^2}{s_n^2} \underset{n \to \infty}{\longrightarrow} 0. \tag{2.6}$$

*Proof.* For every $\varepsilon > 0$, we have

$$\begin{aligned} \sup_{k \le n} \frac{\sigma_{n,k}^2}{s_n^2} &\le \sup_{k \le n} \frac{1}{s_n^2} \mathbb{E}\left[ |X_{n,k}|^2 \, \mathbb{1}_{|X_{n,k}| \le \varepsilon s_n} \right] + \sup_{k \le n} \frac{1}{s_n^2} \mathbb{E}\left[ |X_{n,k}|^2 \, \mathbb{1}_{|X_{n,k}| > \varepsilon s_n} \right] \\ &\le \varepsilon^2 + \sum_{k=1}^{n} \frac{1}{s_n^2} \mathbb{E}\left[ |X_{n,k}|^2 \, \mathbb{1}_{|X_{n,k}| > \varepsilon s_n} \right], \end{aligned}$$

which converges to $\varepsilon^2$ according to (2.3). □

We are now ready to prove Theorem 2.7.2.

*Proof of Theorem 2.7.2.* By independence, for all $n \ge 1$ and all $t \in \mathbb{R}$, we have that

$$\mathbb{E}[\mathrm{e}^{its_n^{-1} \sum_{k=1}^{n} X_{n,k}}] = \mathbb{E}\left[ \prod_{k=1}^{n} \mathrm{e}^{its_n^{-1} X_{n,k}} \right] = \prod_{k=1}^{n} \mathbb{E}[\mathrm{e}^{i(s_n^{-1} t) X_{n,k}}].$$

Recall Theorem 2.6.8, our aim is thus to prove that this converges to $\mathbb{E}[\mathrm{e}^{itZ}] = \mathrm{e}^{-t^2/2}$.

We deduce from Remark 2.7.5 and after taking the expectation that for every $\varepsilon > 0$ and every $t \in \mathbb{R}$,

$$\begin{aligned} \left| \mathbb{E}[\mathrm{e}^{i(s_n^{-1} t) X_{n,k}}] - \left( 1 - \frac{t^2 \sigma_{n,k}^2}{2 s_n^2} \right) \right| &\le \mathbb{E}\left[ \min\left( \frac{t^2 X_{n,k}^2}{s_n^2}, \frac{|t|^3 \, |X_{n,k}|^3}{s_n^3} \right) \right] \\ &\le \frac{t^2}{s_n^2} \mathbb{E}\left[ X_{n,k}^2 \, \mathbb{1}_{|X_{n,k}| > \varepsilon s_n} \right] + \varepsilon |t|^3 \frac{\sigma_{n,k}^2}{s_n^2}. \end{aligned}$$

According to Lemma 2.7.6, we have then for any $t \in \mathbb{R}$,

$$\begin{aligned} \left| \mathbb{E}[\mathrm{e}^{its_n^{-1} \sum_{k=1}^{n} X_{n,k}}] - \prod_{k=1}^{n} \left( 1 - \frac{t^2 \sigma_{n,k}^2}{2 s_n^2} \right) \right| &\le \sum_{k=1}^{n} \left| \mathbb{E}[\mathrm{e}^{i(s_n^{-1} t) X_{n,k}}] - \left( 1 - \frac{t^2 \sigma_{n,k}^2}{2 s_n^2} \right) \right| \\ &\le \sum_{k=1}^{n} \left( \frac{t^2}{s_n^2} \mathbb{E}\left[ X_{n,k}^2 \, \mathbb{1}_{|X_{n,k}| > \varepsilon s_n} \right] + \varepsilon |t|^3 \frac{\sigma_{n,k}^2}{s_n^2} \right) \\ &\le \frac{t^2}{s_n^2} \sum_{k=1}^{n} \mathbb{E}\left[ X_{n,k}^2 \, \mathbb{1}_{|X_{n,k}| > \varepsilon s_n} \right] + \varepsilon |t|^3. \end{aligned}$$

By the assumption (2.3) the last line tends to $|t|^3\varepsilon$ and as $\varepsilon > 0$ is arbitrary, we infer that

$$\left|\mathbb{E}[e^{its_n^{-1}\sum_{k=1}^n X_{n,k}}] - \prod_{k=1}^n\left(1 - \frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)\right| \underset{n\to\infty}{\longrightarrow} 0 \tag{2.7}$$

as soon as (2.3) is satisfied.

Now let $(Z_{n,k})_{1\le k\le n}$ be independent Gaussian r.v.'s with $Z_{n,k} \sim \mathcal{N}(0,\sigma_{n,k}^2)$ respectively. We claim that they satisfy (2.3). Indeed, if $Z \sim \mathcal{N}(0,1)$, then $Z_{n,k}$ has the same law as $\sigma_{n,k}Z$ and so, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\frac{1}{s_n^2}\sum_{k=1}^n \mathbb{E}\left[|Z_{n,k}|^2\,\mathbb{1}_{|Z_{n,k}|>\varepsilon s_n}\right] &= \frac{1}{s_n^2}\sum_{k=1}^n \sigma_{n,k}^2\,\mathbb{E}\left[|Z|^2\,\mathbb{1}_{|Z|>\varepsilon s_n/\sigma_{n,k}}\right]\\ &\le \frac{1}{s_n^2}\sum_{k=1}^n \sigma_{n,k}^2\sqrt{\mathbb{E}[|Z|^4]\,\mathbb{P}(|Z|>\varepsilon s_n/\sigma_{n,k})}\\ &\le \sqrt{\mathbb{E}[|Z|^4]\,\mathbb{P}(|Z|>\varepsilon \inf_{k\le n} s_n/\sigma_{n,k})}.\end{aligned}$$

Recall from Lemma 2.7.7 that $\inf_{k\le n} s_n/\sigma_{n,k} \longrightarrow \infty$, then the last line tends to 0 and so indeed $(Z_{n,k})_{1\le k\le n}$ satisfy (2.3). We infer that they satisfy (2.7) and thus by the triangle inequality:

$$|\,\mathbb{E}[e^{its_n^{-1}\sum_{k=1}^n X_{n,k}}] - \mathbb{E}[e^{its_n^{-1}\sum_{k=1}^n Z_{n,k}}]| \underset{n\to\infty}{\longrightarrow} 0.$$

It remains to observe that since the $Z_{n,k}$'s are independent Gaussian random variables, then $s_n^{-1}\sum_{k=1}^n Z_{n,k}$ has the Gaussian law with mean $s_n^{-1}\sum_{k=1}^n \mathbb{E}[Z_{n,k}] = 0$ and variance $s_n^{-2}\sum_{k=1}^n \mathbb{E}[Z_{n,k}^2] = s_n^{-2}\sum_{k=1}^n \sigma_{n,k}^2 = 1$, that is $s_n^{-1}\sum_{k=1}^n Z_{n,k}$ is a standard Gaussian and so

$$\mathbb{E}[e^{its_n^{-1}\sum_{k=1}^n Z_{n,k}}] = \mathbb{E}[e^{itZ}] = e^{-t^2/2}.$$

This completes the proof. □

**Remark 2.7.8** (Minimal assumption)**.** We have proved that the Lindeberg condition (2.3) implies both the Central Limit Theorem (2.4) and the fact that no one variable dominates the others in the sense that the largest variance is small compared to the sum of the variances (2.6). Feller has proved conversely that (2.4) and (2.6) combined imply the Lindeberg condition (2.3) which is therefore the minimal assumption one can make in order to have a CLT after rescaling by the square-root of the sum of the variances. Let us mention that the convergence to a Gaussian law under a different rescaling may hold, even for i.i.d. r.v.'s with infinite variance (key words are *domain of attraction of a Gaussian law*).

### 2.7.3 Higher dimensions: Gaussian vectors

We now aim at considering CLT's in dimension $d \ge 2$. The first question to address is: what is the analogue of the Gaussian law in higher dimension? From now on, in dimension 1, a constant random variable will be seen as a Gaussian random variable with variance 0, that is we agree that $\mathcal{N}(c,0) = \delta_c$ is the Dirac mass at $c$ for any $c \in \mathbb{R}$.

**Definition 2.7.9.** A random vector $(X_1,\dots,X_d)$ is called a *Gaussian vector* when any linear combination of its coordinates has a Gaussian law in $\mathbb{R}$, i.e. for every $a = (a_1,\dots,a_d)$, we have that $\langle a, X\rangle = \sum_{k=1}^d a_k X_k$ is Gaussian distributed.

By taking $a$ to be a vector in the canonical basis of $\mathbb{R}^d$, we deduce that if $X$ is a Gaussian vector, then each coordinate $X_k$ has a Gaussian law. The converse is not true in general! See the exercise sheet for an example.

**Proposition 2.7.10.** *Suppose that $(X_1,\dots,X_d)$ are independent r.r.v. each with a Gaussian law, then the vector $(X_1,\dots,X_d)$ is a Gaussian vector.*

*Proof.* Fix $a = (a_1, \dots, a_d)$ and $t \in \mathbb{R}$, then by independence, the characteristic function of $\langle a, X\rangle$ at $t$ equals

$$\mathbb{E}\Big[\exp\Big(it\sum_{k=1}^{d} a_k X_k\Big)\Big] = \prod_{k=1}^{d} \exp\Big(ita_k\,\mathbb{E}[X_k] - \frac{a_k^2\,\mathrm{Var}(X_k)}{2}\Big) = \exp\Big(it\,\langle a, \mathbb{E}[X]\rangle - \frac{1}{2}\langle a, Ca\rangle\Big),$$

where we have set $C$ the diagonal matrix whose diagonal coordinates are $C_{k,k} = \mathrm{Var}(X_k)$. This proves that $\langle a, X\rangle$ has the Gaussian law $\mathcal{N}(\langle a, \mathbb{E}[X]\rangle, \langle a, Ca\rangle)$. □

Gaussian vectors whose coordinates are independent standard Gaussian will be the building blocks of the more general ones, in the same way in dimension 1, any Gaussian random variable $X \sim \mathcal{N}(m, \sigma^2)$ can be written in law as $m + \sigma Z$ where $Z = \mathcal{N}(0, 1)$. Let us now characterise Gaussian vectors by their characteristic function.

**Theorem 2.7.11.** *A random vector $X$ is a Gaussian vector if and only if there exists a vector $m \in \mathbb{R}^d$ and a $d \times d$ symmetric positive matrix $C$ such that for all $u \in \mathbb{R}^d$,*

$$\varphi_X(u) = \exp\Big(i\,\langle u, m\rangle - \frac{1}{2}\langle u, Cu\rangle\Big). \tag{2.8}$$

*In this case, $m_k = \mathbb{E}[X_k]$ and $C_{k,\ell} = \mathrm{Cov}(X_k, X_\ell)$ for all $1 \le k, \ell \le d$ and we write $X \sim \mathcal{N}(m, C)$. Finally, for any $m \in \mathbb{R}^d$ and any $d \times d$ symmetric positive matrix $C$, there exists a Gaussian vector $X \sim \mathcal{N}(m, C)$.*

**Some linear algebra.** Recall that a matrix $C$ is said to be *symmetric* when $C^t = C$, and further *positive* when for all $a \in \mathbb{R}^d$, it holds $\langle a, Ca\rangle = a^t Ca \ge 0$. If further $a^t Ca = 0$ only when $a = 0$, then we say that $C$ is *definite positive*. A symmetric positive matrix $C$ has nonnegative eigenvalues and can always be diagonalised in a othornormal basis, i.e. it can be written as $PDP^{-1}$, where $P^{-1} = P^t$ and $D$ is diagonal, with diagonal coordinates given by the eigenvalues of $C$. Then write $\sqrt{D}$ for the diagonal matrix whose entries are the square-root of those of $D$, and let $A = P\sqrt{D}P^{-1}$. Then $A^t = A$ and $A^tA = C$. Finally, $C$ is definite positive if and only if all its eigenvalues are nonzero, which is equivalent to $D$ being invertible, in which case $C$ is and $C^{-1} = PD^{-1}P^t$.

*Proof of Theorem 2.7.11.* Let us start with the last statement and construct for any $m \in \mathbb{R}^d$ and any $d \times d$ symmetric positive matrix $C$ a random vector whose characteristic function is given by the formula (2.8). Let $Y = (Y_1, \dots, Y_d)$ be a vector whose coordinate are i.i.d. standard Gaussian. We have seen in the proof of Proposition 2.7.10 that $Y$ has characteristic function

$$\varphi_Y(u) = \mathbb{E}\big[\mathrm{e}^{i\langle u, Y\rangle}\big] = \exp\Big(-\frac{1}{2}\langle u, u\rangle\Big) = \exp\Big(-\frac{1}{2}|u|^2\Big).$$

Consequently, letting $A$ be the symmetric matrix $A^t = A$ such that $A^2 = C$, we get

$$\varphi_{AY}(u) = \mathbb{E}\big[\mathrm{e}^{i\langle u, AY\rangle}\big] = \mathbb{E}\big[\mathrm{e}^{i\langle A^t u, Y\rangle}\big] = \mathbb{E}\big[\mathrm{e}^{i\langle Au, Y\rangle}\big] = \exp\Big(-\frac{1}{2}|Au|^2\Big) = \exp\Big(-\frac{1}{2}\langle u, Cu\rangle\Big).$$

Then $X = m + AY$ has characteristic function given by (2.8).

This form of characteristic function implies that $X$ is a Gaussian vector since for any $a \in \mathbb{R}^d$ and $t \in \mathbb{R}$,

$$\varphi_{\langle a, X\rangle}(t) = \varphi_X(ta) = \exp\Big(it\,\langle a, m\rangle - \frac{t^2}{2}\langle a, Ca\rangle\Big),$$

so $\langle a, X\rangle \sim \mathcal{N}(\langle a, m\rangle, \langle a, Ca\rangle)$. By taking $a$ to be the $k$'th vector in the canonical basis of $\mathbb{R}^d$, we deduce that $X_k \sim \mathcal{N}(m_k, C_{k,k})$ so $m = \mathbb{E}[X]$ and the diagonal of $C$ is given by $C_{k,k} = \mathrm{Var}(X_k)$. Similarly, by taking $a$ to be the sum of the $k$'th and $\ell$'th vectors in the canonical basis of $\mathbb{R}^d$, we deduce that $Z_k + Z_\ell \sim \mathcal{N}(m_k + m_\ell, C_{k,k} + C_{\ell,\ell} + 2C_{k,\ell})$ and so

$$\mathrm{Cov}(Z_k, Z_\ell) = \frac{\mathrm{Var}(Z_k + Z_\ell) - \mathrm{Var}(Z_k) - \mathrm{Var}(Z_\ell)}{2} = \frac{(C_{k,k} + C_{\ell,\ell} + 2C_{k,\ell}) - C_{k,k} - C_{\ell,\ell}}{2} = C_{k,\ell}.$$

Finally, suppose conversely that $X$ is a Gaussian vector, let $m = \mathbb{E}[X]$ and $C$ denotes its covariance matrix, and let us prove that its characteristic function is given by (2.8). First note that $m$ and $C$ are well-defined since each coordinate has a Gaussian law so is square-integrable. Moreover, for any $u \in \mathbb{R}^d$, the r.v. $\langle u, X\rangle$ has a Gaussian law with mean $m_u = \mathbb{E}[\langle u, X\rangle] = \langle u, \mathbb{E}[X]\rangle = \langle u, m\rangle$ by linearity, and with variance $\sigma_u^2 = \mathrm{Var}(\langle u, X\rangle) = \mathrm{Cov}(\langle u, X\rangle, \langle u, X\rangle) = \langle u, Cu\rangle$ by bilinearity. Therefore

$$\varphi_X(u) = \varphi_{\langle u,X\rangle}(1) = \exp\Big(im_u - \frac{1}{2}\sigma_u^2\Big),$$

which equals the right-hand side of (2.8). □

Recall that Gaussian random variables have a density with respect to the Lebesgue measure, except the degenerate ones with variance 0. This can be generalised to higher dimension.

**Proposition 2.7.12.** *A Gaussian vector $X \sim \mathcal{N}(m, C)$ has a density with respect to the $d$-dimensional Lebesgue measure if and only of $C$ is invertible, and in this case it takes the form: for every $x \in \mathbb{R}^d$,*

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\Big(-\frac{1}{2}\langle x - m, C^{-1}(x - m)\rangle\Big).$$

*Proof.* Recall that we can represent the law of $X$ in the form $m + AY$ where $Y$ is a collection of i.i.d. standard Gaussian r.v.'s. and $A = P\sqrt{D}P^{-1}$ with $P^t = P^{-1}$ and $\sqrt{D}$ is diagonal and made of the square-root of the eigenvalues of $C$. If $C$ is invertible, then so is $A$ so the affine transformation $y \mapsto x = m + AY$ is a diffeomorphism and the change of variable formula yields for any measurable and bounded function $g$:

$$\begin{aligned}
\mathbb{E}[g(m + AY)] &= \int_{\mathbb{R}^d} g(m + Ay) \prod_{k=1}^{d} \Big(\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{y_k^2}{2}\Big)\Big)\, \mathrm{d}y_1 \otimes \cdots \otimes \mathrm{d}y_d \\
&= \int_{\mathbb{R}^d} g(m + Ay) \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{|y|^2}{2}\Big)\, \mathrm{d}y \\
&= \int_{\mathbb{R}^d} g(x) \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{|A^{-1}(x - m)|^2}{2}\Big)\, |\det A^{-1}|\, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} g(x) \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\Big(-\frac{\langle x - m, C^{-1}(x - m)\rangle}{2}\Big)\, \mathrm{d}x.
\end{aligned}$$

Thus when $C$ is invertible, $X$ has indeed the given density.

On the other hand, if $C$ is not invertible, then there exists $a \in \mathbb{R}^d$ such that $Ca = 0$. Consequently, $\mathrm{Var}(\langle a, X\rangle) = \langle a, Ca\rangle = 0$ so $X$ almost surely belongs to the hyperplane $H = a^\perp = \{x \in \mathbb{R}^d : \langle a, x\rangle = 0\}$ which has $d$-dimensional Lebesgue measure 0. In particular $X$ has no density. □

Recall that the covariance between two independent r.r.v.'s is zero, but a null covariance does not imply independence in general. It does for Gaussian vectors!

**Proposition 2.7.13** (Independence)**.** *Let $(X_1, \dots, X_d)$ be a Gaussian vector. Then the variables $(X_1, \dots, X_d)$ are independent if and only if the covariance matrix of $X$ is diagonal.*

*Proof.* The direct implication is known as just recalled. Suppose conversely that the covariance matrix $C$ of $X$ is diagonal. Subtracting the mean $\mathbb{E}[X]$ if necessary, suppose that $\mathbb{E}[X] = 0$. Then we know that for any $u \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\prod_{k=1}^{d} \mathrm{e}^{iu_kX_k}\Big] = \mathbb{E}\Big[\mathrm{e}^{i\langle u,X\rangle}\Big] = \exp\Big(-\frac{1}{2}\langle u, Cu\rangle\Big) = \exp\Big(-\frac{1}{2}\sum_{k=1}^{d} u_k^2 C_{k,k}\Big) = \prod_{k=1}^{d} \mathbb{E}[\mathrm{e}^{iu_kX_k}],$$

which characterises the independence by Corollary 2.6.6. □

**Remark 2.7.14.** More generally, the same proof extends (with more notation) to show that if we partition $X$ into sub-vectors, say $(X^1, \dots, X^k)$ where $X^1 \in \mathbb{R}^{d_1}, \dots, X^k \in \mathbb{R}^{d_k}$, where $d_1 + \cdots + d_k = d$, then the vectors $X^1, \dots, X^k$ are independent if and only if the covariance matrix $C$ is block-diagonal with block sizes $d_1, \dots, d_k$, i.e. if and only if $\mathrm{Cov}(X^i_{i_p}, X^j_{j_q}) = 0$ for any $1 \le i < j \le k$ and any $1 \le i_p \le d_i$ and $1 \le j_q \le d_j$.

We can now easily deduce the CLT for i.i.d. random vectors from the dimension 1 case in Theorem 2.7.1.

**Theorem 2.7.15** (Multivariate CLT)**.** *Let $(X^n)_{n\ge1}$ be i.i.d. random vectors with $\mathbb{E}[X^1] = m \in \mathbb{R}^d$ and $\mathrm{Cov}(X^1) = C$ a symmetric positive definite matrix. Then we have the convergence in distribution:*

$$\frac{X_1 + \cdots + X_n - nm}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} Z \sim \mathcal{N}(0, C).$$

*Proof.* Assume $m = 0$ without loss of generality. Let $u \in \mathbb{R}^d$ and $t = \mathrm{Var}(\langle u, X^1\rangle) = \langle u, Cu\rangle$ by bilinearity of the covariance. We deduce from Proposition 2.6.2 that

$$\varphi_{n^{-1/2}\sum_{k=1}^n X^k}(u) = \varphi_{\langle u, n^{-1/2}\sum_{k=1}^n X^k\rangle}(1) = \varphi_{n^{-1/2}\sum_{k=1}^n \langle u, X^k\rangle}(1).$$

Note that the random variables $\langle u, X^k\rangle$ are i.i.d. with mean 0 and variance $\langle u, Cu\rangle$. Therefore by the CLT in dimension 1, Theorem 2.7.1, the previous characteristic function converges to that of a Gaussian random variable with variance $\langle u, Cu\rangle$ evaluated at 1, namely:

$$\varphi_{n^{-1/2}\sum_{k=1}^n X^k}(u) \xrightarrow[n\to\infty]{} \exp\left(-\frac{\langle u, Cu\rangle^2}{2}\right) = \varphi_Z(u),$$

where $Z \sim \mathcal{N}(0, C)$ and we conclude by Theorem 2.6.8. □

One can get multivariate extensions of Lindeberg's or Lyapunov's CLT in a similar way. For every $n \ge 1$ let $(X^{n,k})_{k\le n}$ be independent random vectors with $\mathbb{E}[X^{n,k}] = 0$ (otherwise subtract the mean) and with covariance matrix $C^{n,k}$ with $\|C^{n,k}\| < \infty$, where we recall the norm $\|\cdot\|$ on symmetric positive matrices given by the largest eigenvalue. Thus $\|C^{n,k}\| < \infty$ if and only if $\langle a, C^{n,k}a\rangle < \infty$ for all $a \in \mathbb{R}^d$. Assume also that at least one of them is invertible and set

$$S^n = \sum_{k=1}^n C^{n,k}.$$

Note that $S^n$ is the covariance matrix of $\sum_{k=1}^n X^{n,k}$ and it is invertible. Then it admits an invertible square-root matrix $\sqrt{S^n}$ and $\|(S^n)^{-1/2}\|^2 < \infty$ equals the inverse of the smallest eigenvalue of $S^n$.

**Theorem 2.7.16** (Lindeberg's multivariate CLT)**.** *Assume the so-called Lindeberg condition: for any $\varepsilon > 0$,*

$$\sum_{k=1}^n \mathbb{E}\left[\|(S^n)^{-1/2}X_{n,k}\|^2 \mathbb{1}_{\|(S^n)^{-1/2}X_{n,k}\|>\varepsilon}\right] \xrightarrow[n\to\infty]{} 0. \tag{2.9}$$

*Then we have the convergence in distribution*

$$(S^n)^{-1/2}(X_{n,1} + \cdots + X_{n,n}) \xrightarrow[n\to\infty]{(d)} Z \sim \mathcal{N}(0, I_d). \tag{2.10}$$

# Part II

# Markov Chains

# Chapter 3

# Discrete Markov Chains

In words a Markov chain is a random process such that, at any given time, the future evolution only depends on the curent position and not on the whole past trajectory. This lack of memory phenomenon appears in many contexts and Markov chains are therefore a very important object in modelisation. They can be studied in fairly general spaces, and this has important applications in probability and statistics in $\mathbb{R}^d$ or more abstract spaces, but this leads to several technicalities involving measure theory, even just to define the basic objects of interest. We shall therefore restrict ourselves to a countable set of values and thus discrete random variables, which eliminates a lot of technicalities (and is already interesting!). We refer the interested student to the books we suggest in the beginning of these notes for more general contexts.

*In this chapter and in the next two, we assume that the random variables take value in a countable set $\mathbb{X}$, equipped with the $\sigma$-algebra $\mathcal{X}$ of all subsets of $\mathbb{X}$.*

**Contents**



In Section 3.1 we first define formally Markov chains by three equivalent formulation of the Markov property. Section 3.2 introduces the main technical tool associated with Markov chains, that is the transition matrix which encodes the one-step displacement probability into an infinite matrix. Section 3.3 shows that a Markov chains can be seen as random dynamical systems, which explains their success in modelising various phenomena. Section 3.4 presents a generalisation of the lack of memory of Markov chain from a fixed time to a random time. The correct notion of random time here being so-called stopping times, which will play a central role. We end by presenting in Section 3.5 an application of this property to solve the discrete Dirichlet problem, related to the question of the first exit or entry point in a given subset, which has many applications from physics to finance.

## 3.1 The Markov property

The term *stochastic process* is meant to describe the evolution of a single random variable as time passes. The formal definition is very simple and does not say much, but we include it since we are going to extensively use this expression, although we shall often drop the adjective "stochastic".

**Definition 3.1.1.** A stochastic process is a sequence of random variables $X = (X_n)_{n\geq 0}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with value in the same measurable space $(\mathbb{X}, \mathcal{X})$.

Recall that for any event $B \in \mathcal{F}$ with nonzero probability, the formula:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

for all $A \in \mathcal{F}$ defines a probability measure $\mathbb{P}(\cdot \mid B)$. Below, we implicitly assume that all the events by which we condition have nonzero probability.

Recall that we will restrict ourselves to stochastic processes $(X_n)_{n\geq 0}$ which take values in a countable space $\mathbb{X}$. Here is our object of interest.

**Theorem 3.1.2.** *Let $(X_n)_n$ be a stochastic process with values in $\mathbb{X}$. The following assertions are equivalent:*

*(i) For every $n \geq 0$, for any $x_0, \dots, x_{n+1} \in \mathbb{X}$, we have:*

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_i = x_i, 0 \leq i \leq n) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

*(ii) For every $n \geq 0$, for any $k \geq 1$ and any $x_0, \dots, x_{n+k} \in \mathbb{X}$, we have:*

$$\mathbb{P}(X_{n+1} = x_{n+1}, \dots, X_{n+k} = x_{n+k} \mid X_i = x_i, 0 \leq i \leq n) = \mathbb{P}(X_{n+1} = x_{n+1}, \dots, X_{n+k} = x_{n+k} \mid X_n = x_n).$$

*(iii) For every $n \geq 1$, for any $k \geq 1$ and any $x_0, \dots, x_{n+k} \in \mathbb{X}$, we have:*

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_{n+1} = x_{n+1}, \dots, X_{n+k} = x_{n+k} \mid X_n = x_n)\\ &\qquad = \mathbb{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1} \mid X_n = x_n) \times \mathbb{P}(X_{n+1} = x_{n+1}, \dots, X_{n+k} = x_{n+k} \mid X_n = x_n).\end{aligned}$$

*These properties are each called the* Markov property *and such a process is called a* Markov chain.

In words, a Markov chain is a process in which the random evolution at the next step $X_{n+1}$ (Property (i)), or all next steps $(X_p)_{p\geq n}$ (Property (ii)) only depends only the current position $X_n$, and all the rest of information from the past is irrelevant. Property (iii) is often stated as "the future and the past are conditionally independent given the present".

*Proof.* Let us first prove that (ii) is equivalent to (iii). Indeed (ii) reads by definition of the conditional expectation:

$$\frac{\mathbb{P}(X_i = x_i, 0 \leq i \leq n+k)}{\mathbb{P}(X_i = x_i, 0 \leq i \leq n)} = \frac{\mathbb{P}(X_i = x_i, n \leq i \leq n+k)}{\mathbb{P}(X_n = x_n)}.$$

After multiplying both sides by $\mathbb{P}(X_i = x_i, 0 \leq i \leq n)/\mathbb{P}(X_n = x_n)$, this is equivalent to:

$$\frac{\mathbb{P}(X_i = x_i, 0 \leq i \leq n+k)}{\mathbb{P}(X_n = x_n)} = \frac{\mathbb{P}(X_i = x_i, 0 \leq i \leq n)}{\mathbb{P}(X_n = x_n)} \frac{\mathbb{P}(X_i = x_i, n \leq i \leq n+k)}{\mathbb{P}(X_n = x_n)},$$

which is the claim (iii).

Next notice that (i) is weaker than (ii) since it corresponds to the case $k = 1$ in the latter. This therefore provides the base case to prove that (i) implies (ii) by induction on $k$. Suppose thus that the claim (ii) holds for some $k \geq 1$. By applying (i) at time $n + k$, we get:

$$\mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_i = x_i, 0 \leq i \leq n+k) = \mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_{n+k} = x_{n+k}),$$

which we can rewrite as:

$$\mathbb{P}(X_i = x_i, 0 \leq i \leq n+k+1) = \mathbb{P}(X_i = x_i, 0 \leq i \leq n+k)\, \mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_{n+k} = x_{n+k}).$$

On the one hand, if we divide both sides by $\mathbb{P}(X_i = x_i, 0 \leq i \leq n)$, then we get:

$$\begin{aligned}&\mathbb{P}(X_i = x_i, n+1 \leq i \leq n+k+1 \mid X_i = x_i, 0 \leq i \leq n)\\ &= \mathbb{P}(X_i = x_i, n+1 \leq i \leq n+k \mid X_i = x_i, 0 \leq i \leq n)\, \mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_{n+k} = x_{n+k})\\ &= \mathbb{P}(X_i = x_i, n+1 \leq i \leq n+k \mid X_n = x_n)\, \mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_{n+k} = x_{n+k})\end{aligned}$$

by the induction hypothesis. On the other hand, if we go back to the preceding display and sum over all values of $x_0, \dots, x_{n-1}$, then we get:

$$\mathbb{P}(X_i = x_i, n \le i \le n+k+1) = \mathbb{P}(X_i = x_i, n \le i \le n+k)\, \mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_{n+k} = x_{n+k}),$$

so, after dividing by $\mathbb{P}(X_n = x_n)$,

$$\begin{aligned}&\mathbb{P}(X_i = x_i, n+1 \le i \le n+k+1 \mid X_n = x_n)\\ &= \mathbb{P}(X_i = x_i, n+1 \le i \le n+k \mid X_n = x_n)\, \mathbb{P}(X_{n+k+1} = x_{n+k+1} \mid X_{n+k} = x_{n+k}),\end{aligned}$$

and we prove that this equals $\mathbb{P}(X_i = x_i, n+1 \le i \le n+k+1 \mid X_i = x_i, 0 \le i \le n)$. □

One can actually drop some indices in the Markov property, except the last one.

**Proposition 3.1.3.** *If $(X_n)_{n\ge 0}$ is a Markov chain, then for every $n \ge 1$ and every subset of indices $\{i_1, \dots, i_k\} \subset \{0, \dots, n-1\}$ we have for any $x_n, x_{n+1} \in \mathbb{X}$,*

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_{i_1} = x_{i_1}, \dots, X_{i_k} = x_{i_k}, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

*Proof.* By the Markov property we have for all $x_0, \dots, x_{n+1} \in \mathbb{X}$:

$$\mathbb{P}(X_j = x_j, 0 \le j \le n+1) = \frac{\mathbb{P}(X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_n = x_n)}\, \mathbb{P}(X_j = x_j, 0 \le j \le n).$$

By summing over all values $x_j$ for $j \in \{0, \dots, n-1\} \setminus \{i_1, \dots, i_k\}$, we obtain:

$$\mathbb{P}(X_{i_1} = x_{i_1}, \dots, X_{i_k} = x_{i_k}, X_n = x_n, X_{n+1} = x_{n+1}) = \frac{\mathbb{P}(X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_n = x_n)}\, \mathbb{P}(X_{i_1} = x_{i_1}, \dots, X_{i_k} = x_{i_k}, X_n = x_n),$$

and the claim follows. □

## 3.2 Transition matrices

The key tool to study Markov chains and the central object in this theory is the transition matrix associated with it.

**Definition 3.2.1.** A *transition matrix* (or *stochastic matrix*, or *transition kernel*) is a measurable function $P : \mathbb{X} \times \mathbb{X} \longrightarrow [0,1]$ such that $P(x, \cdot)$ is a probability on $\mathbb{X}$ for every $x \in \mathbb{X}$.

Recall that a measure $\mu$ on a countable set $\mathbb{X}$ is simply a nonnegative sequence $(\mu(x), x \in \mathbb{X})$. A probability is a measure with $\sum_{x\in\mathbb{X}} \mu(x) = 1$. Hence, $P$ is simply a (possibly infinite) matrix with nonnegative entries, such that the sum over each row equals 1.

**Theorem 3.2.2** (Chapman–Kolmogorov Equation)**.** *A stochastic process $(X_n)_{n\ge 0}$ is a Markov chain if and only if there exist transition matrices $(P_k)_{k\ge 1}$ such that for every $n \ge 0$ and every $x_0, \dots, x_n \in \mathbb{X}$, we have:*

$$\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \mathbb{P}(X_0 = x_0) \prod_{k=1}^{n} P_k(x_{k-1}, x_k).$$

*Proof.* Suppose first that $(X_n)_{n\ge 0}$ is a Markov chain and let us prove the identity by induction. The latter is trivial for $n = 0$, further, we have by the Markov property:

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \dots, X_{n+1} = x_{n+1}) &= \mathbb{P}(X_0 = x_0, \dots, X_n = x_n)\, \mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n)\\ &= \mathbb{P}(X_0 = x_0, \dots, X_n = x_n)\, \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).\end{aligned}$$

Note that $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$ only depends on the joint law of $X_n$ and $X_{n+1}$ as well as on $x_n$ and $x_{n+1}$. Further, for $x_n$ fixed, it defines a probability, thus as a function of both $x_n$ and $x_{n+1}$, it defines a transition matrix $P_{n+1}$ and the claim follows by induction.

Suppose conversely that there exist transition matrices $(P_k)_{k\geq 1}$ such that for every $n \geq 0$ and every $x_0, \dots, x_n \in \mathbb{X}$, we have:

$$\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \mathbb{P}(X_0 = x_0) \prod_{k=1}^{n} P_k(x_{k-1}, x_k).$$

Then

$$\mathbb{P}(X_0 = x_0, \dots, X_n = x_n, X_{n+1} = x_{n+1}) = \mathbb{P}(X_0 = x_0, \dots, X_n = x_n) P_{n+1}(x_n, x_{n+1}).$$

Let us sum over all values $x_0, \dots, x_{n-1}$ to obtain:

$$\mathbb{P}(X_n = x_n, X_{n+1} = x_{n+1}) = \mathbb{P}(X_n = x_n) P_{n+1}(x_n, x_{n+1}).$$

Combining the last two displays, we infer that:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) = P_{n+1}(x_n, x_{n+1}) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n),$$

hence $(X_n)_{n\geq 0}$ is a Markov chain. □

From now on, one should view a function $f$ on $\mathbb{X}$ as a column vector and a measure $\mu$ as a row vector. Then we can define what, when $\mathbb{X}$ is finite, is simply the matrix multiplication as follows.

**Definition 3.2.3.** Let $P, Q$ be transition matrices, let $f : \mathbb{X} \longrightarrow \mathbb{R}$ be a function, and let $\mu$ be a measure on $\mathbb{X}$. Let us define three operations:

- When it makes sense, let $Pf$ be the function given by:

$$Pf(x) = \sum_{y\in\mathbb{X}} P(x, y) f(y) \qquad \text{for all } x \in \mathbb{X},$$

  which is the expectation of $f(Y_x)$ when $Y_x$ has the law $P(x, \cdot)$.

- Let $\mu P$ be the measure given by:

$$\mu P(y) = \sum_{x\in\mathbb{X}} \mu(x) P(x, y) \qquad \text{for all } y \in \mathbb{X}.$$

  When $\mu$ is a probability, say the law of a random variable $Z$, then $\mu P(y)$ is the expectation of $P(Z, y)$.

- Let $PQ$ be the matrix given by:

$$PQ(x, z) = \sum_{y\in\mathbb{X}} P(x, y) Q(y, z) \qquad \text{for all } x, z \in \mathbb{X}.$$

**Exercise 3.2.4.** If $P$ is a transition matrix and $\mu$ is a probability, then so is $\mu P$. If $Q$ is another transition matrix, then so is $PQ$. Consequently $\prod_{k=1}^{n} P_k$ is a transition matrix if the $P_k$'s are.

**Remark 3.2.5.** A consequence of Theorem 3.2.2 is that the law of a Markov chain $(X_n)_n$ is entirely characterised by the transition matrices

$$P_n(x_{n-1}, x_n) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})$$

for every $n \geq 1$ and the *initial distribution*, that is: the law of $X_0$. In particular, for every $n \geq 1$ and $x_0, x_n \in \mathbb{X}$,

$$\begin{aligned}\mathbb{P}(X_n = x_n \mid X_0 = x_0) &= \sum_{x_1,\dots,x_{n-1}} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid X_0 = x_0) \\ &= \sum_{x_1,\dots,x_{n-1}} \prod_{k=1}^{n} P_k(x_{k-1}, x_k) \\ &= \Big(\prod_{k=1}^{n} P_k\Big)(x_0, x_k),\end{aligned}$$

where the product of transition matrices is that defined above. Consequently, if $\pi$ is a probability on $\mathbb{X}$, then for all $x \in \mathbb{X}$, we have

$$\Big(\pi \prod_{k=1}^{n} P_k\Big)(x) = \sum_{x_0 \in \mathbb{X}} \pi(x_0)\Big(\prod_{k=1}^{n} P_k\Big)(x_0, x_k) = \sum_{x_0 \in \mathbb{X}} \pi(x_0)\, \mathbb{P}(X_n = x \mid X_0 = x_0),$$

which equals the probability that $X_n = x$ when $X_0$ has the law $\pi$.

**Definition 3.2.6.** A *homogeneous* Markov chain is a Markov chain in which all the transitions matrices are equal, say $P_n = P$ for all $n$. We then speak of a *$P$-Markov chain.*

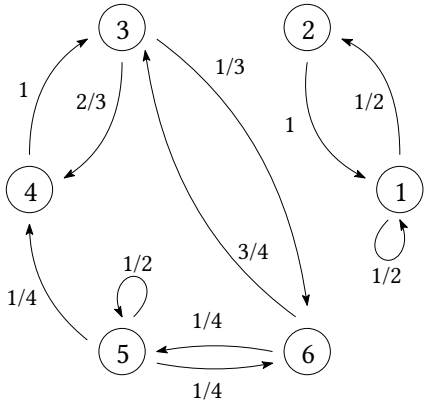
From now on we only consider homogeneous Markov chains. The general case is not more complicated in this chapter and mostly adds more notation, but it becomes intractable in the next chapters. For $x \in \mathbb{X}$, we shall write $\mathbb{P}_x$ to mean that the Markov chain starts from $X_0 = x$ and more generally if $\pi$ is a distribution on $\mathbb{X}$, then we shall write $\mathbb{P}_\pi$ to mean that the Markov chain starts from $X_0$ with the law $\pi$. We then write $\mathbb{E}_x$ and $\mathbb{E}_\pi$ for the associated expectation. Note that in a $P$-Markov chain, we have for every $n \geq 1$ and every initial distribution $\pi$, for every $x, y \in \mathbb{X}$:

$$\mathbb{P}_\pi(X_{n+1} = y \mid X_n = x) = P(x, y) \qquad \text{and} \qquad \mathbb{P}_\pi(X_n = x) = (\pi P^n)(x)$$

Also, for every function $f : \mathbb{X} \longrightarrow \mathbb{R}$ for which the expectations are well-defined, we have

$$\mathbb{E}_\pi[f(X_{n+1}) \mid X_n = x_n] = (Pf)(x_n) \qquad \text{and} \qquad \mathbb{E}_\pi[f(X_n)] = \pi P^n f.$$

Theorem 3.2.2 allows to see any $P$-Markov chain as a random walk on a weighted graph, as in Figure 3.1: draw elements of $\mathbb{X}$ as points, and draw an arrow from $x$ to $y$ with the weight $P(x, y)$ if the latter is nonzero. By Theorem 3.2.2, the probability that the Markov chain follows a given trajectory, conditionally on its starting point, is simply the product of the weights on the arrows along this path.




| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1/2 | 1/2 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 2/3 | 0 | 1/3 |
| 4 | 0 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1/4 | 1/2 | 1/4 |
| 6 | 0 | 0 | 3/4 | 0 | 1/4 | 0 |

Figure 3.1: A weighted graph on the left and the corresponding transition matrix on the right. The probability of any given trajectory is the product of the weights on the corresponding edges.

Theorem 3.2.2 also allows to extend the third formulation of the Markov property in Theorem 3.1.2. Precisely: a $P$-Markov chain is a process such that at any time $n$, conditionally on the value of $X_n$, the futur process $(X_{n+k})_{k\geq 0}$ is also a $P$-Markov chain, started afresh at position $X_n$, independently of the past.

**Corollary 3.2.7** (Restarted process). *Let $(X_n)_{n\geq 0}$ be a $P$-Markov chain, $n \geq 1$, and $x \in \mathbb{X}$. Then conditionally on $\{X_n = x\}$ the process given by $Y_k = X_{n+k}$ for $k \geq 0$ is a $P$-Markov chain started from $Y_0 = x$ and is independent from $(X_0, \dots, X_{n-1})$.*

*Proof.* The conditional independence was proved in Theorem 3.1.2, it only remains to check that $(Y_k)_{k\geq 0}$ is a $P$-Markov chain starting from $x$. By Theorem 3.2.2, for every $y_1, \dots, y_k \in \mathbb{X}$, we have:

$$\begin{aligned}\mathbb{P}(Y_0 = x, Y_1 = y_1, \dots, Y_k = y_k) &= \mathbb{P}(X_n = x, X_{n+1} = y_1, \dots, X_{n+k} = y_k)\\ &= \sum_{x_0,\dots,x_{n-1}} \mathbb{P}(X_0 = x_0, \dots, X_n = x, X_{n+1} = y_1, \dots, X_{n+k} = y_k)\\ &= \sum_{x_0,\dots,x_{n-1}} \mathbb{P}(X_0 = x_0) \prod_{i=1}^{n-1} P(x_{i-1}, x_i) P(x_{n-1}, x) P(x, y_1) \prod_{j=2}^{k} P(y_{j-1}, y_j).\end{aligned}$$

Next notice that Theorem 3.2.2 also yields:

$$\sum_{x_0,\dots,x_{n-1}} \mathbb{P}(X_0 = x_0) \prod_{i=1}^{n-1} P(x_{i-1}, x_i) P(x_{n-1}, x) = \sum_{x_0,\dots,x_{n-1}} \mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \mathbb{P}(X_n = x) = \mathbb{P}(Y_0 = x).$$

We have thus proved that

$$\mathbb{P}(Y_0 = x, Y_1 = y_1, \dots, Y_k = y_k) = \mathbb{P}(Y_0 = x) P(x, y_1) \prod_{j=2}^{k} P(y_{j-1}, y_j),$$

and we conclude from Theorem 3.2.2 again. □

## 3.3 Markov chains as random recursive sequences

Markov chains are in some sense the random analogue of recursive sequences, defined iteratively by $x_{n+1} = f(x_n)$, as shown in the next result. This provides a natural motivation to study Markov chains as well as an easy way to prove that a given process is indeed a Markov chain; it also helps to simulate them in practice.

**Proposition 3.3.1** (Random recursion). *For any $X_0$, any sequence $(\xi_n)_{n\geq 1}$ of i.i.d r.v.'s with values in some space $(E, \mathcal{E})$ and independent of $X_0$, and for any measurable function $f : \mathbb{X} \times E \longrightarrow \mathbb{X}$, the process defined iteratively by:*

$$X_{n+1} = f(X_n, \xi_{n+1})$$

*is a homogeneous Markov chain started from $X_0$, with transition matrix given by $P : (x, y) \mapsto \mathbb{P}(f(x, \xi_1) = y)$. Conversely for any transition matrix $P$ and any random variable $X_0$ in $\mathbb{X}$, there exist such a sequence $(\xi_n)_{n\geq 1}$ and such a measurable function $f : \mathbb{X} \times E \longrightarrow \mathbb{X}$ such that the corresponding Markov chain has transition matrix $P$.*

*Proof.* For every $x_0, \dots, x_n$, since the variables $X_0, \xi_1, \dots, \xi_n$ are independent and $\xi_1, \dots, \xi_n$ have the same law, then we have:

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) &= \mathbb{P}(X_0 = x_0, f(x_0, \xi_1) = x_1, \dots, f(x_{n-1}, \xi_n) = x_n)\\ &= \mathbb{P}(X_0 = x_0) \prod_{k=1}^{n} \mathbb{P}(f(x_{k-1}, \xi_k) = x_k)\\ &= \mathbb{P}(X_0 = x_0) \prod_{k=1}^{n} P(x_{k-1}, x_k).\end{aligned}$$

We conclude from Theorem 3.2.2 that $(X_n)_n$ is a $P$-Markov chain.

Now suppose that we are given a transition matrix $P$. Let us enumerate the state space as $\mathbb{X} = \{x_0, x_1 \dots\}$ and for any $k \geq 0$, let us decompose the interval $[0,1)$ as the disjoint union:

$$[0,1) = \bigcup_{\ell \geq 0} \Big[ \sum_{i=0}^{\ell-1} P(x_k, x_i), \sum_{i=0}^{\ell} P(x_k, x_i) \Big).$$

Further for $u \in [0,1)$, let $f(x_k, u) = x_\ell$ where $\ell \geq 0$ is the unique index such that

$$u \in \Big[ \sum_{i=0}^{\ell-1} P(x_k, x_i), \sum_{i=0}^{\ell} P(x_k, x_i) \Big).$$

Let $(U_n)_{n\geq 1}$ be i.i.d. with the uniform distribution on $[0,1)$ and independent of $X_0$. By the first part of the proof, the sequence defined recursively by $X_{n+1} = f(X_n, U_{n+1})$ is a homogeneous Markov chain with transition matrix given for every $k, \ell \geq 0$ by:

$$\mathbb{P}(f(x_k, U_1) = x_\ell) = \mathbb{P}\Big( U_1 \in \Big[ \sum_{i=0}^{\ell-1} P(x_k, x_i), \sum_{i=0}^{\ell} P(x_k, x_i) \Big) \Big) = P(x_k, x_\ell),$$

hence its transition matrix is indeed $P$. □

In the next chapters, we will be interested in the asymptotic behaviour of a Markov chain. Recall that recursive sequences, of the form $x_{n+1} = f(x_n)$, say with $f$ continuous, if convergent, necessarily converge to a fixed point of the function $f$. The analogue here is given by the notion of *stationary measure*.

**Definition 3.3.2.** A measure $\mu$ on $\mathbb{X}$ is said to be *stationary* or *invariant* for the transition matrix $P$ when for every $y \in \mathbb{X}$,

$$\mu(y) = \sum_{x \in \mathbb{X}} \mu(x) P(x, y),$$

which we shall simply write in the matrix form $\mu = \mu P$.

Every measure $\mu$ on $\mathbb{X}$ will be implicitly assumed to be $\sigma$-finite and non identically equal to zero, which means $\mu(x) < \infty$ for every $x \in \mathbb{X}$ and $\mu(x) > 0$ for at least one $x \in \mathbb{X}$.

**Notation.** We will denote by $\mu$ a stationary *measure*, and by $\pi$ and stationary *probability*, that is a stationary measure with $\pi(\mathbb{X}) = 1$.

The adjective 'stationary' comes from the following observation: if $\pi$ is stationary, that is $\pi = \pi P$, then by iterating this identity we have more generally $\pi = \pi P^n$ for every $n \geq 1$, which is the law of $X_n$ when $X_0$ has the law $\pi$. Hence, we start from a stationary law, then at every time, the Markov chain is distributed as this law. This can actually be strengthened as follows.

**Proposition 3.3.3.** *Let $(X_n)_{n\geq 0}$ be a $P$-Markov chain and suppose that $X_0$ has the law $\pi$. Then $\pi$ is stationary if and only if for any $k \geq 1$, the process $(X_{n+k})_{n\geq 0}$ has the same law as $(X_n)_{n\geq 0}$.*

*Proof.* Suppose first that $\pi$ is stationary, that is $\pi = \pi P = \pi P^k$ for every $k \geq 1$. Then for every $x_{k+1}, \dots, x_{k+n} \in \mathbb{X}$, we have by Theorem 3.2.2:

$$\begin{aligned}
\mathbb{P}_\pi(X_k = x_k, \dots, X_{k+n} = x_{k+n}) &= \sum_{x_1, \dots, x_{k-1}} \mathbb{P}_\pi(X_1 = x_1, \dots, X_{k+n} = x_{k+n}) \\
&= \sum_{x_0, \dots, x_{k-1}} \pi(x_0) \prod_{i=1}^{k} P(x_{i-1}, x_i) \prod_{i=k+1}^{k+n} P(x_{i-1}, x_i) \\
&= \pi P^k(x_k) \prod_{i=k+1}^{k+n} P(x_{i-1}, x_i) \\
&= \pi(x_k) \prod_{i=1}^{n} P(x_{k+i-1}, x_{k+i}).
\end{aligned}$$

Theorem 3.2.2 then shows that $(X_{n+k})_{n\geq 0}$ is a $P$-Markov chain with initial distribution $\pi$, exactly as $(X_n)_{n\geq 0}$.

Conversely, if $(X_{n+1})_{n\geq 0}$ has the same law as $(X_n)_{n\geq 0}$, then in particular $X_1$, which has law $\pi P$, has the same law as $X_0$, which is $\pi$. □

The next result shows that, as fixed points for recursive sequences, stationary distributions are in some sense the only possible limits of a Markov chain.

**Proposition 3.3.4.** *Let $(X_n)_{n\geq 0}$ be a $P$-Markov chain with any initial distribution and let $\pi$ be a probability on $\mathbb{X}$. Suppose that for every $x \in \mathbb{X}$, it holds:*

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}(X_k = x) \underset{n\to\infty}{\longrightarrow} \pi(x).$$

*Then $\pi$ is stationary.*

*Proof.* Let $\pi_n(x) = n^{-1}\sum_{k=0}^{n-1}\mathbb{P}(X_k = x)$, which defines a probability as a convex sum of probabilities. Fix $y \in \mathbb{X}$ and let us prove that $\pi_n P(y)$ converges both to $\pi P(y)$ and to $\pi(y)$ so these quantities are equal.

First, fix $\varepsilon > 0$; since $\pi$ is a probability, then there exists a *finite* set $A \subset \mathbb{X}$ such that $\pi(A^c) = 1 - \pi(A) < \varepsilon$. It follows that $\pi_n(A^c) = 1 - \pi_n(A) \longrightarrow 1 - \pi(A) < \varepsilon$. Suppose that $n$ is large enough so $\pi_n(A^c) < 2\varepsilon$, then:

$$\begin{aligned}|\pi_n P(y) - \pi P(y)| &\leq \sum_{x\in A} |\pi_n(x) - \pi(x)|P(x,y) + \sum_{x\in A^c} \pi_n(x)P(x,y) + \sum_{x\in A^c}\pi(x)P(x,y)\\ &\leq \sum_{x\in A}|\pi_n(x) - \pi(x)|P(x,y) + 3\varepsilon.\end{aligned}$$

The last sum converges to 0 as $n \longrightarrow \infty$ since $A$ is finite. Since $\varepsilon$ is arbitrary, then we conclude that $\pi_n P(y) \longrightarrow \pi P(y)$.

Next, we use the precise form of $\pi_n$. Recall that $P(x,y) = \mathbb{P}(X_{k+1} = y \mid X_k = x)$ for any $k$, then

$$\pi_n P(y) = \frac{1}{n}\sum_{k=0}^{n-1}\sum_{x\in\mathbb{X}}\mathbb{P}(X_k = x)P(x,y) = \frac{1}{n}\sum_{k=0}^{n-1}\sum_{x\in\mathbb{X}}\mathbb{P}(X_k = x, X_{k+1} = y) = \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{P}(X_{k+1} = y).$$

The right-hand side equals precisely

$$\frac{n+1}{n}\pi_{n+1}(y) - \frac{\mathbb{P}(X_0 = y)}{n} = \frac{n+1}{n}\pi_{n+1}(y) - \frac{\mathbb{P}(X_0 = y)}{n}$$

which converges to $\pi(y)$ as $n \longrightarrow \infty$, so indeed $\pi_n P(y) \longrightarrow \pi(y)$. □

Let us note that the assumption is satisfied in each of the following two cases:

(i) If for every $x \in \mathbb{X}$, we have the convergence in probability of the proportion of time spent at $x$, namely:

$$\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{1}_{X_k = x} \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \pi(x).$$

Then $n^{-1}\sum_{k=0}^{n-1}\mathbb{P}(X_k = x) \longrightarrow \pi(x)$ follows from dominated convergence (the sequence is clearly dominated by 1).

(ii) If the law of $X_n$ converges to $\pi$, namely for every $x \in \mathbb{X}$,

$$\mathbb{P}(X_n = x) \underset{n\to\infty}{\longrightarrow} \pi(x).$$

Then $n^{-1}\sum_{k=0}^{n-1}\mathbb{P}(X_k = x) \longrightarrow \pi(x)$ follows from basic calculus.

In the next chapters, we shall provide conditions under which the stationary probability exists and is unique (Corollary 4.2.11), and under which these two cases occur (Corollary 5.1.2 and Theorem 5.2.8 respectively). Proposition 3.3.4 simply shows that the limit has to be a stationary probability.

## 3.4 Stopping times and the strong Markov property

Recall from Corollary 3.2.7 that given a $P$-Markov chain $(X_n)_{n\geq 0}$, for any fixed time $N \geq 1$, the futur process $(X_{N+n})_{n\geq 0}$, conditionally on the value of $X_N$, remains a $P$-Markov chain, started from position $X_N$ and independent of the past $(X_0, \dots, X_{N-1})$. Now imagine that we follow the Markov chain until it reaches a given point $x$ for the first time, it is natural to believe that the futur evolution after this random time is again that of a $P$-Markov chain, started from position $x$ and independent of the past. The good notion of random times to extend to the Markov property is the notion of stopping time.

**Definition 3.4.1.** A *stopping time* relative to a stochastic process $(X_n)_{n\geq 0}$ is a random variable $T$ taking values in $\overline{\mathbb{Z}}_+ = \{0, 1, 2, \dots, \infty\}$ such that for any $n \geq 0$, the event $\{T \leq n\}$ is completely characterised by the random variables $X_0, \dots, X_n$. Formally, for any $n \geq 0$, there exists a measurable function $\varphi_n : \mathbb{X}^{n+1} \to \overline{\mathbb{Z}}_+$ such that

$$\mathbb{1}_{T\leq n} = \varphi_n(X_0, \dots, X_n).$$

In words, a stopping time is a random time which is determined by the past: the trajectory up to the present time is sufficient to tell wether is has already occurred or not yet.

**Exercise 3.4.2.** Prove that if we replace $\{T \leq n\}$ by $\{T = n\}$ then the two definitions coincide.

One can notice that constant random variables $T = N$ for any given $N \in \overline{\mathbb{Z}}_+$ are stopping times: simply take $\varphi_n$ to be equal to 0 for $n < N$ and to 1 for $n \geq N$.

**Example 3.4.3.** Important stopping times are given by the first entry time of the process: fix $A$ a subset of $\mathbb{X}$, then

$$T = \inf\{n \geq 0 : X_n \in A\}$$

is a stopping time, with the convention that $\inf \emptyset = \infty$. Indeed, we have simply:

$$\{T \leq n\} = \bigcup_{k\leq n} \{X_k \in A\},$$

which only depends on $X_0, \dots, X_n$.

It is important to be able to deal with multiple stopping times and we encourage the reader to prove the following elementary results.

**Exercise 3.4.4.** Let $(T_k)_{k\geq 1}$ be stopping times relative to the same stochastic process. Then $\sum_k T_k$, $\inf_k T_k$, $\sup_k T_k$, $\liminf_k T_k$, $\limsup_k T_k$ are all stopping times. In general, the difference is *not*, even in the case $T - 1$ where $T \geq 1$ a.s.

The next extension of the Markov property is very useful since it allows to restart the process afresh at any random stopping time.

**Theorem 3.4.5** (Strong Markov property)**.** *Let $(X_n)_{n\geq 0}$ be a $P$-Markov chain and let $T$ be a stopping time. Fix $x \in \mathbb{X}$. Then conditionally on $\{T < \infty\} \cap \{X_T = x\}$ the process given by $Y_n = X_{T+n}$ for $n \geq 0$ is a $P$-Markov chain started from $Y_0 = x$ and is independent from $(X_0, \dots, X_{T-1})$.*

*Proof.* Recall that if $T = N$ is a deterministic time, then the claim corresponds to Corollary 3.2.7. Now let us split according to these events:

$$\begin{aligned}
&\mathbb{P}(T = N, X_0 = x_0, \dots, X_{N-1} = x_{N-1}, X_N = x, X_{N+1} = y_1, \dots, X_{N+n} = y_n) \\
&= \mathbb{P}(T = N, X_0 = x_0, \dots, X_{N-1} = x_{N-1}, X_N = x)\, \mathbb{P}(X_{N+1} = y_1, \dots, X_{N+n} = y_n \mid X_N = x) \\
&= \mathbb{P}(T = N, X_0 = x_0, \dots, X_{N-1} = x_{N-1}, X_N = x)\, \mathbb{P}(X_1 = y_1, \dots, X_n = y_n \mid X_0 = x).
\end{aligned}$$

Summing over all values of $N$, we obtain for every $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$,

$$\begin{aligned}&\mathbb{P}(T<\infty, X_0=x_0,\dots,X_{T-1}=x_{T-1},X_T=x,X_{T+1}=y_1,\dots,X_{T+n}=y_n)\\&=\mathbb{P}(T<\infty,X_0=x_0,\dots,X_{T-1}=x_{T-1},X_T=x)\,\mathbb{P}(X_1=y_1,\dots,X_n=y_n\mid X_0=x),\end{aligned}$$

and the result follows by dividing by $\mathbb{P}(T<\infty, X_T=x)$. □

**Remark 3.4.6.** The extension of Corollary 3.2.7 to stopping times may seem unnecessary since in each case one can always condition on the value of $T$ and use the simple Markov property, when the time is fixed. However if this is true for discrete-time Markov chains, it is no longer for continuous-time Markov processes (studied next semester).

## 3.5 Harmonic functions and the Dirichlet problem (⋆)

This section relates Markov chains and discrete harmonic functions. We use probabilistic tools to solve the so-called discrete Dirichlet problem. In the same spirit, one can solve partial differential equations using random processes that evolve in continuous time and space. This is a very active topic of research, for its own sake but also for application to physics, biology, epidemiology, finance, etc. that allows to derive theoretical results but also provides numerical schemes for simulations.

**Definition 3.5.1.** Given a transition matrix $P$, a function $h:\mathbb{X}\to\mathbb{R}$ is said to be *harmonic* at $x$ when $Ph(x)=h(x)$. It said to be harmonic on $A\subset\mathbb{X}$ if it is so at every $x\in A$.

We can also define similarly subharmonic ($Ph(x)\geq h(x)$) and superharmonic ($Ph(x)\leq h(x)$) functions, but we shall only consider harmonic functions.

**Remark 3.5.2.** Recall that if $(X_n)_{n\geq 0}$ is a $P$-Markov chain on $\mathbb{X}$, then $Ph(x)=\mathbb{E}_x[h(X_1)]$. A function $h$ is thus harmonic at $x$ when the average value after one step from $x$ is again $h(x)$.

Let $A\subset\mathbb{X}$ be nonempty and let $g:\mathbb{X}\setminus A\to\mathbb{R}$ be a function. The Dirichlet problem raises the question: does there exist a function $h:\mathbb{X}\to\mathbb{R}$ which coincides with $g$ on $\mathbb{X}\setminus A$ and which is harmonic on $A$? if so, is it unique? The harmonicity of such a function $h$ can be rewritten as $(P-I)h=0$ where $I$ is the identity matrix. This can be in many cases seen as a discretised differential equation (as in the finite difference method), as shown in the following example (see also the exercise sheet).

**Example 3.5.3.** Let $\mathbb{X}=\mathbb{Z}$ and let $P(i,j)=\frac{1}{2}\,\mathbb{1}_{|i-j|=1}$. This corresponds to the case where the increments of the Markov chain are i.i.d. with $\mathbb{P}(X_{n+1}=X_n+1)=\mathbb{P}(X_{n+1}=X_n-1)=1/2$. Then $(P-I)h=0$ if and only if

$$-\frac{1}{2}(h(x+1)-2h(x)+h(x-1))=0,$$

which is the discretised heat equation $-\Delta h=0$. The condition $h=g$ on $\mathbb{X}\setminus A$ is interpreted as a boundary condition (here a source of heat) of the equation and we aim at finding the solution in $A$. See Figure 3.2 for a representation.

Although the Dirichlet problem is deterministic, we may solve it using Markov chain theory. Recall indeed from Proposition 3.3.1 that there exists a Markov chain with transition matrix $P$. We shall follow its trajectory until it exits $A$ for the first time. Let us start with the easy case $g=0$.

**Lemma 3.5.4.** *Let $A\subset\mathbb{X}$ be nonempty and finite and let $T_A=\inf\{n\geq 0:\ X_n\in\mathbb{X}\setminus A\}$ be its first exit time. Suppose that for every $x\in A$, we have $\mathbb{P}_x(T_A<\infty)=1$. Then the only function which is $P$-harmonic on $A$ and null on $\mathbb{X}\setminus A$ is the constant null function.*

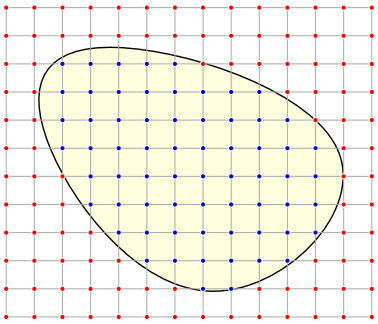


Figure 3.2: The Dirichlet problem consists in finding the values of a harmonic function in the domain $A$ with the given boundary value in $\mathbb{X} \setminus A$.

*Proof.* Clearly the null function is a solution to the problem. Suppose that $h$ is also a solution. We only assume that $A$ is finite to infer that $h$ admits a maximum on $A$: let $x_0 \in A$ be such that $h(x_0) = \max_{y\in A} h(y)$ and suppose that $h(x_0) > 0$. Then $h(x_0) = \max_{y\in\mathbb{X}} h(y)$ and since it is harmonic, then $\sum_{y\in\mathbb{X}} P(x_0, y)(h(y)-h(x_0)) = 0$ which yields $h(y) = h(x_0) = \max h$ for every $y \in \mathbb{X}$ such that $P(x_0, y) > 0$.

Recall that we assume that $\mathbb{P}_{x_0}(T_A < \infty) = 1$, which is equivalent to $\mathbb{P}_{x_0}(T_A \le n) \longrightarrow 1$; in particular this probability is not 0 for some (deterministic) $n$ and thus there exists a path $x_0, x_1, \dots, x_n$ such that $x_1, \dots, x_{n-1} \in A$ and $x_n \in \mathbb{X} \setminus A$ and which has $P(x_{i-1}, x_i) > 0$ for every $1 \le i \le n$. But the previous point then implies by induction that $0 < h(x_0) = h(x_1) = \cdots = h(x_n) = 0$.

We conclude by contradiction that $h(x_0) = 0$ and so $h(x) \le 0$ for every $x \in A$. Notice finally that if $h$ is a solution, then so is $-h$, so the same argument implies $h(x) \ge 0$ for every $x \in A$. □

Let us next turn to the general case; this lemma shall provide the uniqueness argument.

**Theorem 3.5.5.** *Let $A \subset \mathbb{X}$ be nonempty and finite and let $T_A = \inf\{n \ge 0 : X_n \in \mathbb{X} \setminus A\}$ be its first exit time. Suppose that for every $x \in A$, we have $\mathbb{P}_x(T_A < \infty) = 1$. Let $g : \mathbb{X} \setminus A \longrightarrow \mathbb{R}_+$ be a bounded function. Then there exists a unique bounded function $h$ on $\mathbb{X}$ that is $P$-harmonic on $A$ and coincides with $g$ on $\mathbb{X} \setminus A$. It is given by the formula:*

$$h(x) = \mathbb{E}_x[g(X_{T_A})]$$

*for every $x \in \mathbb{X}$.*

From a numerical point of view, the trajectory of a Markov chain is usually easy to implement (recall Proposition 3.3.1); let us simulate a large number, say $K$, of $P$-Markov chains all started at some $x \in A$ and until they first leave $A$, and let us evaluate for each one the function $g$ at their terminal value. Then the Law of Large Numbers shows that the average over these $K$ trajectories converges as $K \longrightarrow \infty$ to $\mathbb{E}_x[g(X_{T_A})] = h(x)$. Let us note that if we assume that $\mathbb{E}_x[T_A] < \infty$, then the Law of Large Numbers also shows that the sum of the length of these $K$ trajectories, i.e. the total number of iterations of the $K$ random recursions, is close to $K \times \mathbb{E}_x[T_A]$ when $K$ is large. The key point that explains the success of this approach is that, as opposed to algebraic schemes, the complexity of the algorithm is quite insensible to the dimension of the space.

*Proof.* Let us start with the uniqueness of the solution: if $h_1$ and $h_2$ are two solution, then $h = h_1 - h_2$ is $P$-harmonic on $A$ and null on $\mathbb{X} \setminus A$, so it is the constant null function by the previous lemma.

Let us next prove that the function $h$ is harmonic on $A$. Fix $x \in A$, and start from $X_0 = x$. There are two possibilities for $X_1$: either $X_1 \in A$, and then the process after time 1 starts from this value and is stopped when exiting $A$, or $X_1 \notin A$ and the process is stopped here. Formally, we infer from applying the Markov

property at time 1 that:

$$\begin{aligned} h(x) &= \mathbb{E}_x[g(X_{T_A})] \\ &= \mathbb{E}_x[g(X_{T_A})\,\mathbb{1}_{X_1\in A}] + \mathbb{E}_x[g(X_{T_A})\,\mathbb{1}_{X_1\notin A}] \\ &= \mathbb{E}_x[\mathbb{E}_{X_1}[g(X_{T_A})]\,\mathbb{1}_{X_1\in A}] + \mathbb{E}_x[g(X_1)\,\mathbb{1}_{X_1\notin A}] \\ &= \mathbb{E}_x[h(X_1)\,\mathbb{1}_{X_1\in A}] + \mathbb{E}_x[g(X_1)\,\mathbb{1}_{X_1\notin A}] \\ &= \mathbb{E}_x[h(X_1)] \\ &= Ph(x). \end{aligned}$$

Thus $h$ is indeed harmonic on $A$. It is also clear that $h = g$ on $\mathbb{X} \setminus A$. □

Let us next present an application of the previous theorem to the so-called first exit side problem. Imagine that we are given two subsets $B$ and $C$, which do not intersect, starting from an arbitrary point $x$, what is the probability that a $P$-Markov chain reaches $B$ before $C$?

**Corollary 3.5.6.** *Let $B, C \subset \mathbb{X}$ be two sets such that $B \cap C = \emptyset$ and $(B \cup C)^c$ is finite and nonempty. Let $\tau_B = \inf\{n \geq 0 : X_n \in B\}$, define $\tau_C$ similarly and assume that at least one of them is almost surely finite for every starting point. Let $g(x) = 1$ for every $x \in B$ and $g(x) = 0$ for every $x \in C$. Then*

$$\mathbb{P}_x(\tau_B < \tau_C) = h(x),$$

*where $h$ is the unique bounded extension of $g$ that is harmonic on $(B \cup C)^c$, given in the previous theorem.*

*Proof.* Since $B \cap C = \emptyset$ then $\tau_B \neq \tau_C$. Then either $\tau_B < \tau_C$ and then $g(X_{\tau_B \wedge \tau_C}) = 1$, or $\tau_B > \tau_C$ and then $g(X_{\tau_B \wedge \tau_C}) = 0$. Thus

$$\mathbb{P}_x(\tau_B < \tau_C) = \mathbb{E}_x[g(X_{\tau_B \wedge \tau_C})].$$

This indeed equals $h(x)$ by the previous theorem, where $A = (B \cup C)^c$, so $\tau_B \wedge \tau_C = T_A$. □

A concrete example, known as the *ruin problem*, is detailed in the exercise sheet. In this problem the Markov chain is simply a random walk on $\mathbb{Z}$, with i.i.d. increments equal to +1 with some fixed probability $p \in (0, 1)$ and equal to −1 with probability −1. We take $B = [K, \infty)$ for some fixed $K \geq 2$ and $C = (-\infty, 0]$, and we let the chain start from some $k \in \{1, \dots, K-1\}$. We imagine this random walk as the fortune of a player betting repeatedly on Head or Tails and who wonders if starting with an initial fortune $k$, they can reach $K$ before getting ruined, see Figure 3.3.

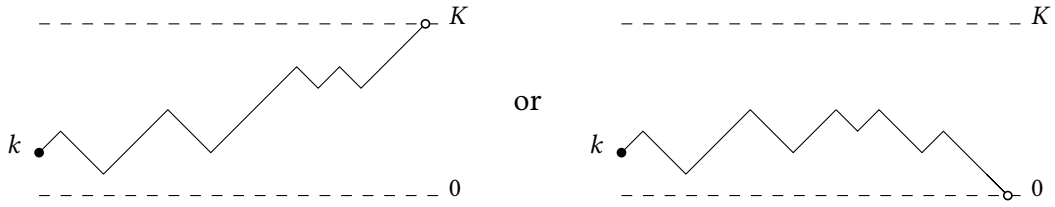



Figure 3.3: The ruin problem: does the player reach the top boundary before the bottom one?

For $p = 1/2$, we mentioned that the harmonicity condition is equivalent to solving:

$$-\frac{1}{2}(h(i+1) - 2h(i) + h(i-1)) = 0, \qquad \text{equivalently:} \qquad h(i+1) - h(i) = h(i) - h(i-1),$$

that is, the increments are all constant. Suppose furthermore that $h(i) = 1$ for $i \geq K$ and $h(i) = 0$ for $i \leq 0$, then the increments $h(i) - h(i-1)$ for $1 \leq i \leq K$ are all equal to $1/K$, that is:

$$h(i) = \frac{i}{K} \qquad 0 \leq i \leq K.$$

Finally, by the previous corollary, for each $1 \leq k < K$, we have:

$$\text{ruin probability} = \mathbb{P}_k(\tau_0 < \tau_K) = 1 - \frac{k}{K}.$$

# Chapter 4

# Classification of states

In this chapter, we consider the following questions: How many times does a (discrete) Markov chain visit a given point? does it always come back to its starting point or not? if so, how long does it take? We shall relate the answer to the problem of existence and uniqueness of a stationary measure or stationary probability which describes the asymptotic behaviour of the chain as shown in the next chapter. We shall also prove a famous result of Pólya: wandering randomly on the ground will always get you back home, but you may get lost in space!

**Contents**



In Section 4.1 we present a first dichotomy: recurrence vs. transience, that is wether a Markov chain always comes back to its starting point or it escapes and leaves forever. This makes an extensive use of the strong Markov property from the previous chapter. Then in Section 4.2 we discuss the existence and uniqueness of the stationary measures and distributions and we distinguish further between two different behaviours for recurrent points. Section 4.3 finally presents an application in the case of the simple random walk in the discrete $d$-dimensional space $\mathbb{Z}^d$: the walk always comes back to its starting point in dimension $d = 1$ or $d = 2$, but not in dimension $d \geq 3$.

## 4.1 Recurrence and Transience

Let us start by introducing the notation we shall use throughout this chapter and the next one.

**Notation.** For each $x \in \mathbb{X}$, let us define inductively the hitting times of $x$ by $H_x^0 = 0$ and for $k \geq 0$:

$$H_x^{k+1} = \inf\{n \geq H_x^k + 1 : X_n = x\}.$$

For $k = 1$, simply write $H_x$ for $H_x^1 = \inf\{n \geq 1 : X_n = x\}$. Note that the starting point does not count and $H_x \geq 1$. Also,

$$H_x^k = \inf\{n \geq 1 : \#\{i \in \{1, \dots, n\} : X_i = x\} = k\}.$$

Let also

$$V_x = \sum_{n \geq 0} \mathbb{1}_{X_n = x}$$

denote the number of visits of $x$. Finally, for $x, y \in \mathbb{X}$, let us denote by

$$\rho_{xy} = \mathbb{P}_x(H_y < \infty) = \mathbb{P}_x(\exists n \geq 1 : X_n = y)$$

the probability to reach $y$ when starting from $x$. When $\rho_{xy} > 0$, we say that $x$ *leads to* $y$ and we denote this by $x \longrightarrow y$. When a Markov chain is described as a walk on a graph, $x \longrightarrow y$ means that there exists a path of arrows from $x$ to $y$. Note that $x$ may not lead to itself.

### 4.1.1 Number of visits

A point is called 'recurrent' when the Markov chain always comes back to it when starting from it and 'transient' otherwise.

**Definition 4.1.1.** For every $x \in \mathbb{X}$, we have the dichotomy:

- either $\rho_{xx} = 1$, then $x$ is said to be *recurrent*,
- or $\rho_{xx} < 1$, then $x$ is said to be *transient*.

The next result shows that the recurrent/transience dichotomy is quite strong: if $x$ is recurrent, then the chain visits $x$ infinitely many times whereas if it is transient, it only visits it a random geometric number of times.

**Proposition 4.1.2.** *The following holds according as wether $x$ is recurrent or transient:*

*(i) If $\rho_{xx} = 1$, then $\mathbb{P}_x(V_x = \infty) = 1$.*

*(ii) If $\rho_{xx} < 1$, then $\mathbb{P}_x(V_x < \infty) = 1$, and precisely $V_x$ has under $\mathbb{P}_x$ the geometric law with mean*

$$\mathbb{E}_x[V_x] = \frac{1}{1-\rho_{xx}} = \frac{1}{\mathbb{P}_x(H_x = \infty)}.$$

The proof is based on the following idea: In order to visit $k$ times the point $y$ when starting from $x$, one first has to reach $y$ and then come back to it $k-1$ times. The next lemma formalises this idea thanks to the strong Markov property.

**Lemma 4.1.3.** *For every $x, y \in \mathbb{X}$ and $k \geq 1$, it holds:*

$$\mathbb{P}_x(H_y^k < \infty) = \mathbb{P}_x(H_y < \infty)\,\mathbb{P}_y(H_y < \infty)^{k-1} = \rho_{xy}\rho_{yy}^{k-1}.$$

*In particular $\mathbb{P}_x(H_x^k < \infty) = \rho_{xx}^k$.*

*Proof.* According to Theorem 3.4.5, under $\mathbb{P}_x$, conditionally on $\{H_y < \infty\}$ and since $X_{H_y} = y$ a.s. the process given by $Y_n = X_{H_y+n}$ for $n \geq 0$ has the same law as $(X_n)_{n\geq 0}$ started from $Y_0 = y$. Moreover, the quantity $H_y^k$ for the chain $(X_n)_n$ equals the quantity $H_y^{k-1}$ for the chain $(Y_n)_n$ and thus:

$$\begin{aligned}\mathbb{P}_x(H_y^k < \infty) &= \mathbb{P}_x(H_y < \infty, H_y^k < \infty)\\ &= \mathbb{P}_x(H_y < \infty)\,\mathbb{P}_x(H_y^k < \infty \mid H_y < \infty)\\ &= \mathbb{P}_x(H_y < \infty)\,\mathbb{P}_y(H_y^{k-1} < \infty).\end{aligned}$$

Taking $x = y$ and $k-1$ instead of $k$, we then get $\mathbb{P}_y(H_y^{k-1} < \infty) = \mathbb{P}_y(H_y < \infty)\,\mathbb{P}_y(H_y^{k-2} < \infty)$ and the claim follows by induction. □

Proposition 4.1.2 then easily follows.

*Proof of Proposition 4.1.2.* According to Lemma 4.1.3 we have for every $k \geq 1$:

$$\mathbb{P}_x(V_x \geq k+1) = \mathbb{P}_x(H_x^k < \infty) = \mathbb{P}_x(H_x < \infty)^k = \rho_{xx}^k.$$

(i) Letting $k \longrightarrow \infty$ in the previous equation, by monotonicity, we conclude that

$$\mathbb{P}_x(V_x = \infty) = \downarrow \lim_{k\to\infty} \mathbb{P}_x(V_x \geq k+1) = 1.$$

(ii) In this case, the identity $\mathbb{P}_x(V_x \geq k+1) = \rho_{xx}^k$ for every $k \geq 1$, with $\rho_{xx} < 1$ shows that $V_x$ has the geometric distribution with parameter $1 - \rho_{xx} > 0$. □

**Remark 4.1.4.** Since $x$ has to be either recurrent or transient, then we see that it is recurrent if and only if

$$\mathbb{E}_x[V_x] = \sum_{n\geq 0} \mathbb{P}_x(X_n = x) = \sum_{n\geq 0} P^n(x,x) = \infty,$$

and in this case we actually have $\mathbb{P}_x(V_x = \infty) = 1$. This criterion is often easy to check in practice given the matrix $P$.

**Remark 4.1.5.** Notice that if $y$ is transient then for any $x \neq y$, we have since $\mathbb{1}_{H_y=\infty} V_y = 0$:

$$\mathbb{E}_x[V_y] = \mathbb{E}_x[\mathbb{1}_{H_y<\infty} V_y] = \mathbb{P}_x(H_y < \infty)\, \mathbb{E}_x[V_y \mid H_y < \infty].$$

By the strong Markov property, since $X_{H_y} = y$, then the last conditional expectation equals $\mathbb{E}_y[V_y]$. We thus have:

$$\mathbb{E}_x[V_y] = \mathbb{P}_x(H_y < \infty)\, \mathbb{E}_y[V_y] \leq \mathbb{E}_y[V_y] < \infty.$$

Hence, whatever the starting point $X_0 = x$, the number of visits of a transient point $y$ has finite expectation (in particular it is finite almost surely).

Let us turn our attention to recurrent points.

**Proposition 4.1.6.** *If $x$ is recurrent and $\rho_{xy} > 0$, then $y$ is recurrent as well and we have $\mathbb{P}_x(V_y = \infty) = \mathbb{P}_y(V_x = \infty) = 1$.*

*Proof.* Fix $x \neq y$. Let us first prove that $\rho_{xy} = 1$. Since $x$ is recurrent, then a.s. we have $H_x^k < \infty$ for all $k \geq 1$ so we may write:

$$\begin{aligned}\mathbb{P}_x(H_y = \infty) &= \mathbb{P}_x\Big(\bigcap_{i\geq 1} \{H_x^i < \infty\} \cap \{y \notin \{X_{H_x^{i-1}+1}, \dots, X_{H_x^i}\}\}\Big) \\ &= \downarrow \lim_{k\to\infty} \mathbb{P}_x\Big(\bigcap_{i=1}^{k} \Big(\{H_x^i < \infty\} \cap \{y \notin \{X_{H_x^{i-1}+1}, \dots, X_{H_x^i}\}\}\Big)\Big).\end{aligned}$$

By induction, applying the strong Markov property at time $H_x^{k-1}$, then $H_x^{k-2}$, etc. since $X_{H_x^i} = x$, we see that the last probability equals

$$\mathbb{P}_x(\{H_x < \infty\} \cap \{H_y > H_x\})^k = \mathbb{P}_x(H_y > H_x)^k.$$

By letting $k \longrightarrow \infty$, we obtain that

$$\mathbb{P}_x(H_y = \infty) = \downarrow \lim_{k\to\infty} \mathbb{P}_x(H_y > H_x)^k,$$

which equals either 0, when $\mathbb{P}_x(H_y > H_x) < 1$, or 1, when $\mathbb{P}_x(H_y > H_x) = 1$. Since we assume that $\rho_{xy} = 1 - \mathbb{P}_x(H_y = \infty) > 0$, then $\rho_{xy} = 1$.

Let next us prove that $\mathbb{P}_y(V_x = \infty) = 1$. Indeed, recall that $\mathbb{P}_x(V_x = \infty) = 1$ from Proposition 4.1.2. Since $\mathbb{P}_x(H_y < \infty) = 1$, then

$$1 = \mathbb{P}_x(H_y < \infty, V_x = \infty) = \mathbb{P}_x\Big(H_y < \infty, \sum_{n\geq H_y} \mathbb{1}_{X_n = x} = \infty\Big) = \mathbb{P}_x\Big(\sum_{n\geq 0} \mathbb{1}_{X_{H_y+n}=x} = \infty \mid H_y < \infty\Big).$$

Let $Y_n = X_{H_y+n}$ for every $n \geq 0$. Then Theorem 3.4.5 states that conditionally on $\{H_y < \infty\}$, since $X_{H_y} = y$, then $(Y_n)_n$ is another Markov chain with the same law as $(X_n)_n$ but started from $y$. Hence the right-hand side above equals $\mathbb{P}_y(V_x = \infty) = 1$.

As a consequence, letting $H_{x,y} = \inf\{n \geq H_x + 1 : X_n = y\}$ denote the first return time to $y$ after visiting $x$, we have:

$$\mathbb{P}_y(H_y < \infty) \geq \mathbb{P}_y(H_x < \infty, H_{x,y} < \infty) = \mathbb{P}_y(H_x < \infty)\,\mathbb{P}_y(H_{x,y} < \infty \mid H_x < \infty).$$

First note that

$$\mathbb{P}_y(H_x < \infty) \geq \mathbb{P}_y(V_x = \infty) = 1.$$

Next, by Theorem 3.4.5,

$$\mathbb{P}_y(H_{x,y} < \infty \mid H_x < \infty) = \mathbb{P}_x(H_y < \infty) = 1.$$

Thus $\mathbb{P}_y(H_y < \infty) = 1$ and $y$ is recurrent. The identity $\mathbb{P}_x(V_y = \infty) = 1$ then follows by exchanging the role of $x$ and $y$. □

### 4.1.2 Communicating classes

Recall the notation $x \longrightarrow y$ when $\rho_{xy} = \mathbb{P}_x(H_y < \infty) > 0$. If both $x \longrightarrow y$ and $y \longrightarrow x$, we write $x \longleftrightarrow y$ and say that *$x$ and $y$ communicate with each other*. We then set $x \sim y$ if $x = y$ or $x \longleftrightarrow y$. As the notation suggests, the relation $\sim$ is an equivalence relation.

**Lemma 4.1.7.** *For every $x, y \in \mathbb{X}$ we have $x \longrightarrow y$ if and only if there exists $n \geq 1$ such that $\mathbb{P}(X_n = y) > 0$. Moreover the relation $\sim$ is an equivalence relation.*

*Proof.* We simply write for $n \geq 1$:

$$\mathbb{P}_x(X_n = y) \leq \mathbb{P}_x\Big(\bigcup_{n\geq 1}\{X_n = y\}\Big) \leq \sum_{n\geq 1}\mathbb{P}_x(X_n = y).$$

Therefore if $x \longrightarrow y$, that is, the probability in the middle is nonzero, then there necessarily exists $n \geq 1$ such that $\mathbb{P}_x(X_n = y) > 0$. The first inequality shows that this is an equivalence. Consequently, if $x \longrightarrow y$ and $y \longrightarrow z$, then there exist $n, m \geq 1$ such that $\mathbb{P}_x(X_n = y) > 0$ and $\mathbb{P}_y(X_m = z) > 0$. We then infer from the Markov property that

$$\mathbb{P}_x(X_{n+m} = z) \geq \mathbb{P}_x(X_n = y, X_{n+m} = z) = \mathbb{P}_x(X_n = y)\,\mathbb{P}_y(X_m = z) > 0,$$

hence $x \longrightarrow z$. This suffices to conclude that $\sim$ is indeed an equivalence relation. □

We may then partition $\mathbb{X}$ into the equivalence classes, which we call the *communicating classes*. Proposition 4.1.6 shows that starting from a recurrent position we can only visit recurrent states. This leads to the following classification.

**Theorem 4.1.8.** *There exists a partition of the space into disjoint subsets:*

$$\mathbb{X} = \mathcal{T} \cup \bigcup_{i\in I}\mathcal{R}_i$$

*such that:*

- *For every $i \in I$ and every $x \in \mathcal{R}_i$, we have $\mathbb{P}_x$-a.s.*

$$V_y = \infty \text{ for all } y \in \mathcal{R}_i \qquad \text{and} \qquad V_y = 0 \text{ for all } y \in \mathbb{X} \setminus \mathcal{R}_i.$$

- *For every $x \in \mathcal{T}$, if $\tau = \inf\{n \geq 1 : X_n \in \bigcup_{i\in I}\mathcal{R}_i\}$, then $\mathbb{P}_x$-a.s.*
  - *either $\tau = \infty$ and $V_y < \infty$ for all $y \in \mathbb{X}$,*
  - *or $\tau < \infty$ and there exists a random index $i \in I$ such that $X_n \in \mathcal{R}_i$ for every $n \geq \tau$.*

*Proof.* The set $\mathcal{T}$ is that of all transient states, whereas $\mathcal{R} = \bigcup_{i\in I} \mathcal{R}_i$ is that of recurrent states. By Proposition 4.1.6, the relation defined on $\mathcal{R}$ by $x \sim y$ if and only if $\rho_{xy} > 0$ is an equivalence relation, then the $\mathcal{R}_i$'s are the corresponding equivalence classes.

Fix such an equivalence class $\mathcal{R}_i$ and $x \in \mathcal{R}_i$. Then Proposition 4.1.6 shows that $\mathbb{P}_x(V_y = \infty) = 1$ for all $y \in \mathcal{R}_i$. On the other hand if $y \notin \mathcal{R}_i$, i.e. if $\rho_{xy} = 0$, then clearly $\mathbb{P}_x(V_y = 0) = 1$. Finally fix $x \in \mathcal{T}$. If $\tau = \infty$ then $V_y = 0$ for every $y \in \mathcal{R}$; also $V_y < \infty$ for all $y \in \mathcal{T}$ by Proposition 4.1.2. Suppose next that $\tau < \infty$, then there exists a random index $i \in I$ such that $X_\tau \in \mathcal{R}_i$ and by the strong Markov property (Theorem 3.4.5), we are back in the first case where the chain starts from a point in $\mathcal{R}_i$. □

**Definition 4.1.9.** If $\rho_{xy} > 0$ for every $x, y \in \mathbb{X}$ then we say that the Markov chain is *irreducible*.

In what follows we will always assume that our chains are irreducible, otherwise we can study each class separately.

**Corollary 4.1.10.** *If the chain is irreducible, then we are in one of the two following situations:*

- *Either every $x \in \mathbb{X}$ is recurrent and $\mathbb{P}_x(V_y = \infty) = 1$ for all $x, y \in \mathbb{X}$,*
- *Or every $x \in \mathbb{X}$ is transient and $\mathbb{E}_x[V_y] < \infty$ for all $x, y \in \mathbb{X}$.*

*Note that if $\mathbb{X}$ is a finite set, then we are necessarily in the first case.*

*Proof.* If there exists $x \in \mathbb{X}$ which is recurrent, then by Proposition 4.1.6 so is every other $y \in \mathbb{X}$ so indeed either every state is recurrent or every state is transient. In the latter case we have then $\mathbb{E}_x[V_y] < \infty$ for all $y \in \mathbb{X}$ by Remark 4.1.5, whereas in the first case we have $V_y = \infty$ for all $y \in \mathbb{X}$ by Proposition 4.1.6 again. Finally,

$$\infty = \sum_{n\geq 0} \sum_{x\in\mathbb{X}} \mathbb{1}_{X_n=x} = \sum_{x\in\mathbb{X}} \sum_{n\geq 0} \mathbb{1}_{X_n=x} = \sum_{x\in\mathbb{X}} V_x.$$

Therefore, if $\mathbb{X}$ is a finite set, then $V_x$ has to be infinite for at least one $x \in \mathbb{X}$. □

**Definition 4.1.11.** If the chain is irreducible, then we say that it is *recurrent* or *transient* according as wether every state is recurrent or transient.

## 4.2 Stationary measures

Recall that Markov chains are random analogues of recursive sequences of the form $x_{n+1} = f(x_n)$, which converge (when they do) to a fixed point of the function $f$ (if the latter is continuous). The analogue of fixed points is given by the notion of *stationary measures* defined in Section 3.3, that is, a measure $\mu$ solution to

$$\mu = \mu P.$$

In this section we analyse the existence and uniqueness of such measures in general, a more precisely of such probabilities $\pi = \pi P$.

Every measure $\mu$ on $\mathbb{X}$ will be implicitly assumed to be $\sigma$-finite and non identically equal to zero, which means $\mu(x) < \infty$ for every $x \in \mathbb{X}$ and $\mu(x) > 0$ for at least one $x \in \mathbb{X}$.

### 4.2.1 An easy subcase: reversibility

The stationarity property will be very important, but in practice it can be hard to check: one has to solve the equation $\mu = \mu P$, that is find a eigenvector associated with the eigenvalue 1 for the transpose matrix $P^t$. Even in a finite set $\mathbb{X}$, but with a large cardinal, this system can be difficult to solve exactly. A simpler, but stronger, property is that of reversibility.

**Definition 4.2.1.** A measure $\mu$ on $\mathbb{X}$ is said to be *reversible* for the transition matrix $P$ when for every $x, y \in \mathbb{X}$,

$$\mu(x)P(x,y) = \mu(y)P(y,x).$$

This condition is also often called "detailed balance condition".

**Proposition 4.2.2.** *A reversible measure is stationary.*

*Proof.* Simply sum the reversibility identity over $x$:

$$\sum_{x\in\mathbb{X}} \mu(x)P(x,y) = \sum_{x\in\mathbb{X}} \mu(y)P(y,x).$$

The left-hand side equals $\mu P(y)$ and the right-hand side $\mu(y)$. □

**Example 4.2.3.** Fix $N \geq 1$ and let $(X_n)_n$ be the following process: let $p \in (0,1)$ and $q = 1-p$, let $X_0 = x_0$ be a fixed value in $\mathbb{X} = \{0, \dots, N-1\}$, and then iteratively set:

$$\mathbb{P}(X_{n+1} = X_n + 1 \bmod N \mid X_n) = p, \qquad \mathbb{P}(X_{n+1} = X_n - 1 \bmod N \mid X_n) = q.$$

Then this process is a Markov chain, with transition matrix

$$P = \begin{pmatrix} 0 & p & 0 & \cdots & 0 & q \\ q & 0 & p & \ddots & \ddots & 0 \\ 0 & q & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & p & 0 \\ 0 & 0 & \ddots & q & 0 & p \\ p & 0 & \cdots & 0 & q & 0 \end{pmatrix}.$$

Let $\pi(j) = 1/N$ for every $j \in \mathbb{X}$ denote the uniform distribution, then

$$\pi(j)P(j,k) = \frac{1}{N}(p\,\mathbb{1}_{k=j+1 \bmod N} + q\,\mathbb{1}_{k=j-1 \bmod N}),$$

for all $j, k \in \mathbb{X}$. Thus on the one hand for any $k \in X$, one has

$$\pi P(k) = \sum_{k=0}^{N-1} \pi(j)P(j,k) = \frac{q+p}{N} = \frac{1}{N} = \pi(k),$$

so $\pi$ is always stationary. On the other hand, one has $\pi(j)P(j,k) = \pi(k)P(k,j)$ only in the case $p = q = 1/2$ so $\pi$ is reversible in this case and when $p \neq q$, there is no reversible distribution.

**Proposition 4.2.4.** *Suppose that $\mu$ is stationary for the transition matrix $P$ and define for every $x, y \in \mathbb{X}$:*

$$P^*(x,y) = \frac{\mu(y)}{\mu(x)}P(y,x).$$

*Then $P^*$ is a transition matrix and $\mu$ is also stationary for $P^*$. In addition $\mu$ is reversible for $P$ if and only if $P = P^*$. Suppose moreover that $\mu$ is a probability and let $(X_n)_n$ denote a $P$–Markov chain with initial distribution $\mu$ and $(X_n^*)_n$ a $P^*$–Markov chain with initial distribution $\mu$. Then for every $n \geq 0$, we have the identity in law:*

$$(X_0^*, \dots, X_n^*) \overset{(d)}{=} (X_n, \dots, X_0).$$

*Finally the probability $\mu$ is reversible for $P$ if and only if we have the identity in law:*

$$(X_0, \dots, X_n) \overset{(d)}{=} (X_n, \dots, X_0)$$

*for every $n \geq 0$.*

This explains the name 'reversible': if one starts with a reversible initial distribution, then the time-reversed process at any time has the same law as the original one.

*Proof.* Clearly $P^*(x,y) \geq 0$ and further, since $\mu = \mu P$, then:

$$\sum_{y\in\mathbb{X}} P^*(x,y) = \sum_{y\in\mathbb{X}} \frac{\mu(y)}{\mu(x)} P(y,x) = \frac{1}{\mu(x)} \mu P(x) = 1.$$

Hence $P^*$ is indeed a transition matrix. Next,

$$\mu P^*(y) = \sum_{x\in\mathbb{X}} \mu(x) P^*(x,y) = \sum_{x\in\mathbb{X}} \mu(y) P(y,x) = \mu(y),$$

so $\mu$ is $P^*$-stationary. Finally,

$$\mu(x)P^*(x,y) = \mu(y)P(y,x) \qquad \text{and} \qquad \mu(y)P^*(y,x) = \mu(x)P(x,y),$$

so $\mu$ is $P^*$-stationary if and only if it is $P$-stationary.

Suppose next that $\mu$ is a probability and let $(X_n)_n$ and $(X_n^*)_n$ be Markov chains with initial distribution $\mu$ and with transition matrix $P$ and $P^*$ respectively. Then by the Chapman–Kolmogorov Equation (twice), we have:

$$\begin{aligned}
\mathbb{P}(X_0^* = x_0, \dots, X_n^* = x_n) &= \mu(x_0) \prod_{k=1}^{n} P^*(x_{k-1}, x_k) \\
&= \mu(x_0) \prod_{k=1}^{n} \frac{\mu(x_k)}{\mu(x_{k-1})} P(x_k, x_{k-1}) \\
&= \mu(x_n) \prod_{k=1}^{n} P(x_k, x_{k-1}) \\
&= \mathbb{P}(X_0 = x_n, \dots, X_n = x_0).
\end{aligned}$$

This proves the identity in law: $(X_0^*, \dots, X_n^*) = (X_n, \dots, X_0)$. Finally, we have shown that $\mu$ is reversible if and only if $P = P^*$, which is equivalent by the previous identity to $(X_0, \dots, X_n) = (X_n, \dots, X_0)$ in law. □

**Example 4.2.5.** Let $V$ be a set, either finite or countable, and let $E \subset \{\{u,v\} : u, v \in V, u \neq v\}$ be a set of unordered pairs of elements in $V$. We call $V$ the vertices, $E$ the edges, and the pair $G = (V,E)$ a graph. Suppose that each edge $e = \{u,v\} \in E$ has a weight $c_e \in (0,\infty)$ and define for every $u \in V$:

$$\mu(u) = \sum_{v\in V} c_{\{u,v\}}.$$

Let us assume that $\mu(u) < \infty$ for every $u \in V$ and define then for $u, v \in V$:

$$P(u,v) = \frac{1}{\mu(u)} c_{\{u,v\}}.$$

This is a transition matrix, and the corresponding Markov chain is called the *random walk on the weighted graph* $G$. In words, at every step, the walk moves from its position $u$ to a neighbour $v$ in $G$ with a probability proportional to the weight $c_{\{u,v\}}$. Since the edges are undirected, then $c_{\{u,v\}} = c_{\{v,u\}}$ and thus:

$$\mu(u)P(u,v) = c_{\{u,v\}} = c_{\{v,u\}} = \mu(v)P(v,u),$$

so $\mu$ is reversible.

*Conversely*, given a transition matrix $P$ and a reversible measure $\mu$, one can consider the graph $G = (V,E)$ with $V = \mathbb{X}$, $E = \{\{x,y\} : P(x,y) > 0\}$, and with the weights $c_{\{x,y\}} = \mu(x)P(x,y)$. Then the random walk on this weighted graph is exactly the $P$-Markov chain.

### 4.2.2 Existence and uniqueness

Recall that we exclude the trivial measure $\mu(x) = 0$ for every $x \in \mathbb{X}$ from the stationary measures. Our first result shows then that if a stationary measure gives a nonzero mass to some point $x$, then it also gives a nonzero mass to any point $y$ where $x$ leads. In particular, in an irreducible chain, any stationary measure gives nonzero mass to every state.

**Lemma 4.2.6.** *Let $\mu$ be a stationary measure, let $x, y \in \mathbb{X}$ be such that $\mu(x) > 0$ and $\rho_{xy} > 0$, then $\mu(y) > 0$.*

*Proof.* Recall that writing:

$$\mathbb{P}_x(H_y < \infty) = \mathbb{P}_x\Big(\bigcup_n \{X_n = y\}\Big) \leq \sum_n P^n(x, y),$$

we infer that if $\mathbb{P}_x(H_y < \infty) > 0$, then there exists $n \geq 1$ such that $P^n(x, y) > 0$. Then for such an index $n$, assuming that $\mu(x) > 0$, we have

$$\mu(y) = \mu P^n(y) = \sum_{z \in \mathbb{X}} \mu(z) P^n(z, y) \geq \mu(x) P^n(x, y),$$

and the right-hand side is nonzero. □

In the exercise sheet, we shall see examples of transient Markov chains which have no stationary measure, or have infinitely many of them. Then next result shows however that it can never have a stationary *probability* measure (with total mass 1).

**Lemma 4.2.7.** *Let $\mu$ be a stationary measure with finite total mass: $\mu(\mathbb{X}) = \sum_x \mu(x) < \infty$. If $x \in \mathbb{X}$ is transient, then $\mu(x) = 0$.*

*Proof.* Let us write:

$$\sum_{n \geq 0} \mu(x) = \sum_{n \geq 0} \sum_{y \in \mathbb{X}} \mu(y) P^n(y, x) = \sum_{y \in \mathbb{X}} \mu(y) \sum_{n \geq 0} P^n(y, x) = \sum_{y \in \mathbb{X}} \mu(y)\, \mathbb{E}_y[V_x].$$

By the strong Markov property, we have:

$$\mathbb{E}_y[V_x] = \rho_{yx}\, \mathbb{E}_x[V_x] \leq \mathbb{E}_x[V_x].$$

Now recall from Proposition 4.1.2 that if $x$ is transient, then

$$\sum_{n \geq 0} \mu(x) \leq \sum_{y \in \mathbb{X}} \mu(y)\, \mathbb{E}_x[V_x] < \infty.$$

This implies that $\mu(x) = 0$. □

As mentioned above, if the chain is irreducible and transient, nothing can be said in general on the existence and uniqueness of stationary measure. This problem can however be solved for recurrent chains. The main theorem of this subsection is the following.

**Theorem 4.2.8.** *If the chain is irreducible and recurrent, then all stationary measures are proportional to each other and they all give nonzero and finite mass to every state.*

The proof takes several intermediate steps and is based on the following key observation: for every $k \geq 1$, for every $x, y \in \mathbb{X}$,

$$\mathbb{P}_x(H_x \geq k, X_{k-1} = z, X_k = y) = \mathbb{P}_x(H_x \geq k, X_{k-1} = z) P(z, y). \tag{4.1}$$

Indeed, this follows by applying the Markov property at time $k - 1$ since the event $\{H_x \geq k, X_{k-1} = z\}$ only depends on $X_0, \dots, X_{k-1}$.

First, given one recurrent state, we can construct explicitly one stationary measure. Note that if the chain is not irreducible and has several recurrent classes (recall Theorem 4.1.8), then Lemma 4.2.9 provides invariant measures which are supported by each disjoint class.

**Lemma 4.2.9.** *Let $x \in \mathbb{X}$ be recurrent, then the measure defined by*

$$\nu_x(y) = \mathbb{E}_x\left[\sum_{k=0}^{H_x-1} \mathbb{1}_{X_k=y}\right] = \mathbb{E}_x\left[\sum_{k=1}^{H_x} \mathbb{1}_{X_k=y}\right] \tag{4.2}$$

*is stationary. Moreover it has $\nu_x(x) = 1$, $\nu_x(\mathbb{X}) = \mathbb{E}_x[H_x]$, and finally $\nu_x(y) > 0$ if and only if $\rho_{xy} > 0$, and in this case $\nu_x(y) < \infty$.*

*Proof.* To prove that $\nu_x$ is stationary, let us write:

$$\nu_x(y) = \mathbb{E}_x\left[\sum_{k\geq 1} \mathbb{1}_{k\leq H_x} \sum_{z\in\mathbb{X}} \mathbb{1}_{X_{k-1}=z}\, \mathbb{1}_{X_k=y}\right] = \sum_{k\geq 1}\sum_{z\in\mathbb{X}} \mathbb{P}_x(k \leq H_x, X_{k-1} = z, X_k = y).$$

Then by (4.1),

$$\nu_x(y) = \sum_{k\geq 1}\sum_{z\in\mathbb{X}} \mathbb{E}_x[\mathbb{1}_{k\leq H_x}\, \mathbb{1}_{X_{k-1}=z}]P(z,y) = \sum_{z\in\mathbb{X}} \mathbb{E}_x\left[\sum_{k=1}^{H_x} \mathbb{1}_{X_{k-1}=z}\right] P(z,y) = \sum_{z\in\mathbb{X}} \nu_x(z)P(z,y).$$

Thus $\nu_x$ is indeed stationary and so $\nu_x = \nu_x P^n$ for every $n \geq 1$.

The properties $\nu_x(x) = 1$ and $\nu_x(\mathbb{X}) = \mathbb{E}_x[H_x]$ are clear, and so is the fact that if $\rho_{xy} = 0$, then $\nu_x(y) \leq \mathbb{E}_x[V_y] = 0$ since in this case $\mathbb{P}_x(V_y = 0) = 1$. Suppose now that $\rho_{xy} > 0$, then there exists $n \geq 1$ such that $P^n(x,y) > 0$. We infer for this $n$ that

$$\nu_x(y) = \sum_{z\in\mathbb{X}} \nu_x(z)P^n(z,y) \geq \nu_x(x)P^n(x,y) = P^n(x,y) > 0.$$

Similarly, we have

$$1 = \nu_x(x) \geq \nu_x(y)P^n(y,x).$$

Recall from Proposition 4.1.6 that if $x$ is recurrent and $\rho_{xy} > 0$, then $\rho_{xy} = 1$ and $\rho_{yx} = 1$, so again there exists $n \geq 1$ with $P^n(y,x) > 0$, which shows that $\nu_x(y) < \infty$. □

We next prove that this particular stationary measure $\nu_x$ is the smallest one that assigns mass 1 to $x$.

**Lemma 4.2.10.** *Let $x \in \mathbb{X}$ be recurrent and let $\nu_x$ be the stationary measure defined in (4.2). If $\mu$ is another stationary measure, then for any $y \in \mathbb{X}$, we have:*

$$\mu(y) \geq \mu(x)\nu_x(y).$$

*Proof.* Let us first prove by induction that for any $n \geq 0$ and any $y \neq x$, we have

$$\mu(y) \geq \mu(x)\sum_{k=0}^{n} \mathbb{P}_x(H_x > k, X_k = y) = \mu(x)\sum_{k=0}^{n} \mathbb{P}_x(X_1 \neq x, \dots, X_k \neq x, X_k = y).$$

For $n = 0$ the right-hand side vanishes for any $y \neq x$. Suppose the identity holds for some $n$, then we can write since $\mu$ is stationary:

$$\mu(y) = \sum_{z\in\mathbb{X}} \mu(z)P(z,y) \geq \sum_{z\in\mathbb{X}} \Big(\mu(x)\sum_{k=0}^{n} \mathbb{P}_x(H_x > k, X_k = z)\Big)P(z,y).$$

On the other hand, by (4.1), we have for each $0 \leq k \leq n$:

$$\mathbb{P}_x(H_x > k, X_{k+1} = y) = \sum_{z\in\mathbb{X}} \mathbb{P}_x(H_x > k, X_k = z, X_{k+1} = y) = \sum_{z\in\mathbb{X}} \mathbb{P}_x(H_x > k, X_k = z)P(z,y).$$

Since $y \neq x$, then $\mathbb{P}_x(H_x > k, X_{k+1} = y) = \mathbb{P}_x(H_x > k+1, X_{k+1} = y)$, so we conclude from the two displays that

$$\begin{aligned}\mu(y) &\geq \mu(x) \sum_{k=0}^{n} \sum_{z \in \mathbb{X}} \mathbb{P}_x(H_x > k, X_k = z) P(z, y) \\ &= \mu(x) \sum_{k=0}^{n} \mathbb{P}_x(H_x > k+1, X_{k+1} = y) \\ &= \mu(x) \sum_{k=1}^{n+1} \mathbb{P}_x(H_x > k, X_k = y).\end{aligned}$$

The sum could as well start from $k = 0$ since the probability vanishes in this case. This completes the induction. Letting $n \longrightarrow \infty$, we conclude that

$$\mu(y) \geq \mu(x) \sum_{k=0}^{\infty} \mathbb{E}_x[\mathbb{1}_{H_x > k} \, \mathbb{1}_{X_k = y}] = \mu(x) \, \mathbb{E}_x\Big[\sum_{k=0}^{H_x - 1} \mathbb{1}_{X_k = y}\Big],$$

and the expectation equals precisely $\nu_x(y)$. □

We can now prove our main result.

*Proof of Theorem 4.2.8.* Let $\mu$ be a stationary measure and let $x \in \mathbb{X}$ be such that $\mu(x) > 0$. For every $y \in \mathbb{X}$, we have $\rho_{xy} > 0$, so Lemma 4.2.6 implies that $\mu(y) > 0$. Recall that Formula (4.2) provides one stationary measure $\nu_x$, which satisfies moreover $0 < \nu_x(y) < \infty$ for every $y \in \mathbb{X}$. By the previous lemma, we have

$$\mu(y) \geq \mu(x) \nu_x(y),$$

for every $y \neq x$, and equality for $y = x$ since $\nu_x(x) = 1$. Then using that both $\mu$ and $\nu_x$ are stationary, we obtain for every $n \geq 1$:

$$\mu(x) = \sum_{y \in \mathbb{X}} \mu(y) P^n(y, x) \geq \sum_{y \in \mathbb{X}} \mu(x) \nu_x(y) P^n(y, x) = \mu(x) \nu_x(x) = \mu(x).$$

This implies that the inequality must be an equality and so

$$\sum_{y \in \mathbb{X}} (\mu(y) - \mu(x) \nu_x(y)) P^n(y, x) = 0,$$

for every $n \geq 1$. Recall that the chain is irreducible, so for any $y \in \mathbb{X}$, there exists $n \geq 1$ such that $P^n(y, x) > 0$, then we must have

$$\mu(y) = \mu(x) \nu_x(y)$$

for the previous sum to vanish. Recall that we chose $x$ so that $\mu(x) > 0$, hence $\mu$ is indeed proportional to $\nu_x$, and since $0 < \nu_x(y) < \infty$ for every $y$, then $0 < \mu(y) < \infty$ for every $y$. □

### 4.2.3 Positive recurrence and null recurrence

By Theorem 4.2.8, in an irreducible and recurrent chain, all stationary measures are proportional so a stationary probability is necessarily unique. We now investigate wether it exists or not.

**Corollary 4.2.11.** *Suppose that the chain is irreducible and recurrent, we have a further dichotomy:*

- *Either all stationary measures have finite total mass and there exists a unique stationary probability $\pi$. The latter never vanishes and moreover for any $x \in \mathbb{X}$, we have the identity:*

$$\mathbb{E}_x[H_x] = \frac{1}{\pi(x)} < \infty.$$

- *Or all stationary measures have infinite total mass and moreover for any $x \in \mathbb{X}$, we have:*

$$\mathbb{E}_x[H_x] = \infty.$$

*When $\mathbb{X}$ is a finite set, we are necessarily in the first case.*

*Proof.* By Theorem 4.2.8, all stationary measures are proportional so either they all have infinite mass, or there exists one with finite mass and then they all do. In the latter case, by rescaling any of them by its total mass we obtain a probability. The latter is necessarily unique since two different probabilities cannot be proportional since they both sum up to 1. Precisely, this unique probability $\pi$ is given for every $x, y \in \mathbb{X}$ by:

$$\pi(y) = \frac{\nu_x(y)}{\nu_x(\mathbb{X})} = \frac{1}{\mathbb{E}_x[H_x]} \mathbb{E}_x\Big[\sum_{k=0}^{H_x-1} \mathbb{1}_{X_k=y}\Big].$$

In particular, since $\nu_x(x) = 1$, then

$$\pi(x) = \frac{1}{\nu_x(\mathbb{X})} = \frac{1}{\mathbb{E}_x[H_x]}.$$

On the other hand, if all stationary measures have infinite mass, then for every $x \in \mathbb{X}$ we have similarly

$$\mathbb{E}_x[H_x] = \nu_x(\mathbb{X}) = \infty.$$

Finally, since we assume that each measure only gives finite mass to every $x \in \mathbb{X}$, then if the latter is a finite set, then each stationary measure must have finite total mass. □

**Definition 4.2.12.** A recurrent irreducible chain is said to be:

- *positive recurrent* if $\mathbb{E}_x[H_x] < \infty$ for all $x \in \mathbb{X}$,
- *null recurrent* if $\mathbb{E}_x[H_x] = \infty$, but still $\mathbb{P}_x(H_x < \infty) = 1$, for all $x \in \mathbb{X}$.

This denomination will be explained by Corollary 5.1.2, we can observe already that a positive recurrent chain has $1/\mathbb{E}_x[H_x] > 0$ for all $x \in \mathbb{X}$ whereas a null recurrent chain has $1/\mathbb{E}_x[H_x] = 0$ for all $x \in \mathbb{X}$.

**Proposition 4.2.13.** *If the chain is irreducible and positive recurrent, then $\mathbb{E}_y[H_x] < \infty$ for all $x, y \in \mathbb{X}$.*

*Proof.* Fix $x, y \in \mathbb{X}$ with $x \neq y$. Observe that for every $n \geq 1$, we have by the Markov property at time $n$:

$$\mathbb{E}_x[H_x] \geq \mathbb{E}_x[H_x \, \mathbb{1}_{H_x>n, X_n=y}] = \mathbb{P}_x(H_x > n, X_n = y)(n + \mathbb{E}_y[H_x]).$$

Since $\mathbb{E}_x[H_x] < \infty$, then it suffices to prove that there exists $n$ with $\mathbb{P}_x(H_x > n, X_n = y) > 0$. Recall from the proof of Proposition 4.1.6 that

$$\mathbb{P}_x(H_y > H_x)^k = \mathbb{P}_x(H_y > H_x^k) \leq \mathbb{P}_x(H_y > k) \underset{k\to\infty}{\longrightarrow} \mathbb{P}_x(H_y = \infty) = 0.$$

Therefore $\mathbb{P}_x(H_y > H_x) < 1$. By taking the complement, we conclude that:

$$0 < \mathbb{P}_x(H_y < H_x) = \mathbb{P}_x\Big(\bigcup_n \{H_x > n, X_n = y\}\Big) \leq \sum_n \mathbb{P}_x(H_x > n, X_n = y),$$

and so at least one term in the sum must be nonzero. □

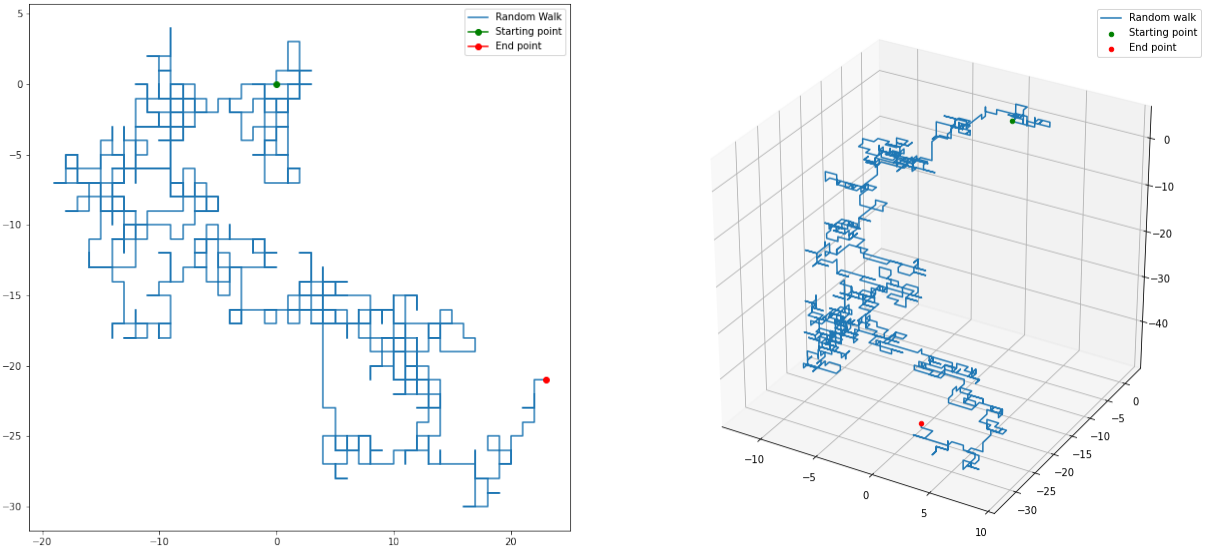



Figure 4.1: The first 1 000 steps of a walk in dimension 2 and in dimension 3.

## 4.3 The Simple Random Walk

The simple random walk in dimension $d \geq 1$ is the Markov chain $(X_n)_n$ on $\mathbb{Z}^d$ started from 0 and with transition matrix $P$ given by:

$$P(x, y) = \frac{1}{2d} \mathbb{1}_{|x-y|=1},$$

for every $x, y \in \mathbb{Z}^d$. In words, at every step, we move from the current position by one in a uniform random direction. The following famous result characterises the asymptotic behaviour of the walk.

**Theorem 4.3.1** (Pólya). *The simple random walk is null recurrent in dimension $d = 1$ and $d = 2$ whereas it is transient in dimension $d \geq 3$.*

*Proof.* It is clear that in any dimension the walk is irreducible. Moreover, for every $x, y \in \mathbb{Z}^d$, we have $P(x, y) = P(y, x)$ so the measure given by $\mu(x) = 1$ for every $x \in \mathbb{Z}^d$ is reversible, hence stationary by Proposition 4.2.2. Since $\mu$ has infinite total mass, then Corollary 4.2.11 implies that the walk cannot be positive recurrent: it is either null recurrent or transient, and this behaviour is dictated by the expected number of visits of 0: according to Proposition 4.1.2, the walk is null recurrent if this expectation is infinite, and it is transient otherwise. By parity, we starting from 0, the walk can only be at 0 at even times, so we may write:

$$\mathbb{E}_0[V_0] = \mathbb{E}_0\left[\sum_{n\geq 0} \mathbb{1}_{X_{2n}=0}\right] = \sum_{n\geq 0} \mathbb{P}_0(X_{2n} = 0).$$

The claim then reduces to check wether this series converges or not.

Let us start with $d = 1$. For every $n \geq 1$ we have $X_{2n} = 0$ if and only if there are $n$ increments equal to $+1$ and $n$ increments equal to $-1$, so, according to Stirling's formula, we have:

$$\mathbb{P}_0(X_{2n} = 0) = \frac{1}{2^{2n}} \binom{2n}{n} = \frac{1}{2^{2n}} \frac{(2n)!}{n!n!} \sim \frac{1}{2^{2n}} \frac{\sqrt{4\pi n}(2n/\mathrm{e})^{2n}}{(\sqrt{2\pi n}(n/\mathrm{e})^n)^2} = \frac{1}{\sqrt{\pi n}}.$$

The corresponding series diverges, so the walk is (null) recurrent.

In dimension $d = 2$, the same reasoning applies: here, $X_{2n} = 0$ if and only if there exists $0 \leq k \leq n$ such that the walk has $k$ increments to the right, $k$ increments to the left, $n - k$ up and $n - k$ down, so now:

$$\mathbb{P}_0(X_{2n} = 0) = \frac{1}{4^{2n}} \sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!} = \frac{1}{4^{2n}} \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k} \binom{n}{n-k} = \frac{1}{4^{2n}} \binom{2n}{n}^2,$$

where the last identity is easily understood: choosing $n$ objects among $2n$ possibilities amounts to choosing $k$ in half of them and $n-k$ in the other half for some $k$. We see that $\mathbb{P}_0(X_{2n}=0)$ in dimension 2 is the square of the same probability in dimension 1! This can easily be seen by rotating the trajectory by 45°: if the walks moves along the diagonals, then we easily see that the two coordinates evolve like two independent walks in dimension 1; moreover the walk in dimension 2 lies at 0 if and only if both coordinate lie at 0 simultaneously. More importantly here, we have $\sum_n \mathbb{P}_0(X_{2n}=0)=\infty$ again so again $\mathbb{E}_0[V_0]=\infty$ and the walk is recurrent.

In dimension 3, we now have:

$$\mathbb{P}_0(X_{2n}=0)=\frac{1}{6^{2n}}\sum_{i+j+k=n}\frac{(2n)!}{i!^2j!^2k!^2}=\frac{1}{12^n}\binom{2n}{n}\sum_{i+j+k=n}\left(\frac{n!}{i!j!k!}\right)^2\frac{1}{3^n}.$$

Roughly speaking, we expect this quantity to be of order $n^{-3/2}$, which is now a convergent series so the walk is now transient. However exact calculations become harder and we will only upper bound this probability (which is sufficient). Let us start by considering the case where $n$ is a multiple of 3. We shall use the following input:

$$\sum_{i+j+k=n}\frac{n!}{i!j!k!}=3^n \qquad \text{and} \qquad \frac{(3\ell)!}{i!j!k!}\le\frac{(3\ell)!}{\ell!^3} \quad \text{for any } i+j+k=3\ell.$$

Indeed for the first one, each summand counts the number of ways to put $n$ objects in three boxes with respectively $i$, $j$, and $k$ in each box, so summing over all possibilities, each object can be put in any box. For the second one, suppose that $i\le j\le k$, otherwise rename them; if $i\le\ell-1$, then $k\ge\ell+1$ so $i+1<k$ and thus $(i+1)!j!(k-1)!<i!j!k!$, therefore the denominator is maximal at $i=j=k=\ell$. Applying Stirling's formula again, we read that:

$$\mathbb{P}_0(X_{6\ell}=0)\le\frac{1}{12^{3\ell}}\binom{6\ell}{3\ell}\frac{(3\ell)!}{\ell!^3}=\frac{1}{12^{3\ell}}\frac{(6\ell)!}{(3\ell)!\ell!^3}\sim\frac{1}{2(\pi\ell)^{3/2}}.$$

It remains to deal with the cases $n=3\ell+1$ and $=3\ell+2$. Notice that if the walk is at 0 at some time $2k$, then makes any move and immediately after its opposite (which has probability 1/6), then it is back at 0 at time $2k+2$. This implies that:

$$\mathbb{P}_0(X_{6\ell}=0)\ge\frac{1}{6}\mathbb{P}_0(X_{6\ell-2}=0) \qquad \text{and} \qquad \mathbb{P}_0(X_{6\ell}=0)\ge\frac{1}{6^2}\mathbb{P}_0(X_{6\ell-4}=0).$$

Thus for any value $n$, the probability $\mathbb{P}_0(X_{2n}=0)$ is asymptotically bounded by some constant times $n^{-3/2}$ so indeed 0 is transient.

Finally in dimension $d\ge4$, one could extend the previous reasoning. We propose another approach by comparison with the case $d=3$. Let $X_n=(X_{1,n},\dots,X_{d,n})$ and let $Y_n=(X_{1,n},X_{2,n},X_{3,n})$ denote the path that follows only the first three coordinates. At each step, one coordinate of $X_n$ chosen uniformly at random changes by either +1 of −1 uniformly at random and independently of the choice of the coordinate. If this coordinate is one of the first three, then the corresponding coordinate of $Y_n$ changes accordingly, as for the walk in dimension 3. However if the coordinate of $X_n$ that changes is not one of the first three, then $Y_{n+1}=Y_n$. Hence $(Y_n)_n$ has the law of the walk in dimension 3 with additional independent random delays, that is time-intervals during which it stays constant. The lengths of these intervals are i.i.d. geometric distributed (on $\mathbb{Z}_+$) with parameter $3/d$. Since we know that the three dimensional walk only visits 0 finitely many times, and since the delays are all finite, then $(Y_n)_n$ also only visits 0 finitely many times. This implies that $(X_n)_n$ only visits 0 finitely many times. □

# Chapter 5

# Convergence of Markov Chains

We study more specifically in this chapter the asymptotic behaviour of Markov chains. One of the reason to introduce them was to extend the Law of Large Numbers and the Central Limit Theorem when the increments are not independent or identically distributed. The so-called ergodic theorem stipulates that under suitable assumptions the Markov chain forgets its starting point and it also describes the limit, in several different senses, in terms of the transition matrix. We shall describe precisely this type of result and finish with some applications to numerical simulations.

**Contents**



In Section 5.1 we present analogues of the Law of Large Numbers and the Central Limit Theorem for a sequence $(f(X_n))_{n\geq 0}$ where $f$ is a real-valued function. We shall see that the empirical average of $f$ along the trajectory of the chain converges almost surely to its expectation with respect to the stationary distribution (when it exists), and with Gaussian fluctuations. Then Section 5.2 discusses another aspect of the ergodic theorem, which shows that the stationary distribution is indeed the limit law of $X_n$ as $n \longrightarrow \infty$; we also discuss there the speed of the convergence. Finally Section 5.3 presents some numerical applications and introduces the Monte Carlo method.

## 5.1 Law of Large Numbers & Central Limit Theorem

Recall from Remark 4.1.5 that if $x \in \mathbb{X}$ is transient, then whatever the starting point, the chain a.s. will never visit $x$ again after a long time. We then ask about the behaviour when $x$ is recurrent. Thanks to Theorem 4.1.8 we may assume that the chain is irreducible, otherwise we work in the class that contains $x$ and we ignore the other states. We will be interested in two different aspects. Here we first consider the average of a real-valued function along the trajectory of the chain, and extend in this context the LLN and CLT. In the next section, we shall discuss the convergence of the law of $X_n$.

Recall that for a measure $\mu$ on $\mathbb{X}$ and a function $f$ for which the integral is well-defined, we write $\mu(f)$ for the integral $\int f \, \mathrm{d}\mu = \sum_{x\in\mathbb{X}} f(x)\mu(x)$.

### 5.1.1 Almost sure convergence

Let us state straight away the main result of this section. Recall from Theorem 4.2.8 that all stationary measure are proportional to each other, so the limit below is unique and does not depend on the choice of $\mu$.

**Theorem 5.1.1.** *Suppose that the chain is irreducible and recurrent and let $\mu$ be any stationary measure. Let $g : \mathbb{X} \longrightarrow [0,\infty)$ with $0 < \mu(g) < \infty$ and let $f : \mathbb{X} \longrightarrow \mathbb{R}$ with either $f \geq 0$ or $\mu(|f|) < \infty$. Then for any initial*

*distribution, we have:*

$$\frac{\sum_{k=0}^{n-1} f(X_k)}{\sum_{k=0}^{n-1} g(X_k)} \xrightarrow[n\to\infty]{a.s.} \frac{\mu(f)}{\mu(g)}.$$

Before proving this result, let us derive an immediate corollary that shows that the stationary distribution arises as the asymptotic proportion of time spent at each state.

**Corollary 5.1.2.** *Suppose that the chain is irreducible and recurrent and let $f$ and $g$ be as above.*

*(i) If the chain is positive recurrent and if $\pi$ denotes its unique stationary probability, then for every $x \in \mathbb{X}$, we have $\mathbb{P}_x$-a.s.*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to\infty]{} \pi(f).$$

*In particular, for every $x, y \in \mathbb{X}$, we have $\mathbb{P}_x$-a.s.*

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{X_k=y} \xrightarrow[n\to\infty]{} \pi(y) = \frac{1}{\mathbb{E}_y[H_y]}.$$

*(ii) If the chain is null recurrent then any stationary measure necessarily has infinite mass and for every $x \in \mathbb{X}$, we have $\mathbb{P}_x$-a.s.*

$$\frac{1}{n}\sum_{k=0}^{n-1} g(X_k) \xrightarrow[n\to\infty]{} 0.$$

*In particular, for every $x, y \in \mathbb{X}$, we have $\mathbb{P}_x$-a.s.*

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{X_k=y} \xrightarrow[n\to\infty]{} 0.$$

*Proof.* Simply apply Theorem 5.1.1 to $g = 1$ or $f = 1$ respectively. □

**Remark 5.1.3.** More generally, if the chain is not irreducible and if $y$ is positive recurrent, then we have:

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{X_k=y} \xrightarrow[n\to\infty]{a.s.} \frac{1}{\mathbb{E}_y[H_y]}\,\mathbb{1}_{H_y<\infty}.$$

Moreover, the left-hand side always lies between 0 and 1 so we can take the expectation by dominated convergence, which reads:

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}_x(X_k = y) \xrightarrow[n\to\infty]{} \frac{\mathbb{P}_x(H_y < \infty)}{\mathbb{E}_y[H_y]}.$$

With the convention that $1/\infty = 0/\infty = 0$, this still holds for null recurrent $y$'s, as well as for transient ones since then $\mathbb{E}_y[H_y] = \infty$ and we have seen that $\sum_{k=0}^{\infty} \mathbb{P}_x(X_k = y) < \infty$.

*Proof of Theorem 5.1.1.* Let us fix the starting point $X_0 = x$. If $X_0$ is random, then we simply apply the result to any fixed $x$ and then average with respect to the law of $X_0$.

The idea is to cut the trajectory at every visit of $x$. By successive applications of the strong Markov property at each time $H_x^j$, the random variables defined by:

$$Y_k = \sum_{i=H_x^{k-1}}^{H_x^k - 1} g(X_i)$$

for $k \geq 1$ are i.i.d. Recall from Theorem 4.2.8 that all stationary measures are proportional, and precisely every such measure $\mu$ takes the form $\mu(y) = \mu(x)\nu_x(y)$, where $\nu_x$ is the only stationary measure that has $\nu_x(x) = 1$ and is given by (4.2). Then

$$\mathbb{E}_x[Y_1] = \mathbb{E}_x\left[\sum_{i=0}^{H_x-1} g(X_i)\right] = \mathbb{E}_x\left[\sum_{i=0}^{H_x-1}\sum_{y\in\mathbb{X}} g(y)\,\mathbb{1}_{X_i=y}\right] = \sum_{y\in\mathbb{X}} g(y)\nu_x(y) = \sum_{y\in\mathbb{X}} g(y)\frac{\mu(y)}{\mu(x)} = \frac{\mu(g)}{\mu(x)}.$$

Recall that $\mu(x) > 0$ and that we assume that $\mu(g) < \infty$, then the $Y_k$'s are integrable and by the usual Law of Large Numbers we have under $\mathbb{P}_x$:

$$\frac{1}{n}\sum_{i=0}^{H_x^n-1} g(X_i) = \frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow[n\to\infty]{a.s.} \frac{\mu(g)}{\mu(x)}.$$

We aim at comparing the sum on the left with that up to a fix number of steps. Let us denote by $V_x(n) = \sum_{i=0}^{n} \mathbb{1}_{X_i=x}$ the number of visits of $x$ up to time $n$ and observe that:

$$H_x^{V_x(n)-1} \le n < H_x^{V_x(n)}.$$

Since $g \ge 0$, then we infer that

$$\frac{1}{V_x(n)}\sum_{i=0}^{H_x^{V_x(n)-1}-1} g(X_i) \le \frac{1}{V_x(n)}\sum_{i=0}^{n-1} g(X_i) \le \frac{1}{V_x(n)}\sum_{i=0}^{H_x^{V_x(n)}-1} g(X_i).$$

Recall that $x$ is recurrent, so $V_x(n) \uparrow_n V_x = \infty$ a.s. Then combined with the above LLN for the $Y_k$'s, we infer that both the lower and upper bound converge a.s. to $\mu(g)/\mu(x)$ and so

$$\frac{1}{V_x(n)}\sum_{i=0}^{n-1} g(X_i) \xrightarrow[n\to\infty]{a.s.} \frac{\mu(g)}{\mu(x)}.$$

If $f$ is a nonnegative function with $\mu(f) < \infty$, then the same holds with $f$ in place of $g$, from which we conclude that

$$\frac{\sum_{k=0}^{n-1} f(X_k)}{\sum_{k=0}^{n-1} g(X_k)} \xrightarrow[n\to\infty]{a.s.} \frac{\mu(f)}{\mu(g)}.$$

If $f$ is not necessarily nonnegative but has $\mu(|f|) < \infty$, we may decomposing as $f = f^+ - f^-$, apply the above convergence to $f^+$ and to $f^-$ separately, and conclude by linearity.

Finally, if $f \ge 0$ and $\mu(f) = \infty$, then we can apply the previous result to a sequence $(f_N)_{N\ge1}$ of nonnegative functions that satisfy $\mu(f_N) < \infty$ and $f_N \uparrow_N f$ and conclude by comparison. Such functions can be explicitly given e.g. by taking $(x_i)_{i\ge1}$ an enumeration of $\mathbb{X}$ and setting $f_N(x_i) = (f(x_i) \wedge N)\,\mathbb{1}_{i\le N}$ (recall indeed that $\mu(x) < \infty$ for every given $x$ so $\mu(f_N) < \infty$ for each $N$). □

### 5.1.2 A Central Limit Theorem

Recall from Corollary 5.1.2 that when the chain is irreducible and admits a stationary probability $\pi$, for any function $f : \mathbb{X} \to \mathbb{R}$ integrable for $\pi$, it holds whatever the starting point $X_0$:

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to\infty]{a.s.} \pi(f).$$

The CLT below shows that the deviations away from this limit are asymptotically Gaussian. For simplicity, we restrict here to *finite* state spaces. In this case, any irreducible chain has a unique stationary probability $\pi$ and any function $f$ is $\pi$-integrable.

**Theorem 5.1.4** (Markov chain's CLT). *Suppose that $\mathbb{X}$ is a finite set and that the chain is irreducible, with stationary probability $\pi$. Let $f : \mathbb{X} \to \mathbb{R}$ and define*

$$u(x) = \sum_{k\ge0} P^k f(x) \qquad \text{and then} \qquad \sigma^2 = \pi\big(Pu^2 - (Pu)^2\big).$$

*Then for any starting point of the chain, we have the convergence in distribution:*

$$\frac{1}{\sqrt{n\sigma^2}}\sum_{k=0}^{n}\big(f(X_k) - \pi(f)\big) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

The proof of this CLT relies on a CLT for martingales, proved in Chapter 9. It is therefore differed to Section 9.7.2. Let us here only discuss the variance $\sigma^2$ which appears in the statement.

Recall that $Pu(x) = \mathbb{E}[u(X_1) \mid X_0 = x]$; define then the *conditional variance* by:

$$\begin{aligned}\mathrm{Var}(u(X_1) \mid X_0 = x) &= \mathbb{E}[(u(X_1) - \mathbb{E}[u(X_1) \mid X_0 = x])^2 \mid X_0 = x] \\ &= \mathbb{E}[u(X_1)^2 \mid X_0 = x] - \mathbb{E}[u(X_1) \mid X_0 = x]^2 \\ &= Pu^2(x) - (Pu(x))^2.\end{aligned}$$

Then the constant $\sigma^2$ in the theorem equals

$$\sigma^2 = \sum_{x \in \mathbb{X}} \pi(x) \,\mathrm{Var}(u(X_1) \mid X_0 = x),$$

that is, the expectation of this conditional variance when $X_0$ has the law $\pi$. Beware that this is not equal to the variance of $u(X_1)$, for which we need to add $\mathrm{Var}_\pi(Pu(X_0))$, which corresponds to the variance of the conditional expectation $\mathbb{E}[u(X_1) \mid X_0 = x]$ when $X_0$ has law $\pi$. Notice that if the $X_k$'s are i.i.d. with law $\pi$, then this additional term vanishes and indeed $\sigma^2 = \mathrm{Var}(u(X_1))$. Let us mention that expressing $\sigma^2$ is not simple in general, and one often approximates it numerically.

## 5.2 Convergence to the equilibrium

As we observed in Remark 5.1.3, if the chain is irreducible and positive recurrent with stationary distribution $\pi$, then we have the convergence in Cesàro mean: for every $x \in \mathbb{X}$,

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{P}(X_k = x) \underset{n \to \infty}{\longrightarrow} \pi(x),$$

for any initial distribution. We now aim at the convergence of $\mathbb{P}(X_n = x)$ for every $x$, that is the convergence in distribution of $X_n$ to the law $\pi$. There is however a simple technical issue that can prevent such a convergence called *periodic* behaviour that we first discuss.

### 5.2.1 Aperiodicity

Take a simple random walk on a cycle of length 4, that is $X_{n+1} = X_n \pm 1 \bmod 4$ with probability 1/2 and 1/2. If $X_0 = 0$, then $X_{2n} \in \{0, 2\}$ and $X_{2n+1} \in \{1, 3\}$ so the sequence cannot converge in distribution. In this example, the chain can only come back to its starting point after an even number of steps. This motivates the next definition.

**Definition 5.2.1.** A Markov chain or a transition matrix is said to be *aperiodic* when for every $x, y \in \mathbb{X}$, there exists $m \geq 1$ such that for every $n \geq m$, we have $P^n(x, y) = \mathbb{P}_x(X_n = y) > 0$.

The aperiodicity condition is stronger than irreducibility which only asks the existence for each pair $x, y$ of one index $n$ with $P^n(x, y) > 0$. Here we require that they all work except finitely many. When the chain is irreducible, it suffices to check the definition above with a single point $x$ for it to be aperiodic.

**Lemma 5.2.2.** *A Markov chain is aperiodic if and only if it is irreducible and there exists $x \in \mathbb{X}$ such that there exists $m \geq 1$ with $P^n(x, x) = \mathbb{P}_x(X_n = x) > 0$ for every $n \geq m$.*

*Proof.* The direct implication is clear, let us only prove the converse one. Let $y, z \in \mathbb{X}$. Since the chain is irreducible, then there exist $i, k \geq 1$ such that $P^i(y, x) > 0$ and $P^k(x, z) > 0$. Moreover for any $j \geq m$ we have $P^j(x, x) > 0$. We infer from the Markov property applied at time $i + j$ first and then at time $i$ that:

$$\begin{aligned}\mathbb{P}_y(X_{i+j+k} = z) &\geq \mathbb{P}_y(X_i = x, X_{i+j} = x, X_{i+j+k} = z) \\ &= \mathbb{P}_y(X_i = x, X_{i+j} = x)\, \mathbb{P}_x(X_k = z) \\ &= \mathbb{P}_y(X_i = x)\, \mathbb{P}_x(X_j = x)\, \mathbb{P}_x(X_k = z),\end{aligned}$$

which equals $P^i(y, x) P^j(x, x) P^k(x, z) > 0$. □

**Remark 5.2.3.** A simple case is when the assumption holds with $m = 1$, that is $P(x, x) > 0$ for some $x \in \mathbb{X}$. A way to force such a behaviour consists in adding random 'delays', for example, suppose that at each time with probability 1/2 we move according to the chain, and with probability 1/2 we stay at the current position. This defines a new Markov chain, whose transition matrix is $(P + I)/2$, where $I$ is the identity matrix. Note that a measure is stationary for $(P + I)/2$ if and only if it is for $P$ so the new chain has the same asymptotic behaviour as the original one, simply slowed down roughly by a factor 2.

Our definition of aperiodicity is not the usual one but is the one that is useful here. The link between the two definitions uses some arithmetics. Recall that if $A \subset \mathbb{N}$ is a nonempty and finite set, we let GCD $A$ denote the greatest integer $d \geq 1$ such that each element of $A$ is a multiple of $d$. If $A \subset \mathbb{N}$ is infinite, then let $d_n = \mathrm{GCD}(A \cap \{1, \dots, n\})$ for every $n \geq \min A$ and observe that $d_{n+1} \leq d_n$. Hence $(d_n)_n$ converges to some $d \geq 1$ and, since they are integers, we have actually $d_n = d$ for every $n$ large enough; we set GCD $A = d$.

**Definition 5.2.4.** For each $x \in \mathbb{X}$, let

$$I(x) = \{n \geq 1 : \ P^n(x, x) > 0\} \qquad \text{and} \qquad d(x) = \mathrm{GCD}\ I(x).$$

If $I(x) \neq \emptyset$, then $d(x) \geq 1$ is called the *period* of $x$.

In our simple example with four states, each point has period 2.

**Proposition 5.2.5.** *Suppose that the chain is irreducible. First for every $x, y \in \mathbb{X}$, we have $d(x) = d(y)$. Moreover, the chain is aperiodic if and only if $d(x) = 1$.*

*Proof.* Fix $x, y \in \mathbb{X}$. Since the chain is irreducible, then there exists $n_1, n_2 \geq 1$ such that both $P^{n_1}(x, y) > 0$ and $P^{n_2}(y, x) > 0$. As in the previous proof, we infer from the Markov property that

$$P^{n_2+n_1}(y, y) \geq P^{n_2}(y, x)P^{n_1}(x, y) > 0$$

so $n_2 + n_1 \in I(y)$. Similarly, for every $n \in I(x)$, we have

$$P^{n_2+n+n_1}(y, y) \geq P^{n_2}(y, x)P^n(x, x)P^{n_1}(x, y) > 0$$

so $n_2 + n + n_1 \in I(y)$. Thus $d(y)$ divides both $n_2 + n_1$ and $n_2 + n + n_1$ and thus divides $n$ for every $n \in I(x)$. Therefore $d(y)$ divides $d(x)$. By a symmetric argument, $d(x)$ divides $d(y)$ and so $d(x) = d(y)$.

Next, if the chain is aperiodic, then for every $x \in \mathbb{X}$ there exists $n(x, x)$ such that $n \in I(x)$ for every $n \geq n(x, x)$, which implies that $d(x) = 1$ (since for example $I(x)$ contains two prime numbers). Suppose finally that $d(x) = 1$. With the same argument as above, $I(x)$ is stable under addition, namely if $n, m \in I(x)$, then $n + m \in I(x)$ since $P^{n+m}(x, x) \geq P^n(x, x)P^m(x, x)$. Thus the claim follows by combining Lemma 5.2.2 above and Lemma 5.2.6 below. □

**Lemma 5.2.6.** *Suppose that $A \subset \mathbb{N}$ is an infinite set stable under addition: if $n, m \in A$, then $n + m \in A$.*

*(i) If $A$ contains two consecutive integers, say $a, a + 1 \in A$, then $A$ contains all the integers $n \geq a^2$.*

*(ii) If $A$ has* GCD $A = 1$ *then it contains two consecutive integers.*

*Hence, if $A$ is stable under addition and has* GCD $A = 1$*, then it contains all the integers but finitely many.*

*Proof.* (i) Suppose $a, a + 1 \in A$. Since $A$ is stable under addition, then every multiple of $a$ belongs to $A$; in particular $ka^2 + \ell a \in A$ for all $k \geq 1$ and $\ell \geq 0$. More generally, any integer $n \geq a^2$ can be written uniquely as

$$n = ka^2 + r = ka^2 + \ell a + s = (ka + \ell - s)a + s(a + 1)$$

with $k \geq 1$, $0 \leq \ell \leq a - 1$, and $0 \leq s \leq a - 1$. Indeed take first the Euclidean division of $n$ by $a^2$ and then that of the rest $r$ by $a$. Notice then that $ka + \ell - s \geq 1$; since $a, a + 1 \in A$ and $A$ is stable under addition, then the right-hand side belongs to $A$ which therefore contains all the integers $n \geq a^2$.

(ii) Suppose GCD $A = 1$ and take two elements $a < b$ in $A$. If $b = a+1$ we are done so suppose $k = b-a \geq 2$. Since GCD $A = 1$, then there necessarily exists $c \in A$ which is not a multiple of $k$, as otherwise $k$ divides $A$. Write the Euclidean division $c = ik + r$ with $i \geq 1$ and $1 \leq r \leq k-1$. Since $A$ is stable under addition and $b \in A$, then $b' = (i+1)b \in A$, and since both $a, c \in A$, then $a' = (i+1)a + c \in A$ so we found two elements $a' < b'$ in $A$ whose difference is:

$$b' - a' = (i+1)k - c = k - r \leq k - 1.$$

If $k-r \geq 2$ we can iterate this argument and construct $a'' < b''$ in $A$ whose difference is $b''-a'' \leq k-r-1$. After at most $k-1$ iterations, we find two consecutive elements of $A$. □

### 5.2.2 Periodic chains (⋆)

Recall that for any $d \geq 2$, the process $(X_{nd})_{n\geq 0}$ is a Markov chain with transition matrix $P^d$. If the original chain is irreducible and has period $d$, then $(X_{nd})_{n\geq 0}$ is almost aperiodic. Actually it may not be irreducible so the precise statement is the following. To ease notation, if $d \geq 2$ and $k \in \{0, \dots, d-1\}$, we let $k_d = (k+1) \bmod d$, that is precisely $k_d = k+1$ when $0 \leq k \leq d-2$ and $k_d = 0$ when $k = d-1$.

**Proposition 5.2.7.** *Let $d \geq 2$ and let $P$ be the transition matrix of an irreducible and $d$-periodic chain. Then there exists a partition $\mathbb{X} = \mathbb{X}_0 \cup \cdots \cup \mathbb{X}_{d-1}$ such that for every $k \in \{0, \dots, d-1\}$, for every $y \in \mathbb{X}$, we have $y \in \mathbb{X}_{k_d}$ if and only if there exists $x \in \mathbb{X}_k$ such that $P(x,y) > 0$. Moreover for every $k \in \{0, \dots, d-1\}$, the matrix $(P^d(x,y))_{x,y\in\mathbb{X}_k}$ is irreducible and aperiodic.*

*Proof.* Step 1: the partition. Fix $x \in \mathbb{X}$ and for every $k \in \{0, \dots, d-1\}$, let

$$\mathbb{X}_k = \{y \in \mathbb{X} : \exists n \geq 0 \text{ such that } P^{k+nd}(x,y) > 0\}.$$

Since $P$ is irreducible, then for every $y \in \mathbb{X}$, there exists $m \geq 1$ such that $P^m(x,y) > 0$. Writing the Euclidean division $m = nd + n$, we obtain that $\bigcup_{k=1}^{d-1} \mathbb{X}_k = \mathbb{X}$. We then claim that these sets are disjoint. Indeed, suppose that there exists $y \in \mathbb{X}_k \cap \mathbb{X}_\ell$. Then there exist $n_k, n_\ell \geq 0$ such that both $P^{k+n_k d}(x,y) > 0$ and $P^{\ell+n_\ell d}(x,y) > 0$. By irreducibility, there also exists $m \geq 0$ such that $P^m(y,x) > 0$ so by concatenating the paths, we infer from the Chapman–Kolmogorov equations that both $k + n_k d + m \in I(x)$ and $\ell + n_\ell d + m \in I(x)$. Recall that $d = \text{GCD } I(x)$ so both $k + n_k d + m$ and $\ell + n_\ell d + m$ are multiple of $d$ and thus so is their difference $k - \ell + (n_k - n_\ell)d$ so finally $k - \ell$ is a multiple of $d$. However recall that $k, \ell \in \{0, \dots, d-1\}$ so $0 \leq |k-\ell| \leq d-1$ and the only possibility that this is a multiple of $d$ is $k - \ell = 0$. Thus $\mathbb{X}_k \cap \mathbb{X}_\ell \neq \emptyset \implies k = \ell$.

Step 2: the equivalence. Fix $0 \leq k \leq d-1$ and fix $z \in \mathbb{X}$. We aim at proving that there exists $\exists y \in \mathbb{X}_k$ such that $P(y,z) > 0$ if and only if $z \in \mathbb{X}_{k_d}$. Recall the definition of $k_d$ and $\mathbb{X}_j$, then $z \in \mathbb{X}_{k_d}$ means that there exists $n \geq 0$ such that $P^{k+1+nd}(x,z) > 0$. Notice then that for every $0 \leq k \leq d-1$ and $n \geq 0$, we have:

$$P^{k+1+nd}(x,z) = (P^{k+nd}P)(x,z) = \sum_{y\in\mathbb{X}} P^{k+nd}(x,y)P(y,z).$$

Hence $P^{k+1+nd}(x,z) > 0$ if and only if there exists $y \in \mathbb{X}$ such that both $P^{k+nd}(x,y) > 0$ and $P(y,z) > 0$, namely if and only if there exists $y \in \mathbb{X}_k$ such that $P(y,z) > 0$.

Step 3: the position at time $n$. Fix $0 \leq k \leq d-1$, $y \in \mathbb{X}_k$, and $z \in \mathbb{X}$, we show by induction that for every $n \geq 0$,

$$P^n(y,z) > 0 \implies z \in \mathbb{X}_{(n+k) \bmod d}. \tag{5.1}$$

Hence, starting from $X_0 \in \mathbb{X}_k$, we have $X_n \in \mathbb{X}_{(n+k) \bmod d}$ almost surely for every $n \geq 0$. Indeed, for $n = 0$ we have $P^n(y,z) = \mathbb{1}_{y=z}$, so this is clear. For $n = 1$, this is also a consequence of the previous step since $(1+k) \bmod d = k_d$. Suppose that this holds for some $n \geq 0$, then as before, we have $P^{n+1}(y,z) = \sum_{v\in\mathbb{X}} P^n(y,v)P(v,z)$ which is positive if and only if there exists $v \in \mathbb{X}$ such that both $P^n(y,v) > 0$ and $P(v,z) > 0$. By the induction hypothesis $P^n(y,v) > 0$ implies that $v \in \mathbb{X}_{(n+k) \bmod d}$. Then by Step 2, the fact that there exists $v \in \mathbb{X}_{n \bmod d}$ such that $P(v,z) > 0$ is equivalent to $z \in \mathbb{X}_{(n+k \bmod d)+1 \bmod d} = \mathbb{X}_{(n+k+1) \bmod d}$.

Step 4: the restricted matrices. Fix $0 \le k \le d-1$ and let us prove that the matrix $(P^d(y,z))_{y,z\in\mathbb{X}_k}$ is an irreducible and aperiodic transition matrix. Obviously the entries are nonnegative. For every $y \in \mathbb{X}_k$, we infer from the previous step that $P^d(y,z) = 0$ for every $z \notin \mathbb{X}_k$, thus:

$$\sum_{z\in\mathbb{X}_k} P^d(y,z) = \sum_{y\in\mathbb{X}} P^d(x,z) = 1.$$

Hence $(P^d(y,z))_{y,z\in\mathbb{X}_k}$ is a transition matrix.

Next fix $y, z \in \mathbb{X}_k$, since $P$ is irreducible then there exists $n \ge 1$ such that $P^n(y,z) > 0$. Let us write the Euclidean division $n = md + r$ with $0 \le r \le d-1$. Then by the Chapman–Kolmogorov equations, we have

$$P^{md+r}(y,z) = \sum_{u_1,\dots,u_m\in\mathbb{X}} P^d(y,u_1)\prod_{i=1}^{m-1} P^d(u_i,u_{i+1})P^r(u_m,z).$$

Since the left-hand side is positive, then there exist $u_1,\dots,u_m \in \mathbb{X}$ such that each matrix entry on the right is positive. By the previous step, since $y \in \mathbb{X}_k$ then $P^d(y,u_1) > 0$ implies $u_1 \in \mathbb{X}_k$ as well and this further implies by induction that $u_i \in \mathbb{X}_k$ for each $1 \le i \le m$. Hence $u_m \in \mathbb{X}_k$ and $P^r(u_m,z) > 0$ which implies that $z \in \mathbb{X}_{(k+r) \bmod d}$. But since $z \in \mathbb{X}_k$ and these sets are disjoint, then necessarily $(k+r) \bmod d = k$, namely $r$ is a multiple of $d$ and since $0 \le r \le d-1$ then $r = 0$. We have thus proved that $n = md$ satisfies $P^n(y,z) = (P^d)^m(y,z) > 0$ so indeed $(P^d(y,z))_{y,z\in\mathbb{X}_k}$ is irreducible.

It remains to prove that it is aperiodic. Fix $y \in \mathbb{X}_k$ and recall that $P$ is $d$-periodic, that is $d = \mathrm{GCD}\, I(y)$. The latter is defined as $\lim_N \mathrm{GCD}(I(y) \cap N)$, where, since $\mathrm{GCD}(I(y) \cap N)$ is integer-valued, the limit is achieved and $d = \mathrm{GCD}(I(y) \cap N)$ for every $N$ large enough. Therefore there exist $j \ge 1$ and integers $n_1,\dots,n_j \in I(y)$ such that $\mathrm{GCD}(n_1,\dots,n_j) = d$. Let us write $n_i = m_i d$ for each $1 \le i \le n$, then $(P^d)^{m_i}(y,y) > 0$ and $\mathrm{GCD}(m_1,\dots,m_j) = 1$, hence $y$ has period 1 for $P^d$. □

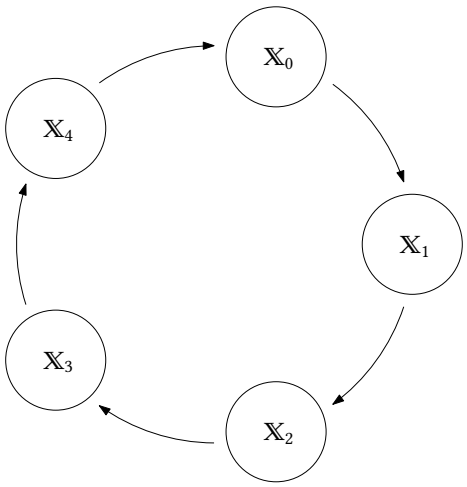



Figure 5.1: Illustration of an irreducible 5-periodic chain: the space $\mathbb{X}$ is partitioned into 5 disjoint subspaces along which the chain rotates as shown in (5.1). If we flash the chain every 5 steps, it always falls into the same subspace and it defines an irreducible and aperiodic chain.

### 5.2.3 Convergence to the stationary distribution

The next theorem proves that periodicity of an (irreducible positive recurrent) chain is the only issue that can prevent its convergence in distribution. The limit is the stationary probability as we saw in Proposition 3.3.4.

**Theorem 5.2.8.** *Suppose that the chain is irreducible, positive recurrent, and aperiodic. Let $\pi$ denote its unique stationary probability measure. Then for any initial distribution, we have:*

$$\sum_{x\in\mathbb{X}} |\mathbb{P}(X_n = x) - \pi(x)| \underset{n\to\infty}{\longrightarrow} 0.$$

*In particular, whatever the initial distribution, $(X_n)_n$ converges in distribution to $\pi$.*

**Remark 5.2.9.** The convergence above is stronger than convergence in distribution, it is called convergence in *total variation.* The total variation is a notion of distance between two probability measures on $\mathbb{X}$, say $\pi$ and $\pi'$, given by $\frac{1}{2}\sum_{x\in\mathbb{X}} |\pi(x) - \pi'(x)|$.

The proof of Theorem 5.2.8 is based on the *coupling* of two independent Markov chains. Let us refer to Section 5.2.6 for some discussion relating this notion and the total variation distance. Let us split the proof into several intermediate results.

**Lemma 5.2.10.** *Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be two independent Markov chains, with transition matrix $P_X$ and $P_Y$ respectively.*

*(i) The pair $((X_n, Y_n))_{n\geq 0}$ is a Markov chain on $\mathbb{X}^2$.*

*(ii) If $(X_n)_n$ and $(Y_n)_n$ are both irreducible and aperiodic, then so is $(X_n, Y_n)_n$.*

*(iii) If moreover $(X_n)_n$ and $(Y_n)_n$ are both positive recurrent, then so is $(X_n, Y_n)_n$.*

*Proof.* (i) Fix any possible trajectories $x_0, \dots, x_n$ and $y_0, \dots, y_n$, then by independence:

$$\begin{aligned}\mathbb{P}_{x_0,y_0}\Big(\bigcap_{i=1}^{n}\{(X_i, Y_i) = (x_i, y_i)\}\Big) &= \mathbb{P}_{x_0}\Big(\bigcap_{i=1}^{n}\{X_i = x_i\}\Big)\,\mathbb{P}_{y_0}\Big(\bigcap_{i=1}^{n}\{Y_i = y_i\}\Big)\\ &= \prod_{i=1}^{n} P_X(x_{i-1}, x_i)\prod_{i=1}^{n} P_Y(y_{i-1}, y_i)\\ &= \prod_{i=1}^{n}\big(P_X(x_{i-1}, x_i)P_Y(y_{i-1}, y_i)\big).\end{aligned}$$

One easily checks that

$$(P_X \otimes P_Y)((x, y), (x', y')) = P_X(x, x')P_Y(y, y')$$

is a transition matrix on $\mathbb{X}^2$, so indeed the pair $((X_n, Y_n))_{n\geq 0}$ is a $P_X \otimes P_Y$-Markov chain.

(ii) Suppose $P_X$ and $P_Y$ irreducible and aperiodic, then for any $x_1, x_2, y_1, y_2$, there exist $i, j \geq 1$ such that for any $n \geq i$, we have $P_X^n(x_1, x_2) > 0$ and for any $n \geq j$, we have $P_Y^n(y_1, y_2) > 0$, therefore for any $n \geq \max(i, j)$, we have $(P_X \otimes P_Y)^n((x_1, y_1)(x_2, y_2)) > 0$ and $P_X \otimes P_Y$ is thus irreducible and aperiodic.

(iii) If the chains are positive recurrent, then there exist a $P_X$-stationary probability measure $\pi_X$ and a $P_Y$-stationary probability measure $\pi_Y$ and one easily checks that the product probability measure $(\pi_X \otimes \pi_Y)(x, y) = \pi_X(x)\pi_Y(y)$ is then $P_X \otimes P_Y$-stationary so the pair is positive recurrent. □

Observe that the fact that each chain is aperiodic is crucial to deduce that the pair is even irreducible, as otherwise it may be the case that $\{X_n = x\} \cap \{Y_n = y\} = \emptyset$ for all $n$. For a concrete example, take again two independent walks on the cycle of length 4, started at $(1, 1)$, it will never reach $(1, 2)$.

**Lemma 5.2.11.** *Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be two independent $P$-Markov chains and define their* coupling time*:*

$$T = \inf\{n \geq 0 : X_n = Y_n\}. \tag{5.2}$$

*Define also for every $n \geq 0$,*

$$Z_n = X_n \mathbb{1}_{n<T} + Y_n \mathbb{1}_{n\geq T}.$$

*Then $(Z_n)_{n\geq 0}$ is a $P$-Markov chain with same initial position as $(X_n)_{n\geq 0}$, so they have the same law.*

*Proof.* By the previous lemma the pair $(X_n, Y_n)_n$ is a Markov chain. Observe that

$$T = \inf\{n \geq 0 : \ (X_n, Y_n) \in \{(x, x), x \in \mathbb{X}\}\}$$

is a stopping time for this process. Fix $n \geq 1$ and $z_0, \dots, z_n \in \mathbb{X}$ and let us write:

$$\mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n) = \mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n, T \geq n) + \sum_{k=0}^{n-1} \mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n, T = k).$$

When $T \geq n$ we have $Z_i = X_i$ for every $i \leq n$, including $i = n$ so the first probability on the right equals $\mathbb{P}(X_0 = z_0, \dots, X_n = z_n, T \geq n)$. Next, for $0 \leq k \leq n-1$ fixed, the summand on the right equals:

$$\mathbb{P}(X_0 = z_0, \dots, X_k = z_k, Y_0 \neq z_0, \dots, Y_{k-1} \neq z_{k-1}, Y_k = z_k, \dots, Y_n = z_n),$$

which, by independence of the chains, can be split as:

$$\mathbb{P}(X_0 = z_0, \dots, X_k = z_k)\, \mathbb{P}(Y_0 \neq z_0, \dots, Y_{k-1} \neq z_{k-1}, Y_k = z_k, \dots, Y_n = z_n).$$

By applying the Markov property to the chain $(Y_n)_n$ at time $k$, the very last probability equals

$$\mathbb{P}(Y_0 \neq z_0, \dots, Y_{k-1} \neq z_{k-1}, Y_k = z_k)\, \mathbb{P}_{z_k}(Y_1 = z_{k+1}, \dots, Y_{n-k} = z_n)$$

and since the chains $(X_n)_n$ and $(Y_n)_n$ have the same transition matrix $P$, then appealing e.g. to Theorem 3.2.2, we have:

$$\mathbb{P}_{z_k}(Y_1 = z_{k+1}, \dots, Y_{n-k} = z_n) = \mathbb{P}_{z_k}(X_1 = z_{k+1}, \dots, X_{n-k} = z_n).$$

Wrapping up, we infer that

$$\begin{aligned}
&\mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n, T = k) \\
&= \mathbb{P}(X_0 = z_0, \dots, X_k = z_k)\, \mathbb{P}(Y_0 \neq z_0, \dots, Y_{k-1} \neq z_{k-1}, Y_k = z_k)\, \mathbb{P}_{z_k}(X_1 = z_{k+1}, \dots, X_{n-k} = z_n) \\
&= \mathbb{P}(X_0 = z_0, \dots, X_n = z_n)\, \mathbb{P}(Y_0 \neq z_0, \dots, Y_{k-1} \neq z_{k-1}, Y_k = z_k) \\
&= \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, Y_0 \neq z_0, \dots, Y_{k-1} \neq z_{k-1}, Y_k = z_k) \\
&= \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, T = k).
\end{aligned}$$

Gathering our findings, we conclude that

$$\begin{aligned}
\mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n) &= \mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n, T \geq n) + \sum_{k=0}^{n-1} \mathbb{P}(Z_0 = z_0, \dots, Z_n = z_n, T = k) \\
&= \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, T \geq n) + \sum_{k=0}^{n-1} \mathbb{P}(X_0 = z_0, \dots, X_n = z_n, T = k) \\
&= \mathbb{P}(X_0 = z_0, \dots, X_n = z_n),
\end{aligned}$$

and the claim follows from Theorem 3.2.2. □

We can now easily derive Theorem 5.2.8.

*Proof of Theorem 5.2.8.* Suppose the chain is aperiodic and positive recurrent, with stationary probability $\pi$. Let us use the previous notation. Let $(X_n)_{n\geq 0}$ start from an arbitrary distribution and independently let $(Y_n)_{n\geq 0}$ start from the stationary distribution $\pi$. Then $Y_n$ has the law $\pi$ for every $n \geq 0$. Since $Z_n$ has the same law as $X_n$ for every $n \geq 0$, then

$$\begin{aligned}
|\mathbb{P}(X_n = x) - \pi(x)| &= |\mathbb{P}(Z_n = x) - \mathbb{P}(Y_n = x)| \\
&= |\mathbb{P}(X_n = x, T > n) + \mathbb{P}(Y_n = x, T \leq n) - \mathbb{P}(Y_n = x)| \\
&= |\mathbb{P}(X_n = x, T > n) - \mathbb{P}(Y_n = x, T > n)| \\
&\leq \mathbb{P}(X_n = x, T > n) + \mathbb{P}(Y_n = x, T > n).
\end{aligned}$$

Then summing over all possible values of $x$, we obtain:

$$\sum_{x\in\mathbb{X}} |\mathbb{P}(X_n = x) - \pi(x)| \le 2\,\mathbb{P}(T > n).$$

Now recall that the chain $(X_n, Y_n)_n$ is irreducible and recurrent, so a.s. for every initial distribution, we have $T = \inf\{n \ge 0 :\ (X_n, Y_n) \in \{(x,x), x \in \mathbb{X}\}\} < \infty$ and thus $\mathbb{P}(T > n) \longrightarrow 0$ as $n \longrightarrow \infty$. □

### 5.2.4 Null recurrent chains (⋆)

In the transient or null recurrent case, the limit in the previous theorem is simply 0.

**Proposition 5.2.12.** *If $(X_n)_n$ is aperiodic and either transient or null recurrent, then for every $x, y \in \mathbb{X}$ we have $\mathbb{P}_x(X_n = y) \longrightarrow 0$ as $n \longrightarrow \infty$.*

*Proof.* First if the chain is transient, then we know from Remark 4.1.5 that the expected number of visit of any point is finite, namely, for all $x, y \in \mathbb{X}$,

$$\sum_{n\ge 0} \mathbb{P}_x(X_n = y) = \mathbb{E}_x[V_y] < \infty,$$

so in particular $\mathbb{P}_x(X_n = y) \longrightarrow 0$ as $n \longrightarrow \infty$.

Suppose henceforth that the chain $(X_n)_n$ is null recurrent. From the previous proof, we know that if we start another independent chain $(Y_n)_n$ with the same transition matrix $P$, then the pair $(X_n, Y_n)$ has transition matrix $P \otimes P$ which is irreducible but now can be either transient or recurrent. Let us note that it cannot be positive recurrent. Indeed since $(X_n)_n$ is null recurrent than it admits stationary measures and they all have infinite total mass. Now if $\mu$ is such a measure, then $\mu \otimes \mu$ is stationary for $P \otimes P$, and it also have infinite mass so the claim follows from Corollary 4.2.11.

If the pair is transient, then we infer as above that

$$\mathbb{P}_{(x,x)}((X_n, Y_n) = (y, y)) \underset{n\to\infty}{\longrightarrow} 0.$$

On the other hand, by independence, the left-hand side equals $\mathbb{P}_x(X_n = y)^2$ which therefore converges to 0.

Suppose henceforth that both $P$ and $P \otimes P$ are null recurrent. Then exactly as in the previous proof, whatever the initial distribution of $(X_0, Y_0)$, the coupling time $T = \inf\{n \ge 0 :\ (X_n, Y_n) \in \{(x,x), x \in \mathbb{X}\}\}$ is finite almost surely (by recurrence of the pair) and thus

$$|\,\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| \le 2\,\mathbb{P}(T > n) \underset{n\to\infty}{\longrightarrow} 0$$

for all initial distributions of $(X_0, Y_0)$. Taking $X_0 = x_0$ and $Y_0 = y_0$, we read:

$$P^n(x_0, y) - P^n(y_0, y) \underset{n\to\infty}{\longrightarrow} 0. \tag{5.3}$$

Recall that our aim is to prove that $P^n(x, y) \longrightarrow 0$ for every $x, y \in \mathbb{X}$. Fix $x \in \mathbb{X}$ and let us enumerate $\mathbb{X}$ as $\{y_1, y_2, \dots\}$. Notice that $P^n(x, y_1) \in [0, 1]$ so there exists a subsequence $(n_k)_k$ such that $P^{n_k}(x, y_1) \longrightarrow \rho(y_1) \in [0, 1]$ as $k \longrightarrow \infty$. Similarly we can then extract from $(n_k)_k$ a subsequence $(n_{k_j})_j$ such that $P^{n_{k_j}}(x, y_2) \longrightarrow \rho(y_2) \in [0, 1]$ as $j \longrightarrow \infty$, and of course $P^{n_{k_j}}(x, y_1) \longrightarrow \rho(y_1)$. By induction (this is Cantor's diagonal extraction argument), we obtain that there exists a subsequence, say, $(m_i)_i$ such that $P^{m_i}(x, y)$ converges to a limit $\rho(y) \in [0, 1]$ for all $y \in \mathbb{X}$. Combined with (5.3) we may replace $x$ by any other point, namely

$$P^{m_i}(u, y) \longrightarrow \rho(y)$$

for every $u, y \in \mathbb{X}$.

We claim that $\rho$ is a finite stationary measure. Indeed:

$$\begin{aligned}\rho P(z) &= \sum_{y\in\mathbb{X}} \rho(y)P(y,z)\\ &= \sum_{y\in\mathbb{X}} \Big(\lim_{i\to\infty} P^{m_i}(x,y)\Big)P(y,z)\\ &\le \liminf_{i\to\infty} \sum_{y\in\mathbb{X}} P^{m_i}(x,y)P(y,z),\end{aligned}$$

where the inequality follows from Fatou's lemma (Theorem 1.3.12) applied to the measure $P(\cdot, z)$. Now observe that

$$\sum_{y\in\mathbb{X}} P^{m_i}(x,y)P(y,z) = P^{m_i+1}(x,z) = \sum_{y\in\mathbb{X}} P(x,y)P^{m_i}(y,z).$$

Recall that $P^{m_i}(y,z) \longrightarrow \rho(z)$ for every $y, z \in \mathbb{X}$ so by dominated convergence, for every $z \in \mathbb{X}$, we have:

$$\rho P(z) \le \liminf_{i\to\infty} \sum_{y\in\mathbb{X}} P(x,y)P^{m_i}(y,z) = \sum_{y\in\mathbb{X}} P(x,y)\rho(z) = \rho(z).$$

Let us next sum over $z$ the left-hand side:

$$\sum_{z\in\mathbb{X}} \rho P(z) = \sum_{z\in\mathbb{X}}\sum_{y\in\mathbb{X}} \rho(y)P(y,z) = \sum_{y\in\mathbb{X}}\sum_{z\in\mathbb{X}} \rho(y)P(y,z) = \sum_{y\in\mathbb{X}} \rho(y).$$

Hence $\rho P(z) \le \rho(z)$ for every $z \in \mathbb{X}$ and the sum over $z$ of both sides are equal. This implies that $\rho P(z) = \rho(z)$ for every $z \in \mathbb{X}$ so $\rho$ is indeed stationary. By Fatou's lemma again, the total mass of $\rho$ is:

$$\sum_{y\in\mathbb{X}} \rho(y) = \sum_{y\in\mathbb{X}} \Big(\lim_{i\to\infty} P^{m_i}(x,y)\Big) \le \liminf_{i\to\infty} \sum_{y\in\mathbb{X}} P^{m_i}(x,y) = 1.$$

Hence $\rho$ is a stationary measure with finite mass. Note that it could be the constant null measure.

To conclude, if there exists $y \in \mathbb{X}$ such that $P^n(x,y)$ does not converge to 0, then it has a subsequence with a positive limit. Then by starting our diagonal argument with this one, we get $\rho(y) > 0$ for this particular value, and hence $\rho$ is a nontrivial stationary measure with finite mass, which contradicts the fact that $P$ is null recurrent. Hence $\rho$ is the constant null measure and $P^n(x,y) \longrightarrow 0$ along any subsequence, hence $P^n(x,y) \longrightarrow 0$ as we claimed. □

### 5.2.5 Speed of convergence

As always, in practice, a convergence result such as in Theorem 5.2.8 is not enough since $n$ will not tend to infinity, and quantifying how far from the limit we are at a given $n$ is crucial. This is not an easy question and often there are no universal response. Let us give a criterion due to Döblin which implies an exponential rate of convergence; notice the power of this result which provides an explicit bound that applies uniformly over all starting points and any time $n$.

**Theorem 5.2.13.** *Suppose that the chain is irreducible and aperiodic and that it satisfies the Döblin condition: there exist an integer $k \ge 1$, a real number $\delta > 0$, as well as a probability measure $\nu$ on $\mathbb{X}$ such that:*

$$\mathbb{P}_x(X_k = y) \ge \delta\nu(y) \qquad \text{for every } x, y \in \mathbb{X}. \tag{5.4}$$

*Then the chain has a stationary probability $\pi$ and it satisfies: for every $n \ge 1$,*

$$\sup_{x_0\in\mathbb{X}} \sum_{x\in\mathbb{X}} |\mathbb{P}_{x_0}(X_n = x) - \pi(x)| \le 2(1-\delta)^{\lfloor n/k\rfloor}.$$

Since the chain is irreducible, then for every pair $x, y \in \mathbb{X}$, there exists $k \geq 1$ such that $\mathbb{P}_x(X_k = y) > 0$; in (5.4) we require a uniform lower bound on this probability. Notice that if $\mathbb{X}$ is finite, then this condition always holds for an irreducible and aperiodic chain. Indeed in this case, there exists $k \geq 1$ such that $\mathbb{P}_x(X_k = y) > 0$ for all pairs $x, y \in \mathbb{X}$ and we may set:

$$\delta = \sum_{y \in \mathbb{X}} \min_{x \in \mathbb{X}} \mathbb{P}_x(X_k = y) > 0 \qquad \text{and then} \qquad \nu(y) = \frac{1}{\delta} \min_{x \in \mathbb{X}} \mathbb{P}_x(X_k = y).$$

This provides a good starting point, but in practice the rate of convergence is often obtained by a specific analysis of the model which often allows to obtain a better bound than in Theorem 5.2.13.

The proof of Theorem 5.2.13 takes three steps that we separate: we first prove the existence of $\pi$, then we prove that it suffices to consider the case $k = 1$, and finally we prove the upper bound when $k = 1$.

*Proof of existence of $\pi$.* Let us first prove that the assumptions ensure the existence of a stationary probability. Recall that this is equivalent to the existence of a positive recurrent state. Fix henceforth $y \in \mathbb{X}$ such that $\nu(y) > 0$, which exists since $\sum_x \nu(x) = 1$, and let us prove that $\mathbb{E}_y[H_y] < \infty$. This is an application of the "what can happen will happen" principle discussed in the exercises: since the Döblin condition (5.4) stipulates that, whatever the current position, there is a probability at least $\delta\nu(y)$ to lie at $y$ $k$ steps later, then this will occur after at most a renom geometric number of trials.

Precisaly, by applying the Markov property at time $(n-1)k$, we have:

$$\begin{aligned}
\mathbb{P}_y(H_y > nk) &\leq \sum_{x \neq y} \mathbb{P}_y(H_y > (n-1)k, X_{(n-1)k} = x, H_y > nk) \\
&= \sum_{x \neq y} \mathbb{P}_y(H_y > (n-1)k, X_{(n-1)k} = x)\, \mathbb{P}_x(H_y > k) \\
&\leq \sum_{x \neq y} \mathbb{P}_y(H_y > (n-1)k, X_{(n-1)k} = x)\, \mathbb{P}_x(X_k \neq y) \\
&\leq \sum_{x \neq y} \mathbb{P}_y(H_y > (n-1)k, X_{(n-1)k} = x) \cdot (1 - \delta\nu(y)) \qquad \text{by (5.4)} \\
&= \mathbb{P}_y(H_y > (n-1)k) \cdot (1 - \delta\nu(y)).
\end{aligned}$$

We infer by induction that $\mathbb{P}_y(H_y > nk) \leq (1 - \delta\nu(y))^n$ for every $n \geq 1$ and thus:

$$\mathbb{E}_y[H_y] = \sum_{n \geq 0} \mathbb{P}_y(H_y > n) \leq k \sum_{n \geq 0} \mathbb{P}_y(H_y > nk) < \infty.$$

Therefore $y$ is positive recurrent and so is the entire chain by irreducibility, so it admits a unique stationary distribution $\pi$. □

*Proof of the exponential bound.* Fix any distribution $\rho_0$ on $\mathbb{X}$, let $X_0$ be distributed as $\rho_0$, and then let us denote by $\rho_n$ the law of $X_n$ for every $n \geq 1$. By the Markov property at time $nk$, we have:

$$\rho_{(n+1)k}(x) = \sum_{z \in \mathbb{X}} \rho_{nk}(z)\, \mathbb{P}_z(X_k = x).$$

In the particular case $\rho_0 = \pi$ is the stationary distribution, we know that $\rho_n = \pi$ for every $n \geq 1$ so

$$\pi(x) = \sum_{z \in \mathbb{X}} \pi(z)\, \mathbb{P}_z(X_k = x).$$

Notice that since both $\rho_{nk}$ and $\pi$ are probability measures, then

$$\sum_{z \in \mathbb{X}} (\rho_{nk}(z) - \pi(z))\delta\nu(x) = \Big( \sum_{z \in \mathbb{X}} \rho_{nk}(z) - \sum_{z \in \mathbb{X}} \pi(z) \Big) \delta\nu(x) = (1-1)\delta\nu(x) = 0.$$

Combining these remarks, we infer that

$$\begin{aligned}\sum_{x\in\mathbb{X}} |\rho_{(n+1)k}(x) - \pi(x)| &= \sum_{x\in\mathbb{X}} \Big| \sum_{z\in\mathbb{X}} (\rho_{nk}(z) - \pi(z))(\mathbb{P}_z(X_k = x) - \delta\nu(x)) \Big| \\ &\leq \sum_{x,z\in\mathbb{X}} |\rho_{nk}(z) - \pi(z)| \cdot |\,\mathbb{P}_z(X_k = x) - \delta\nu(x)| \\ &= \sum_{x,z\in\mathbb{X}} |\rho_{nk}(z) - \pi(z)| \cdot (\mathbb{P}_z(X_k = x) - \delta\nu(x)) \qquad \text{by (5.4)} \\ &= \sum_{z\in\mathbb{X}} |\rho_{nk}(z) - \pi(z)| \sum_{x\in\mathbb{X}} (\mathbb{P}_z(X_k = x) - \delta\nu(x)) \\ &= \sum_{z\in\mathbb{X}} |\rho_{nk}(z) - \pi(z)| \cdot (1 - \delta).\end{aligned}$$

We infer by induction that

$$\sum_{x\in\mathbb{X}} |\rho_{nk}(x) - \pi(x)| \leq (1-\delta)^n \sum_{x\in\mathbb{X}} |\rho_k(x) - \pi(x)| \leq 2(1-\delta)^n,$$

since $\sum_{x\in\mathbb{X}} |\rho_k(x) - \pi(x)| \leq \sum_{x\in\mathbb{X}} (\rho_k(x) - \pi(x)) = 2$.

Finally, if $m \geq 1$ is any integer, then we may write the Euclidean division $m = nk + r$ with $0 \leq r \leq k-1$, and the similarly as above:

$$\begin{aligned}\sum_{x\in\mathbb{X}} |\rho_m(x) - \pi(x)| &= \sum_{x\in\mathbb{X}} \Big| \sum_{z\in\mathbb{X}} (\rho_{nk}(z) - \pi(z))\, \mathbb{P}_z(X_r = x) \Big| \\ &\leq \sum_{x,z\in\mathbb{X}} |\rho_{nk}(z) - \pi(z)| \cdot \mathbb{P}_z(X_r = x) \\ &= \sum_{z\in\mathbb{X}} |\rho_{nk}(z) - \pi(z)|.\end{aligned}$$

We conclude from the previous case. □

### 5.2.6 Coupling and total variation distance (⋆)

As we mentioned already, Theorem 5.2.8 and Theorem 5.2.13 control the total variation distance between the law of $X_n$ and the stationary distribution $\pi$ and the proof relies on a coupling argument. Let us here discuss more the relation between these two notions.

**Definition 5.2.14.** Let $\pi$ and $\nu$ be two probability measures on $\mathbb{X}$, we define

$$\|\pi - \nu\|_{TV} = \frac{1}{2} \sum_{x\in\mathbb{X}} |\pi(x) - \nu(x)|,$$

which is called the *total variation distance* between $\pi$ and $\nu$.

**Exercise 5.2.15.** Prove that the total variation distance is a distance between probability measures on $\mathbb{X}$.

This notion of distance is stronger than the convergence in distribution: if one thinks of a probability as a function from the subsets of $\mathbb{X}$ to $[0,1]$, then the convergence in distribution is a pointwise convergence, whereas the convergence for the total variation distance is a uniform convergence as we next prove.

**Proposition 5.2.16.** *Let $\pi$ and $\nu$ be two probability measures on $\mathbb{X}$, then*

$$\|\pi - \nu\|_{TV} = \sup_{A\subset\mathbb{X}} |\pi(A) - \nu(A)|,$$

*where we recall that $\pi(A) = \sum_{x\in A} \pi(x)$.*

*Proof.* Let us first prove that the left-hand side is smaller than or equal to the right-hand side by providing one subset $A$ which realises the total variation distance. Precisely, let $A = \{x \in \mathbb{X} : \ \pi(x) \geq \nu(x)\}$, then

$$\begin{aligned}\|\pi - \nu\|_{TV} &= \frac{1}{2}\sum_{x\in A}|\pi(x) - \nu(x)| + \frac{1}{2}\sum_{x\in A^c}|\pi(x) - \nu(x)| \\ &= \frac{1}{2}\sum_{x\in A}(\pi(x) - \nu(x)) - \frac{1}{2}\sum_{x\in A^c}(\pi(x) - \nu(x)) \\ &= \frac{\pi(A) - \pi(A^c) - \nu(A) + \nu(A^c)}{2}.\end{aligned}$$

Since $\pi$ and $\nu$ are probability measures, then we have $\pi(A) + \pi(A^c) = \nu(A) + \nu(A^c) = 1$ and thus

$$\frac{\pi(A) - \pi(A^c) - \nu(A) + \nu(A^c)}{2} = \pi(A) - \nu(A) - \frac{1}{2}\underbrace{(\pi(A) + \pi(A^c) - \nu(A) - \nu(A^c))}_{=0},$$

and similarly:

$$\frac{\pi(A) - \pi(A^c) - \nu(A) + \nu(A^c)}{2} = \nu(A^c) - \pi(A^c) + \frac{1}{2}\underbrace{(\pi(A) + \pi(A^c) - \nu(A) - \nu(A^c))}_{=0}.$$

Thus, for this choice of $A$, we have

$$\|\pi - \nu\|_{TV} = \pi(A) - \nu(A) = \nu(A^c) - \pi(A^c). \tag{5.5}$$

Next for any subset $B \subset \mathbb{X}$, we have since $\pi \leq \nu$ on $A^c$ and $\pi \geq \nu$ on $A$:

$$\begin{aligned}\pi(B) - \nu(B) &= \pi(B \cap A) + \pi(B \cap A^c) - \nu(B \cap A) - \nu(B \cap A^c) \\ &\leq \pi(B \cap A) - \nu(B \cap A) \\ &\leq \pi(B \cap A) + \pi(B^c \cap A) - \nu(B \cap A) - \nu(B^c \cap A) \\ &= \pi(A) - \nu(A),\end{aligned}$$

and similarly $\nu(B) - \pi(B) \leq \nu(A^c) - \pi(A^c)$ so

$$|\pi(B) - \nu(B)| \leq \pi(A) - \nu(A) = \nu(A^c) - \pi(A^c) = \|\pi - \nu\|_{TV}$$

for all $B \subset \mathbb{X}$. □

Let us turn to the notion of coupling.

**Definition 5.2.17.** Let $\pi$ and $\nu$ be two probability measures on $\mathbb{X}$. A *coupling* of $\pi$ and $\nu$ is a probability measure $\rho$ on $\mathbb{X}^2$ such that if $(X, Y)$ has the law $\rho$, then $X$ has the law $\pi$ and $Y$ has the law $\nu$. We shall denote by $\mathcal{C}(\pi, \nu)$ the set of all their couplings.

**Example 5.2.18.** If $\pi = \nu$ is the Bernoulli law with parameter $1/2$, we can take either $X$ and $Y$ independent with this law, or $X = Y$, or $X = 1 - Y$. This provides three different couplings.

Couplings relate to the total variation distance as follows.

**Proposition 5.2.19.** *Let $\pi$ and $\nu$ be two probability measures on $\mathbb{X}$. Then*

$$\|\pi - \nu\|_{TV} = \min_{\rho\in\mathcal{C}(\pi,\nu)} \rho(X \neq Y),$$

*where $\rho(X \neq Y)$ is the probability that $X$ differs from $Y$ when the pair $(X, Y)$ has the law $\rho$.*

*Proof.* Again, let us first prove that the left-hand side is smaller than or equal to the right-hand side. If $X$ has the law $\pi$ and $Y$ the law $\nu$, then for any coupling and any subset $A \subset \mathbb{X}$ it holds

$$\begin{aligned}
\mathbb{P}(X \neq Y) &\geq \mathbb{P}(X \in A, Y \in A^c) \\
&\geq \mathbb{P}(X \in A, Y \in A^c) - \mathbb{P}(X \in A^c, Y \in A) \\
&= \mathbb{P}(X \in A, Y \in A) + \mathbb{P}(X \in A, Y \in A^c) - \mathbb{P}(X \in A, Y \in A) - \mathbb{P}(X \in A^c, Y \in A) \\
&= \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\
&= \pi(A) - \nu(A).
\end{aligned}$$

Recall from the previous proposition that $\|\pi - \nu\|_{TV}$ is the supremum over $A$ of these quantities, hence

$$\mathbb{P}(X \neq Y) \geq \|\pi - \nu\|_{TV}$$

for every coupling. To prove the equality, it remains to find one optimal coupling. Let $A = \{x \in \mathbb{X} : \ \pi(x) \geq \nu(x)\}$ and then

$$p = 1 - \sum_{x \in \mathbb{X}} \min(\pi(x), \nu(x)) = 1 - (\nu(A) + \pi(A^c)) = \pi(A) - \nu(A) = \|\pi - \nu\|_{TV},$$

by (5.5). Let then $\xi$ have the Bernoulli law with parameter $p$. If $\xi = 0$, then let $X = Y$ be distributed according to

$$\mathbb{P}(X = x \mid \xi = 0) = \frac{1}{1-p} \min(\pi(x), \nu(x)) = \frac{1}{1-p} (\pi(x) \mathbb{1}_{x \in A^c} + \nu(x) \mathbb{1}_{x \in A}).$$

If $\xi = 1$, then let $X$ and $Y$ be independent and sampled respectively from:

$$\mathbb{P}(X = x \mid \xi = 1) = \frac{\pi(x) - \nu(x)}{\|\pi - \nu\|_{TV}} \mathbb{1}_{x \in A} \qquad \text{and} \qquad \mathbb{P}(Y = y \mid \xi = 1) = \frac{\nu(y) - \pi(y)}{\|\pi - \nu\|_{TV}} \mathbb{1}_{y \in A^c},$$

which are indeed probabilities by (5.5).

Let us check that this defines a coupling in that $X$ has the law $\pi$ and $Y$ the law $\nu$: since $p = \|\pi - \nu\|_{TV}$, then simply

$$\begin{aligned}
\mathbb{P}(X = x) &= p \ \mathbb{P}(X = x \mid \xi = 1) + (1 - p) \ \mathbb{P}(X = x \mid \xi = 0) \\
&= (\pi(x) - \nu(x)) \mathbb{1}_{x \in A} + (\pi(x) \mathbb{1}_{x \in A^c} + \nu(x) \mathbb{1}_{x \in A}) \\
&= \pi(x),
\end{aligned}$$

and similarly

$$\begin{aligned}
\mathbb{P}(Y = y) &= p \ \mathbb{P}(Y = y \mid \xi = 1) + (1 - p) \ \mathbb{P}(Y = y \mid \xi = 0) \\
&= (\nu(y) - \pi(y)) \mathbb{1}_{y \in A^c} + (\pi(y) \mathbb{1}_{y \in A^c} + \nu(y) \mathbb{1}_{y \in A}) \\
&= \nu(y).
\end{aligned}$$

Finally, if $\xi = 0$ then $X = Y$ and if $\xi = 1$, then $X \in A$ and $Y \in A^c$ so $X \neq Y$ if and only if $\xi = 1$, which occurs with probability $p = \|\pi - \nu\|_{TV}$. □

## 5.3 Monte–Carlo simulations

Up to now, we assumed in this chapter that we had a Markov chain, coming from a modelisation, and we studied its behaviour. One can conversely use Markov chains to study, and precisely here simulate, exactly or approximately, a given distribution. This concept is called MCMC for "Markov chain Monte–Carlo". Indeed, suppose we have a finite, but large, set $\mathbb{X}$ and a probability measure $\pi$ on this set. Even in a simple

setting, say if $\pi$ is the uniform distribution on $\mathbb{X}$, it may not be easy to simulate in practice a random variable with the law $\pi$, or close to it.

However in many cases we are able to construct a Markov chain $(X_n)_{n\geq 0}$ on $\mathbb{X}$ that has $\pi$ as stationary probability and we are able to simulate it, using e.g. the representation as a random recursion. According to Corollary 5.1.2, if we simulate one trajectory $X_0, \dots, X_n$ for a large $n$, then the average amount of time $n^{-1}\sum_{k=0}^{n-1} \mathbb{1}_{X_k=x}$ spent at a given $x$ approximates $\pi(x)$. More generally, the average $n^{-1}\sum_{k=0}^{n-1} f(X_k)$ of a function converges to its integral $\pi(f) = \sum_{x\in\mathbb{X}} f(x)\pi(x)$, and Theorem 5.1.4 provides asymptotic confident intervals, just in the same way we use the CLT for i.i.d. random variables. If one is interested in numerically computing this limit integral, then this can provide a more efficient way than deterministic schemes whose complexity grows with the dimension. In another direction, one can sample a large number $N$ of i.i.d. trajectories $(X_0^i, \dots, X_n^i)$ for $1 \leq i \leq N$, and then by the usual Law of Large Numbers, the average $N^{-1}\sum_{i=1}^{N} \mathbb{1}_{X_n^i=x}$ approximates $\mathbb{P}(X_n = x)$ which itself approximates $\pi(x)$ by Theorem 5.2.8, and with an exponential speed of convergence as shown by Theorem 5.2.13.

In the next subsection we describe an algorithm to run such a Markov chain, which we first apply to particular laws called Gibbs measures. Finally we relate these measures to the problem of minimising a cost function.

Throughout this section, we assume that $\mathbb{X}$ is a finite (but very large) set. One can think of a discretised compact subset in $\mathbb{R}^d$ with a small mesh size, or to a large finite network for example.

### 5.3.1 The Metropolis–Hastings algorithm

Let $\pi$ denote a probability measure on a finite set $\mathbb{X}$ and assume that $\pi(x) > 0$ for every $x \in \mathbb{X}$ (otherwise simply remove all the points $x$ where $\pi(x) = 0$). Let $h : (0,\infty) \longrightarrow (0,1]$ be a nondecreasing function that satisfies $h(u) = uh(1/u)$ for every $u > 0$; two usual examples are:

$$h(u) = \min\{u, 1\} \qquad \text{as well as} \qquad h(u) = \frac{u}{u+1}.$$

Let $P_0$ be an irreducible transition matrix on $\mathbb{X}$ that has for any $x, y \in \mathbb{X}$:

$$P_0(x,y) > 0 \iff P_0(y,x) > 0.$$

This transition matrix is called a *proposal* matrix. Let us then define the *rejection probability*:

$$R(x,y) = h\Big(\frac{\pi(y)P_0(y,x)}{\pi(x)P_0(x,y)}\Big),$$

which is well-defined for $x \neq y$ such that $P_0(x,y) \neq 0$; when $P_0(x,y) = 0$, we simply put $R(x,y) = 0$. Finally let us set:

$$P(x,y) = P_0(x,y)R(x,y) \qquad \text{for } x \neq y \text{ and then} \qquad P(x,x) = 1 - \sum_{y\neq x} P(x,y). \tag{5.6}$$

Recall the notion of reversibility from Definition 4.2.1.

**Theorem 5.3.1.** *The matrix $P$ from* (5.6) *is an irreducible transition matrix and the law $\pi$ is reversible for $P$. Finally $P$ is aperiodic as soon as either $h < 1$ or $P_0$ is aperiodic.*

*Proof.* Clearly $\sum_y P(x,y) = 1$ and $P(x,y) \geq 0$ if $x \neq y$. For $x = y$, we have since $h \leq 1$:

$$P(x,x) = 1 - \sum_{y\neq x} P_0(x,y)R(x,y) \geq 1 - \sum_{y\neq x} P_0(x,y) = P_0(x,x) \geq 0$$

since $P_0$ is a transition matrix. Also, since $h > 0$, then irreducibility of $P$ is inherited from that of $P_0$: for every $x, y \in \mathbb{X}$, there exists $n \geq 1$ such that $P_0^n(x,y)$, and thus $P^n(x,y) > 0$. The reversibility of $\pi$ follows by

the property $h(u) = uh(1/u)$, namely for $x \neq y$:

$$\begin{aligned}\pi(x)P(x,y) &= \pi(x)P_0(x,y)h\Big(\frac{\pi(y)P_0(y,x)}{\pi(x)P_0(x,y)}\Big)\\ &= \pi(x)P_0(x,y)\frac{\pi(y)P_0(y,x)}{\pi(x)P_0(x,y)}h\Big(\frac{\pi(x)P_0(x,y)}{\pi(y)P_0(y,x)}\Big)\\ &= \pi(y)P_0(y,x)h\Big(\frac{\pi(x)P_0(x,y)}{\pi(y)P_0(y,x)}\Big)\\ &= \pi(y)P(y,x).\end{aligned}$$

Let us finally focus on aperiodicity of $P$. Recall from Remark 5.2.3 that an easy case is when there exists $x \in \mathbb{X}$ such that $P(x,x) > 0$. Writing again $P(x,x) = 1 - \sum_{y\neq x} P_0(x,y)R(x,y)$, this is the case as soon as there exists $y \neq x$, such that both $P_0(x,y) > 0$ and $R(x,y) < 1$. In particular this holds as soon as $h < 1$. Next, if $R(x,y) = 1$ for every $x \neq y$ such that $P_0(x,y) > 0$, then $P(x,y) = P_0(x,y) > 0$ for all such pairs, and $P(x,y) = 0 = P_0(x,y)$ for the other pairs, so finally $P = P_0$ which is thus aperiodic if we suppose that $P_0$ is aperiodic (!). □

Suppose that we are able to generate a Markov chain with transition matrix $P_0$, say using the representation $P_0(x,y) = \mathbb{P}(f(x,\xi) = y)$, then we can generate a Markov chain $(X_n)_{n\geq 0}$ with transition matrix $P$ by running the following algorithm:

- Initialise with some $X_0$
- For $k$ from 0 to $n-1$, do:
  - Sample $Y$ from $\mathbb{P}(Y = y) = \mathbb{P}(f(X_k,\xi) = y)$
  - Sample $U$ with the uniform distribution on $[0,1]$
  - If $U < R(X_k, Y)$, then set $X_{k+1} = Y$, else set $X_{k+1} = X_k$
- Return $(X_0, \dots, X_n)$

According to Theorem 5.3.1 this Markov chain $(X_n)_{n\geq 0}$ has stationary distribution $\pi$ and since the state space $\mathbb{X}$ is finite, then Theorem 5.2.13 applies so there exist $\delta > 0$ and $k \geq 1$ such that for every $n \geq 1$, it holds:

$$\sum_{x\in\mathbb{X}} |\mathbb{P}(X_n = x) - \pi(x)| \leq 2(1-\delta)^{\lfloor n/k\rfloor},$$

uniformly for all initial distributions. Hence this algorithm allows to generate $X_n$ with a law close to $\pi$, with a control of the error, which decays exponentially fast to 0.

**Remark 5.3.2.** In the first version of this algorithm, the function $h$ was precisely $h(u) = \min\{u, 1\}$ and the matrix $P_0$ was symmetric in that $P_0(x,y) = P_0(y,x)$ for all $x, y$. In this case we have simply:

$$R(x,y) = \begin{cases} 1 & \text{when } \pi(y) \geq \pi(x),\\ \frac{\pi(y)}{\pi(x)} & \text{when } \pi(y) < \pi(x).\end{cases}$$

**Remark 5.3.3.** It may be interesting in some cases to take a transition matrix of the form $P_0(x,y) = P_0(y)$, that is, sample a proposal move $Y_{n+1}$ at every step that is independent of $X_n$. But then the acceptance of this proposal still depends on $X_n$ through the function $R$.

### 5.3.2 Gibbs measures

A very useful particularity of the previous algorithm is that it only depends on $\pi$ through ratios of the form $\pi(y)/\pi(x)$. In particular this can be used to approximate $\pi$ when the latter is only known up to a

multiplicative constant. This is very well suited to study *Gibbs measures* which come from statistical physics. Let again $\mathbb{X}$ be a large but finite set and let $V : \mathbb{X} \longrightarrow \mathbb{R}$ be a function which we call a potential. For any $T > 0$, we define a probability measure on $\mathbb{X}$ by setting for every $x \in \mathbb{X}$:

$$\pi_T(x) = \frac{1}{Z_T} \exp\Big(-\frac{V(x)}{T}\Big) \qquad \text{where} \qquad Z_T = \sum_{x \in \mathbb{X}} \exp\Big(-\frac{V(x)}{T}\Big). \tag{5.7}$$

In many cases the potential $V(x)$ can be computed for any given $x \in \mathbb{X}$, but computing $Z_T$ requires to compute $V(x)$ *for all* $x \in \mathbb{X}$, which cannot be done in practice when $\mathbb{X}$ is too large. The Metropolis–Hastings algorithm allows to approximate $\pi_T$ without computing $Z_T$!

Let us describe here one historical example of application to the Ising model, which is a simplified version of a magnetic system. Take a large number $N$ of particles placed on a regular grid, say for example a rectangle in $\mathbb{Z}^2$, which represents a piece of metal; each particle $i$ possesses a spin $s_i \in \{-1, +1\}$, which corresponds to its orientation. A configuration of spins is then an element $\mathbf{s} = (s_i)_{1 \le i \le N} \in \{-1, +1\}^N$. Let us write $i \sim j$ when the particles $i$ and $j$ are neighbours in the grid. We then consider the potential:

$$V(\mathbf{s}) = -\sum_{i \sim j} s_i s_j.$$

Two configurations minimise the energy (the "fundamental states"): $s_i = +1$ for every $i$ and $s_i = -1$ for every $i$. More generally, the potential is small when the spins of neighbours tend to align with each other, and therefore these configurations are given a higher probability in the corresponding Gibbs measure $\pi_T$.

Computing $V(\mathbf{s})$ for any given configuration $\mathbf{s}$ takes a linear complexity, of order $N$, but computing the normalising constant $Z_T$ requires to compute $V(\mathbf{s})$ for all the $2^N$ configurations! However the Metropolis–Hastings algorithm can be easily implemented here. As proposal $P_0$, given a spin configuration, choose one particle uniformly at random and replace its spin by its opposite. Formally: for $\mathbf{s} \in \{-1, +1\}^N$ and $i \in \{1, \dots, N\}$, let $\mathbf{s}^{(i)} \in \{-1, +1\}^N$ be given by $s_j^{(i)} = s_j$ for $j \neq i$ and $s_i^{(i)} = -s_i$. Then set

$$P_0(\mathbf{s}, \mathbf{s}^{(i)}) = \frac{1}{N} \qquad \text{for every } \ 1 \le i \le N.$$

For two such configurations $\mathbf{s}$ and $\mathbf{s}^{(i)}$, we have

$$V(\mathbf{s}^{(i)}) - V(\mathbf{s}) = 2 s_i \sum_{j \sim i} s_j \qquad \text{and so} \qquad \frac{\pi_T(\mathbf{s}^{(i)})}{\pi_T(\mathbf{s})} = \exp\Big(-\frac{2}{T} s_i \sum_{j \sim i} s_j\Big).$$

Note that $P_0(\mathbf{s}^{(i)}, \mathbf{s}) = P_0(\mathbf{s}, \mathbf{s}^{(i)})$; take $h(u) = \min\{u, 1\}$, then by Remark 5.3.2, the Metropolis–Hastings algorithm works as follows:

- Initialise with some $X_0 = \mathbf{s}$
- For $k$ from 0 to $n - 1$, do:
  - Let $X_{k+1} = X_k$
  - Sample $I$ uniformly at random in $\{1, \dots, N\}$ and compute $Z = 2X_k(I) \sum_{j \sim I} X_k(j)$
  - If $Z \le 0$, then set $X_{k+1}(I) := -X_{k+1}(I)$,
  - Else, sample $U$ uniformly at random in $[0, 1]$, if $U < \exp(-2Z/T)$, then set $X_{k+1}(I) := -X_{k+1}(I)$
- Return $(X_0, \dots, X_n)$

We see in the loop that every time changing the random spin by its opposite reduces the total energy, we accept this change so we tend to decrease the energy as time goes by. On the other hand we also allow randomly to move to a state with higher energy, so we do not get trapped in a local minimum of energy. Let us push further this idea in the next problem.

### 5.3.3 Optimisation problem and simulated annealing

Let $V$ be a general potential on a large finite set $\mathbb{X}$. Our aim is to find, algorithmically, a way to minimise $V$. If $V$ is a convex function, there is a well-known method, called the gradient descent, which is described in Section 9.8, which defines a recursive sequence that converges to the unique minimiser. However if $V$ has several local minima which are not global minima, then this algorithm may converge to one of them and completely miss the global minimum.

We shall circumvent this issue by means of the Gibbs measure $\pi_T$ associated with $V$, which can be approximated by the Metropolis–Hastings algorithm. The parameter $T > 0$ is interpreted as the temperature. When $T$ is high, then so are the random fluctuations: in the previous algorithm, the threshold $\exp(-2Z/T)$ when $Z > 0$ is close to 1 so we accept most of the proposals. However when $T$ is small, proposals which increase the energy are more often rejected so the configurations with minimal energy are preferred (in the Ising model, the spins tend to align more with each other). Formally, given any potential $V$ on $\mathbb{X}$, let

$$\mathcal{M}(V) = \operatorname{argmin} V = \{x \in \mathbb{X} : \ V(x) = \min_{y \in \mathbb{X}} V(y)\}$$

denote the set of minimisers of $V$.

**Lemma 5.3.4.** *The Gibbs measure $\pi_T$ converges to the uniform distribution on $\mathcal{M}(V)$ as $T \longrightarrow 0$, namely for every $x \in \mathbb{X}$, we have:*

$$\lim_{T \to 0} \pi_T(x) = \begin{cases} \operatorname{Card}(\mathcal{M}(V))^{-1} & \text{if } x \in \mathcal{M}(V), \\ 0 & \text{if } x \notin \mathcal{M}(V). \end{cases}$$

*Proof.* Let $V^* = \min V$ denote the minimum value of $V$, then for every $x \in \mathbb{X}$, we have after multiplying the numerator and denominator by $\exp(V^*/T)$:

$$\pi_T(x) = \frac{1}{\sum_{y \in \mathbb{X}} \exp(-(V(y) - V^*)/T)} \exp\Big(-\frac{V(x) - V^*}{T}\Big).$$

Now on the right-hand side, the term in the exponential vanishes when $x \in \mathcal{M}(V)$, whereas it tends to $-\infty$ as $T \longrightarrow 0$ when $x \notin \mathcal{M}(V)$. Thus indeed:

$$\lim_{T \to 0} \pi_T(x) = \mathbb{1}_{x \in \mathcal{M}(V)} \Bigg( \sum_{y \in \mathcal{M}(V)} \mathbb{1}_{y \in \mathcal{M}(V)} \Bigg)^{-1} = \mathbb{1}_{x \in \mathcal{M}(V)} \frac{1}{\operatorname{Card}(\mathcal{M}(V))},$$

and the proof is complete. □

We can then use this property to solve our optimisation problem. The *simulated annealing* consists in running the Metropolis–Hastings algorithm to approximate $\pi_T$ but letting $T = T_n$ vary at each step. Notice that the Markov chain is then inhomogeneous in time. The idea is to have $T_n$ relatively large at first, so $\pi_{T_n}$ fluctuates a lot and the Markov chain moves a lot and visits many states, and slowly let $T_n$ tend to 0 so the Markov chain stabilises on the minimum. The speed of convergence of $T_n$ to 0 is then crucial. We will not prove the following result.

**Theorem 5.3.5.** *For any potential $V$ on a finite set $\mathbb{X}$ and any proposal transition matrix $P_0$, there exists a constant $C > 0$ which depends on both $V$ and $P_0$ such that the Metropolis–Hastings algorithm run with $T_n = C(\log n)^{-1}$ satisfies*

$$\mathbb{P}(X_n \in \mathcal{M}(V)) \underset{n \to \infty}{\longrightarrow} 1.$$

*In words the algorithm stabilises on a minimiser of $V$ with arbitrarily high probability.*

As a last example of application, let us consider the travelling salesman problem. Let $N$ points $z_1, \dots, z_N$ in $\mathbb{R}^2$ which we think as locations that our salesman has to visit while minimising the total travel distance.

A configuration is here an order on the $z_i$'s, or equivalently a permutation $\sigma$ of $\{1, \dots, N\}$, and the potential of a permutation $\sigma = (\sigma(1), \dots, \sigma(N))$ is the total length:

$$V(\sigma) = \sum_{i=1}^{N} \|z_{\sigma(i+1)} - z_{\sigma(i)}\|,$$

where we put $\sigma(N+1) = \sigma(1)$ so the salesman ends the journey by coming back to the starting point. Here again the size $N!$ of the set of configurations does not allow to compute $V(\sigma)$ for every permutation $\sigma$, but we can use the simulated annealing. Indeed, for a permutation $\sigma$ and two indices $i \neq j$, let $\sigma^{(i,j)}$ denote the permutation obtained from $\sigma$ by simply exchanging $\sigma(i)$ and $\sigma(j)$. Then we can use as proposal transitions the matrix:

$$\forall i \neq j, \quad P_0(\sigma, \sigma^{(i,j)}) = \frac{1}{N(N-1)} \qquad \text{and otherwise} \qquad P_0(\sigma, \sigma') = 0.$$

In words, similarly to the Ising model, we pick two different locations uniformly at random and exchange their order in the tour.

# Part III

# Martingales

# Chapter 6

# Conditional Expectation

This chapter introduces the notion of conditional expectation of a random variable given any other one, which generalises the supposedly known case of conditioning with respect to a discrete random variable, or when the pair has a joint density. This short and technical chapter is the foundation of the theory of martingales (and Markov chains in continuous spaces) developed subsequently.

**Contents**



We start by presenting in Section 6.1 with a probabilistic vocabulary an actually general notion of orthogonal projection in a Hilbert space, which we first apply to the so-called linear regression, namely solving the problem to find the affine combination of known random variables that best approximates in the least mean square sense an unknown one. Then in Section 6.2 we construct the conditional expectation by roughly speaking extending the orthogonal projection from $L^2$ to $L^1$. In Section 6.3 we relate this new abstract notion to the two familiar cases of conditioning a real random variable with respect to another one when either the latter is discrete, or when the pair has a joint density. In Section 6.4 we present all basic key properties of the conditional expectation that are used all the time: first, properties that extend the usual ones of the expectation, then some specific ones such as the tower property, and then the relation with independence. Section 6.6 discusses the case of Gaussian vectors for which conditional expectation actually coincides with the linear regression problem and can be easily calculated. Finally Section 6.7 mentions some developments that the curious reader may have in mind about the notion of conditional probability but which are beyond the scope of this course.

## 6.1 Orthogonal projection in $L^2$

Recall from Section 2.2 for $p \geq 1$ the spaces $L^p$ of random variables $X$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}$ and such that $\mathbb{E}[|X|^p] < \infty$, in which two random variables that are equal almost surely are seen as the same object. This space when $p = 2$ is equipped with a scalar product: $X \cdot Y = \mathbb{E}[XY]$, whose associated norm $\|\cdot\|_2$ is complete, hence it is a *Hilbert space*. Such spaces are very close to Euclidean spaces. The next theorem considers the orthogonal projection on a complete subspace.

**Theorem 6.1.1** (Orthogonal projection)**.** *Let $\mathcal{K}$ be a complete vector subspace of $L^2$ and let $X \in L^2$. Then there exists $\widehat{X} \in \mathcal{K}$ which satisfies the following two equivalent properties:*

*(i)* $\|X - \widehat{X}\|_2 = \inf\{\|X - Z\|_2 : Z \in \mathcal{K}\}$.

*(ii)* $X - \widehat{X} \perp Z$, *i.e.* $\mathbb{E}[(X - \widehat{X})Z] = 0$, *for every* $Z \in \mathcal{K}$.

*Moreover, if $\widehat{X}' \in \mathcal{K}$ is another random variable satisfying these properties, then $\|\widehat{X} - \widehat{X}'\|_2 = 0$ so $\widehat{X} = \widehat{X}'$ a.s. Finally, the projection is linear in the sense that for $X, Y \in L^2$, if $\widehat{X}$ and $\widehat{Y}$ denote respectively their orthogonal projection, then the orthogonal projection of $X + Y$ equals $\widehat{X} + \widehat{Y}$ a.s.*

*Proof.* Let $\Delta = \inf\{\|X - Z\|_2 : Z \in \mathcal{K}\}$ and let $(Z_n)_{n\geq 1}$ be a sequence in $\mathcal{K}$ such that $\|X - Z_n\|_2 \longrightarrow \Delta$. Note the parallelogram identity:

$$\begin{aligned}
&\|X - Z_n\|_2^2 + \|X - Z_m\|_2^2 \\
&= \mathbb{E}[(X - Z_n)^2] + \mathbb{E}[(X - Z_m)^2] \\
&= \mathbb{E}\left[\left(\left(X - \frac{Z_n + Z_m}{2}\right) - \left(\frac{Z_n - Z_m}{2}\right)\right)^2\right] + \mathbb{E}\left[\left(\left(X - \frac{Z_n + Z_m}{2}\right) + \left(\frac{Z_n - Z_m}{2}\right)\right)^2\right] \\
&= 2\,\mathbb{E}\left[\left(X - \frac{Z_n + Z_m}{2}\right)^2\right] + 2\,\mathbb{E}\left[\left(\frac{Z_n - Z_m}{2}\right)^2\right] \\
&\geq 2\Delta^2 + \frac{1}{2}\|Z_n - Z_m\|_2^2,
\end{aligned}$$

since $(Z_n + Z_m)/2 \in \mathcal{K}$. Hence

$$\|Z_n - Z_m\|_2^2 \leq 2\big(\|X - Z_n\|_2^2 + \|X - Z_m\|_2^2 - 2\Delta^2\big) \underset{n,m\to\infty}{\longrightarrow} 0,$$

so $(Z_n)_{n\geq 1}$ is a Cauchy sequence. Since $\mathcal{K}$ is a complete then $(Z_n)_{n\geq 1}$ converges in $L^2$ to some $\widehat{X} \in \mathcal{K}$. Now by the Minkowski inequality, we have

$$\Delta \leq \|X - \widehat{X}\|_2 \leq \|X - Z_n\|_2 + \|Z_n - \widehat{X}\|_2 \underset{n\to\infty}{\longrightarrow} \Delta,$$

thus (i) holds.

For every $Z \in \mathcal{K}$ and $t \in \mathbb{R}$ we have:

$$\mathbb{E}[(X - \widehat{X} - tZ)^2] = \mathbb{E}[(X - \widehat{X})^2] + t^2\,\mathbb{E}[Z^2] - 2t\,\mathbb{E}[(X - \widehat{X})Z].$$

Therefore, for every $Z \in \mathcal{K}$,

$$\min_{t\in\mathbb{R}} \mathbb{E}[(X - \widehat{X} - tZ)^2] = \mathbb{E}[(X - \widehat{X})^2] - \frac{\mathbb{E}[(X - \widehat{X})Z]^2}{\mathbb{E}[Z^2]}.$$

Since every element of $\mathcal{K}$ can be written as $\widehat{X} + tZ$, then $\widehat{X}$ satisfies (i) if and only if it satisfies (ii).

Next, if $\widehat{X}' \in \mathcal{K}$ satisfies (ii), then by expanding the expectations, we infer that $\mathbb{E}[\widehat{X}Z] = \mathbb{E}[\widehat{X}'Z]$ for any $Z \in \mathcal{K}$; in particular, $\|\widehat{X} - \widehat{X}'\|_2^2 = \mathbb{E}[\widehat{X}^2] + \mathbb{E}[(\widehat{X}')^2] - 2\,\mathbb{E}[\widehat{X}\widehat{X}'] = 0$.

Finally, for any $Z \in \mathcal{K}$ we have

$$\mathbb{E}[(X + X' - (\widehat{X} + \widehat{X}'))Z] = \mathbb{E}[(X - \widehat{X})Z] + \mathbb{E}[(X' - \widehat{X}')Z] = 0$$

so $\widehat{X} + \widehat{X}' \in \mathcal{K}$ satisfies (ii) for $X + X'$ so it must be a.s. equal to its orthogonal projection. □

Let us apply Theorem 6.1.1 to a particular space $\mathcal{K}$. Let $X, Y_1, \dots, Y_n$ be real random variables in $L^2$. The *linear regression* of $X$ over $Y = (Y_1, \dots, Y_n)$ is the affine combination of the $Y_k$'s that minimises the $L^2$ distance to $X$, that is, provided it exists, the vector $(\alpha_0, \dots, \alpha_n) \in \mathbb{R}^{n+1}$ such that:

$$\mathbb{E}\left[\left(X - \alpha_0 - \sum_{k=1}^n \alpha_k Y_k\right)^2\right] = \min_{(\beta_0,\dots,\beta_n)\in\mathbb{R}^n} \mathbb{E}\left[\left(X - \beta_0 - \sum_{k=1}^n \beta_k Y_k\right)^2\right]. \tag{6.1}$$

We shall assume that no $Y_k$ is an affine combination of the other ones, which is equivalent to assuming that their covariance matrix $C_Y = (\mathrm{Cov}(Y_i, Y_j))_{1\le i,j\le n}$ is invertible.

One can solve the case $n = 0$ and $n = 1$ by hand.

**Exercise 6.1.2.** Suppose that $X \in L^2$. Prove that $\mathbb{E}[X]$ is the best approximation of $X$ by a constant in the sense:

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{c\in\mathbb{R}} \mathbb{E}[(X - c)^2].$$

You may simply expend $\mathbb{E}[(X - c)^2]$ to get a simple function of $c$ that you now very well.

**Exercise 6.1.3.** Suppose that $X, Y \in L^2$ with $\mathrm{Var}(Y) > 0$. Prove that the best approximation of $X$ of the form $a + bY$ with $a, b \in \mathbb{R}$ is given by:

$$a = \mathbb{E}[X] - b\,\mathbb{E}[Y], \qquad \text{and then} \qquad b = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]}{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}.$$

Thus the minimiser $a + bY$ is given by:

$$\mathbb{E}[X] + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}[Y]).$$

As previously you may simply expend $\mathbb{E}[(X - a - bY)^2]$ to get a quadratic function of $a$ and $b$.

The general problem can be solved using the orthogonal projection.

**Corollary 6.1.4.** *When $C_Y = (\mathrm{Cov}(Y_i, Y_j))_{1\le i,j\le n}$ is invertible there is a unique solution to* (6.1)*, which is given by $\alpha_0 = \mathbb{E}[X] - \sum_{k=1}^n \alpha_k\,\mathbb{E}[Y_k]$ and $\alpha = (\alpha_1, \dots, \alpha_n)$ is $\alpha = C_Y^{-1}\,\mathrm{Cov}(X, Y)$. The best affine approximation of $X$ by the $Y_k$'s is thus given by:*

$$\alpha_0 + \sum_{k=1}^n \alpha_k Y_k = \mathbb{E}[X] + \left(C_Y^{-1}\,\mathrm{Cov}(X, Y)\right)^t (Y - \mathbb{E}[Y]).$$

*Proof of Corollary 6.1.4.* Let $\mathcal{K}$ denote the linear space spanned by $1, Y_1, \dots, Y_n$ and let $\widehat{X}$ denote the orthogonal projection of $X$ on $\mathcal{K}$. Since $\widehat{X} \in \mathcal{K}$ then there exists $\lambda_0, \dots, \lambda_n$ such that:

$$\widehat{X} = \lambda_0 + \sum_{k=1}^n \lambda_k (Y_k - \mathbb{E}[Y_k]),$$

and we know from Theorem 6.1.1 that it solves (6.1). By orthogonality, we have $\mathbb{E}[(X - \widehat{X})Z] = 0$ for any $Z \in \mathcal{K}$. In particular, for $Z = 1$, we infer that

$$\mathbb{E}[X] = \mathbb{E}[\widehat{X}] = \lambda_0.$$

Further, for any $1 \le \ell \le n$, we have $\mathbb{E}[(X - \widehat{X})(Y_\ell - \mathbb{E}[Y_\ell])] = 0$ which is equivalent to

$$\mathrm{Cov}(X, Y_\ell) = \mathrm{Cov}(\widehat{X}, Y_\ell) = \sum_{k=1}^n \lambda_k\,\mathrm{Cov}(Y_k, Y_\ell).$$

Conversely, if the $\lambda_k$'s form such a solution, then the random variable $\widehat{X} = \lambda_0 + \sum_{k=1}^n \lambda_k (Y_k - \mathbb{E}[Y_k])$ belongs to $\mathcal{K}$ and one easily shows that $\mathbb{E}[(X - \widehat{X})Z] = 0$ for any $Z \in \mathcal{K}$ so it coincides with the orthogonal projection which minimises the square distance. □

## 6.2 The conditional expectation

The previous result provides the best approximation of a random variable $X$ as an affine combination of another one $Y = (Y_1, \dots, Y_n)$. However it may exist better approximations, that use non linear functions. If $\mathbb{E}[X^2] < \infty$, this relies on a more abstract orthogonal projection. It can actually be extended assuming only $\mathbb{E}[|X|]$ and this is formalised in the notion of *conditional expectation*. Below we try to give both the picture of orthogonal projection and that of prediction of $X$ given the information provided by the random variable $Y$.

**Notation.** Throughout this chapter, we denote by $X$ a random variable with values either in $[0, \infty]$ or in $\mathbb{R}$ with $\mathbb{E}[|X|] < \infty$ in the latter case, and we denote by $Y$ a random variable with values in a general measured space.

**Theorem 6.2.1** (Conditional Expectation)**.** *Let $X$ be a random variable with values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and let $Y$ be any random variable. Suppose that either $X \in [0, \infty]$ a.s. or $\mathbb{E}[|X|] < \infty$. Then there exists a measurable real-valued function $\Psi$ satisfying the following properties:*

*(i) $\Psi(Y) \in [0, \infty]$ a.s. or $\mathbb{E}[|\Psi(Y)|] < \infty$ respectively,*

*(ii) For any function $h$ either nonnegative or bounded respectively, we have:*

$$\mathbb{E}[Xh(Y)] = \mathbb{E}[\Psi(Y)h(Y)].$$

*Moreover, if $\Phi$ is another such function, then $\Psi(Y) = \Phi(Y)$ a.s.*

We call $\Psi(Y)$ a version of the *conditional expectation of $X$ given $Y$* and denote it by $\mathbb{E}[X \mid Y]$. In everyday use we do not distinguish several almost surely equal versions and speak of *the* conditional expectation.

*Proof.* Existence in the $L^2$ case. Let us suppose first that $\mathbb{E}[|X|^2] < \infty$. Let $L^2(Y)$ denote the space of random variables of the form $g(Y)$ with $\mathbb{E}[|g(Y)|^2] < \infty$. This subspace of $L^2$ is complete so by Theorem 6.1.1 there exists an a.s. unique orthogonal projection of $X$ onto $L^2(Y)$, which takes the form $\widehat{X} = \Psi(Y)$ with $\mathbb{E}[|\Psi(Y)|^2] < \infty$ and satisfies the orthogonality property:

$$\mathbb{E}[(X - \Psi(Y))h(Y)] = 0, \qquad \text{equivalently} \qquad \mathbb{E}[Xh(Y)] = \mathbb{E}[\Psi(Y)h(Y)],$$

for every measurable function $h$ such that $\mathbb{E}[|h(Y)|^2] < \infty$.

Existence in the nonnegative case. Suppose next that $X \geq 0$ a.s. For any $n \geq 1$, let $X_n = \min\{X, n\} \in L^2$ and let $\widehat{X}_n = \Psi_n(Y)$ denote its orthogonal projection on $L^2(Y)$. Then $\mathbb{1}_{\{\Psi_n(Y)<0\}} \in L^2(Y)$ as well, so by the above orthogonality property, we have:

$$0 \leq \mathbb{E}[X_n \, \mathbb{1}_{\{\Psi_n(Y)<0\}}] = \mathbb{E}[\Psi_n(Y) \, \mathbb{1}_{\{\Psi_n(Y)<0\}}] \leq 0.$$

Thus the nonnegative random variable $\Psi_n(Y) \, \mathbb{1}_{\{\Psi_n(Y)<0\}}$ has expectation 0, so it equals 0 a.s. and so $\Psi_n(Y) \geq 0$ a.s. The same argument applied to $X_{n+1} - X_n \geq 0$ combined with linearity of the projection shows that $0 \leq \Psi_n(Y) \leq \Psi_{n+1}(Y)$ a.s. so we can define its a.s. limit $\Psi(Y) = \uparrow \lim_n \Psi_n(Y) \in [0, \infty]$. Now fix any measurable function $h \geq 0$ and let $h_n = \min\{h, n\}$ so $\mathbb{E}[h_n(Y)^2] < \infty$. Then we have by monotone convergence:

$$\mathbb{E}[Xh(Y)] = \uparrow \lim_{n\to\infty} \mathbb{E}[X_n h_n(Y)] = \uparrow \lim_{n\to\infty} \mathbb{E}[\Psi_n(Y)h_n(Y)] = \mathbb{E}[\Psi(Y)h(Y)].$$

Note that by taking $h = 1$ we infer that $\mathbb{E}[\Psi(Y)] = \mathbb{E}[X]$.

Existence in the integrable case. Finally, if $\mathbb{E}[|X|] < \infty$ but $X$ is not necessarily nonnegative, write $X = X^+ - X^-$ with $X^+ = \max(X, 0) \geq 0$ and $X^- = -\min(X, 0) = \max(-X, 0) \geq 0$, so that $|X| = |X^+| + |X^-|$. Construct $\Psi^+(Y)$ and $\Psi^-(Y)$ as above, which have $\mathbb{E}[\Psi^+(Y)] = \mathbb{E}[X^+] < \infty$ and $\mathbb{E}[\Psi^-(Y)] = \mathbb{E}[X^-] < \infty$. Define then $\Psi = \Psi^+(Y) - \Psi^-(Y)$, which has $\mathbb{E}[|\Psi(Y)|] \leq \mathbb{E}[\Psi^+(Y)] + \mathbb{E}[\Psi^-(Y)] < \infty$. Let $h$ be a bounded

function and decompose similarly $h(Y) = h(Y)^+ - h(Y)^-$. By linearity, we infer from the case of nonnegative random variables that

$$\begin{aligned}\mathbb{E}[\Psi(Y)h(Y)] &= \mathbb{E}[\Psi^+(Y)h(Y)^+] - \mathbb{E}[\widehat{X}^-(Y)h(Y)^+] - \mathbb{E}[\Psi^+(Y)h(Y)^-] + \mathbb{E}[\Psi^-(Y)h(Y)^-] \\ &= \mathbb{E}[X^+h(Y)^+] - \mathbb{E}[X^-h(Y)^+] - \mathbb{E}[X^+h(Y)^-] + \mathbb{E}[X^-h(Y)^-] \\ &= \mathbb{E}[Xh(Y)].\end{aligned}$$

This completes the proof of existence of $\Psi(Y)$.

Uniqueness. Suppose $\Psi(Y)$ and $\Phi(Y)$ both satisfy the theorem, then $h(Y) = \mathbb{1}_{\Psi(Y)>\Phi(Y)}$ is bounded so

$$\begin{aligned}\mathbb{E}[(\Psi(Y) - \Phi(Y))\,\mathbb{1}_{\Psi(Y)>\Phi(Y)}] &= \mathbb{E}[\Psi(Y)\,\mathbb{1}_{\Psi(Y)>\Phi(Y)}] - \mathbb{E}[\Phi(Y)\,\mathbb{1}_{\Psi(Y)>\Phi(Y)}] \\ &= \mathbb{E}[X\,\mathbb{1}_{\Psi(Y)>\Phi(Y)}] - \mathbb{E}[X\,\mathbb{1}_{\Psi(Y)>\Phi(Y)}] \\ &= 0.\end{aligned}$$

Hence the nonnegative random variable $(\Psi(Y)-\Phi(Y))\,\mathbb{1}_{\Psi(Y)>\Phi(Y)}$ must be 0 a.s. which means that $\Psi(Y) \le \Phi(Y)$ a.s. By a symmetric argument we also have $\Psi(Y) \ge \Phi(Y)$ a.s. □

**Remark 6.2.2.** The restriction to $h$ either nonnegative or bounded ensures that $\mathbb{E}[\Psi(Y)h(Y)]$ and $\mathbb{E}[Xh(Y)]$ are well-defined but the identity $\mathbb{E}[\Psi(Y)h(Y)] = \mathbb{E}[Xh(Y)]$ extends as soon as both sides make sense by similar approximations as in the proof.

**Example 6.2.3.** Let us consider a few extreme examples. In both cases, one simply checks that the given candidate satisfies the properties of the conditional expectation and conclude by uniqueness.

(i) If $X = f(Y)$ is a measurable function of $Y$, then $\mathbb{E}[f(Y) \mid Y] = f(Y)$ a.s. In words, if we are given all the possible information about $Y$, then $X = f(Y)$ is determined so the best prediction is $f(Y)$ itself; put differently, we want to project a vector on a subspace where it already lives, so it doesn't move anywhere.

(ii) If $Y$ is constant a.s. then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ a.s. Here we are given no information at all, so our prediction $\Psi(Y)$ is a constant, and the best constant is $\mathbb{E}[X]$.

(iii) More generally if $X$ and $Y$ are independent, then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ a.s. Again, here we are given irrelevant information, so our prediction $\Psi(Y)$ is a constant.

To be explicit, since we shall frequently condition a random variable $X$ with respect to several random variables $Y_1, \dots, Y_n$, this amounts to condition with respect to $Y = (Y_1, \dots, Y_n)$, namely, when $X$ is either nonnegative or integrable, we have

$$\mathbb{E}[X \mid Y_1, \dots, Y_n] = \Psi(Y_1, \dots, Y_n),$$

where $\Psi$ is a measurable function characterised by Property (ii), namely:

$$\mathbb{E}[Xh(Y_1, \dots, Y_n)] = \mathbb{E}[\Psi(Y_1, \dots, Y_n)h(Y_1, \dots, Y_n)]$$

for any measurable function $h$ either nonnegative (when $\Psi$ is) or bounded (when $\Psi$ is integrable).

## 6.3 Two familiar cases

Let us compare this notion of conditional expectation with the familiar ones of conditioning $X$ with respect to $Y$ when either $Y$ is a discrete r.v. or when the pair $(X, Y)$ has a density.

### 6.3.1 Discrete case

Fix an event $B \in \mathcal{F}$ with nonzero probability, then it is well-known that the formula

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

for all $A \in \mathcal{F}$ defines a probability measure $\mathbb{P}(\cdot \mid B)$. Then one can construct the expectation with respect to this probability, which takes the form

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X\, \mathbb{1}_B]}{\mathbb{P}(B)}$$

for any r.v. either nonnegative or integrable. Indeed by definition it holds true for any $X$ of the form $\mathbb{1}_A$ with $A \in \mathcal{F}$, and then as usual, it extends to simple r.v.'s by linearity, and further to nonnegative r.v.'s by monotone convergence, and finally to integrable r.v.'s by decomposing $X = X^+ - X^-$.

Let $Y$ be a discrete random variable, taking its values in a countable set $\{y_n : n \geq 1\}$ and assume that $\mathbb{P}(Y = y_n) \in (0, 1)$ for all $n \geq 1$. Then we can partition $\Omega$ into the disjoint subsets $\{Y = y_n\}$ for $n \geq 1$. For each $n \geq 1$, one can define $\mathbb{E}[\cdot \mid Y = y_n]$ as above.

**Lemma 6.3.1.** *For any $X$ either nonnegative or integrable, we have a.s.*

$$\mathbb{E}[X \mid Y] = \Psi(Y) \qquad \text{where for each } n \geq 1, \qquad \Psi(y_n) = \mathbb{E}[X \mid Y = y_n].$$

*Proof.* Define the function $\Psi$ as the right-hand side, which is nonnegative if $X$ is and notice that:

$$|\Psi(y_n)| = \Big|\frac{\mathbb{E}[X\, \mathbb{1}_{Y=y_n}]}{\mathbb{P}(Y = y_n)}\Big| \leq \frac{\mathbb{E}[|X|\, \mathbb{1}_{Y=y_n}]}{\mathbb{P}(Y = y_n)},$$

hence if $X$ is integrable, then

$$\mathbb{E}[|\Psi(Y)|] = \sum_{n\geq 1} |\Psi(y_n)|\, \mathbb{P}(Y = y_n) \leq \sum_{n\geq 1} \mathbb{E}[|X|\, \mathbb{1}_{Y=y_n}] = \mathbb{E}[|X|] < \infty.$$

Next take any measurable function $h$ either nonnegative if $X$ is or bounded if $X$ is integrable, then similarly,

$$\begin{aligned}
\mathbb{E}[\Psi(Y)h(Y)] &= \sum_{n\geq 1} \Psi(y_n)h(y_n)\, \mathbb{P}(Y = y_n) \\
&= \sum_{n\geq 1} \mathbb{E}[X\, \mathbb{1}_{Y=y_n}]h(y_n) \\
&= \mathbb{E}\Big[\sum_{n\geq 1} Xh(y_n)\, \mathbb{1}_{Y=y_n}\Big] \\
&= \mathbb{E}[Xh(Y)].
\end{aligned}$$

The claim then follows by uniqueness in Theorem 6.2.1. □

If $Y$ only takes values in $\{y_n : n \geq 1\}$ a.s. then for definiteness we set $\Psi(y) = 0$ or any other arbitrary value for all $y \notin \{y_n : \mathbb{P}(Y = y_n) > 0\}$.

### 6.3.2 Density case

Suppose that $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ are such that the pair $(X, Y)$ has a density $f_{(X,Y)}$ with respect to the Lebesgue measure in the sense that for any measurable and nonnegative function $g : \mathbb{R}^{n+m} \to \mathbb{R}$,

$$\mathbb{E}[g(X, Y)] = \int_{\mathbb{R}^n \times \mathbb{R}^m} g(x, y) f_{(X,Y)}(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

Then in particular for $g : \mathbb{R}^m \to \mathbb{R}$ measurable and nonnegative, by Fubini's Theorem,

$$\mathbb{E}[g(Y)] = \int_{\mathbb{R}^m} g(y) f_{(X,Y)}(x, y)\, \mathrm{d}x\, \mathrm{d}y = \int_{\mathbb{R}^m} g(y) \Big( \int_{\mathbb{R}^n} f_{(X,Y)}(x, y)\, \mathrm{d}x \Big)\, \mathrm{d}y,$$

so $f_Y : y \mapsto \int_{\mathbb{R}^n} f_{(X,Y)}(x,y)\,\mathrm{d}x$ is a density for $Y$. Note that if $f_Y(y) = 0$, then $f_{(X,Y)}(x,y) = 0$ for almost all $x$, thus, if $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ is another measurable and nonnegative function, then

$$\begin{aligned}\mathbb{E}[h(X)g(Y)] &= \int_{\mathbb{R}^n\times\mathbb{R}^m} h(x)g(y)f_{(X,Y)}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^m} g(y)\Big(\int_{\mathbb{R}^n} h(x)f_{(X,Y)}(x,y)\,\mathrm{d}x\,\mathbb{1}_{f_Y(y)\neq 0}\Big)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^m} g(y)\Big(\int_{\mathbb{R}^n} h(x)\frac{f_{(X,Y)}(x,y)}{f_Y(y)}\,\mathrm{d}x\,\mathbb{1}_{f_Y(y)\neq 0}\Big)f_Y(y)\,\mathrm{d}y.\end{aligned}$$

Let us therefore set

$$\Psi(y) = \int_{\mathbb{R}^n} h(x)\frac{f_{(X,Y)}(x,y)}{f_Y(y)}\,\mathbb{1}_{f_Y(y)\neq 0}\,\mathrm{d}x,$$

then we see that

$$\mathbb{E}[h(X)g(Y)] = \int_{\mathbb{R}^m} g(y)\Psi(y)f_Y(y)\,\mathrm{d}y = \mathbb{E}[\Psi(Y)g(y)],$$

hence $\Psi(Y)$ is a version of the conditional expectation of $h(X)$ given $Y$.

The function defined for any $y \in \mathbb{R}^m$ fixed by

$$f_{X|Y=y} : x \mapsto \frac{f_{(X,Y)}(x,y)}{f_Y(y)}\,\mathbb{1}_{f_Y(y)\neq 0}$$

is called the *conditional density* of $X$ given $Y = y$. The function $\Psi$ is often denoted by

$$\mathbb{E}[h(X) \mid Y = y] = \Psi(y).$$

This allows to write, analogously to the discrete case,

$$\mathbb{E}[h(X) \mid Y] = \Psi(Y) \qquad \text{a.s. where for } y \in \mathbb{R}^m, \qquad \Psi(y) = \int_{\mathbb{R}^n} h(x)f_{X|Y=y}\,\mathrm{d}x.$$

Beware this is just a notation since $\mathbb{P}(Y = y) = 0$ for any given $y$!

## 6.4 Similarities with the usual expectation

Let us start with some easy (but used all the times) properties. Some of them have been partly proved during the course of the proof of Theorem 6.2.1.

**Lemma 6.4.1.** *The conditional expectation $\mathbb{E}[\,\cdot \mid Y]$ enjoys the following properties. Assume that either $X, X' \in L^1$ or $X, X' \geq 0$ a.s.*

*(i)* $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$. *(Very useful!)*

*(ii) Positivity: If $X \geq 0$ then $\mathbb{E}[X \mid Y] \geq 0$ a.s. and moreover if $\mathbb{E}[X \mid Y] = 0$ a.s. then $X = 0$ a.s.*

*(iii) Linearity: $\mathbb{E}[aX + bX' \mid Y] = a\,\mathbb{E}[X \mid Y] + b\,\mathbb{E}[X' \mid Y]$ a.s. for all $a, b \in \mathbb{R}$ if $X, X' \in L^1$ and $a, b \geq 0$ if $X, X' \geq 0$.*

*(iv) Monotonicity: If $X \leq X'$ a.s. then $\mathbb{E}[X \mid Y] \leq \mathbb{E}[X' \mid Y]$ a.s.*

*(v) If $X = f(Y)$, then $\mathbb{E}[X \mid Y] = X$ a.s. This holds in particular for constants.*

*(vi) $|\mathbb{E}[X \mid Y]| \leq \mathbb{E}[|X| \mid Y]$ a.s. Consequently $\mathbb{E}[|\mathbb{E}[X \mid Y]|] \leq \mathbb{E}[|X|]$ a.s.*

*Proof.* It mostly is a matter of checking Property (ii) in Theorem 6.2.1 and using uniqueness.

(i) Take $h(Y) = 1$ in (ii) of Theorem 6.2.1.

(ii) We already proved that if $X \geq 0$ then $\mathbb{E}[X \mid Y] \geq 0$ a.s. Suppose now that $\mathbb{E}[X \mid Y] = 0$ a.s. Then by the first point $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = 0$ so $X = 0$ a.s.

(iii) It is clear that $a\,\mathbb{E}[X \mid Y] + b\,\mathbb{E}[X' \mid Y]$ is integrable, and by linearity of the usual (!) expectation, for any measurable and bounded function $h$,

$$\begin{aligned}\mathbb{E}[(a\,\mathbb{E}[X \mid Y] + b\,\mathbb{E}[X' \mid Y])h(Y)] &= a\,\mathbb{E}[\mathbb{E}[X \mid Y])h(Y)] + b\,\mathbb{E}[\mathbb{E}[X' \mid Y])h(Y)] \\ &= a\,\mathbb{E}[Xh(Y)] + b\,\mathbb{E}[X'h(Y)] \\ &= \mathbb{E}[(aX + bX')h(Y)].\end{aligned}$$

(iv) Monotonicity follows by linearity and positivity (write $X' = X + Z$ with $Z \geq 0$).

(v) $X$ verifies the two properties in Theorem 6.2.1.

(vi) Let $X = X^+ - X^-$ and $|X| = X^+ + X^-$, then by linearity, and since both $\mathbb{E}[X^\pm \mid Y] \geq 0$,

$$|\mathbb{E}[X \mid Y]| = |\mathbb{E}[X^+ \mid Y] - \mathbb{E}[X^- \mid Y]| \leq \mathbb{E}[X^+ \mid Y] + \mathbb{E}[X^- \mid Y] = \mathbb{E}[|X| \mid Y],$$

which proves the claim. □

**Exercise 6.4.2.** Suppose $\mathbb{E}[X^2] < \infty$ and define the *conditional variance* by:

$$\mathrm{Var}(X \mid Y) := \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y] = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2.$$

Show the identity:

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

Finally the conditional expectation also satisfies the same convergence theorems (monotone, Fatou, dominated) and inequalities (Jensen, Hölder) as the usual expectation.

**Lemma 6.4.3.** *The conditional expectation enjoys the following properties.*

*(i) If $0 \leq X_n \leq X_{n+1}$ a.s. then $\mathbb{E}[\uparrow\lim_n X_n \mid Y] = \uparrow\lim_n \mathbb{E}[X_n \mid Y]$ a.s.*

*(ii) If $X_n \geq 0$ a.s. for all $n$, then $\mathbb{E}[\liminf_n X_n \mid Y] \leq \liminf_n \mathbb{E}[X_n \mid Y]$ a.s.*

*(iii) If $X_n \longrightarrow X$ a.s. and there exists $Z \in L^1$ such that $|X_n| \leq Z$ for all $n$, then $X \in L^1$ and $\mathbb{E}[X_n \mid Y] \longrightarrow \mathbb{E}[X \mid Y]$ a.s. and in $L^1$.*

*(iv) Let $\phi$ be a convex function from an open interval $I$ to $\mathbb{R}$ and let $X \in L^1$ be a random variable such that $X \in I$ a.s. Then $\mathbb{E}[\phi(X) \mid Y] \geq \phi(\mathbb{E}[X \mid Y])$ a.s.*

*(v) $\|\mathbb{E}[X \mid Y]\|_p \leq \|X\|_p$ for any $p \geq 1$.*

*(vi) If $p, q > 1$ satisfy $1/p + 1/q = 1$, then $\mathbb{E}[|X_1X_2| \mid Y] \leq \mathbb{E}[|X_1|^p \mid Y]^{1/p}\,\mathbb{E}[|X_2|^q \mid Y]^{1/q}$.*

*Proof.* (i) This was somehow proved in the proof of Theorem 6.2.1. By monotonicity, the sequence $\Psi_n(Y) = \mathbb{E}[X_n \mid Y]$ is a.s. nondecreasing so we can define a.s. $0 \leq \Psi(Y) = \uparrow\lim_n \mathbb{E}[X_n \mid Y]$. Let $h \geq 0$, then according to Theorem 6.2.1 in the nonnegative case, we have:

$$\mathbb{E}[\Psi_n(Y)h(Y)] = \mathbb{E}[X_nh(Y)]$$

for every $n \geq 0$. We then infer from the usual monotone convergence applied to both sides that

$$\mathbb{E}[\uparrow\lim_{n\to\infty} \Psi_n(Y)h(Y)] = \uparrow\lim_{n\to\infty} \mathbb{E}[\Psi_n(Y)h(Y)] = \uparrow\lim_{n\to\infty} \mathbb{E}[X_nh(Y)] = \mathbb{E}[\uparrow\lim_{n\to\infty} X_nh(Y)].$$

This characterises $\mathbb{E}[\uparrow\lim_n X_n \mid Y]$ as $\uparrow\lim_n \mathbb{E}[X_n \mid Y]$.

(ii) Apply the previous point to the nondecreasing sequence $(\inf_{k\geq n} X_k)_n$ to get

$$\mathbb{E}[\liminf_{n\to\infty} X_n \mid Y] = \mathbb{E}[\uparrow \lim_{n\to\infty} \inf_{k\geq n} X_k \mid Y] = \uparrow \lim_{n\to\infty} \mathbb{E}[\inf_{k\geq n} X_k \mid Y] \qquad \text{a.s.}$$

Now for every $n$, we have by monotonicity, $\mathbb{E}[\inf_{k\geq n} X_k \mid Y] \leq \inf_{k\geq n} \mathbb{E}[X_k \mid Y]$ and the claim follows.

(iii) Apply the previous point to $Z + X_n \geq 0$ to get

$$\begin{aligned}\mathbb{E}[Z \mid Y] + \mathbb{E}[X \mid Y] &= \mathbb{E}[Z + X \mid Y] \\ &= \mathbb{E}[\liminf_{n\to\infty}(Z + X_n) \mid Y] \\ &\leq \liminf_{n\to\infty} \mathbb{E}[Z + X_n \mid Y] \\ &\leq \mathbb{E}[Z \mid Y] + \liminf_{n\to\infty} \mathbb{E}[X_n \mid Y]\end{aligned}$$

a.s. and similarly, with $Z - X_n \geq 0$ instead,

$$\begin{aligned}\mathbb{E}[Z \mid Y] - \mathbb{E}[X \mid Y] &= \mathbb{E}[Z - X \mid Y] \\ &= \mathbb{E}[\liminf_{n\to\infty}(Z - X_n) \mid Y] \\ &\leq \liminf_{n\to\infty} \mathbb{E}[Z - X_n \mid Y] \\ &\leq \mathbb{E}[Z \mid Y] - \limsup_{n\to\infty} \mathbb{E}[X_n \mid Y]\end{aligned}$$

a.s. Recall that $Z \in L^1$ so $\mathbb{E}[Z \mid Y] \in L^1$ and thus is a.s. finite, then by subtracting this term, we infer that

$$\mathbb{E}[X \mid Y] \leq \liminf_{n\to\infty} \mathbb{E}[X_n \mid Y] \leq \limsup_{n\to\infty} \mathbb{E}[X_n \mid Y] \leq \mathbb{E}[X \mid Y],$$

a.s. hence $\mathbb{E}[X_n \mid Y] \longrightarrow \mathbb{E}[X \mid Y]$ a.s.

Moreover, $|\mathbb{E}[X_n \mid Y]| \leq \mathbb{E}[|X_n| \mid Y] \leq \mathbb{E}[|Z| \mid Y] \in L^1$ so by the usual dominated convergence theorem, $\mathbb{E}[X_n \mid Y] \longrightarrow \mathbb{E}[X \mid Y]$ in $L^1$.

(iv) Let us recall that $\phi$ being convex, if we set $A_\phi = \{(a, b) \in \mathbb{R}^2 : \ ax + b \leq \phi(x) \text{ for all } x \in I\}$, then for any $x \in \mathbb{R}$, we have

$$\phi(x) = \sup\{ax + b : \ (a, b) \in A_\phi\} = \sup\{ax + b : \ (a, b) \in A_\phi \cap \mathbb{Q}^2\}.$$

For any $(a, b) \in A_\phi \cap \mathbb{Q}^2$ we have a.s.

$$\mathbb{E}[\phi(X) \mid Y] \geq \mathbb{E}[aX + b \mid Y] = a\,\mathbb{E}[X \mid Y] + b.$$

Since $A_\phi \cap \mathbb{Q}^2$ is countable, this property actually holds a.s. simultaneously for all pairs $(a, b) \in A_\phi \cap \mathbb{Q}^2$ so we can take the supremum and conclude that a.s.

$$\mathbb{E}[\phi(X) \mid Y] \geq \sup\{a\,\mathbb{E}[X \mid Y] + b : \ (a, b) \in A_\phi \cap \mathbb{Q}^2\} = \phi(\mathbb{E}[X \mid Y]).$$

(v) By convexity of $|\cdot|^p$ we infer from the previous point that $\mathbb{E}[|X|^p \mid Y] \geq |\mathbb{E}[X \mid Y]|^p$ a.s. By further taking the expectation we find $\mathbb{E}[|X|^p] \geq \mathbb{E}[|\mathbb{E}[X \mid Y]|^p]$.

(vi) Recall from the proof of Hölder's inequality (Theorem 2.2.4) the a.s. inequality (Young): for any integrable random variables $U$ and $V$,

$$|UV| \leq \frac{|U|^p}{p} + \frac{|V|^q}{q} \qquad \text{so} \qquad \mathbb{E}[|UV| \mid Y] \leq \frac{1}{p}\,\mathbb{E}[|U|^p \mid Y] + \frac{1}{q}\,\mathbb{E}[|V|^q \mid Y].$$

For $U = X_1/\mathbb{E}[|X_1|^p \mid Y]^{1/p}$ and $V = X_2/\mathbb{E}[|X_2|^q \mid Y]^{1/q}$, we read

$$\mathbb{E}\Big[\frac{|X_1X_2|}{\mathbb{E}[|X_1|^p \mid Y]^{1/p}\,\mathbb{E}[|X_2|^q \mid Y]^{1/q}} \,\Big|\, Y\Big] \le \frac{1}{p}\,\mathbb{E}\Big[\frac{|X_1|^p}{\mathbb{E}[|X_1|^p \mid Y]} \,\Big|\, Y\Big] + \frac{1}{q}\,\mathbb{E}\Big[\frac{|X_2|^q}{\mathbb{E}[|X_2|^q \mid Y]} \,\Big|\, Y\Big]$$

a.s. We then infer from Lemma 6.5.2 below that a.s.

$$\frac{\mathbb{E}[|X_1X_2| \mid Y]}{\mathbb{E}[|X_1|^p \mid Y]^{1/p}\,\mathbb{E}[|X_2|^q \mid Y]^{1/q}} \le \frac{1}{p} + \frac{1}{q} = 1,$$

and the claim follows by rearranging the terms. □

## 6.5 Properties of the conditional expectation

Let us see in this section some specific properties of the conditional expectation which are very useful.

### 6.5.1 Two key tools

The first property says that projecting on a subspace and then on a subsubsapce in two steps amounts to directly project on the smallest one. In terms of quantity of information, it also means that restricting further the information amounts to directly take the least amount of information.

**Lemma 6.5.1** (Tower property)**.** *For any nonnegative or integrable random variable $X$, it holds a.s.*

$$\mathbb{E}\big[\mathbb{E}[X \mid Y_1] \mid Y_1, Y_2\big] = \mathbb{E}[X \mid Y_1] = \mathbb{E}\big[\mathbb{E}[X \mid Y_1, Y_2] \mid Y_1\big].$$

*Proof.* The first equality follows from Lemma 6.4.1 since $\mathbb{E}[X \mid Y_1]$ is a measurable function of $Y_1$ and thus of the pair $(Y_1, Y_2)$, a function which only depends on the first coordinate. For the second equality, similarly $\mathbb{E}[X \mid Y_1, Y_2]$ takes the form $\Psi(Y_1, Y_2)$, and for $h$ either nonnegative or bounded, we have that $h(Y_1)$ is a function of $(Y_1, Y_2)$ and thus, using Property (ii) in Theorem 6.2.1 twice,

$$\mathbb{E}\big[\mathbb{E}\big[\mathbb{E}[X \mid Y_1, Y_2] \mid Y_1\big]h(Y_1)\big] = \mathbb{E}\big[\mathbb{E}[X \mid Y_1, Y_2]h(Y_1)\big] = \mathbb{E}[Xh(Y_1)].$$

Therefore $\mathbb{E}[\mathbb{E}[X \mid Y_1, Y_2] \mid Y_1]$ is a version of $\mathbb{E}[X \mid Y_1]$. □

The second lemma extends the well-known property $\mathbb{E}[cX] = c\,\mathbb{E}[X]$ where $c$ is constant.

**Lemma 6.5.2** (Taking out what is known)**.** *Let $X$ and $f(Y)$ be two random variables such that either both are nonnegative or both $X \in L^1$ and $Xf(Y) \in L^1$. Then a.s.*

$$\mathbb{E}[Xf(Y) \mid Y] = \mathbb{E}[X \mid Y]f(Y).$$

*Proof.* Suppose that both $X, f(Y) \ge 0$ a.s. so $Xf(Y) \ge 0$ and then $\mathbb{E}[X \mid Y]f(Y) \ge 0$ a.s. Also $\mathbb{E}[X \mid Y]f(Y)$ is a measurable function of $Y$ so it remains to prove Property (ii) in Theorem 6.2.1. Fix $h \ge 0$ measurable, then $f(Y)h(Y) \ge 0$ is a measurable function of $Y$, so by this very property,

$$\mathbb{E}\big[\mathbb{E}[X \mid Y]f(Y) \times h(Y)\big] = \mathbb{E}\big[\mathbb{E}[X \mid Y] \times f(Y)h(Y)\big] = \mathbb{E}[X \times f(Y)h(Y)] = \mathbb{E}[Xf(Y) \times h(Y)].$$

Therefore $\mathbb{E}[X \mid Y]f(Y)$ is a version of $\mathbb{E}[Xf(Y) \mid Y]$.

In the case $X, Xf(Y) \in L^1$, we have $\mathbb{E}[X \mid Y]f(Y) \in L^1$ since, by Lemma 6.4.1,

$$\mathbb{E}[|\mathbb{E}[X \mid Y]f(Y)|] \le \mathbb{E}[\mathbb{E}[|X| \mid Y]\,|f(Y)|] = \mathbb{E}[|X|\,|f(Y)|] < \infty.$$

Let $h$ be bounded, the preceding argument fails here because $f(Y)h(Y)$ is not necessarily bounded so we cannot apply Property (ii) as directly. However, decomposing $X = X^+ - X^-$, $f(Y) = f(Y)^+ - f(Y)^-$, and $h(Y) = h(Y)^+ - h(Y)^-$, we can deduce the result from the preceding case. □

### 6.5.2 Conditioning and Independence

We have seen that $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$ is a constant when $X$ is independent from $Y$. The next result extends this identity by showing that adding irrelevant information does not change the prediction. From a geometric point of view, when projecting a vector on a subspace, one can forget any direction that is orthogonal to the original vector.

**Lemma 6.5.3.** *Let $X$ be either nonnegative or integrable, let $Y$ and $Z$ be random variables and assume that $Z$ is independent of the pair $(X, Y)$. Then a.s.*

$$\mathbb{E}[X \mid Y, Z] = \mathbb{E}[X \mid Y].$$

*Proof.* Let us suppose that $X \geq 0$ and that $\mathbb{E}[X] < \infty$. The random variable $Y$ takes value in some space $(E_1, \mathscr{E}_1)$ and $Z$ in $(E_2, \mathscr{E}_2)$. Fix two events $A \in \mathscr{E}_1$ and $B \in \mathscr{E}_2$, then by independence twice,

$$\mathbb{E}[\mathbb{E}[X \mid Y]\, \mathbb{1}_{Y\in A}\, \mathbb{1}_{Z\in B}] = \mathbb{E}[\mathbb{E}[X \mid Y]\, \mathbb{1}_{Y\in A}]\, \mathbb{E}[\mathbb{1}_{Z\in B}] = \mathbb{E}[X\, \mathbb{1}_{Y\in A}]\, \mathbb{E}[\mathbb{1}_{Z\in B}] = \mathbb{E}[X\, \mathbb{1}_{Y\in A}\, \mathbb{1}_{Z\in B}].$$

Define two measures on the product space $(E_1 \times E_2, \mathscr{E}_1 \otimes \mathscr{E}_2)$ by

$$\mu(C) = \mathbb{E}[\mathbb{E}[X \mid Y]\, \mathbb{1}_{(Y,Z)\in C}] \qquad \text{and} \qquad \nu(C) = \mathbb{E}[X\, \mathbb{1}_{(Y,Z)\in C}]$$

respectively. Then we have shown that they agree on the set $\pi = \{A \cap B : A \in \mathscr{E}_1, B \in \mathscr{E}_2\}$. This is a $\pi$-system and the measures have the same finite total mass $\mathbb{E}[X]$ so they agree on $\sigma(\pi) = \mathscr{E}_1 \otimes \mathscr{E}_2$ by Theorem 1.1.13. This proves:

$$\mathbb{E}[\mathbb{E}[X \mid Y]h(Y, Z)] = \mathbb{E}[Xh(Y, Z)]$$

for any function $h$ of the form $h(Y, Z) = \mathbb{1}_{(Y,Z)\in C}$ with $C \in \mathscr{E}_1 \otimes \mathscr{E}_2$. We then extend the identity to any measurable nonnegative or integrable functions by the usual approximation by simple functions and linearity of expectation, see Section 1.4 for details. If $X$ is integrable but can be negative, then apply the result to $X^+$ and $X^-$ and us linearity of the conditional expectation. If $X \geq 0$ but $\mathbb{E}[X] = \infty$, then apply this result to $\min(X, n)$ and use the conditional monotone convergence. In any case, we see that $\mathbb{E}[X \mid Y]$ satisfies the two properties that characterise $\mathbb{E}[X \mid Y, Z]$ in Theorem 6.2.1 and we conclude by uniqueness. □

Our last result is also very useful for calculations.

**Theorem 6.5.4.** *Let $X$ and $Y$ be two independent random variable, not necessarily real-valued, and let $g$ be a real-valued measurable function, either nonnegative or integrable. Then a.s.*

$$\mathbb{E}[g(X, Y) \mid Y] = \Psi_g(Y) \qquad \textit{where} \qquad \Psi_g(y) = \mathbb{E}[g(X, y)].$$

*Proof.* The random variable $\Psi_g(Y)$ is indeed $\sigma(Y)$-measurable and either nonnegative or integrable. Further, for any measurable function $h$, either nonnegative or bounded, we have by independence and then Fubini's theorem:

$$\begin{aligned}
\mathbb{E}[g(X, Y)h(Y)] &= \int g(x, y)h(y)\, \mathbb{P}_X(\mathrm{d}x)\, \mathbb{P}_Y(\mathrm{d}y) \\
&= \int \left( \int g(x, y)h(y)\, \mathbb{P}_X(\mathrm{d}x) \right) \mathbb{P}_Y(\mathrm{d}y) \\
&= \int \Psi_g(y)h(y)\, \mathbb{P}_Y(\mathrm{d}y) \\
&= \mathbb{E}[\Psi_g(Y)h(Y)],
\end{aligned}$$

so $\Psi_g(Y)$ is a version of $\mathbb{E}[g(X, Y) \mid Y]$. □

## 6.6 Gaussian vectors and linear regression (⋆)

Recall the notion of Gaussian vectors from Section 2.7. We already saw that they naturally appear in the CLT, and that they simplify the independence by making it equivalent to null covariance. They also simplify the conditional expectation and allow to explicit compute it.

Indeed, recall the linear regression from Corollary 6.1.4, which is the best approximation of a random variable $X$ amongst all affine combinations of given random variables $Y_1, \dots, Y_n$. It is in general less precise (in the $L^2$ norm) than the conditional expectation. On the other hand it is much simpler to compute. In the case of Gaussian vectors it turns out that the conditional expectation matches the linear regression, so we have the best approximation which is fairly simple to compute!

**Theorem 6.6.1.** *Let $(X, Y_1, \dots, Y_n)$ be a Gaussian vector in dimension $n+1$ with mean $0$ (which we can always assume by subtracting the mean). Then there exist real numbers $\alpha_1, \dots, \alpha_n$ such that*

$$\mathbb{E}[X \mid Y_1, \dots, Y_n] = \sum_{k=1}^{n} \alpha_k Y_k \qquad \text{a.s.}$$

*Moreover, let*

$$\widehat{X} = \sum_{k=1}^{n} \alpha_k Y_k \qquad \text{and} \qquad \sigma^2 = \mathbb{E}\big[(X - \widehat{X})^2\big],$$

*then for any measurable function $h$ either nonnegative or such that $h(X)$ is integrable, it holds:*

$$\mathbb{E}[h(X) \mid Y_1, \dots, Y_n] = \int_{\mathbb{R}} h(x) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\widehat{X})^2}{2\sigma^2}\Big)\, \mathrm{d}x \qquad \text{a.s.}$$

*Proof.* Note that all random variables belong to the space $L^2$, then let $\widehat{X} = \sum_{k=1}^{n} \alpha_k Y_k$ denote the orthogonal projection of $X$ onto the vector space spanned by $(1, Y_1, \dots, Y_n)$. By orthogonality, we have:

$$\mathrm{Cov}(X - \widehat{X}, Y_j) = \mathbb{E}[(X - \widehat{X}) Y_j] = 0,$$

for every $1 \le j \le n$. Note that the vector $(X - \widehat{X}, Y_1, \dots, Y_n)$ is a Gaussian vector since every linear combination of its coordinates is a linear combination of $(X, Y_1, \dots, Y_n)$. Then by Proposition 2.7.13 or rather its extension in Remark 2.7.14, we infer that $X - \widehat{X}$ is independent from $(Y_1, \dots, Y_n)$. Since $\widehat{X} \,\textcircled{m}\, \sigma(Y_1, \dots, Y_n)$ and all the random variables are centred, then a.s.

$$\mathbb{E}[X \mid Y_1, \dots, Y_n] = \underbrace{\mathbb{E}[X - \widehat{X} \mid Y_1, \dots, Y_n]}_{= \mathbb{E}[X - \widehat{X}] = 0} + \underbrace{\mathbb{E}[\widehat{X} \mid Y_1, \dots, Y_n]}_{= \widehat{X}},$$

which proves the first claim. For the second one, recall that $X - \widehat{X}$ is independent from $(Y_1, \dots, Y_n)$ and has a Gaussian law with mean 0, and variance $\sigma^2$. Then by Theorem 6.5.4, we have:

$$\mathbb{E}[h(X) \mid Y_1, \dots, Y_n] = \mathbb{E}\Big[h\Big(X - \widehat{X} + \sum_{k=1}^{p} \alpha_k y_k\Big) \,\Big|\, Y_1, \dots, Y_n\Big] = \Psi(Y_1, \dots, Y_n)$$

a.s. where $\Psi$ is defined as follows:

$$\Psi(y_1, \dots, y_n) = \int_{\mathbb{R}} h\Big(z + \sum_{k=1}^{p} \alpha_k y_k\Big) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{z^2}{2\sigma^2}\Big)\, \mathrm{d}z.$$

The claim then follows by a change of variables. □

## 6.7 Regular conditional probabilities (⋆)

Let us give here some pointers to a difficult question that we will not answer. Recall that given an event $B \in \mathcal{F}$ with nonzero probability, one can define the conditional probability given $B$ by $\mathbb{P}(\,\cdot \mid B) = \mathbb{P}(\,\cdot \cap B)/\,\mathbb{P}(B)$. This indeed defines a probability measure on $\mathcal{F}$. In the context of conditioning with respect to a random variable instead, we define conditional probabilities as follows.

**Definition 6.7.1.** For any event $A \in \mathcal{F}$, set

$$\mathbb{P}(A \mid Y) = \mathbb{E}[\mathbb{1}_A \mid Y].$$

Beware that it is a random variable, which takes the form $\Psi(Y)$ for some measurable function $Y$.

By linearity and conditional monotone convergence, given any sequence $(A_n)_{n\geq 1}$ of disjoint events, we have a.s.

$$\mathbb{P}\Big(\bigcup_{n\geq 1} A_n \,\Big|\, Y\Big) = \uparrow \lim_{N\to\infty} \sum_{n\leq N} \mathbb{P}(A_n \mid Y) = \sum_{n\geq 1} \mathbb{P}(A_n \mid Y).$$

We are thus tempted to believe that $\mathbb{P}(\,\cdot \mid Y)$ defines a.s. a random probability measure. However for this, the above display should hold a.s. simultaneously for all sequences of events $(A_n)_{n\geq 1}$ and in general there are uncountably many of such sequences. Thus, we may speak in general about the conditional expectation of an (integrable) random variable, but not about its conditional law. The notion we are looking for is the following.

**Definition 6.7.2.** Let $X$ be a random variable with values in a measurable space $(E, \mathcal{E})$ and let $Y$ be another random variable. A function $\nu : \Omega \times \mathcal{E} \longrightarrow [0,1]$ is called a *regular conditional law of $X$ given $Y$* when it satisfies:

(i) $\nu(\omega, \cdot)$ defines a probability measure on $(E, \mathcal{E})$ for $\mathbb{P}$-a.e. $\omega \in \Omega$,

(ii) $\nu(\cdot, B)$ is a version of $\mathbb{P}(X \in B \mid Y)$ for every $B \in \mathcal{E}$.

In particular, when $(E, \mathcal{E}) = (\Omega, \mathcal{F})$ and $X$ is the identity, then such a map $\nu$ is called a *regular conditional probability given $Y$*.

The usefulness of regular conditional laws is that they allow to extend the usual expectation in a very straightforward way. Let us illustrate this.

**Proposition 6.7.3.** *If $\nu$ is a regular conditional law of $X$ given $Y$, then for any measurable function $f$ either nonnegative or integrable, we have a.s.*

$$\mathbb{E}[f(X) \mid Y](\omega) = \int_{\mathbb{R}} f(x)\nu(\omega, \mathrm{d}x).$$

*Proof.* If $f$ is the indicator of a set $A \in \mathcal{F}$, then this reads a.s.

$$\mathbb{P}(X \in A \mid Y)(\omega) = \nu(\omega, A),$$

which is the definition of the regular conditional law. As usual, this extends to simple functions by linearity, and further to nonnegative functions by monotone convergence, and finally to integrable functions by decomposing $f = f^+ - f^-$. This also shows that $\omega \mapsto \int_{\mathbb{R}} f(x)\nu(\omega, \mathrm{d}x)$ is measurable. □

Regular conditional laws is the notion that is needed to consider Markov chains on a general space. They generalise in this context the transition matrices that we used in countable spaces. Such regular conditional laws do not always exist, but quite often in practice, and rather explicitly. Indeed, recall the conditional expectation with respect to a discrete random variable $Y$, then the map $\nu$ in Definition 6.7.2 is given by:

$$\nu(\omega, B) = \Phi(Y(\omega), B) \qquad \text{where} \qquad \Phi(y, B) = \frac{\mathbb{P}(X \in B,\, Y = y)}{\mathbb{P}(Y = y)}.$$

Similarly, when the pair $(X, Y)$ has a density $f_{(X,Y)}$ with respect to the Lebesgue measure, then one can define the conditional density $f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} \mathbb{1}_{f_Y(y)\neq 0}$ and then, with the previous notation, we have:

$$\nu(\omega, B) = \Phi(Y(\omega), B) \qquad \text{where} \qquad \Phi(y, B) = \int_B f_{X|Y=y}(x)\,\mathrm{d}x.$$

# Chapter 7

# Some generalities on stochastic processes

Recall that the term *stochastic process* simply refers to a sequence of random variables $X = (X_n)_{n\geq 0}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with values in the same measurable space $(E, \mathcal{E})$, but we shall think of such a sequence as describing the evolution of a single random variable as time passes. In this very short chapter, we introduce the notion of filtrations, which formalise the evolution of time, as well as that of stopping times, which are random times which do not provide information about the future. We also present a generalisation of the conditional expectation with respect not to a random variable, but rather a $\sigma$-algebra.

**Contents**



In Section 7.1 we mainly introduce basic definitions about stochastic processes, especially the notion of filtrations which formalise the accumulation of information as time goes by, and the notion of stopping times which is the "correct" notion of random times, which cannot see the future. These generalise the notions we used for discrete Markov chains and which will be used for martingales. In Section 7.2 we discuss more precisely the notion of a stochastic process seen up to a stopping time. Finally in Section 7.3 we present the conditional expectation with respect to a $\sigma$-algebra which will be used in the subsequent chapters.

**Notation.** In this chapter, all the random variables are real-valued and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. From now on, for two real numbers $s$ and $t$, we write:

$$s \wedge t = \min(s, t) \qquad \text{and} \qquad s \vee t = \max(s, t).$$

Also, in order to lighten the notation, we usually drop the "a.s" mention when considering relations between random variables.

## 7.1 Filtrations & Stopping times

Recall that a $\sigma$-algebra $\mathcal{F}$ on $\Omega$ is a collection of subsets of $\Omega$ that has the following three property:

$$\Omega \in \mathcal{F}, \qquad A \in \mathcal{F} \implies A^c \in \mathcal{F}, \qquad A_n \in \mathcal{F} \text{ for all } n \geq 1 \implies \bigcup_{n\geq 1} A_n \in \mathcal{F}.$$

From now on we will be working with several $\sigma$-algebras on $\Omega$.

**Definition 7.1.1.** We say that $\mathcal{G}$ is a sub-$\sigma$-algebra of $\mathcal{F}$, which we simply write as $\mathcal{G} \subset \mathcal{F}$, if it is a $\sigma$-algebra on $\Omega$ and if it is contained in $\mathcal{F}$ in that $A \in \mathcal{F}$ for every set $A \in \mathcal{G}$.

In words $\mathcal{F}$ describes all the possible events, and $\mathcal{G}$ is a sub-collection of events, which we can view as a quantity of information in the sense that the knowledge of $\mathcal{G}$ allows us to say wether any event $A \in \mathcal{G}$ occurs or not. The formal description of the time evolution is then given by a *filtration*.

**Definition 7.1.2.** A *filtration* on $\Omega$ is an nondecreasing sequence $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ of sub-$\sigma$-algebras. We also define $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n) \subset \mathcal{F}$. The quadruple $(\Omega, \mathcal{F}, (\mathcal{F}_n)_n, \mathbb{P})$ is called a *filtered space*. When needed, we agree that $\mathcal{F}_{-1} = \{\emptyset, \Omega\}$ is the trivial $\sigma$-algebra, with no information.

In the analogy with geometry used in the previous chapter, one can imagine $\mathcal{F}$ as an infinite dimensional space like the spaces $\ell^1$ or $\ell^2$ on real-valued sequences and each $\mathcal{F}_n$ as $\mathbb{R}^n$. From the point of view of sub-$\sigma$-algebras as partial information, the $\sigma$-algebra $\mathcal{F}_n$ represents all the information available at time $n$; note that we accumulate more and more without forgetting past information.

Recall next that a random variable $X$ with values in some space $(E, \mathcal{E})$ is a measurable function that is, a function $X : \Omega \longrightarrow E$ which satisfies:

$$\text{for any } B \in \mathcal{E}, \qquad \{X \in B\} \in \mathcal{F},$$

where $\{X \in B\}$ stands for the set $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. The question of measurability with respect to a sub-$\sigma$-algebra of $\mathcal{F}$ will play a crucial role.

**Definition 7.1.3.** We say that $X$ is measurable with respect to $\mathcal{G} \subset \mathcal{F}$, or is simply '$\mathcal{G}$-measurable', when for every $B \in \mathcal{E}$ we have $\{X \in B\} \in \mathcal{G}$.

In words $X$ is $\mathcal{G}$-measurable when the information contained in $\mathcal{G}$ characterises entirely $X$. From a geometric point of view, one can figure a vector belonging to a subspace.

**Notation.** *Personal notation, not standard outside this course*: $X \,\textcircled{m}\, \mathcal{G}$ to mean that $X$ is $\mathcal{G}$-measurable.

Given a random variable $X$, there usually exist many sub-$\sigma$-algebras $\mathcal{G} \subset \mathcal{F}$ such that $X \,\textcircled{m}\, \mathcal{G}$. Taking the intersection of them, we define $\sigma(X)$ the smallest sub-$\sigma$-algebra that makes $X$ measurable. More generally, one define $\sigma(X_1, \dots, X_n)$ as the smallest sub-$\sigma$-algebra that makes each $X_k$ measurable for $k \leq n$. This $\sigma$-algebra is said to be "generated by $X_1, \dots, X_n$".

The filtrations we shall encounter in this course will be of this form: we have a certain stochastic process $X = (X_n)_{n\geq 0}$ and we consider the so-called *natural filtration* given by:

$$\mathcal{F}_n^X = \sigma(X_k, k \leq n), \tag{7.1}$$

for every $n \geq 0$.

**Definition 7.1.4.** A stochastic process $(Y_n)_{n\geq 0}$ is said to be:

- *adapted* to the filtration $(\mathcal{F}_n)_{n\geq 0}$ when $Y_n \,\textcircled{m}\, \mathcal{F}_n$ for every $n \geq 0$.
- *predictable* for the filtration $(\mathcal{F}_n)_{n\geq 0}$ when $Y_n \,\textcircled{m}\, \mathcal{F}_{n-1}$ for every $n \geq 0$.

In words, a predictable process is a process in which the value at any given time is completely determined by the information at the previous step, we shall see it as a parameter that we can tune before the next step. On the contrary, the issue of an adapted process is not entirely determined at the previous step, and still remains random, and is only revealed at the next step.

**Remark 7.1.5.** If $\mathcal{F}_n = \mathcal{F}_n^X = \sigma(X_k, k \leq n)$ is the natural filtration of another process $(X_n)_n$, then $(Y_n)_{n\geq 0}$ is adapted when it takes the form $Y_n = g(X_0, \dots, X_n)$ for some measurable function $g$; it is predictable when $Y_n = g(X_0, \dots, X_{n-1})$.

Recall that we extensively considered Markov chains up to a finite random time. The general definition of a stopping time extends that used previously, which referred to the natural filtration of the process.

**Definition 7.1.6.** A *stopping time* relative to a filtration $(\mathcal{F}_n)_n$ is a random variable $T$ taking values in $\overline{\mathbb{Z}}_+ = \{0, 1, 2, \dots, \infty\}$ such that:

$$\{T = n\} \in \mathcal{F}_n \text{ for any } n, \qquad \text{or equivalently,} \qquad \{T \leq n\} \in \mathcal{F}_n \text{ for any } n.$$

In other words, $T$ is a stopping time when $(\mathbb{1}_{T=n})_{n\geq 0}$ is adapted, or equivalently when $(\mathbb{1}_{T\leq n})_{n\geq 0}$ is adapted.

The equivalence is easily checked by writing $\{T \leq n\} = \bigcup_{k\leq n}\{T = k\}$ for one implication and $\{T = n\} = \{T \leq n\} \setminus \{T \leq n-1\}$ for the other one. In words, a stopping time is a random time which is determined by the past: the information of the present is sufficient to tell wether it has already occurred or not yet. One can notice that constant random variables $T = k$ for any given $k \in \overline{\mathbb{Z}}_+$ are stopping times.

**Example 7.1.7.** Important stopping times are given by the first entry time of an adapted process: take $X = (X_n)_n$ adapted to $(\mathcal{F}_n)_n$ and fix any measurable set $A$, then

$$T = \inf\{n : X_n \in A\}$$

is a stopping time. Indeed,

$$\{T > n\} = \bigcap_{k=0}^{n}\{X_k \in A^c\} \in \mathcal{F}_n.$$

Throughout this course we assume $\inf \emptyset = \infty$. It is thus important that $T$ may take value $\infty$.

Let us observe that the definition of a stopping time also applies to $n = \infty$. Indeed, recall that we set $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$, thus if $T$ is a stopping time, then

$$\{T = \infty\} = \Big(\bigcup_{n\geq 0}\{T \leq n\}\Big)^c \in \mathcal{F}_\infty.$$

It is important to be able to deal with multiple stopping times and we encourage the reader to prove the following elementary results.

**Exercise 7.1.8.** Let $(T_k)_{k\geq 1}$ be stopping times, then $\sum_k T_k$, $\inf_k T_k$, $\sup_k T_k$, $\liminf_k T_k$, $\limsup_k T_k$ are all stopping times. In general, the difference is *not*, even in the case $T - 1$ where $T \geq 1$ a.s.

## 7.2 Stopped $\sigma$-algebras and stopped processes

Let $T$ be a stopping time relative to a filtration $(\mathcal{F}_n)_n$; the information available at this random time is encoded into the following collection of subsets:

$$\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T = n\} \in \mathcal{F}_n \text{ for all } n \geq 0\}. \tag{7.2}$$

The notation creates no conflict since if $T$ is constant equal to some fixed $k$, then for $n \neq k$, we have $A \cap \{T = n\} = \emptyset$ and for $n = k$, we have $A \cap \{T = k\} = A$, so $\mathcal{F}_T = \{A \in \mathcal{F} : A \in \mathcal{F}_k\} = \mathcal{F}_k$ in this case.

**Proposition 7.2.1.** *For any stopping time $T$, the collection of sets $\mathcal{F}_T$ is a sub-$\sigma$-algebra of $\mathcal{F}$. Moreover, it can be equivalently defined by*

$$\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T \leq n\} \in \mathcal{F}_n \text{ for all } n \geq 0\}.$$

*Finally, for any real-valued random variable $X$, we have $X$ $\textcircled{m}$ $\mathcal{F}_T$ if and only if $X\,\mathbb{1}_{T=n}$ $\textcircled{m}$ $\mathcal{F}_n$ for every $n \geq 0$. In this case, we have $X\,\mathbb{1}_{T=\infty}$ $\textcircled{m}$ $\mathcal{F}_\infty$.*

*Proof.* Clearly $\emptyset \cap \{T = n\} = \emptyset \in \mathcal{F}_n$ for any $n \geq 0$ so $\emptyset \in \mathcal{F}_T$. Also, if $A_k \in \mathcal{F}_T$ for all $k \geq 0$, that is if $A_k \cap \{T = n\} \in \mathcal{F}_n$ for any $n, k \geq 0$, then

$$\Big(\bigcup_k A_k\Big) \cap \{T = n\} = \Big(\bigcup_k A_k \cap \{T = n\}\Big) \in \mathcal{F}_n.$$

The (slightly) tricky part is to show that if $A \in \mathcal{F}_T$, then $A^c \in \mathcal{F}_T$. For this, observe that

$$A^c \cap \{T = n\} = (A \cup \{T \neq n\})^c = ((A \cap \{T = n\}) \cup \{T \neq n\})^c.$$

If $A \in \mathcal{F}_T$, then $(A \cap \{T = n\}) \in \mathcal{F}_n$, also $\{T \neq n\} \in \mathcal{F}_n$, hence $(A^c \cap \{T = n\}) \in \mathcal{F}_n$ for every $n$. This concludes the proof that $\mathcal{F}_T$ is a sub-$\sigma$-algebra of $\mathcal{F}$.

Next let us prove that $\mathcal{F}_T = \mathcal{G}_T$, which we define as:

$$\mathcal{G}_T = \{A \in \mathcal{F} : \ A \cap \{T \leq n\} \in \mathcal{F}_n \text{ for all } n \geq 0\}.$$

If $A \in \mathcal{F}_T$, then for every $n \geq k \geq 0$, we have $A \cap \{T = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ so

$$A \cap \{T \leq n\} = A \cap \Big(\bigcup_{k \leq n} \{T = k\}\Big) = \bigcup_{k \leq n} (A \cap \{T = k\}) \in \mathcal{F}_n.$$

Thus $A \in \mathcal{G}_T$ and we have shown that $\mathcal{F}_T \subset \mathcal{G}_T$. Conversely, let $A \in \mathcal{G}_T$, then

$$A \cap \{T = n\} = A \cap (\{T \leq n\} \cap \{T \leq n-1\}^c) = (A \cap \{T \leq n\}) \cap \{T \leq n-1\}^c \in \mathcal{F}_n.$$

Thus $\mathcal{G}_T \subset \mathcal{F}_T$ and the proof is complete.

Finally, let $X$ be a random variable and suppose first that $X \,\textcircled{m}\, \mathcal{F}_T$. Let us prove that $X \mathbb{1}_{T=n} \,\textcircled{m}\, \mathcal{F}_n$, that is $\{X \mathbb{1}_{T=n} \in B\} \in \mathcal{F}_n$ for any measurable set $B$. Indeed:

$$\{X \mathbb{1}_{T=n} \in B\} = (\{X \in B\} \cap \{T = n\}) \cup (\{0 \in B\} \cap \{T \neq n\}).$$

By definition, if $X \,\textcircled{m}\, \mathcal{F}_T$, then $\{X \in B\} \in \mathcal{F}_T$ and so $\{X \in B\} \cap \{T = n\} \in \mathcal{F}_n$. On the other hand we also have $\{T \neq n\} = \{T = n\}^c \in \mathcal{F}_n$ and $\{0 \in B\}$ has nothing to do with $\mathcal{F}_n$: it is either $\Omega$ of $\emptyset$ according as wether $0 \in B$ or not. Therefore $\{X \mathbb{1}_{T=n} \in B\} \in \mathcal{F}_n$ as we wanted and thus $X \mathbb{1}_{T=n} \,\textcircled{m}\, \mathcal{F}_n$.

Suppose conversely that $X \mathbb{1}_{T=n} \,\textcircled{m}\, \mathcal{F}_n$ for every $n$ and let us prove that $X \,\textcircled{m}\, \mathcal{F}_T$, that is $\{X \in B\} \in \mathcal{F}_T$ for any measurable set $B$. The latter is equivalent to $\{X \in B\} \cap \{T = n\} \in \mathcal{F}_n$ for any measurable set $B$ and any $n$ and we write now:

$$\{X \in B\} \cap \{T = n\} = \{X \mathbb{1}_{T=n} \in B\} \cap \{T = n\} \in \mathcal{F}_n$$

since each term on the right belongs to $\mathcal{F}_n$.

To conclude, notice that for every $n \geq 0$, we have $X \mathbb{1}_{T \leq n} = \sum_{k \leq n} X \mathbb{1}_{X=k} \,\textcircled{m}\, \mathcal{F}_n$ when $X \mathbb{1}_{X=k} \,\textcircled{m}\, \mathcal{F}_k$ for each $k$. Further, if $X \geq 0$, then $X \mathbb{1}_{T=\infty} = \uparrow\lim_n X \mathbb{1}_{T \leq n}$ which is then $\mathcal{F}_\infty$-measurable. In general, we may write $X = X^+ - X^-$ with $X^+ \geq 0$ and $X^- \geq 0$ to infer from the nonnegative case that $X \mathbb{1}_{T=\infty} \,\textcircled{m}\, \mathcal{F}_\infty$. □

Recall from the above exercise that the minimum of two stopping times is again a stopping time. Let us compare their associated sub-$\sigma$-algebras. Recall the notation $s \wedge t = \min(s, t)$.

**Lemma 7.2.2.** *Let $S$ and $T$ be two stopping times, then so is $S \wedge T$ and we have*

$$\mathcal{F}_S \cap \mathcal{F}_T = \mathcal{F}_{S \wedge T}.$$

*The latter contains the events $\{S \leq T\}$, $\{S \geq T\}$, and $\{S = T\}$.*

It follows that if we know that $S \leq T$, then $\mathcal{F}_S \subset \mathcal{F}_T$.

*Proof.* Let us prove both inclusions. Suppose first that $A \in \mathcal{F}$ satisfies for every $n$ both $A \cap \{S \le n\} \in \mathcal{F}_n$ and $A \cap \{T \le n\} \in \mathcal{F}_n$, then

$$A \cap \{S \wedge T \le n\} = A \cap (\{S \le n\} \cup \{T \le n\}) = (A \cap \{S \le n\}) \cup (A \cap \{T \le n\}) \in \mathcal{F}_n.$$

By Proposition 7.2.1, this shows that $\mathcal{F}_S \cap \mathcal{F}_T \subset \mathcal{F}_{S \wedge T}$.

Conversely, suppose that $A \in \mathcal{F}$ satisfies $A \cap \{S \wedge T \le n\} \in \mathcal{F}_n$ for every $n$, then

$$A \cap \{S \le n\} = (A \cap (\{S \le n\} \cup \{T \le n\})) \cap \{S \le n\} \in \mathcal{F}_n,$$

so $\mathcal{F}_{S \wedge T} \subset \mathcal{F}_S$. The same argument applies to $T$ and so $\mathcal{F}_{S \wedge T} \subset \mathcal{F}_S \cap \mathcal{F}_T$.

Finally, let us prove that $\{S \le T\} \in \mathcal{F}_S \cap \mathcal{F}_T$, the proof for $\{S \ge T\}$ is similar, and the case of $\{S = T\}$ follows by taking their intersection. On the one hand, for every $n \ge 0$, we have $\{S \le T\} \cap \{T = n\} = \{S \le n\} \cap \{T = n\} \in \mathcal{F}_n$ since both $\{S \le n\} \in \mathcal{F}_n$ and $\{T = n\} \in \mathcal{F}_n$, hence $\{S \le T\} \in \mathcal{F}_T$. On the other hand, $\{S \le T\} \cap \{S = n\} = \{T \ge n\} \cap \{S = n\} \in \mathcal{F}_n$ since $\{T \ge n\} = \{T \le n-1\}^c \in \mathcal{F}_{n-1} \subset \mathcal{F}_n$ and $\{S = n\} \in \mathcal{F}_n$, hance $\{S \le T\} \in \mathcal{F}_S$ as well. □

Recall that we motivated the notion of stopping time by the will to follow a process until such a time.

**Definition 7.2.3.** For a process $(X_n)_{n \ge 0}$ and a random time $T \in \overline{\mathbb{Z}}_+$, we define the stopped process $X^T = (X_n^T)_{n \ge 0}$ by:

$$X_n^T = X_{n \wedge T} = X_n \mathbb{1}_{n \le T} + X_T \mathbb{1}_{T < n}.$$

In words, the process $X^T$ simply follows the trajectory of $X$, but if we reach $T$ (on the event $\{T < \infty\}$, otherwise we simply continue forever), then it after this time it remains constant. We are next concerned with this terminal value.

**Lemma 7.2.4.** *Let $(X_n)_n$ be an adapted process and let $X_\infty$ be some random variable with $X_\infty$ (m) $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$. Let $T$ be a stopping time, then*

$$X_T = \sum_{n \ge 0} X_n \mathbb{1}_{T=n} + X_\infty \mathbb{1}_{T=\infty}$$

*is a random variable and $X_T$ (m) $\mathcal{F}_T$. Also, the stopped process $(X_n^T)_{n \ge 0}$ is adapted.*

*Proof.* The first claim follows from Proposition 7.2.1 since $X_T \mathbb{1}_{T=n} = X_n \mathbb{1}_{T=n}$ (m) $\mathcal{F}_n$. For the second claim, simply observe that for any measurable set and any $n \ge 0$, we have since $X_T$ (m) $\mathcal{F}_T$:

$$\{X_{n \wedge T} \in B\} = (\underbrace{\{X_n \in B\}}_{\in \mathcal{F}_n} \cap \underbrace{\{n < T\}}_{\in \mathcal{F}_n}) \cup \underbrace{(\{X_T \in B\} \cap \{T \le n\})}_{\in \mathcal{F}_n} \in \mathcal{F}_n$$

and the stopped process $(X_n^T)_{n \ge 0}$ is therefore adapted as we claimed. □

We stress that the notation $X_T$ does not make sense if $T$ can be infinite and $X_\infty$ is not defined! Usually when $X_n$ converges a.s. as $n \to \infty$, we let $X_\infty$ denote its limit, otherwise we may set $X_\infty = 0$ so $X_T = \sum_{n \ge 0} X_n \mathbb{1}_{T=n}$ is well-defined in any case.

## 7.3 Conditioning with respect to a $\sigma$-algebra

Recall the conditional expectation of a real-valued random variable $X$ given any random variable $Y$ defined in Theorem 6.2.1. The good notion of conditional expectation is actually with respect to the $\sigma$-algebra generated by $Y$ rather than the random variable itself, in the following sense.

**Lemma 7.3.1.** *If $Y$ and $Y'$ are two random variables such that $\sigma(Y) = \sigma(Y')$, then $\mathbb{E}[X \mid Y'] = \mathbb{E}[X \mid Y]$ a.s. for any real-valued random variable $X$, either nonnegative or integrable.*

*Proof.* It suffices to prove that $\mathbb{E}[X \mid Y']$ satisfies the two characteristic properties of $\mathbb{E}[X \mid Y]$ from Theorem 6.2.1. First, by this very theorem, we have $\mathbb{E}[X \mid Y'] = \Psi(Y')$ for some measurable function $\Psi$, either nonnegative or integrable. By Lemma 1.2.8 we infer that $\mathbb{E}[X \mid Y']$ is $\sigma(Y')$-measurable, and since $\sigma(Y) = \sigma(Y')$, then by this lemma again, there exists a measurable function $\Phi$ such that $\mathbb{E}[X \mid Y'] = \Phi(Y)$, which proves Property (i) of the theorem. Similarly, for every $h$ measurable, the random variable $h(Y)$ also takes the form $h'(Y')$ for some measurable function $h'$ and Property (ii) follows. □

We can then extend the definition of conditional expectation with respect to any sub-$\sigma$-algebra.

**Definition 7.3.2.** Let $\mathcal{G} \subset \mathcal{F}$ be a sub-$\sigma$-algebra and let $X$ be a real-valued random variable such that either $X \in [0, \infty]$ a.s. or $\mathbb{E}[|X|] < \infty$. Then there exists a real-valued random variable $Z$ satisfying the following properties:

(i) $Z \,\textcircled{m}\, \mathcal{G}$ and either $Z \in [0, \infty]$ a.s. or $\mathbb{E}[|Z|] < \infty$ respectively,

(ii) For any random variable $W \,\textcircled{m}\, \mathcal{G}$ either nonnegative or bounded respectively, we have:

$$\mathbb{E}[XW] = \mathbb{E}[ZW].$$

Moreover, if $Z'$ is another such random variable, then $Z = Z'$ a.s.

The proof is exactly the same as for Theorem 6.2.1 which, by the previous lemma, considered in fact the special case $\mathcal{G} = \sigma(Y)$. Actually, this particular case is not restrictive in the sense that if $\mathcal{G} \subset \mathcal{F}$ is a sub-$\sigma$-algebra and if one considers the identity random variable $Y(\omega) = \omega$ but seen as a measurable function $Y : (\Omega, \mathcal{F}) \longrightarrow (\Omega, \mathcal{G})$, then we have

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X \mid Y]$$

a.s. by Lemma 1.2.8 again. All the properties of the conditional expectation from the previous chapter then extend readily to conditioning with respect to a sub-$\sigma$-algebra.

Let us consider the particular case when $\mathcal{G} = \mathcal{F}_T$ is the stopped $\sigma$-algebra.

**Proposition 7.3.3.** *Let $X$ be a random variable, either nonnegative or integrable and let $T$ be a stopping time, then*

$$\mathbb{E}[X \mid \mathcal{F}_T] = \sum_{n \geq 0} \mathbb{E}[X \mid \mathcal{F}_n] \, \mathbb{1}_{T=n} + \mathbb{E}[X \mid \mathcal{F}_\infty] \, \mathbb{1}_{T=\infty}.$$

*Proof.* Let $Z$ denote the right-hand side in the claim. We simply check that it satisfies the two characteristic properties of $\mathbb{E}[X \mid \mathcal{F}_T]$. Let us first consider the case when $X \geq 0$, so $Z \geq 0$. It follows from Lemma 7.2.4 that $Z \,\textcircled{m}\, \mathcal{F}_T$. Fix then $W \,\textcircled{m}\, \mathcal{F}_T$ a nonnegative random variable. Then by Proposition 7.2.1 we have $\mathbb{1}_{T=n} W \,\textcircled{m}\, \mathcal{F}_n$ for every $n \in \{0, 1, \dots\} \cup \{\infty\}$, hence

$$\begin{aligned}
\mathbb{E}[ZW] &= \sum_{n \geq 0} \mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{F}_n] \, \mathbb{1}_{T=n} W\Big] + \mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{F}_\infty] \, \mathbb{1}_{T=\infty} W\Big] \\
&= \sum_{n \geq 0} \mathbb{E}[X \, \mathbb{1}_{T=n} W] + \mathbb{E}[X \, \mathbb{1}_{T=\infty} W] \\
&= \mathbb{E}[XW].
\end{aligned}$$

This shows that indeed $Z = \mathbb{E}[X \mid \mathcal{F}_T]$ when $X \geq 0$. In the integrable case, let us write $X = X^+ - X^-$, where both $X^+ \geq 0$ and $X^- \geq 0$, so

$$\mathbb{E}[X^+ \mid \mathcal{F}_T] = \sum_{n \geq 0} \mathbb{E}[X^+ \mid \mathcal{F}_n] \, \mathbb{1}_{T=n} + \mathbb{E}[X^+ \mid \mathcal{F}_\infty] \, \mathbb{1}_{T=\infty},$$

and the same holds with $X^-$. Subtracting these two identities yields our claim. □

Recall the tower property from Lemma 6.5.1; combined with Lemma 7.2.2 on stopped filtrations, we can extend it as follows.

**Lemma 7.3.4.** *Let $S$ and $T$ be two stopping times and let $X$ be a random variable, either nonnegative or integrable. We have:*

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_S] \mid \mathcal{F}_T\big] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_T] \mid \mathcal{F}_S\big] = \mathbb{E}[X \mid \mathcal{F}_{S\wedge T}].$$

*Proof.* Let $A \in \mathcal{F}_{S\wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$, then using successively that $A \in \mathcal{F}_S$ and then $A \in \mathcal{F}_T$, we obtain using the characteristic property of conditional expectation:

$$\mathbb{E}\big[\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_T] \mid \mathcal{F}_S\big]\, \mathbb{1}_A\big] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_T]\, \mathbb{1}_A\big] = \mathbb{E}[X\, \mathbb{1}_A].$$

Hence, the random variable $Z = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_T] \mid \mathcal{F}_S\big]$ $\textcircled{m}$ $\mathcal{F}_{S\wedge T}$ satisfies

$$\mathbb{E}[X\,W] = \mathbb{E}[Z\,W]$$

for every random variable $W = \mathbb{1}_A$ with $A \in \mathcal{F}_{S\wedge T}$. We extend then this identity to general random variables $W$ $\textcircled{m}$ $\mathcal{F}_{S\wedge T}$ by the usual approximation of measurable functions by simple functions and linearity of expectation, see Section 1.4 for details. □

# Chapter 8

# Martingales & Stopping times

Martingales have been introduced as generalisations of sums of independent zero-mean random variables by only assuming that the increments have a null conditional expectation (in the sense of Chapter 6) given the past. They are often presented as modeling the evolution of the fortune of player who is betting on a fair game, and we shall also stick to this picture. It turns out that martingales form a very rich class of stochastic processes and their seemingly very simple definition will actually allow us to derive many strong results which make them a very important tool in modern probability and statistics. We focus in this first chapter on the stopping problem, that is evaluating a martingale at a random stopping time, which is particularly useful to study random walks and more generally Markov chains. We shall also discuss the optimal stopping problem as an application of this theory.

**Contents**



In Section 8.1 we first define (super- and sub-) martingales and study some natural transformations. A first simple but important result proved in Section 8.2 is that a (super- and sub-) martingale stopped at a random stopping time remains a (super- and sub-) martingale. With some extra argument, this enables us to compute the expectation of a martingale evaluated at a stopping time, which has important applications. Section 8.3 presents some decompositions of martingales, the first one which will be used in the subsequent chapter. Section 8.4 discusses the relation between martingales, Markov chains, and harmonic functions. Finally Section 8.5 considers the optimal stopping problem that is: how can you try to maximise your (mean) gain in a random game? The question a priori does not concern martingales, but its solution does (and the stopping theorem).

## 8.1 Martingales & first properties

Let us fix an underlying filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ and an adapted real-valued stochastic process $(M_n)_{n\geq 0}$ as in Definition 7.1.4. We often consider the natural filtration $\mathcal{F}_n = \sigma(M_k, k \leq n)$, but sometimes we may have more information encoded into $\mathcal{F}_n$. The situation is the following: at any time $n$ we have the information of the past up to time $n$, encoded in $\mathcal{F}_n$, and we try using this knowledge to predict $M_{n+1}$. Informally, our best guest is given by the conditional expectation:

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n],$$

which exists as soon as $\mathbb{E}[|M_{n+1}|] < \infty$. If $\mathcal{F}_n = \mathcal{F}_n^M = \sigma(M_0, \dots, M_n)$ is the natural filtration, then recall that:

$$\mathbb{E}[M_{n+1} \mid M_0, \dots, M_n] = \Psi(M_0, \dots, M_n)$$

for some measurable (deterministic) function $\Psi$.

**Definition 8.1.1.** A stochastic process $(M_n)_{n\geq 0}$ is said to be *integrable*, or more generally *in $L^p$* for some $p \geq 1$ when for each given $n$, we have $\mathbb{E}[|M_n|^p] < \infty$. We stress that no uniformity in $n$ is required and we may have $\mathbb{E}[|M_n|^p] \longrightarrow \infty$.

Here is the definition of a martingale.

**Definition 8.1.2.** An adapted and integrable stochastic process $(M_n)_n$ is called a *(sub/super-)martingale* when it satisfies the characteristic property: For every $n \geq 0$,

$$\begin{aligned}
&\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] \geq M_n \qquad \text{(submartingale)},\\
&\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] \leq M_n \qquad \text{(supermartingale)},\\
&\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n \qquad \text{(martingale)}.
\end{aligned}$$

Notice that $(M_n)_n$ is a (sub/super-)martingale if and only if $(M_n - M_0)_n$ is and $M_0 \in L^1$. We shall therefore often forget about the initial value $M_0$ and simply take equal to 0. Also let us already note that $(M_n)_n$ is submartingale if and only if $(-M_n)_n$ is supermartingale and that $(M_n)_n$ is martingale if and only if it is both a supermartingale and a submartingale. Hence, properties for one model are easily transferred to another model such as the next easy one.

**Lemma 8.1.3.** *If $(M_n)_n$ is a submartingale, then for every $n > m$, it holds:*

$$\mathbb{E}[M_n \mid \mathcal{F}_m] \geq M_m.$$

*The converse inequality holds for supermartingales and an equality for martingales.*

*Proof.* Recall the tower property in Lemma 6.5.1:

$$\mathbb{E}[M_n \mid \mathcal{F}_m] = \mathbb{E}[\mathbb{E}[\cdots \mathbb{E}[M_n \mid \mathcal{F}_{n-1}] \cdots \mid \mathcal{F}_{m+1}] \mid \mathcal{F}_m].$$

The claim then easily follows by induction. □

By taking the expectation, we deduce that the sequence $(\mathbb{E}[M_n])_n$ is monotone, namely, for a submartingale we have for every pair $n > m$,

$$\mathbb{E}[M_n] \geq \mathbb{E}[M_m] \geq \mathbb{E}[M_0].$$

The converse inequalities hold for supermartingales and equalities for martingales.

**Lemma 8.1.4.** *Let $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ be a convex function and let $(M_n)_n$ be an adapted process. Suppose that $(\phi(M_n))_n$ is integrable.*

*(i) If $(M_n)_n$ is a martingale, then $(\phi(M_n))_n$ is a submartingale.*

*(ii) If $(M_n)_n$ is a submartingale and $\phi$ is nondecreasing, then $(\phi(M_n))_n$ is a submartingale.*

*Proof.* A direct application of the conditional Jensen inequality (recall Lemma 6.4.3) yields:

$$\mathbb{E}[\phi(M_{n+1}) \mid \mathcal{F}_n] \geq \phi(\mathbb{E}[M_{n+1} \mid \mathcal{F}_n]).$$

If $(M_n)_n$ is a martingale, then the right-hand side is $\phi(M_n)$, whereas it is larger than or equal to $\phi(M_n)$ if both $(M_n)_n$ is a submartingale and $\phi$ is nondecreasing. □

**Example 8.1.5.** Here are some convex functions that often appear in martingale theory: $|x|$, $x^2$, $e^{cx}$ and $e^{-cx}$ with $c > 0$, as well as $\max(x, c)$ and $-\min(x, c)$ with $c \in \mathbb{R}$. Notice that for $c = 0$, the last two amount to $x^+$ and $x^-$ respectively.

We often interpret martingales as the fortune of a gambler as time passes: the game is fair for martingales, and (un)favourable for (super/sub)martingales. In the basic model the gambler bets 1 unit of money at every step, and at the $n$'th game they win (or loose) $M_n - M_{n-1}$. Now suppose they can choose to bet $H_n$ for the $n$'th game, then the algebraic gain becomes $H_n(M_n - M_{n-1})$. Of course, the choice of $H_n$ should not depend on the result of the $n$'th game: it has to be made before and therefore it can only depend on the previous history. Formally $(H_n)_n$ must be predictable in the sense of Definition 7.1.4.

One can now wonder: is there a possibility to turn a unfavourable game into a favourable game by betting appropriately? The answer is of course no for otherwise casinos would not exist. The proof is a simple exercise but this result will shortly have a very important application.

**Lemma 8.1.6** (You cannot trick the game)**.** *Let $M = (M_n)_{n\geq 0}$ be an adapted process and $H = (H_n)_{n\geq 1}$ be a predictable process. Define a new process $H \bullet M$ by $(H \bullet M)_0 = 0$ and for $n \geq 1$,*

$$(H \bullet M)_n = \sum_{k=1}^{n} H_k(M_k - M_{k-1})$$

*and suppose that it is integrable.*

*(i) If $M$ is a martingale, then so is $H \bullet M$.*

*(ii) If $M$ is a submartingale (resp. supermartingale), then so is $H \bullet M$ if in addition and $H_n \geq 0$ for all $n$.*

To ensure that $H \bullet M$ is integrable we typically assume either that $H$ is bounded or that both $M_n, H_n \in L^2$ for every $n$.

*Proof.* Note that $H \bullet M$ is adapted since for $k \leq n$, each $H_k, M_k \textcircled{m} \mathcal{F}_n$. Then, for every $n \geq 1$, let us write:

$$\mathbb{E}[(H \bullet M)_n \mid \mathcal{F}_{n-1}] = (H \bullet M)_{n-1} + \mathbb{E}[H_n(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}].$$

Recall that $M_{n-1}, H_n \textcircled{m} \mathcal{F}_{n-1}$, then by first taking out what is known (Lemma 6.5.2), we have:

$$\mathbb{E}[H_n(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}] = H_n\,\mathbb{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] = H_n(\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] - M_{n-1}).$$

For a martingale, the term in parenthesis vanishes; for a submartingale, it is nonnegative, so assuming that $H_n \geq 0$, the right-hand side is nonnegative; it is similarly nonpositive in the case of a supermartingale. □

**Remark 8.1.7.** Next semester you will study processes $(M_t)_{t\in[0,\infty)}$ that evolve in continuous time. The analogue of the transformation $(H \bullet M)_n = \sum_{k=1}^{n} H_k \Delta M_k$ becomes $(H \bullet M)_t = \int_0^t H_s dM_s$. This object, basically constructed by a limit of Riemann sums, is called the *stochastic integral* and is a fundamental object in the study of continuous-time stochastic processes.

## 8.2 The stopping theorem

Let us continue in the hope of winning at an unfavourable game. To complete our strategy, in addition to be able to freely choose the amount of money we bet, we can also decide when to leave the game. As for betting, the decision to leave at time $n$ must only depend on the information up to time $n$, and thus must be formally a *stopping time* in the sense of Definition 7.1.6. Recall also the stopped process from Definition 7.2.3. By betting one unit until we decide to stop, we obtain the following result.

**Lemma 8.2.1.** *Let $(M_n)_{n\geq 0}$ be a (sub/super)martingale and let $T$ be a stopping time. Then the stopped process $(M_{n\wedge T})_{n\geq 0}$ is a (sub/super)martingale.*

*Proof.* For every $n \geq 1$, we have $\{T \geq n\} = \{T \leq n-1\}^c \in \mathcal{F}_{n-1}$ so the process defined by $H_n = \mathbb{1}_{T \geq n}$ is predictable and obviously bounded and nonnegative. Notice that $M_{n \wedge T} = M_0 + (H \bullet M)_n$. The claim then follows from Lemma 8.1.6. □

Consequently, for any stopping time $T$ and any $n \geq 0$, we have $\mathbb{E}[M_{n\wedge T}] \geq \mathbb{E}[M_0]$ for a submartingale, the converse inequality for a supermartingale, and an equality for a martingale. Suppose that $T < \infty$ and recall the random variable $M_T$ from Lemma 7.2.4. Let us already note that in general these inequalities do not extend to $M_T$. As a concrete example, the simple random walk on $\mathbb{Z}$ is a martingale and we saw in Theorem 4.3.1 that it was recurrent, so $T = \inf\{n : M_n = -1\} < \infty$ a.s. and here $M_T = -1 \neq M_0 = 0$.

Now we would very much like to know wether $\mathbb{E}[M_T] \geq \mathbb{E}[M_0]$ for a submartingale. This is certainly the case if $T$ is bounded, i.e. there exists a deterministic integer $N$ such that $T \leq N$ a.s. since then $M_{n\wedge T} = M_T$ for all $n > N$. In this setting of bounded stopping times, we can be more precise and extend the identity $\mathbb{E}[M_n \mid \mathcal{F}_m] \geq M_m$ valid for all deterministic $n \geq m$ (recall Lemma 8.1.3).

**Theorem 8.2.2.** *Let $(M_n)_{n\geq 0}$ be a submartingale and let $S$ and $T$ be two bounded stopping times satisfying $S \leq T$. Then $M_S, M_T \in L^1$ and we have*

$$\mathbb{E}[M_T \mid \mathcal{F}_S] \geq M_S \qquad \text{and so} \qquad \mathbb{E}[M_T] \geq \mathbb{E}[M_S] \geq \mathbb{E}[M_0].$$

*The converse inequalities hold for supermartingales and equalities for martingales.*

*Proof.* Suppose that $S \leq T \leq N$ where $N$ is deterministic. Then $|M_T| = \sum_{n=0}^N |M_n| \mathbb{1}_{T=n} \leq \sup_{n\leq N} |M_n| \in L^1$ and similarly for $M_S$. Fix $A \in \mathcal{F}_S$ and define $H_n = \mathbb{1}_A \mathbb{1}_{S<n\leq T}$ for every $n \geq 1$. By Proposition 7.2.1, we have $A \cap \{S < n\} \in \mathcal{F}_{n-1}$ and since $\{T \geq n\} \in \mathcal{F}_{n-1}$ as well, then $(H_n)_n$ is predictable. It is obviously bounded, and thus we infer from Lemma 8.1.6 that $(H \bullet M)$ is again a submartingale, started at 0. In particular,

$$0 = \mathbb{E}[(H \bullet M)_0] \leq \mathbb{E}[(H \bullet M)_N] = \mathbb{E}\Big[\sum_{n=1}^N \mathbb{1}_A \mathbb{1}_{S<n\leq T}(M_n - M_{n-1})\Big] = \mathbb{E}[\mathbb{1}_A(M_T - M_S)].$$

We conclude that $\mathbb{E}[\mathbb{E}[M_T - M_S \mid \mathcal{F}_S]\, \mathbb{1}_A] \geq 0$ for every $A \in \mathcal{F}_S$; taking $A = \{\mathbb{E}[M_T - M_S \mid \mathcal{F}_S] < 0\}$, we obtain a nonpositive random variable with a nonnegative expectation, so it vanishes a.s. namely $\mathbb{E}[M_T - M_S \mid \mathcal{F}_S] \geq 0$. □

Let us mention a converse statement that provides a useful characterisation of (sub/super-)martingales.

**Corollary 8.2.3.** *Let $(M_n)_{n\geq 0}$ be an adapted and integrable process. Then it is a submartingale if and only if for every bounded stopping times $T \geq S$ we have $M_S, M_T \in L^1$ and*

$$\mathbb{E}[M_T] \geq \mathbb{E}[M_S].$$

*The same holds for supermartingales with the converse inequality and for martingales with an equality.*

**Remark 8.2.4.** In the case of martingales, since we have an equality, then an adapted and integrable process $(M_n)_n$ is a martingale if and only if for all bounded stopping times $T$, we have $M_T \in L^1$ and $\mathbb{E}[M_T] = \mathbb{E}[M_0]$. This is not true for submartingales (with $\geq$), for otherwise every nonnegative deterministic sequence would be nondecreasing!

*Proof.* The direct implication follows from the previous theorem; let us henceforth suppose that for every bounded stopping times $S \leq T$ we have $M_S, M_T \in L^1$ and $\mathbb{E}[M_T] \geq \mathbb{E}[M_S]$. Fix $n \geq 0$ and $A \in \mathcal{F}_n$ and define

$$T = (n+1)\, \mathbb{1}_{A^c} + n\, \mathbb{1}_A.$$

Then $T \leq n+1$ is a stopping time since $\{T = n\} = A \in \mathcal{F}_n$ and $\{T = n+1\} = A^c \in \mathcal{F}_n \subset \mathcal{F}_{n+1}$, and $\{T = k\} = \emptyset \in \mathcal{F}_k$ otherwise. Note that $M_T = M_{n+1}\, \mathbb{1}_{A^c} + M_n\, \mathbb{1}_A = M_n + (M_{n+1} - M_n)\, \mathbb{1}_{A^c}$. Since $n+1$ is also a bounded stopping time, then by our assumption:

$$\mathbb{E}[M_{n+1}] \geq \mathbb{E}[M_T] = \mathbb{E}[M_n] + \mathbb{E}[(M_{n+1} - M_n)\, \mathbb{1}_{A^c}].$$

We infer that

$$\mathbb{E}[(M_{n+1} - M_n)\,\mathbb{1}_A] \geq 0,$$

for all $A \in \mathcal{F}_n$, which, as in the previous proof, implies $\mathbb{E}[M_{n+1} - M_n \mid \mathcal{F}_n] \geq 0$. Since $n$ is arbitrary, then $(M_n)_n$ is a submartingale. □

Now what can be said for unbounded stopping times? Here are two useful cases in practice. Notice that the less restrictive assumptions we make on $T$, the more restrictive we make on $M$.

**Proposition 8.2.5.** *Let $(M_n)_{n\geq 0}$ be a submartingale and let $T$ be a stopping time. Suppose we are in one of the following two cases:*

*(i) either $\mathbb{E}[T] < \infty$ and $(M_{n\wedge T})_n$ has bounded increments, i.e. there exists a deterministic $C < \infty$ such that $|M_{n\wedge T} - M_{(n-1)\wedge T}| \leq C$ for all $n$ a.s.*

*(ii) or $T < \infty$ a.s. and $(M_{n\wedge T})_n$ is bounded, i.e. there exists a deterministic $C < \infty$ such that $|M_{n\wedge T}| \leq C$ for all $n$ a.s.*

*Then in both cases $M_T \in L^1$ and we have $\mathbb{E}[M_T] \geq \mathbb{E}[M_0]$. If $M$ is instead a supermartingale, then $\mathbb{E}[M_T] \leq \mathbb{E}[M_0]$, and finally if $M$ is a martingale, then $\mathbb{E}[M_T] = \mathbb{E}[M_0]$.*

*Proof.* By Lemma 8.2.1 we know that $\mathbb{E}[M_{n\wedge T}] \geq \mathbb{E}[M_0]$ for all $n \geq 0$. Moreover, if $T < \infty$ a.s. then $M_{n\wedge T} \longrightarrow M_T$ a.s. Then the second case follows from dominated convergence. As for the first claim, under the bounded increment assumption we have:

$$|M_{n\wedge T}| = \left|M_0 + \sum_{k=1}^{n\wedge T}(M_k - M_{k-1})\right| \leq |M_0| + \sum_{k=1}^{n\wedge T}|M_k - M_{k-1}| \leq |M_0| + C(n\wedge T) \leq |M_0| + CT.$$

If $T$ is integrable, then we can again apply the dominated convergence theorem. □

**Remark 8.2.6.**

- These cases are just a suggestion and in practice, one can safely apply Lemma 8.2.1 and try to pass to limit depending on the situation. We used here the dominated convergence theorem, but monotone convergence can be useful as well.
- This proposition is sometimes stated with the more restricted assumption that $M$ or its increments are bounded, not the stopped process. In practice it is very often the case, especially in the second one, that the whole process is unbounded, but the stopped process is.

## 8.3 Some decompositions (⋆)

Since gambling doesn't work, let us give another, more mathematical, motivation to consider such objects. The first result is sometimes called the Doob–Meyer decomposition. It shows that martingales and predictable processes naturally appear in random processes. It will play an important role in the next chapter. The other results of this section will not be used in the sequel.

**Lemma 8.3.1.** *Let $(X_n)_n$ be an integrable process adapted to a filtration $(\mathcal{F}_n)_n$. Then:*

*(i) There exist a predictable process $(A_n)_n$ and a martingale $(M_n)_n$, both for the filtration $(\mathcal{F}_n)_n$ and both null at 0, such that for every $n \geq 0$,*

$$X_n = X_0 + M_n + A_n.$$

*(ii) If $(A'_n)_n$ and $(M'_n)_n$ is another such pair, then $A'_n = A_n$ and $M'_n = M_n$ for all $n$.*

*(iii) Finally $(X_n)_n$ is a submartingale, resp. supermartingale, resp. martingale, if and only if $(A_n)_n$ is nondecreasing, resp. nonincreasing, resp. constant (null).*

*Proof.* Suppose first there exists such a decomposition. Since $A$ is predictable and $M$ a martingale, then we must have for all $n \geq 1$:

$$\mathbb{E}[X_n - X_{n-1} \mid \mathscr{F}_{n-1}] = A_n - A_{n-1}, \qquad \text{hence} \qquad A_n = \sum_{k=1}^{n} \mathbb{E}[X_k - X_{k-1} \mid \mathscr{F}_{k-1}].$$

A posteriori, this process is indeed predictable and thus is the only possible one null at 0. Next let us set $M_n = X_n - X_0 - A_n$, we obtain:

$$\mathbb{E}[M_n - M_{n-1} \mid \mathscr{F}_{n-1}] = \mathbb{E}[X_n - X_{n-1} - \mathbb{E}[X_n - X_{n-1} \mid \mathscr{F}_{n-1}] \mid \mathscr{F}_{n-1}] = 0,$$

so it is a martingale null at zero. The uniqueness of $M$ follows then from that of $A$. The last part of the statement is now clear. □

**Remark 8.3.2.** Recall that if $(M_n)_n$ is a martingale then $(M_n^2)_n$ is a submartingale provided integrability. The nondecreasing predictable process in the associated decomposition plays an important role, see Section 9.6.2.

Next, we are used to decompose a random variable as $X = X^+ - X^-$. Recall from Lemma 8.1.4 that if $(X_n)_n$ is martingale, then $(X_n^+)_n$ and $(X_n^-)_n$ are submartingales. The next result, sometimes called the Krickeberg decomposition, shows that we can decompose it as the difference of two *martingales* under an optimal assumption.

**Lemma 8.3.3.** *Let $(X_n)_n$ be a martingale. It satisfies $\sup_n \mathbb{E}[|X_n|] < \infty$ if and only if there exist two nonnegative martingales $(M_n)_n$ and $(N_n)_n$ such that:*

$$X_n = M_n - N_n.$$

*Moreover in this case, there exists a unique such decomposition which satisfies:*

$$\sup_{n \geq 0} \mathbb{E}[|X_n|] = \mathbb{E}[M_0] + \mathbb{E}[N_0],$$

*and $(M_n)_n$ and $(N_n)_n$ are the smallest nonnegative martingales which bound $(X_n)_n$ and $(-X_n)_n$ above respectively.*

*Proof.* Note that if $(M_n)_n$ and $(N_n)_n$ are nonnegative martingales, then indeed their difference remains a martingale, and moreover, for every $n \geq 0$, we have

$$\mathbb{E}[|M_n - N_n|] \leq \mathbb{E}[|M_n|] + \mathbb{E}[|N_n|] = \mathbb{E}[M_n] + \mathbb{E}[N_n] = \mathbb{E}[M_0] + \mathbb{E}[N_0],$$

so the left-hand side is bounded uniformly in $n$.

Conversely, suppose that $(X_n)_n$ is a martingale with $\sup_n \mathbb{E}[|X_n|] < \infty$. Fix $n \geq 0$ and define two sequences $(M_k^{(n)})_k$ and $(N_k^{(n)})_k$ by $M_k^{(n)} = \mathbb{E}[X_{n+k}^+ \mid \mathscr{F}_n]$ and $N_k^{(n)} = \mathbb{E}[X_{n+k}^- \mid \mathscr{F}_n]$ respectively, so by the tower property,

$$X_n = \mathbb{E}[X_{n+k} \mid \mathscr{F}_n] = M_k^{(n)} - N_k^{(n)},$$

for every $k \geq 0$. Let us focus on $(M_k^{(n)})_k$, as the other sequence satisfies similar properties. First, we claim that it is nondecreasing. Indeed by the tower property and convexity, since $(X_{n+k})_n$ is a martingale, then

$$M_{k+1}^{(n)} = \mathbb{E}[X_{n+k+1}^+ \mid \mathscr{F}_n] = \mathbb{E}[\mathbb{E}[X_{n+k+1}^+ \mid \mathscr{F}_{n+k}] \mid \mathscr{F}_n] \geq \mathbb{E}[X_{n+k}^+ \mid \mathscr{F}_n] = M_k^{(n)}.$$

Thus $(M_k^{(n)})_k$ converges a.s. to a limit, say $M_\infty^{(n)} \in [0, \infty]$. It actually is finite, and even integrable by Fatou's lemma:

$$\mathbb{E}[M_\infty^{(n)}] \leq \liminf_{k \to \infty} \mathbb{E}[M_k^{(n)}] \leq \sup_{k \geq 0} \mathbb{E}[M_k^{(n)}] = \sup_{k \geq 0} \mathbb{E}[X_{n+k}^+] \leq \sup_{k \geq 0} \mathbb{E}[|X_k|] < \infty.$$

We claim that letting $n$ vary, the integrable process $(M_\infty^{(n)})_n$ is a martingale. Indeed, by the tower property again, we have

$$\mathbb{E}[M_k^{(n+1)} \mid \mathscr{F}_n] = \mathbb{E}[\mathbb{E}[X_{n+1+k}^+ \mid \mathscr{F}_{n+1}] \mid \mathscr{F}_n] = \mathbb{E}[X_{n+1+k}^+ \mid \mathscr{F}_n] = M_{k+1}^{(n)}.$$

Letting $k \longrightarrow \infty$, we conclude by conditional monotone convergence (Lemma 6.4.3) that

$$\mathbb{E}[M_\infty^{(n+1)} \mid \mathscr{F}_n] = \mathbb{E}[\uparrow\lim_k M_k^{(n+1)} \mid \mathscr{F}_n] = \uparrow\lim_k \mathbb{E}[M_k^{(n+1)} \mid \mathscr{F}_n] = \uparrow\lim_k M_{k+1}^{(n)} = M_\infty^{(n)}.$$

The exact same argument applies to $N_k^{(n)} = \mathbb{E}[X_{n+k}^- \mid \mathscr{F}_n]$ and its nonnegative martingale limit $N_\infty^{(n)}$. We can thus let $k \longrightarrow \infty$ in our decomposition $X_n = M_k^{(n)} - N_k^{(n)}$ to obtain:

$$X_n = M_\infty^{(n)} - N_\infty^{(n)}.$$

Finally, for $k \geq 0$, we have:

$$M_k^{(0)} + N_k^{(0)} = \mathbb{E}[X_k^+ \mid \mathscr{F}_0] + \mathbb{E}[X_k^- \mid \mathscr{F}_0] = \mathbb{E}[|X_k| \mid \mathscr{F}_0].$$

Take the expectation of both sides and use monotone convergence to conclude that

$$\mathbb{E}[M_\infty^{(0)}] + \mathbb{E}[N_\infty^{(0)}] = \sup_n \mathbb{E}[|X_n|].$$

Suppose now that there is another such decomposition $X_n = M_n' - N_n'$ as differences of nonnegative martingales. Notice that $M_n' \geq X_n^+$ and $N_n' \geq X_n^-$ for every $n$, then by taking conditional expectations, we get for every $n \geq 0$:

$$M_n' = \mathbb{E}[M_{n+k}' \mid \mathscr{F}_n] \geq \mathbb{E}[X_{n+k}^+ \mid \mathscr{F}_n] = M_k^{(n)} \underset{k\longrightarrow\infty}{\longrightarrow} M_\infty^{(n)}.$$

Hence $M_n' \geq M_\infty^{(n)}$ and similarly $N_n' \geq N_\infty^{(n)}$ for every $n \geq 0$. If this new decomposition also satisfies $\sup_n \mathbb{E}[|X_n|] = \mathbb{E}[M_0'] + \mathbb{E}[N_0']$, then necessarily, $\mathbb{E}[M_n'] = \mathbb{E}[M_0'] = \mathbb{E}[M_\infty^{(0)}] = \mathbb{E}[M_\infty^{(n)}]$ and $\mathbb{E}[N_n'] = \mathbb{E}[N_0'] = \mathbb{E}[N_\infty^{(0)}] = \mathbb{E}[N_\infty^{(n)}]$ so combined with the previous bounds we get $M_n' = M_\infty^{(n)}$ and $N_n' = N_\infty^{(n)}$, hence the uniqueness of the decomposition.

Suppose $(Y_n)_n$ is a nonnegative martingale which satisfies $Y_n \geq X_n$. Then $Y_n \geq X_n^+$ and the previous argument shows that $Y_n \geq M_\infty^{(n)}$. Similarly, if instead $Y_n \leq -X_n$, then $Y_n \geq N_\infty^{(n)}$, hence the minimality property of $M_\infty^{(n)}$ and $N_\infty^{(n)}$. □

Here is a last result known as the Riesz decomposition. Recall that if $(X_n)_n$ is a submartingale then $(\mathbb{E}[X_n])_n$ is nondecreasing.

**Lemma 8.3.4.** *Suppose $(X_n)_n$ is a submartingale with $\sup_n \mathbb{E}[X_n] < \infty$. Then there exists a unique decomposition*

$$X_n = M_n - Y_n$$

*where $(M_n)_n$ is a martingale and $(Y_n)$ is a nonnegative supermartingale with $\mathbb{E}[Y_n] \longrightarrow 0$. The process $(M_n)_n$ is the smallest supermartingale bounded below by $(X_n)_n$.*

*Proof.* We proceed with ideas similar to those of the previous proof. Fix $n$ and let $M_k^{(n)} = \mathbb{E}[X_{n+k} \mid \mathscr{F}_n]$, then

$$X_n \leq \mathbb{E}[X_{n+k} \mid \mathscr{F}_n] \leq \mathbb{E}[\mathbb{E}[X_{n+k+1} \mid \mathscr{F}_{n+k}] \mid \mathscr{F}_n] = \mathbb{E}[X_{n+k+1} \mid \mathscr{F}_n].$$

Thus the sequence $(M_k^{(n)})_k$ is nondecreasing and lower bounded by $X_n$; we let $M_\infty^{(n)} \in [X_n, \infty]$ denote its limit as $k \longrightarrow \infty$. By monotone convergence, applied to the nonnegative and nondecreasing sequence $(M_k^{(n)} - X_n)_k$), and since $X_n$ is integrable, we obtain:

$$\mathbb{E}[M_\infty^{(n)}] = \mathbb{E}[X_n] + \mathbb{E}[M_\infty^{(n)} - X_n] = \mathbb{E}[X_n] + \uparrow\lim_{k\longrightarrow\infty} \mathbb{E}[M_k^{(n)} - X_n] = \uparrow\lim_{k\longrightarrow\infty} \mathbb{E}[M_k^{(n)}] = \uparrow\lim_{k\longrightarrow\infty} \mathbb{E}[X_k] < \infty.$$

Thus the sequence $(M_\infty^{(n)})_n$ is integrable. Next,

$$\mathbb{E}[M_k^{(n+1)} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X_{n+1+k} \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[X_{n+1+k} \mid \mathcal{F}_n] = M_{k+1}^{(n)}.$$

We then let $k \longrightarrow \infty$, using conditional monotone convergence on the left, to obtain that $(M_\infty^{(n)})_n$ is a martingale. Consequently, the difference $Y_n = M_\infty^{(n)} - X_n$ defines a supermartingale; it is nonnegative by construction and we have seen that $\mathbb{E}[M_\infty^{(n)}] = \lim_k \mathbb{E}[X_k]$, so $\mathbb{E}[Y_n] \longrightarrow 0$ as $n \longrightarrow \infty$. Now let $(Z_n)_n$ be a supermartingale with $Z_n \geq X_n$ for every $n$. Then

$$Z_n \geq \mathbb{E}[Z_{n+k} \mid \mathcal{F}_n] \geq \mathbb{E}[X_{n+k} \mid \mathcal{F}_n] = M_k^{(n)}.$$

Letting $k \longrightarrow \infty$, we obtain that $Z_n \geq M_\infty^{(n)}$ which is therefore the smallest such supermartingale.

Suppose finally that there exists another such decomposition as $X_n = M'_n + Y'_n = M_\infty^{(n)} + Y_n$. Then the process $Z_n = M'_n - M_\infty^{(n)} = Y'_n - Y_n$ is both a martingale as well as the difference of two nonnegative supermartingales whose expectation tends to 0. Since it is a martingale, then by convexity,

$$\mathbb{E}[|Z_n|] \leq \mathbb{E}[|Z_{n+k}|] \leq \mathbb{E}[Y_{n+k}] + \mathbb{E}[Y'_{n+k}] \underset{k\to\infty}{\longrightarrow} 0$$

hence $Z_n = 0$ for every $n$ and the decomposition is indeed unique. □

## 8.4 Martingales and Markov chains (⋆)

There is a deep connection between martingales and Markov chains, in relation also with harmonic functions discussed in Section 3.5. Let $(X_n)_n$ denote a Markov chain with values in a countable set $\mathbb{X}$, with transition matrix $P$. Recall that for every function $f : \mathbb{X} \longrightarrow \mathbb{R}$ for which the expectation is well-defined, we have for every $x \in \mathbb{X}$

$$P^n f(x) = \sum_{y\in\mathbb{X}} P^n(x,y) f(y) = \sum_{y\in\mathbb{X}} \mathbb{P}_x(X_n = y) f(y) = \mathbb{E}_x[f(X_n)].$$

We let $I$ denote the identity matrix on $\mathbb{X}$.

**Theorem 8.4.1.** *Let $(X_n)_n$ be a stochastic process with values in $\mathbb{X}$ and let $P$ be a transition matrix. Then $(X_n)_n$ is a $P$-Markov chain if and only if for every measurable and bounded function $f : \mathbb{X} \longrightarrow \mathbb{R}$, the process given by:*

$$M_n^f = f(X_n) - f(X_0) - \sum_{k=0}^{n-1} (P - I) f(X_k)$$

*is a martingale null at 0.*

*Proof.* Let $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$. The process $(M_n^f)_n$ is adapted and integrable for any bounded function $f$ since then $Pf$ and hence $(P - I)f$ are also bounded. Moreover we have

$$M_{n+1}^f - M_n^f = f(X_{n+1}) - f(X_n) - (P - I)f(X_n) = f(X_{n+1}) - Pf(X_n),$$

hence

$$\mathbb{E}[M_{n+1}^f \mid \mathcal{F}_n] = M_n^f \qquad \text{if and only if} \qquad \mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] = Pf(X_n),$$

and the right-hand sides holds if and only if $(X_n)_n$ is a $P$-Markov chain.

Indeed $(X_n)_n$ is a $P$-Markov chain when for every $x_0, \dots, x_{n+1}$ we have:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) = P(x_n, x_{n+1}).$$

If this holds, then for $f$ bounded, we have:

$$\begin{aligned}
\mathbb{E}[f(X_{n+1}) \mid X_0 = x_0, \dots, X_n = x_n] &= \sum_{x_{n+1}\in\mathbb{X}} f(x_{n+1})\, \mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) \\
&= \sum_{x_{n+1}\in\mathbb{X}} f(x_{n+1}) P(x_n, x_{n+1}) \\
&= Pf(x_n).
\end{aligned}$$

Recall that $\mathbb{E}[f(X_{n+1}) \mid X_0, \dots, X_n] = \Psi(X_0, \dots, X_n)$ where $\Psi(x_0, \dots, x_n) = \mathbb{E}[f(X_{n+1}) \mid X_0 = x_0, \dots, X_n = x_n]$, hence $\mathbb{E}[f(X_{n+1}) \mid X_0, \dots, X_n] = Pf(X_n)$ for a $P$-Markov chain. Conversely, this identity applied to $f(y) = \mathbb{1}_{y=x_{n+1}}$ shows that:

$$\begin{aligned}
\mathbb{P}(X_0 = x_0, \dots, X_{n+1} = x_{n+1}) &= \mathbb{E}\left[f(X_{n+1}) \prod_{k=0}^{n} \mathbb{1}_{X_k = x_k}\right] \\
&= \mathbb{E}\left[\mathbb{E}[f(X_{n+1}) \mid X_0, \dots, X_n] \prod_{k=0}^{n} \mathbb{1}_{X_k = x_k}\right] \\
&= \mathbb{E}\left[Pf(X_n) \prod_{k=0}^{n} \mathbb{1}_{X_k = x_k}\right] \\
&= Pf(x_n)\, \mathbb{P}(X_0 = x_0, \dots, X_n = x_n) \\
&= P(x_n, x_{n+1})\, \mathbb{P}(X_0 = x_0, \dots, X_n = x_n),
\end{aligned}$$

which shows that $(X_n)_n$ is a $P$-Markov chain. □

**Remark 8.4.2.** Recall that a function $h$ is $P$-harmonic when $Ph = h$. Thus, if $(X_n)_n$ is a $P$-Markov chain, then $(h(X_n))_n$ is a martingale if and only if $h$ is $P$-harmonic. More generally, it is a submartingale (resp. supermartingale) if and only if $h$ is subharmonic (resp. superharmonic).

**Remark 8.4.3.** If $(X_n)_n$ is a Markov chain, then the martingale $(M_n^f)_n$ is that in the Doob–Meyer decomposition of the process $(f(X_n))_n$ provided by Lemma 8.3.1. The sum $\sum_{k=0}^{n-1}(P - I)f(X_k)$ indeed corresponds to the predictable part.

We can also consider functions of both the position and the time.

**Theorem 8.4.4.** *Let $(X_n)_n$ be a $P$-Markov chain and let $f : \mathbb{Z}_+ \times \mathbb{X} \longrightarrow \mathbb{R}$. Let $M_n^f = f(n, X_n)$ and assume that $\mathbb{E}[|M_n^f|] < \infty$ and that*

$$Pf(n+1, x) = \sum_{y \in \mathbb{X}} P(x, y) f(n+1, y) = f(n, x).$$

*Then $(M_n^f)_n$ is a martingale.*

*Proof.* The process $(M_n^f)_n$ is adapted to $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ and we assume that it is integrable. Next, by the Markov property,

$$\mathbb{E}[f(n+1, X_{n+1}) \mid X_0 = x_0, \dots, X_n = x_n] = \mathbb{E}[f(n+1, X_{n+1}) \mid X_n = x_n] = Pf(n+1, x_n),$$

which equals $f(n, x_n)$ by our assumption. We conclude as in the previous proof. □

**Example 8.4.5.** Let $(S_n)_n$ denote the simple random walk on $\mathbb{Z}$, corresponding to $P(i, i-1) = P(i, i+1) = 1/2$. The harmonic functions $h = Ph$ are the solutions to:

$$h(j) = \frac{1}{2}(h(j+1) + h(j-1)), \quad \text{equivalently} \quad h(j+1) - h(j) = h(j) - h(j-1).$$

The increments are constant, so the solutions are easily found to be

$$h(k) = k(h(1) - h(0)) + h(0).$$

Then Theorem 8.4.1 shows that for any value $a = h(0)$ and $b = h(1)$, we have that

$$(b - a)S_n + a \quad \text{is a martingale.}$$

The functions $f$ that satisfy the identity in Theorem 8.4.4 are solution to:

$$f(n, i) = \frac{1}{2}(f(n+1, i-1) - f(n+1, i+1)).$$

This gets more complicated, but one easily checks that $f(n, i) = i^2 - n$ is a solution, so

$$S_n^2 - n \quad \text{is a martingale.}$$

These results allow to use martingale techniques to study Markov chains. As an example, one can give an alternative treatment of the Dirichlet problem presented in Section 3.5 by relying on the stopping theorem for the martingale $(M_n^f)_n$. We shall solve in the exercises the ruin problem in this way. The recurrence/transience of a Markov chain can also be studied by means of such martingales decomposition.

## 8.5 Optimal stopping problem with finite horizon

Consider the following game. You are given a fixed finite horizon $N \geq 2$, and for every $n = 1, \dots, N$, a certain amount of money is proposed to you and you can decide:

- either to take it and stop without knowing the future proposals,
- or to refuse it and hear the next proposal.

At time $n = N$, if you have refused all the previous offers, then you get the last one. The question is then to try to design a strategy to maximise the probability that the offer you accept is the overall best one. This problem is also known under the name of the 'secretary problem' in which one can imagine auditioning candidates one after the others for an open secretary position, and trying to hire the best person, or also 'marriage problem' in which you try to get the best partner... These problems are also very standard in financial mathematics. Let us formalise mathematically the problem.

**Definition 8.5.1** (Finite Horizon Optimal Stopping Problem)**.** Fix an integer $N$, a finite filtration $(\mathcal{F}_n)_{0\leq n\leq N}$, and an adapted and integrable process $(X_n)_{0\leq n\leq N}$. Let $\mathbb{T}_N$ denote the set of all stopping times with values in $\{0, \dots, N\}$. An *optimal stopping time* is a stopping time $\tau \in \mathbb{T}_N$ that satisfies:

$$\mathbb{E}[X_\tau] = \sup_{T\in\mathbb{T}_N} \mathbb{E}[X_T],$$

which may not exist nor be unique. The questions are: do they exist and can we find them explicitly?

We shall start the process $(X_n)_n$ rather at time $n = 1$, and take $\mathcal{F}_0 = \{\emptyset, \Omega\}$ the trivial $\sigma$-algebra. Also, for definiteness, let $X_n = X_N$ and $\mathcal{F}_n = \mathcal{F}_N$ for $n > N$.

**Remark 8.5.2.** Actually, in our motivation problem (see Subsection 8.5.2) and this is often the case, the sequence $(X_n)_n$ is not adapted to $(\mathcal{F}_n)_n$, that is, the knowledge of $\mathcal{F}_n$ does not entirely determine $X_n$ and some randomness remains. In this case, define $\widetilde{X}_n = \mathbb{E}[X_n \mid \mathcal{F}_n]$, which is adapted, and observe that for any stopping time $\tau \in \mathbb{T}_N$, we have since $\{\tau = n\} \in \mathcal{F}_n$,

$$\mathbb{E}[X_\tau] = \sum_{n=0}^{N} \mathbb{E}[X_n \mathbb{1}_{\tau=n}] = \sum_{n=0}^{N} \mathbb{E}[\widetilde{X}_n \mathbb{1}_{\tau=n}] = \mathbb{E}[\widetilde{X}_\tau].$$

Therefore a stopping time is optimal for $(X_n)_n$ if and only if it is optimal for $(\widetilde{X}_n)_n$, and also the maximal expected gain satisfies:

$$\sup_{\tau\in\mathbb{T}_N} \mathbb{E}[X_\tau] = \sup_{\tau\in\mathbb{T}_N} \mathbb{E}[\widetilde{X}_\tau].$$

We can thus always come back to the adapted case.

### 8.5.1 Solution via the Snell enveloppe

We are going to solve this problem using martingale theory via the notion of *Snell enveloppe*. To get the intuition consider the case $N = 2$ in our problem: you are proposed an amount $X_1$, do you take it or refuse it to get $X_2$? The answer depends on what you expect $X_2$ to be, given the information $\mathcal{F}_1$ (the amount $X_1$). Precisely: compute the conditional expectation $\mathbb{E}[X_2 \mid \mathcal{F}_1]$, then you accept $X_1$ if the latter is larger than $\mathbb{E}[X_2 \mid \mathcal{F}_1]$ and refuse it otherwise. This motivates the following definition.

**Definition 8.5.3** (Snell enveloppe)**.** Under the preceding notation, the sequence $(S_n)_{0\le n\le N}$ defined by the backward recursion:

$$S_N = X_N \qquad \text{and} \qquad S_n = \max(X_n,\, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]) \quad \text{for } 0 \le n \le N-1,$$

is called the *Snell enveloppe* of $(X_n)_{0\le n\le N}$.

As before, we extend $S_n = S_N = X_N$ for $n > N$. Here is another way of understanding the Snell enveloppe.

**Lemma 8.5.4.** *The Snell enveloppe of $(X_n)_n$ is the smallest $(\mathcal{F}_n)_n$-supermartingale above $(X_n)_n$ in the sense that $S_n \ge X_n$ for every $n$.*

*Proof.* Notice that since both $X_n$ and $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$ are $\mathcal{F}_n$-measurable and integrable, then so is $S_n$; moreover $S_n \ge \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$ for $n \le N$; after $N$, both $(S_n)_{n\ge N}$ and $(\mathcal{F}_n)_{n\ge N}$ are constant so $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[S_N \mid \mathcal{F}_N] = S_N = S_n$, hence it is a supermartingale. The bound $S_n \ge X_n$ is also obvious from the construction. Let us prove that it is the smallest such supermartingale. Let $(Y_n)_n$ be another one, then first $Y_n \ge X_n = S_n$ for $n \ge N$. We use then a backward induction: let us assume that for some $1 \le n \le N$ we have $Y_n \ge S_n$, and let us prove that $Y_{n-1} \ge S_{n-1}$. Since $(Y_n)_n$ is a supermartingale, then

$$Y_{n-1} \ge \mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] \ge \mathbb{E}[S_n \mid \mathcal{F}_{n-1}].$$

Since in addition $Y_{n-1} \ge X_{n-1}$ by assumption, then actually:

$$Y_{n-1} \ge \max(X_{n-1}, \mathbb{E}[S_n \mid \mathcal{F}_{n-1}]) = S_{n-1}.$$

We conclude by a backward induction. □

As a consequence, we can upper bound the maximal expected gain in the optimal stopping problem.

**Corollary 8.5.5.** *If $(S_n)_{0\le n\le N}$ is the Snell enveloppe of $(X_n)_{0\le n\le N}$, then*

$$\sup_{\tau\in\mathbb{T}_N} \mathbb{E}[X_\tau] \le \mathbb{E}[S_0].$$

*Proof.* Fix a stopping time $\tau \in \mathbb{T}_N$. Since $(S_n)_n$ is a supermartingale, then so is $(S_{n\wedge\tau})_n$ and since $(S_n)_n$ is above $(X_n)_n$, then

$$\mathbb{E}[X_{n\wedge\tau} \mid \mathcal{F}_0] \le \mathbb{E}[S_{n\wedge\tau} \mid \mathcal{F}_0] \le S_0$$

for any $n$. Taking $n = N$, we obtain $\mathbb{E}[X_\tau \mid \mathcal{F}_0] \le S_0$ and we conclude by taking the expectation on both sides. □

Let us next consider two special stopping times given by:

$$\tau_\star = \inf\{n \ge 0 :\ S_n = X_n\} \qquad \text{and} \qquad \tau^\star = \inf\{n \ge 0 :\ S_n > \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]\}. \tag{8.1}$$

Note that the backward recursion $S_n = \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n])$ for $n \le N-1$ shows both that $\tau_\star \le N$ and that it is equivalently given by $\tau_\star = \inf\{n \ge 0 :\ S_n \ge \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]\}$. The strict inequality required in $\tau^\star$ may not occur, so it can be infinite.

**Lemma 8.5.6.** *Both stopped processes $(S_{n\wedge\tau_\star})_n$ and $(S_{n\wedge\tau^\star})_n$ are martingales.*

*Proof.* The key is to note that first if $\tau$ is a stopping time, then $S_\tau\, \mathbb{1}_{\tau\le n} = \sum_{k\le n} S_k\, \mathbb{1}_{\tau=k}$ Ⓜ $\mathcal{F}_n$, and second that both stopping times $\tau_\star$ and $\tau^\star$ have the property that if $\tau \ge n+1$, then $S_n = \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$. Indeed, we know that $S_n = \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]) \ge X_n$ so if $n < \tau_\star$, then $S_n > X_n$ and thus $S_n = \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$. Similarly,

if $n < \tau^\star$, then $S_n \le \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$ and thus an equality holds. If $\tau$ is any stopping time with this property, then we have:

$$\begin{aligned}\mathbb{E}[S_{(n+1)\wedge\tau} \mid \mathcal{F}_n] &= \mathbb{E}[S_{n+1}\,\mathbb{1}_{\tau\ge n+1} \mid \mathcal{F}_n] + \mathbb{E}[S_\tau\,\mathbb{1}_{\tau\le n} \mid \mathcal{F}_n] \\ &= \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]\,\mathbb{1}_{\tau\ge n+1} + S_\tau\,\mathbb{1}_{\tau\le n} \\ &= S_n\,\mathbb{1}_{\tau\ge n+1} + S_\tau\,\mathbb{1}_{\tau\le n} \\ &= S_{n\wedge\tau}.\end{aligned}$$

Thus in this case $(S_{n\wedge\tau})_n$ is a martingale. □

The next theorem shows that the upper bound from Corollary 8.5.5 is achieved, it also characterises the optimal stopping times using the Snell enveloppe and shows that the two previous ones are respectively the smallest and largest ones.

**Theorem 8.5.7.** *Let $(S_n)_{0\le n\le N}$ be the Snell enveloppe of $(X_n)_{0\le n\le N}$: The following holds.*

*(i) A stopping time $\tau \in \mathbb{T}_N$ is optimal if and only if $X_\tau = S_\tau$ and $(S_{n\wedge\tau})_n$ is a martingale.*

*(ii) It holds $\mathbb{E}[S_0] = \sup_{\tau\in\mathbb{T}_N} \mathbb{E}[X_\tau]$.*

*(iii) A stopping time $\tau$ is optimal if and only if $\tau_\star \le \tau \le \tau^\star$ and $X_\tau = S_\tau$, where $\tau_\star$ and $\tau^\star$ are defined in* (8.1)*. In particular $\tau_\star$ is always optimal and $\tau^\star$ is as soon as it is not infinite.*

*Proof.* Let us start with the converse implication in (i). Fix $\tau \in \mathbb{T}_N$ and suppose that $X_\tau = S_\tau$ and $(S_{n\wedge\tau})_n$ is a martingale. Then by the stopping theorem, since $\tau$ is bounded,

$$S_0 = \mathbb{E}[S_\tau \mid \mathcal{F}_0] = \mathbb{E}[X_\tau \mid \mathcal{F}_0].$$

In particular, taking the expectation, we infer from the previous corollary that:

$$\sup_{\tau\in\mathbb{T}_N} \mathbb{E}[X_\tau] \le \mathbb{E}[S_0] = \mathbb{E}[X_\tau].$$

Hence $\tau$ is optimal and moreover we have:

$$\mathbb{E}[S_0] = \sup_{\tau\in\mathbb{T}_N} \mathbb{E}[X_\tau].$$

At this point, we have proved that if a stopping time $\tau \in \mathbb{T}_N$ satisfies both $X_\tau = S_\tau$ and $(S_{n\wedge\tau})_n$ is a martingale, then it is optimal and (ii) holds. Since $\tau_\star \le n$ satisfies these two conditions by Lemma 8.5.6, then it is optimal and (ii) always holds. Similarly $\tau^\star$ is optimal as soon as it is smaller than or equal to $N$.

Now let us prove the direct implication in (i), that is let $\tau \in \mathbb{T}_N$ be optimal and let us prove that necessarily $X_\tau = S_\tau$ and $(S_{n\wedge\tau})_n$ is a martingale. Recall that $(S_n)_n$ is a supermartingale, so $\mathbb{E}[S_\tau] \le \mathbb{E}[S_0]$, and that it satisfies $S_n \ge X_n$ for every $n$ and so $S_\tau \ge X_\tau$. In addition, since $\tau$ is optimal and (ii) holds as we just proved, then $\mathbb{E}[X_\tau] = \mathbb{E}[S_0]$. Hence $S_\tau \ge X_\tau$ and $\mathbb{E}[S_\tau] \le \mathbb{E}[X_\tau]$, which implies that $S_\tau = X_\tau$. Similarly, we know that $(S_{n\wedge\tau})_n$ is a supermartingale, so $\mathbb{E}[S_{(n+1)\wedge\tau} \mid \mathcal{F}_n] \le S_{n\wedge\tau}$ and further the expectation is nonincreasing so:

$$\mathbb{E}[S_\tau] = \mathbb{E}[S_{N\wedge\tau}] \le \mathbb{E}[S_{n\wedge\tau}] \le \mathbb{E}[S_{0\wedge\tau}] = \mathbb{E}[S_0].$$

Since $\tau$ is optimal, then $\mathbb{E}[S_\tau] = \mathbb{E}[S_0]$, so $\mathbb{E}[S_{n\wedge\tau}]$ is constant; in particular $\mathbb{E}[S_{(n+1)\wedge\tau} \mid \mathcal{F}_n] \le S_{n\wedge\tau}$ have the same expectations and thus are equal: the stopped process $(S_{n\wedge\tau})_n$ is a martingale.

It remains to prove (iii). First, if $\tau$ is optimal, then it has $X_\tau = S_\tau$ so $\tau \ge \tau_\star$. In addition $(S_{n\wedge\tau})_n$ is a martingale so on the event $n < \tau$ we have $S_n = \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$ and thus $n < \tau^\star$. This shows that $\tau \le \tau^\star$. Conversely, if $\tau \le \tau^\star \wedge N$ has $X_\tau = S_\tau$, since $(S_{n\wedge\tau^\star})_n$ is a martingale by Lemma 8.5.6, then the stopped process $S_{n\wedge\tau} = S_{n\wedge\tau^\star\wedge\tau}$ is a martingale as well, so $\tau$ is optimal by the first item. □

### 8.5.2 An explicit calculation

Let us come back to our original game or 'secretary problem' and apply Theorem 8.5.7 to design a strategy that maximises the probability to get the overall best offer.

**Modelisation.** Let us formalise the problem: $N \geq 2$ values $a_1 < \cdots < a_N$ are fixed in advanced, but unknown to us, and are presented to us one at a time in a random order, given by a uniform random permutation $\sigma$ of $\{1, \dots, N\}$. Let $X_n = 1$ if $a_{\sigma(n)} = \max_{0 \leq i \leq N} a_i$ and $X_n = 0$ otherwise, that is $X_n = 1$ when the $n$'th offer is the overall best one. Let $\mathbb{T}_N$ denote the set of stopping times less than or equal to $N$. We aim at finding $\tau \in \mathbb{T}_N$ that solves:

$$\mathbb{E}[X_\tau] = \sup_{T \in \mathbb{T}_N} \mathbb{E}[X_T] = \sup_{\tau \in \mathbb{T}_N} \mathbb{P}(a_{\sigma(\tau)} = \max a_i).$$

This situation is a typical example of what we explained in Remark 8.5.2: at time $n$, the information that we have is about the values of the first $n$ numbers that appeared, that is $\mathcal{F}_n$ is generated by $a_{\sigma(1)}, \dots, a_{\sigma(n)}$, whereas $X_n$ uses the information of all the numbers $a_1, \dots, a_N$, given by $\mathcal{F}_N$. Then as we explained, it is equivalent to solve the problem with $X_n$ replaced by $\widetilde{X}_n = \mathbb{E}[X_n \mid \mathcal{F}_n]$.

Let us first express this random variable $\widetilde{X}_n$. Let $A_n = \{a_{\sigma(n)} > \max_{k \leq n-1} a_{\sigma(k)}\}$ be the event that the $n$'th offer is better than all the previous ones, so $A_n \in \mathcal{F}_n$. By symmetry, the $A_n$'s are independent and $\mathbb{P}(A_n) = 1/n$ respectively. Then $\widetilde{X}_n$ equals the conditional probability given $a_{\sigma(1)}, \dots, a_{\sigma(n)}$ of $A_n \cap A_{n+1}^c \cap \cdots \cap A_N^c$, that is, by independence:

$$\widetilde{X}_n = \mathbb{1}_{A_n} \prod_{k=n+1}^{N} \frac{k-1}{k} = \frac{n}{N} \mathbb{1}_{A_n}.$$

Let us solve the optimal stopping problem for this sequence.

**The Snell enveloppe.** For $1 \leq n \leq N-1$, define the rest of the harmonic sum:

$$r_n = \sum_{k=n}^{N-1} \frac{1}{k},$$

define also $r_0 = \infty$ and $r_N = 0$. Note that $(r_n)_n$ decreases and has $r_1 \geq 1$ so there exists a unique index $n_\star \in \{1, \dots, N-1\}$ such that:

$$r_N < \cdots < r_{n_\star} \leq 1 < r_{n_\star - 1} < \cdots < r_0.$$

Recall that the Snell enveloppe is defined by the backward recursion:

$$S_N = \widetilde{X}_N \qquad \text{and for } 1 \leq n \leq N-1, \qquad S_n = \max(\widetilde{X}_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]).$$

By a backward induction, we can show that:

$$S_n = \frac{n}{N} \mathbb{1}_{A_n} + \frac{n}{N} r_n \mathbb{1}_{A_n^c} \quad \text{for } n_\star \leq n \leq N \qquad \text{and} \qquad S_n = \frac{n_\star - 1}{N} r_{n_\star - 1} \quad \text{for } 1 \leq n < n_\star. \tag{8.2}$$

Indeed, this holds true for $n = N$ since $r_N = 0$ and $\widetilde{X}_n = \frac{n}{N} \mathbb{1}_{A_n}$. Then for $n_\star \leq n < N$, if this holds for $n+1$, then we get since the $A_k$'s are independent Bernoulli with parameter $1/k$:

$$\begin{aligned}
\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}\left[ \frac{n+1}{N} \mathbb{1}_{A_{n+1}} + \frac{n+1}{N} r_{n+1} \mathbb{1}_{A_{n+1}^c} \mid \mathcal{F}_n \right] \\
&= \frac{n+1}{N} \frac{1}{n+1} + \frac{n+1}{N} r_{n+1} \frac{n}{n+1} \\
&= \frac{1}{N} + \frac{n}{N} \left( r_n - \frac{1}{n} \right) \\
&= \frac{n}{N} r_n.
\end{aligned}$$

Consequently, since $\widetilde{X}_n = \frac{n}{N}\,\mathbb{1}_{A_n}$ and $r_n \le 1$ for $n \ge n_\star$, then

$$S_n = \max(\widetilde{X}_n,\, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]) = \frac{n}{N}\max(\mathbb{1}_{A_n}, r_n) = \frac{n}{N}\,\mathbb{1}_{A_n} + \frac{n}{N} r_n\,\mathbb{1}_{A_n^c},$$

which concludes the induction in this case and this identity holds for all $n \ge n_\star$.

Next take $n = n_\star - 1$, then the first two equalities in the last display still hold true, the difference is that now $r_{n_\star - 1} > 1 \ge \mathbb{1}_{A_{n_\star-1}}$ and so we have:

$$S_{n_\star-1} = \frac{n_\star - 1}{N}\max(\mathbb{1}_{A_{n_\star-1}}, r_{n_\star-1}) = \frac{n_\star - 1}{N} r_{n_\star-1},$$

which initialises the backward induction for the second part of (8.2). Next note that since $r_{n_\star-1} > 1$, then for any $n < n_\star - 1$ it holds:

$$\widetilde{X}_n = \frac{n}{N}\,\mathbb{1}_{A_n} < \frac{n_\star - 1}{N} r_{n_\star-1}.$$

Therefore if the right-hand side equals $S_{n+1}$, then

$$S_n = \max(\widetilde{X}_n,\, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]) = \frac{n_\star - 1}{N} r_{n_\star-1}$$

as well, concluding the induction.

**The solution.** Note that in addition to (8.2), we just saw that $S_n > \widetilde{X}_n$ for $n < n_\star$, whereas for $n \ge n_\star$, we have $S_n = \widetilde{X}_n + \frac{n}{N} r_n\,\mathbb{1}_{A_n^c}$ which equals $\widetilde{X}_n$ if and only if $A_n$ holds true or $n = N$. We infer from Theorem 8.5.7 that the smallest optimal stopping time $\tau_\star = \inf\{n \ge 1 :\ S_n = X_n\}$ is given by:

$$\tau_\star = \inf\{n \ge n_\star :\ A_n \text{ holds true}\} \wedge N.$$

The strategy thus consists in rejecting arbitrarily the first $n_\star$ offers that are presented to you and then after that, accepting the first offer that comes up and which is the best one so far, or taking the last one if the best offer was before time $n_\star$. Theorem 8.5.7 proves that the probability to end up with the overall best offer with this strategy equals:

$$p_{\max} = \mathbb{E}[S_1] = \frac{n_\star - 1}{N} r_{n_\star-1}.$$

Furthermore, if other strategies can lead to this same probability, this one is the quickest and no strategy can do better.

**Large $N$ limit.** Let us finally recall that $n_\star$ was defined as the only index between 1 and $N$ satisfying:

$$\sum_{k=n_\star}^{N-1} \frac{1}{k} = r_{n_\star} \le 1 < r_{n_\star-1} = \sum_{k=n_\star-1}^{N-1} \frac{1}{k}.$$

Note that $n_\star$ depends on $N$, and so does $p_{\max}$. We aim at finding their limit behaviour as $N \longrightarrow \infty$. Basic calculus shows that if $n/N \longrightarrow a$ for some $a \in [0,1]$, then, with the convention $\log(1/0) = \infty$,

$$r_n = \frac{1}{N}\sum_{k=n}^{N-1} \frac{1}{k/N} \underset{N\to\infty}{\longrightarrow} \int_a^1 \frac{1}{x}\,\mathrm{d}x = \log(1/a).$$

Since the ratio $n_\star/N$ lies between 0 and 1, it admits subsequential limits as $N \longrightarrow \infty$ and the previous display combined with the condition $r_{n_\star} \le 1 < r_{n_\star-1}$ imply that that $\mathrm{e}^{-1}$ is the only possible subsequential limit. We infer that:

$$\frac{n_\star}{N} \underset{N\to\infty}{\longrightarrow} \mathrm{e}^{-1} \quad \text{which implies} \quad r_{n_\star} \underset{N\to\infty}{\longrightarrow} 1 \quad \text{so finally} \quad p_{\max} = \frac{n_\star - 1}{N} r_{n_\star-1} \underset{N\to\infty}{\longrightarrow} \mathrm{e}^{-1}.$$

Hence, when $N$ is large, one should first reject the $n_\star \sim \mathrm{e}^{-1} N \approx 0{,}37 \times N$ first offers, then accept the first one larger than all the previous ones, which provides a probability $p_{\max} \sim \mathrm{e}^{-1} \approx 37\%$ to get the overall best.

## 8.6 Optimal stopping problem with infinite horizon ($\star$)

Let us generalise in this section the optimal stopping problem when the horizon $N = \infty$.

**Definition 8.6.1.** Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration and fix an adapted stochastic process $(X_n)_{n\geq 0}$ satisfying the integrability condition:

$$\mathbb{E}\Big[\sup_{n\geq 0} |X_n|\Big] < \infty. \tag{8.3}$$

Let $\mathbb{T}$ denote the set of all finite stopping times. An *optimal stopping time* is a stopping time $\tau \in \mathbb{T}$ that satisfies:

$$\mathbb{E}[X_\tau] = \sup_{T\in\mathbb{T}} \mathbb{E}[X_T].$$

Note that the integrability condition (8.3) ensures that $\mathbb{E}[X_T]$ is well-defined for all $T \in \mathbb{T}$.

The case of finite horizon $N$ corresponds to taking $X_n = X_N$ and $\mathcal{F}_n = \mathcal{F}_N$ for every $n \geq N$, in which case an optimal stopping time, if any, can always be taken less than or equal to $N$. Moreover, in this case, the integrability condition (8.3) is equivalent to simply requiring $\mathbb{E}[|X_n|] < \infty$ for each $n$ as we did. Finally, Remark 8.5.2 applies as previously: we can always replace $X_n$ by $\mathbb{E}[X_n \mid \mathcal{F}_n]$ so the assumption that $(X_n)_n$ is adapted can be dropped.

Let us start with an example which can be solved by hand.

### 8.6.1 An explicit example

Suppose that you possess a car that you do not use and that you want to sell: you are getting random offers for it and want the highest one. We neglect here that the value tends to diminish with time and suppose that the offers are i.i.d. However every month you have to pay the insurance, possible the parking, etc. so you also want to sell it as quick as possible. Let us formalise the problem mathematically.

**Modelisation.** Let $(U_k)_{k\geq 1}$ be i.i.d. random variables such that $U_1 > 0$ a.s. and $\mathbb{E}[U_1^2] < \infty$. They represent the offers proposed to you, say once every week. Let $c > 0$ be a real number which represents the fixed cost per week of the car. Let $V_n = \sup_{k\leq n} U_k$; we are interested in maximising the quantity:

$$X_n = V_n - cn,$$

which represents your possible gain at time $n$ if you are not forced to answer an offer right away, so they are not limited in time and you can choose the highest one you have received so far. We let $\mathcal{F}_n = \sigma(U_k, k \leq n)$.

**Optimal stopping time.** Let $M = \sup\{x \geq 0 : \ \mathbb{P}(U_1 \geq x) > 0\}$ denote the supremum of the support of the law of $U_1$, which is infinite if $U_1$ is unbounded. For $x \in \mathbb{R}$, let $f(x) = \mathbb{E}[\max(0, U_1 - x)] = \mathbb{E}[(U_1 - x)^+]$, which is a continuous and decreasing function on $(-\infty, M)$ which converges to 0 at $M$ (and is constant equal to 0 after $M$ if the latter is finite). In particular for $c > 0$, there is a unique solution $\gamma \in (-\infty, M)$ of $f(\gamma) = c$. Define then the stopping time:

$$T_\star = \inf\{n \geq 1 : \ f(V_n) \leq c\} = \inf\{n \geq 1 : \ V_n \geq \gamma\}.$$

It is an easy exercise to show that $V_n$ increases and converges to $M$ almost surely, so $f(V_n)$ decreases and converges towards 0. In particular $T_\star$ is finite almost surely and we have $f(V_n) \geq c$ for every $n < T_\star$ and $f(V_n) \leq c$ for every $n \geq T_\star$.

The function $f$ and the stopping time $T_\star$ appear as follows: since the $U_k$'s are i.i.d. then,

$$\begin{aligned}
\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[\max(V_n, U_{n+1}) - (c+1)n \mid \mathcal{F}_n] \\
&= V_n + \mathbb{E}[\max(0, U_{n+1} - V_n) \mid \mathcal{F}_n] - (c+1)n \\
&= V_n + f(V_n) - (c+1)n \\
&= X_n + f(V_n) - c.
\end{aligned}$$

Hence $T_\star$ is the instant such that $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \geq X_n$ for every $n < T_\star$ and $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leq X_n$ for every $n \geq T_\star$, that is, at time $T_\star$, the process $(X_n)_n$ transitions from a submartingale to a supermartingale. We claim that this time is optimal, provided $X_{T_\star}$ is integrable. Indeed, iterating these inequalities, we get $X_n \mathbb{1}_{n<T_\star} \leq \mathbb{E}[X_{T_\star} \mathbb{1}_{n<T_\star} \mid \mathcal{F}_n]$ and $\mathbb{E}[X_n \mathbb{1}_{n\geq T_\star} \mid \mathcal{F}_{T_\star}] \leq X_{T_\star} \mathbb{1}_{n\geq T_\star}$. Let $T$ be another stopping time, then, provided $X_T$ is integrable as well, we can extend these inequalities to $n = T$, namely:

$$X_T \mathbb{1}_{T<T_\star} \leq \mathbb{E}[X_{T_\star} \mathbb{1}_{T<T_\star} \mid \mathcal{F}_T] \qquad \text{and} \qquad \mathbb{E}[X_T \mathbb{1}_{T\geq T_\star} \mid \mathcal{F}_{T_\star}] \leq X_{T_\star} \mathbb{1}_{T\geq T_\star}.$$

Taking the expectation and summing the two inequalities, we obtain:

$$\mathbb{E}[X_T] \leq \mathbb{E}[X_{T_\star}].$$

**Optimal gain.** Let us next compute $\mathbb{E}[X_{T_\star}]$. We have the equality of events: $\{V_n \geq c\} \cap \bigcap_{k<n}\{V_k < c\} = \{U_n \geq c\} \cap \bigcap_{k<n}\{U_k < c\}$ from which follows:

$$T_\star = \inf\{n \geq 1 : \ U_n \geq \gamma\}.$$

Hence $T_\star$ has the geometric law with mean $1/\mathbb{P}(U_n \geq \gamma)$. Further, for every $n \geq 1$, it holds:

$$\mathbb{E}[U_1 \mathbb{1}_{U_1\geq\gamma}] = \mathbb{E}[(U_1 - \gamma) \mathbb{1}_{U_1\geq\gamma}] + \gamma \mathbb{P}(U_1 \geq \gamma) = f(\gamma) + \gamma \mathbb{P}(U_1 \geq \gamma) = c + \gamma \mathbb{P}(U_1 \geq \gamma).$$

Using that the $U_k$'s are i.i.d. we infer that:

$$\mathbb{E}[U_n \mathbb{1}_{T_\star=n}] = \mathbb{E}\Big[U_n \mathbb{1}_{U_n\geq\gamma} \prod_{k=1}^{n-1} \mathbb{1}_{U_k<\gamma}\Big] = \mathbb{E}[U_1 \mathbb{1}_{U_1\geq\gamma}]\, \mathbb{P}(U_1 < \gamma)^{n-1} = (c + \gamma \mathbb{P}(U_1 \geq \gamma))\ \mathbb{P}(U_1 < \gamma)^{n-1}.$$

Summing over all $n \geq 1$, we get:

$$\mathbb{E}[U_{T_\star}] = \sum_{n\geq 1} \mathbb{E}[U_n \mathbb{1}_{T_\star=n}] = (c + \gamma \mathbb{P}(U_1 \geq \gamma)) \sum_{n\geq 1} \mathbb{P}(U_1 < \gamma)^{n-1} = \frac{c + \gamma \mathbb{P}(U_1 \geq \gamma)}{\mathbb{P}(U_1 \geq \gamma)},$$

so finally, since $U_{T_\star} = V_{T_\star}$:

$$\sup_{T\in\mathbb{T}} \mathbb{E}[X_T] = \mathbb{E}[X_{T_\star}] = \mathbb{E}[U_{T_\star}] - c\,\mathbb{E}[T_\star] = \frac{c + \gamma \mathbb{P}(U_1 \geq \gamma)}{\mathbb{P}(U_1 \geq \gamma)} - \frac{c}{\mathbb{P}(U_1 \geq \gamma)} = \gamma,$$

which we recall is the solution to $\mathbb{E}[(U_1 - \gamma)^+] = c$.

### 8.6.2 Essential supremum

We cannot define the Snell enveloppe by the backward recursion in infinite horizon as we are not able to initialise it; we shall provide another, more robust, definition. Let us go back to the case of finite horizon $N$ first. Recall that the Snell enveloppe $(S_{n\wedge N})_n$ is the smallest supermartingale above $(X_{n\wedge N})_n$. Consequently, if we let $\mathbb{T}_{n,N}$ denote the set of stopping times with values in $\{n, \dots, N\}$, then for any such $T \in \mathbb{T}_{n,N}$, we have:

$$S_n \geq \mathbb{E}[S_T \mid \mathcal{F}_n] \geq \mathbb{E}[X_T \mid \mathcal{F}_n].$$

Furthermore, for

$$\tau_{\star,n} = \inf\{k \in \{n, \dots, N\} : \ S_k = X_k\} \in \mathbb{T}_{n,N},$$

we have $S_{\tau_{\star,n}} = X_{\tau_{\star,n}}$, which implies as in the proof of Lemma 8.5.6 that $(S_{k\wedge\tau_{\star,n}})_k$ is actually a martingale. Hence, for this particular choice, we have:

$$S_n = \mathbb{E}[S_{\tau_{\star,n}} \mid \mathcal{F}_n] = \mathbb{E}[X_{\tau_{\star,n}} \mid \mathcal{F}_n].$$

Combining the two displays, we obtain:

$$S_n = \mathbb{E}[X_{\tau_{\star,n}} \mid \mathcal{F}_n] \geq \mathbb{E}[X_T \mid \mathcal{F}_n]$$

for every stopping time $T \in \mathbb{T}_{n,N}$. In a sense $\tau_{\star,n}$ is thus the stopping time $T \in \mathbb{T}_{n,N}$ that maximises $\mathbb{E}[X_T \mid \mathcal{F}_n]$, and this maximal value is $S_n$. This extends Theorem 8.5.7 which concerns the case $n = 0$. However there is a measurability issue here: the set $\mathbb{T}_{n,N}$ is uncountable, and in general the supremum of uncountably many measurable functions is not measurable, that is $\sup_{T\in\mathbb{T}_{n,N}} \mathbb{E}[X_T \mid \mathcal{F}_n]$ is not a well-defined random variable a priori.

**Example 8.6.2.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the interval $[0,1]$ equipped with the Borel $\sigma$-algebra $\mathcal{B}([0,1])$ and the Lebesgue measure. Let $A \subset [0,1]$ be your favorite non-Borel set and for every $a \in A$ and $t \in [0,1]$, let $X_a(t) = \mathbb{1}_{t=a}$. Then $X_a$ is indeed measurable but $\sup_{a\in A} X_a = \mathbb{1}_A$ is not.

The correct notion is that of *essential supremum.*

**Lemma 8.6.3.** *Let $I$ be any set and $(X_i, i \in I)$ a collection of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with values in $\mathbb{R} \cup \{-\infty, \infty\}$. There exists a random variable $X$ on this space that has:*

*(i) For every $i \in I$, we have $X \geq X_i$ almost surely,*

*(ii) If $Y$ also satisfies (i), then $Y \geq X$ almost surely.*

*(iii) If $Z$ satisfies (i) and (ii), then $X = Z$ almost surely.*

*(iv) There exists a* countable *subset $J \subset I$ such that $X = \sup_{j\in J} X_j$ almost surely.*

*This a.s. unique random variable $X$ is denoted by*

$$X = \operatorname*{esssup}_{i\in I} X_i,$$

*and called the* essential supremum *of $(X_i, i \in I)$.*

Continuing the previous example, one has $X_a = 0$ almost surely for each $a \in A$ and therefore $\operatorname{esssup}_{a\in A} X_a = 0$ almost surely.

*Proof.* Note that uniqueness in (iii) is a direct consequence of (ii) since the symmetric argument shows that $X \leq Y$ a.s. Let us first construct $X$ and prove the last claim. Recall that $\mathbb{R} \cup \{-\infty, \infty\}$ can be mapped on $[0,1]$ by an increasing bijection (e.g. $x \mapsto 1/2 + \arctan(x)/\pi$ extended to $\pm\infty$); applying this bijection to our random variables, we may, and shall, assume that they all take value in $[0,1]$. For any countable subset $J \subset I$, let us set $X^J = \sup_{j\in J} X_j$, which is a well defined random variable. Define then:

$$\alpha = \sup\{\mathbb{E}[X^J] : J \subset I \text{ countable}\}.$$

As a supremum of real numbers, there exists a sequence of countable sets of indices $(J_n)_n$ such that $\alpha = \lim_n \mathbb{E}[X^{J_n}]$. The set $\bigcup_n J_n$ is countable so we associated with it $X = X^{\bigcup_n J_n} = \sup_{j\in\bigcup_n J_n} X_j$.

Let us prove that $X$ satisfies the first claim. By construction, we have both $\mathbb{E}[X] \geq \mathbb{E}[X^{J_n}]$ for every $n$ and $\mathbb{E}[X] \leq \alpha$. Since $\mathbb{E}[X^{J_n}] \longrightarrow \alpha$, then we infer that $\mathbb{E}[X] = \alpha$. Now fix any $i \in I$ and consider the countable set $\{i\} \cup \bigcup_n J_n$ and the associated random variable $X^{\{i\}\cup\bigcup_n J_n} = \sup_{j\in\{i\}\cup\bigcup_n J_n} X_j = \max(X_i, X \geq X)$. Then similarly, we have both $\mathbb{E}[X^{\{i\}\cup\bigcup_n J_n}] \leq \alpha$ and $\mathbb{E}[X^{\{i\}\cup\bigcup_n J_n}] \geq \mathbb{E}[X] = \alpha$, hence $\mathbb{E}[X^{\{i\}\cup\bigcup_n J_n}] = \alpha = \mathbb{E}[X]$. Combining with the (a.s.) bound $X^{\{i\}\cup\bigcup_n J_n} \geq X$, we infer that $X = X^{\{i\}\cup\bigcup_n J_n} \geq X_i$ a.s. wich proves the first claim. Finally, if $Y$ also satisfies this bound, then in particular with probability 1 we have $Y \geq X_j$ simultaneously for every $j \in \bigcup_n J_n$, and thus $Y \geq \sup_{j\in\bigcup_n J_n} X_j = X$. □

Following the discussion before this lemma, the Snell enveloppe $(S_n)_{0\leq n\leq N}$ of $(X_n)_{0\leq n\leq N}$ is given by:

$$S_n = \operatorname*{esssup}_{T\in\mathbb{T}_{n,N}} \mathbb{E}[X_T \mid \mathcal{F}_n]. \tag{8.4}$$

Note that this is an alternative way to define $S_n$.

### 8.6.3 Optimal stopping with infinite horizon

We may now define the Snell enveloppe in general. The integrability condition (8.3) allows to define $\mathbb{E}[X_T \mid \mathcal{F}_n]$ for any finite stopping time.

**Definition 8.6.4.** The *Snell enveloppe* $(S_n)_{n\geq 0}$ of $(X_n)_{n\geq 0}$ is defined by:

$$S_n = \operatorname*{esssup}_{T\in\mathbb{T}_n} \mathbb{E}[X_T \mid \mathcal{F}_n],$$

where $\mathbb{T}_n$ denotes the set of finite stoping times with values in $\{n, n+1, \dots\}$.

The next lemma links this definition with the backward recursion that we used in the finite horizon case.

**Lemma 8.6.5.** *Assume the integrability condition* (8.3). *The* Snell enveloppe $(S_n)_{n\geq 0}$ *of* $(X_n)_{n\geq 0}$ *solves the backward recursion: for every* $n \geq 0$,

$$S_n = \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]). \tag{8.5}$$

*Moreover, it is the smallest* $(\mathcal{F}_n)_n$*-supermartingale above* $(X_n)_n$ *in the sense that* $S_n \geq X_n$ *for every* $n$.

*Proof.* Let us prove successively that:

(i) $S_n \geq \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n])$, this immediately implies both the supermartingale property and the bound $S_n \geq X_n$ for every $n$.

(ii) $S_n = \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n])$,

(iii) $(S_n)_n$ is the smallest supermartingale above $(X_n)_n$.

Step 1: proof of $S_n \geq \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n])$. The integrability condition (8.3) provides domination that shows that $(S_n)_n$ is integrable, so we can make sense of $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$. Since $S_n \geq \mathbb{E}[X_T \mid \mathcal{F}_n]$ for any $T \in \mathbb{T}_n$, then in particular for $T = n$ we have $S_n \geq X_n$. Let us next argue that $S_n \geq \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]$. By Lemma 8.6.3, for each $n$ there is a countable subset $\{\theta^N_{n+1}, N \geq 1\} \subset \mathbb{T}_{n+1}$ such that $S_{n+1} = \sup_{N\geq 1} \mathbb{E}[X_{\theta^N_{n+1}} \mid \mathcal{F}_{n+1}]$. Let us transform the $\theta^N_{n+1}$ so the conditional expectation are nondecreasing: for each $N \geq 1$, let $k^N_{n+1}$ be any index $k \leq N$ such that:

$$\mathbb{E}[X_{\theta^k_{n+1}} \mid \mathcal{F}_{n+1}] = \max_{j\leq N} \mathbb{E}[X_{\theta^j_{n+1}} \mid \mathcal{F}_{n+1}]$$

and set

$$T^N_{n+1} = \theta^{k^N_{n+1}}_{n+1} \qquad \text{so now} \qquad S_{n+1} = \uparrow \lim_{N\to\infty} \mathbb{E}[X_{T^N_{n+1}} \mid \mathcal{F}_{n+1}].$$

Using (8.3) we may apply the dominated convergence theorem under the conditional expectation $\mathbb{E}[\,\cdot \mid \mathcal{F}_n]$ (recall Lemma 6.4.3) and deduce from the tower property that:

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = \lim_{N\to\infty} \mathbb{E}[\mathbb{E}[X_{T^N_{n+1}} \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \lim_{N\to\infty} \mathbb{E}[X_{T^N_{n+1}} \mid \mathcal{F}_n].$$

Since $T^N_{n+1} \in \mathbb{T}_{n+1}$, then each conditional expectation on the right is upper bounded by $\operatorname{esssup}_{T\in\mathbb{T}_{n+1}} \mathbb{E}[X_T \mid \mathcal{F}_n]$. Letting $N \to \infty$, we infer that $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] \leq \operatorname{esssup}_{T\in\mathbb{T}_{n+1}} \mathbb{E}[X_T \mid \mathcal{F}_n]$. On the other hand, for any $T \in \mathbb{T}_{n+1}$ we have $S_{n+1} \geq \mathbb{E}[X_T \mid \mathcal{F}_{n+1}]$, so $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] \geq \mathbb{E}[X_T \mid \mathcal{F}_n]$ by the tower property again. By Lemma 8.6.3, this implies:

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = \operatorname*{esssup}_{T\in\mathbb{T}_{n+1}} \mathbb{E}[X_T \mid \mathcal{F}_n] \leq \operatorname*{esssup}_{T\in\mathbb{T}_n} \mathbb{E}[X_T \mid \mathcal{F}_n] = S_n.$$

We have thus proved that $S_n \geq \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n])$.

Step 2: proof of $S_n = \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n])$. For any $T \in \mathbb{T}_n$, we have $T \vee (n+1) \in \mathbb{T}_{n+1}$ so by Lemma 8.6.3:

$$\begin{aligned}
\mathbb{E}[X_T \mid \mathcal{F}_n] &= \mathbb{E}[X_n \mathbb{1}_{T=n} + X_{T\vee(n+1)} \mathbb{1}_{T\geq n+1} \mid \mathcal{F}_n] \\
&= X_n \mathbb{1}_{T=n} + \mathbb{E}[X_{T\vee(n+1)} \mid \mathcal{F}_n] \mathbb{1}_{T\geq n+1} \\
&\leq \max\big(X_n, \mathbb{E}[X_{T\vee(n+1)} \mid \mathcal{F}_n]\big) \\
&\leq \max\Big(X_n, \operatorname*{esssup}_{T\in\mathbb{T}_{n+1}} \mathbb{E}[X_T \mid \mathcal{F}_n]\Big) \\
&\leq \max(X_n, \mathbb{E}[S_{n+1} \mid \mathcal{F}_n]),
\end{aligned}$$

hence the last line is larger than or equal to $\operatorname{esssup}_{T\in\mathbb{T}_n} \mathbb{E}[X_T \mid \mathcal{F}_n] = S_n$. This shows that $(S_n)_n$ solves the backward recursion relations.

Step 3: minimality of $(S_n)_n$. Finally, if $(Y_n)_n$ is another supermartingale above $(X_n)_n$, then for any stopping time $T \in \mathbb{T}_n$ and any $N \geq n$ we have

$$Y_n \geq \mathbb{E}[Y_{T\wedge N} \mid \mathcal{F}_n] \geq \mathbb{E}[X_{T\wedge N} \mid \mathcal{F}_n].$$

Letting $N \longrightarrow \infty$ and applying the dominated convergence theorem under the conditional expectation $\mathbb{E}[\,\cdot \mid \mathcal{F}_n]$, using again the assumption (8.3), we infer that $Y_n \geq \mathbb{E}[X_T \mid \mathcal{F}_n]$. As this holds for any $T \in \mathbb{T}_n$, we conclude that

$$Y_n \geq \operatorname*{esssup}_{T\in\mathbb{T}_n} \mathbb{E}[X_T \mid \mathcal{F}_n] = S_n,$$

and the proof is complete. □

Let $(S_n)_n$ denote the Snell enveloppe of $(X_n)_n$ and define:

$$\tau_{\star,n} = \inf\{\, m \geq n : \; S_m = X_m\} \in \mathbb{T}_n,$$

which is a stopping time for every $n \geq 0$. It generalises $\tau_\star = \tau_{\star,0}$ from (8.1).

**Lemma 8.6.6.** *Assume the integrability condition* (8.3). *For every $n \geq 0$ fixed, the stopped process $(S_{m\wedge\tau_{\star,n}})_{m\geq n}$ is a martingale. Moreover, if $\tau_{\star,n}$ is finite almost surely, then it realises the essential supremum:*

$$S_n = \mathbb{E}[X_{\tau_{\star,n}} \mid \mathcal{F}_n] \geq \mathbb{E}[X_T \mid \mathcal{F}_n]$$

*for every $T \in \mathbb{T}_n$.*

*Proof.* Let us decompose: for every $m \geq n$,

$$S_{(m+1)\wedge\tau_{\star,n}} = S_{\tau_{\star,n}} \mathbb{1}_{\tau_{\star,n}\leq m} + S_{m+1} \mathbb{1}_{\tau_{\star,n}\geq m+1}.$$

Observe that $S_{\tau_{\star,n}} \mathbb{1}_{\tau_{\star,n}\leq m} = \sum_{k\leq m} S_k \mathbb{1}_{\tau_{\star,n}=k}$ Ⓜ $\mathcal{F}_m$ and since $\mathbb{1}_{\tau_{\star,n}\geq m+1}$ Ⓜ $\mathcal{F}_m$ as well, then:

$$\mathbb{E}[S_{(m+1)\wedge\tau_{\star,n}} \mid \mathcal{F}_m] = S_{\tau_{\star,n}} \mathbb{1}_{\tau_{\star,n}\leq m} + \mathbb{E}[S_{m+1} \mid \mathcal{F}_m] \mathbb{1}_{\tau_{\star,n}\geq m+1}.$$

On the other hand, if $\tau_{\star,n} \geq m+1$, then $X_m \neq S_m$, and since $S_m = \max(X_m, \mathbb{E}[S_{m+1} \mid \mathcal{F}_m]) \geq X_m$, then this means that $S_m > X_m$ and so in fact $S_m = \mathbb{E}[S_{m+1} \mid \mathcal{F}_m]$. Consequently,

$$\mathbb{E}[S_{(m+1)\wedge\tau_{\star,n}} \mid \mathcal{F}_m] = S_{\tau_{\star,n}} \mathbb{1}_{\tau_{\star,n}\leq m} + S_m \mathbb{1}_{\tau_{\star,n}\geq m+1} = S_{m\wedge\tau_{\star,n}}.$$

Thus $(S_{m\wedge\tau_{\star,n}})_{m\geq n}$ is a martingale. In particular $S_n = \mathbb{E}[S_{m\wedge\tau_{\star,n}} \mid \mathcal{F}_n]$ for every $m \geq n$. Suppose that $\tau_{\star,n} < \infty$ almost surely, so we can make sense of $X_{\tau_{\star,n}} = S_{\tau_{\star,n}} = \lim_m S_{m\wedge\tau_{\star,n}}$. Moreover we have $|S_k| \leq \mathbb{E}[\sup_k |X_k| \mid \mathcal{F}_n]$ which has finite mean by the integrability condition (8.3). We may thus apply the dominated convergence theorem under the conditional expectation $\mathbb{E}[\,\cdot \mid \mathcal{F}_n]$ (recall Lemma 6.4.3) and deduce that:

$$\mathbb{E}[X_{\tau_{\star,n}} \mid \mathcal{F}_n] = \mathbb{E}[S_{\tau_{\star,n}} \mid \mathcal{F}_n] = \lim_{m\longrightarrow\infty} \mathbb{E}[S_{m\wedge\tau_{\star,n}} \mid \mathcal{F}_n] = S_n.$$

The lower bound $S_n \geq \mathbb{E}[X_T \mid \mathcal{F}_n]$ for every $T \in \mathbb{T}_n$ follows from the definition of $S_n$. □

We may now solve the optimal stopping problem. Let us actually generalise the original problem as follows: fix $n \geq 0$ and search for $\tau \in \mathbb{T}_n$, the set of finite stopping times larger than or equal to $n$, that has:

$$\mathbb{E}[X_\tau] = \sup_{T \in \mathbb{T}_n} \mathbb{E}[X_T]. \tag{8.6}$$

The original problem is for $n = 0$. Recall the stopping time $\tau_{\star,n} = \inf\{m \geq n : S_m = X_m\} \in \mathbb{T}_n$, then a consequence of Lemma 8.6.6 is that

$$\sup_{T \in \mathbb{T}_n} \mathbb{E}[X_T] \leq \mathbb{E}[X_{\tau_{\star,n}}] = \mathbb{E}[S_n].$$

The next theorem shows that this upper bound is achieved, so $\tau_{\star,n}$ is an optimal stopping time, and is actually the smallest one.

**Theorem 8.6.7** (General case)**.** *Assume the integrability condition* (8.3)*. For every $n \geq 0$ fixed, the following holds as soon as $\tau_{\star,n}$ is finite almost surely:*

*(i) A stopping time $\tau \in \mathbb{T}_n$ solves* (8.6) *if and only if $S_\tau = X_\tau$ and $(S_{m\wedge\tau})_{m\geq n}$ is a martingale.*

*(ii) It holds $\mathbb{E}[S_n] = \sup_{T\in\mathbb{T}_n} \mathbb{E}[X_T]$.*

*(iii) The stopping time $\tau_{\star,n}$ solves* (8.6) *and any other such solution $\tau \in \mathbb{T}_n$ has $\tau \geq \tau_{\star,n}$.*

*If $\tau_{\star,n}$ can be infinite with positive probability, then there is no solution to* (8.6)*.*

*Proof.* Fix $n \geq 0$ and suppose first that $\tau_{\star,n}$ is finite almost surely. Let $\tau \in \mathbb{T}_n$ be such that both $S_\tau = X_\tau$ and $(S_{m\wedge\tau})_{m\geq n}$ is a martingale. In particular $\mathbb{E}[S_n] = \mathbb{E}[S_{m\wedge\tau}]$ for every $m \geq n$ and $S_{m\wedge\tau} \longrightarrow S_\tau$ as $m \longrightarrow \infty$. As in the previous proof, the integrability condition (8.3) allows us to apply the dominated convergence theorem and get:

$$\mathbb{E}[S_\tau] = \lim_{m\longrightarrow\infty} \mathbb{E}[S_{m\wedge\tau}] = \mathbb{E}[S_n].$$

We observed already that $\mathbb{E}[S_n] \geq \sup_{T\in\mathbb{T}_n} \mathbb{E}[X_T] \geq \mathbb{E}[S_\tau] = \mathbb{E}[S_n]$ so these inequalities are equalities and $\tau$ therefore solves (8.6). In addition, in this case (ii) holds as well.

Recall from Lemma 8.6.6 that $\tau_{\star,n}$ has that $S_{\tau_{\star,n}} = X_{\tau_{\star,n}}$ and $(S_{m\wedge\tau_{\star,n}})_{m\geq n}$ is a martingale. Hence it solves (8.6) and (ii) always holds.

Suppose conversely that $\tau \in \mathbb{T}_n$ solves (8.6) and let us prove that necessarily $S_\tau = X_\tau$ and $(S_{m\wedge\tau})_{m\geq n}$ is a martingale. On the one hand the latter is a supermartingale by Lemma 8.6.5, so in particular $\mathbb{E}[S_n] \geq \mathbb{E}[S_{n\wedge\tau}] = \mathbb{E}[S_{m\wedge\tau}]$ for any $m \geq n$; letting $m \longrightarrow \infty$, by dominated convergence again, we obtain $\mathbb{E}[S_n] \geq \mathbb{E}[S_\tau] \geq \mathbb{E}[X_\tau]$. On the other hand $\tau$ solves (8.6) and (ii) holds as we just proved so finally:

$$\mathbb{E}[X_\tau] = \sup_{T\in\mathbb{T}_n} \mathbb{E}[X_T] = \mathbb{E}[S_n] \geq \mathbb{E}[S_\tau] \geq \mathbb{E}[X_\tau].$$

Thus all these inequalities are equalities. Recall that $S_m \geq X_m$ for every $m$ so $S_\tau \geq X_\tau$ and since their expectation are equal then actually $S_\tau = X_\tau$. Similarly, the stopped process $(S_{m\wedge\tau})_{m\geq n}$ is a supermartingale: $\mathbb{E}[S_{(m+1)\wedge\tau} \mid \mathcal{F}_m] \leq S_{m\wedge\tau}$ for every $m$ and we claim that their expectations are equal, so again the random variables are equal. Indeed, the expectation of a supermartingale is nonincreasing so for every $\ell \geq m \geq n$ we have by dominated convergence again:

$$\mathbb{E}[S_n] = \mathbb{E}[S_{n\wedge\tau}] \geq \mathbb{E}[S_{m\wedge\tau}] \geq \mathbb{E}[S_{\ell\wedge\tau}] \underset{\ell\longrightarrow\infty}{\longrightarrow} \mathbb{E}[S_\tau] = \mathbb{E}[S_n],$$

so $\mathbb{E}[S_{m\wedge\tau}] = \mathbb{E}[S_n]$ for every $m$ so $(S_{m\wedge\tau})_{m\geq n}$ is not only a supermartingale but a martingale. This proves (i).

Recall the consequence of Lemma 8.6.6 that $\sup_{T\in\mathbb{T}_n} \mathbb{E}[X_T] \leq \mathbb{E}[X_{\tau_{\star,n}}] = \mathbb{E}[S_n]$. Since the extremities are equal by (ii), then the inequality is an equality, that is: $\tau_{\star,n}$ solves (8.6). On the other hand, any other solution $\tau \in \mathbb{T}_n$ has $S_\tau = X_\tau$ by (i), so necessarily $\tau \geq \tau_{\star,n}$ by definition of the latter. This also proves that if there is a solution $\tau$ to (8.6), which is finite by definition of a solution, then $\tau_{\star,n} \leq \tau < \infty$, so by contraposition, if $\tau_{\star,n} = \infty$ with positive probability, then there is no solution to (8.6). □

# Chapter 9

# Convergence of martingales

This chapter is dedicated to the study of the asymptotic behaviour of martingales and their convergence in different senses. The fundamental result is an almost sure convergence which, in addition to the Borel–Cantelli lemma, is basically the only general tool to prove such a convergence in general, besides bare-hand study of the model. We then discuss the convergence in $L^1$ and in $L^p$. We finish with an extension of the Central Limit Theorem which was one of the first concerns in this theory: how to generalise this result without independence between the increments? This CLT will be applied to prove a CLT for Markov chains. We shall also mention some applications to numerical simulations.

**Contents**



In Section 9.1 we first discuss almost sure convergence of martingales. Let us stress already the mild assumptions that are used, for example being nonnegative suffices! Section 9.2 further discusses the convergence in $L^1$, which actually holds if and only if the martingale takes the form of successive conditioning of a fixed random variable. Section 9.3 presents some further developments with the notion of uniform integrability. In Section 9.4 we prove very useful bounds on the maximum of a martingale and derive $L^p$ convergence results with $p > 1$. Section 9.5 shows in particular that a martingale with bounded increments either converges to a finite limit or it oscilles between $+\infty$ and $-\infty$, but it cannot converge to infinity! Sections 9.6 and 9.7 develop extensions of the LLN and CLT in the context of martingales in $L^2$. Finally Section 9.8 presents an application to the stochastic gradient descent that allows to numerically approximate the minimiser of a function.

## 9.1 Almost sure convergence

Martingales are one of the few tools we have to prove almost sure convergence, besides the Borel–Cantelli lemma. This is based on the following result.

**Theorem 9.1.1.** *Let $(M_n)_n$ be a (sub/super-)martingale bounded in $L^1$, i.e. such that*

$$\sup_{n\geq 0} \mathbb{E}[|M_n|] < \infty.$$

*Then $M_n$ converges a.s. to some $M_\infty$ which has $\mathbb{E}[|M_\infty|] < \infty$.*

Let us stress right away that the convergence may not hold in $L^1$! This will be our next topic. A classical proof of Theorem 9.1.1 is based on the idea of counting 'upcrossings' of an interval. Fix $a < b$ and a sequence $(x_n)_n$ on $\mathbb{R}$ and define two sequences $0 = t_0 \le s_1 \le t_1 \le s_2 \le \dots$ inductively by:

$$s_k = \inf\{j \ge t_{k-1} : x_j \le a\} \qquad \text{and} \qquad t_k = \inf\{j \ge s_k : x_j \ge b\}.$$

Then write $U_n^{a,b} = \sup\{k \ge 0 : t_k \le n\}$ for the number of upcrossings of $[a, b]$ by $x$ up to time $n$. We also let $U_\infty^{a,b} = \uparrow\lim_n U_n^{a,b}$ denote the total upcrossing number.

**Lemma 9.1.2.** *The sequence $(x_n)_n$ converges in $[-\infty, \infty]$ if and only if $U_\infty^{a,b} < \infty$ for all rational numbers $a < b$.*

*Proof.* Notice that $U_\infty^{a,b} < \infty$ if and only if there exists $k \ge 0$ such that the corresponding upcrossing times satisfy $t_k < \infty = t_{k+1}$, namely $x_n > a$ for every $n > t_k$ or $x_n < b$ for every $n > t_k$. Recall that $(x_n)_n$ converges in $[-\infty, \infty]$ if and only if $\liminf_n x_n = \limsup_n x_n$.

If $(x_n)_n$ does not converge, then $\liminf_n x_n < \limsup_n x_n$ and so there exist two rational numbers $a < b$ such that $\liminf_n x_n < a < b < \limsup_n x_n$, which implies $U_\infty^{a,b} = \infty$ for this pair. Suppose conversely that there exist two rational numbers $a < b$ such that $U_\infty^{a,b} = \infty$, so each $s_k, t_k$ is finite and we have:

$$\liminf_n x_n \le \liminf_k x_{s_k} \le a < b \le \limsup_k x_{t_k} \le \limsup_n x_n,$$

so $(x_n)_n$ does not converge. □

The proof of Theorem 9.1.1 then mostly relies on checking that the number of upcrossings of any interval is almost surely finite. This is based on the following result.

**Lemma 9.1.3.** *Let $(M_n)_n$ be a supermartingale and let $a < b$. Then the mean upcrossing number satisfies:*

$$\mathbb{E}[U_n^{a,b}] \le \frac{\mathbb{E}[(M_n - a)^-]}{b - a} \le \frac{|a| + \mathbb{E}[|M_n|]}{b - a}.$$

Once again, we can explain the proof in terms of a strategy, see Figure 9.1. Imagine $(M_n)_n$ representing the price of an asset on the stock market. If, as the author of these lines, you know nothing about finance, a naive way to try to make money is to wait until the price is low, here below $a$, then buy one unit at this time, so now your fortune follows the same evolution as the price, then wait until the price gets above $b$ to sell and thus freeze your fortune until the price gets below $a$ again, etc. Note that every time you sell, you earned at least $b - a$, which corresponds to the denominator in the lemma. The numerator appears to take into account that at time $n$, you may be engaged since the price went below $a$ again, but not yet above $b$, so you may be loosing some money. This explains why you should only use this strategy to prove theorems: after going below $a$ the price might not go up again!

*Proof.* Let us write $0 = T_0 \le S_1 \le T_1 \le S_2 \le \dots$ for the random times defined as above for $M$. Since $M$ is adapted, then these are stopping times and so the process defined for $n \ge 1$ by:

$$H_n = \sum_{k \ge 1} \mathbb{1}_{S_k < n \le T_k}$$

is predictable. It is obviously nonnegative and it is bounded since at most one indicator can take value 1. The process $(H \bullet M)$ corresponds to our strategy presented above. Let us decompose the trajectory

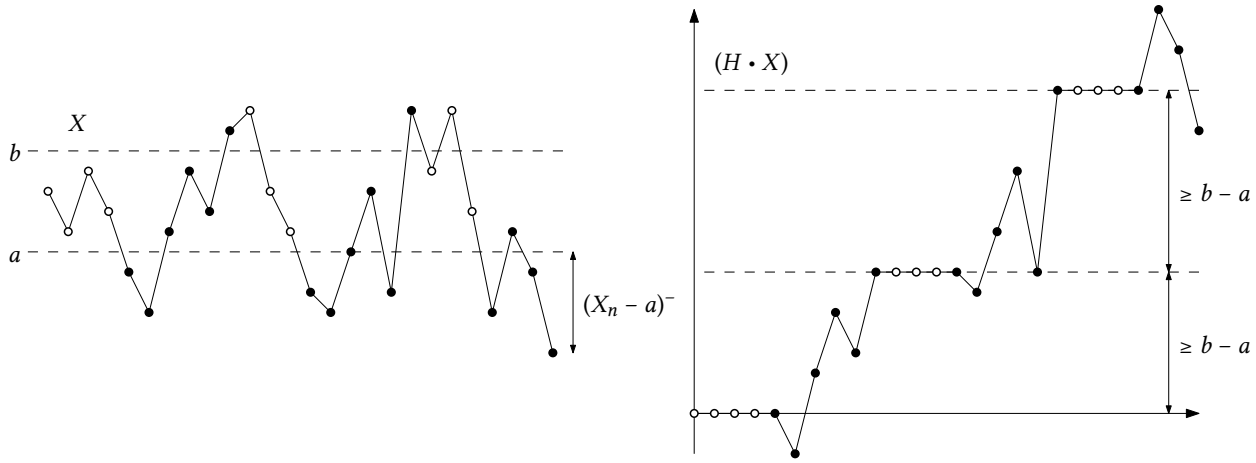



Figure 9.1: The proof of Lemma 9.1.3 explained.

according to the upcrossings, and notice that the total increment along each one is $M_{T_j} - M_{S_j} \geq b - a$, so:

$$\begin{aligned}(H \bullet M)_N &= \sum_{n=1}^{N} \sum_{k\geq 1} \mathbb{1}_{S_k < n \leq T_k}(M_n - M_{n-1}) \\ &= \sum_{k\geq 1} \sum_{n=S_k+1}^{N\wedge T_k} (M_n - M_{n-1}) \\ &= \sum_{j=1}^{U_N^{a,b}} (M_{T_j} - M_{S_j}) + \mathbb{1}_{S_{U_N+1}\leq N}(M_N - M_{S_{U_N+1}}) \\ &\geq (b-a)U_N^{a,b} - (M_N - a)^-,\end{aligned}$$

where the last term accounts for a possible final incomplete upcrossing, in which case we simply throw away a temporary positive gain, but we cannot forget a temporary loss. According to Lemma 8.1.6 the process $H \bullet M$ remains a supermartingale, and so $\mathbb{E}[(H \bullet M)_N] \leq 0$, which implies the first inequality. The second one follows since $(x - a)^- \leq |a| + |x|$. □

Theorem 9.1.1 now easily follows.

*Proof of Theorem 9.1.1.* Combining Lemma 9.1.3 and monotone convergence, we have

$$\mathbb{E}[U_\infty^{a,b}] \leq \frac{1}{b-a}\Big(|a| + \sup_{n\geq 1} \mathbb{E}[|M_n|]\Big) < \infty.$$

Consequently, $U_\infty^{a,b} < \infty$ a.s. and this holds in fact a.s. simultaneously for all pairs of rational numbers $a < b$ since there are countably many of them. We infer from Lemma 9.1.2 that $(M_n)_n$ converges to some limit $M_\infty \in [-\infty, \infty]$. Finally by Fatou's lemma,

$$\mathbb{E}[|M_\infty|] \leq \liminf_{n\to\infty} \mathbb{E}[|M_n|] \leq \sup_{n\geq 1} \mathbb{E}[|M_n|] < \infty,$$

so $M_\infty$ is integrable (and thus finite). □

We can derive a particularly useful, and extraordinary at first sight, result.

**Corollary 9.1.4.** *Let $c \geq 0$ and let $(M_n)_n$ be a supermartingale that has $\inf_n M_n \geq -c$ a.s. Then $(M_n)_n$ converges a.s. to some $M_\infty$ which has $\mathbb{E}[|M_\infty|] < \infty$ and moreover $M_n \geq \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ for all $n$.*

*Proof.* Let us set $Y_n = M_n + c \geq 0$. This remains a supermartingale, and it is bounded in $L^1$ since $\mathbb{E}[|Y_n|] = \mathbb{E}[Y_n] \leq \mathbb{E}[Y_0]$. We can then apply Theorem 9.1.1 and deduce the a.s. convergence to some $Y_\infty \in L^1$. Moreover,

$$\mathbb{E}[Y_\infty \mid \mathcal{F}_n] \leq \liminf_{k\to\infty} \mathbb{E}[Y_{n+k} \mid \mathcal{F}_n] \leq Y_n,$$

by the conditional Fatou lemma and Lemma 8.1.3. The claim follows by subtracting $c$. □

Note that a priori even for martingales, we only have the inequality $M_n \geq \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ because of the conditional Fatou lemma. The next subsection discusses when we have equality, and also when $M_n$ converges to $M_\infty$ in $L^1$.

## 9.2 Closed martingales and $L^1$ convergence

A particular case of martingales are those defined by taking successive conditional expectations of a fixed random variable.

**Definition 9.2.1.** Let $\xi \in L^1$, then a sequence defined by $M_n = \mathbb{E}[\xi \mid \mathcal{F}_n]$ for every $n \geq 0$ is called a *closed martingale*.

Integrability of this process follows from Lemma 6.4.1; the martingale property then comes from the tower property:

$$\mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[\xi \mid \mathcal{F}_n].$$

As shown in the next result, these martingales are the generic case of martingales that converge in $L^1$. Notice also the surprising fact that a martingale that converges in $L^1$ necessarily converges almost surely. Recall that $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n) \subset \mathcal{F}$ is the limit of the filtration.

**Theorem 9.2.2.** *Let $(M_n)_n$ be a martingale. The following assertions are equivalent:*

*(i) It is closed: there exists $\xi \in L^1$ such that $M_n = \mathbb{E}[\xi \mid \mathcal{F}_n]$ for every $n \geq 0$.*

*(ii) It converges almost surely and in $L^1$ to some $M_\infty$.*

*(iii) It converges in $L^1$ to some $M_\infty$.*

*Moreover, when this holds we have $M_\infty = \mathbb{E}[\xi \mid \mathcal{F}_\infty]$ and $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$.*

By the very last assertion of this theorem, we do not need to specify $\xi$ when we speak of a closed martingale since we may always take $\xi = M_\infty$.

*Proof.* (i) $\Longrightarrow$ (ii): Suppose first that $M_n = \mathbb{E}[\xi \mid \mathcal{F}_n]$ for every $n \geq 0$, with $\xi \in L^1$. Then $\mathbb{E}[|M_n|] \leq \mathbb{E}[|\xi|] < \infty$ so $(M_n)_n$ is bounded in $L^1$. By Theorem 9.1.1 it therefore converges a.s. to some $M_\infty \in L^1$. Then for every $\varepsilon, K > 0$, we have:

$$\begin{aligned}
\mathbb{E}[|M_n - M_\infty|] &\leq \mathbb{E}[|M_n - M_\infty| \, \mathbb{1}_{|M_n - M_\infty| \leq \varepsilon}] + \mathbb{E}[|M_\infty| \, \mathbb{1}_{|M_n - M_\infty| > \varepsilon}] + \mathbb{E}[|M_n| \, \mathbb{1}_{|M_n - M_\infty| > \varepsilon}] \\
&\leq \varepsilon + \mathbb{E}[|M_\infty| \, \mathbb{1}_{|M_n - M_\infty| > \varepsilon}] + \mathbb{E}[|M_n| \, \mathbb{1}_{|M_n| \leq K} \, \mathbb{1}_{|M_n - M_\infty| > \varepsilon}] + \mathbb{E}[|M_n| \, \mathbb{1}_{|M_n| > K} \, \mathbb{1}_{|M_n - M_\infty| > \varepsilon}] \\
&\leq \varepsilon + \mathbb{E}[|M_\infty| \, \mathbb{1}_{|M_n - M_\infty| > \varepsilon}] + K \, \mathbb{P}(|M_n - M_\infty| > \varepsilon) + \mathbb{E}[|M_n| \, \mathbb{1}_{|M_n| > K}].
\end{aligned}$$

The first expectation tends to 0 by dominated convergence, so does the probability, and for the last expectation, note that since $M_n = \mathbb{E}[\xi \mid \mathcal{F}_n]$ and since $\{|M_n| > K\} \in \mathcal{F}_n$, then

$$\mathbb{E}[|M_n| \, \mathbb{1}_{|M_n| > K}] \leq \mathbb{E}[\mathbb{E}[|\xi| \mid \mathcal{F}_n] \, \mathbb{1}_{|M_n| > K}] = \mathbb{E}[|\xi| \, \mathbb{1}_{|M_n| > K}].$$

By dominated convergence, as $n \longrightarrow \infty$, the right-hand side converges to $\mathbb{E}[|\xi| \, \mathbb{1}_{|M_\infty| > K}]$. Thus for every $\varepsilon, K > 0$, we have:

$$\limsup_{n\to\infty} \mathbb{E}[|M_n - M_\infty|] \leq \varepsilon + \mathbb{E}[|\xi| \, \mathbb{1}_{|M_\infty| > K}].$$

Letting further $K \longrightarrow \infty$ (by dominated convergence again) and $\varepsilon \longrightarrow 0$, we conclude that: indeed $M_n \longrightarrow M_\infty$ in $L^1$ and thus (i) $\Longrightarrow$ (ii).

Obviously (ii) $\Longrightarrow$ (iii); let us prove that (iii) $\Longrightarrow$ (i). Suppose thus that $M_n$ converges to $M_\infty$ in $L^1$. Since $(M_n)_n$ is a martingale, then for any $n, k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[|M_n - \mathbb{E}[M_\infty \mid \mathcal{F}_n]|\big] &= \mathbb{E}\big[|\mathbb{E}[M_{n+k} \mid \mathcal{F}_n] - \mathbb{E}[M_\infty \mid \mathcal{F}_n]|\big] \\ &\leq \mathbb{E}\big[\mathbb{E}[|M_{n+k} - M_\infty| \mid \mathcal{F}_n]\big] \\ &= \mathbb{E}[|M_{n+k} - M_\infty|] \underset{k\to\infty}{\longrightarrow} 0.\end{aligned}$$

Hence $\mathbb{E}[|M_n - \mathbb{E}[M_\infty \mid \mathcal{F}_n]|] = 0$ and so $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ a.s. This proves (i) and the last identity in the claim.

It only remains to prove that if (i) holds, then $M_\infty = \mathbb{E}[\xi \mid \mathcal{F}_\infty]$. First observe that each $M_n$ is $\mathcal{F}_n \subset \mathcal{F}_\infty$-measurable, so their limit $M_\infty$ is $\mathcal{F}_\infty$-measurable. Next, since $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ for every $n \geq 0$, then for every event $A \in \mathcal{F}_n$, we have

$$\mathbb{E}[\xi\, \mathbb{1}_A] = \mathbb{E}[M_n\, \mathbb{1}_A] = \mathbb{E}[M_\infty\, \mathbb{1}_A].$$

In other words, the two measures on $(\Omega, \mathcal{F})$ defined by $\mu(A) = \mathbb{E}[\xi\, \mathbb{1}_A]$ and $\nu(A) = \mathbb{E}[M_\infty\, \mathbb{1}_A]$ agree on the $\pi$-system $\bigcup_n \sigma(\mathcal{F}_k, k \leq n)$ and they have the same finite total mass $\mathbb{E}[\xi] = \mathbb{E}[M_\infty] < \infty$, hence they agree on $\sigma(\bigcup_n \sigma(\mathcal{F}_k, k \leq n)) = \mathcal{F}_\infty$ by Theorem 1.1.13. This means that for every event $A \in \mathcal{F}_\infty$, we have $\mathbb{E}[\xi\, \mathbb{1}_A] = \mathbb{E}[M_\infty\, \mathbb{1}_A]$ and thus $M_\infty = \mathbb{E}[\xi \mid \mathcal{F}_\infty]$. □

This corollary can be used to give an extension of the celebrated Kolmogorov 0-1 law in Theorem 2.1.16.

**Corollary 9.2.3** (Lévy's 0-1 law)**.** *For any $A \in \mathcal{F}_\infty$, we have*

$$\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] \underset{n\to\infty}{\longrightarrow} \mathbb{1}_A \quad \textit{a.s. and in } L^1.$$

*Proof.* The sequence defined by $M_n = \mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n]$ is a closed martingale, which therefore converges a.s. and in $L^1$ to some $M_\infty$ which satisfies $M_\infty = \mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_\infty] = \mathbb{1}_A$ since $A \in \mathcal{F}_\infty$. □

**Remark 9.2.4.** This indeed extends Theorem 2.1.16. Recall from Example 2.1.12 that the grouping property shows that if $(M_n)_{n\geq 1}$ are independent random variables then the $\sigma$-algebras $\mathcal{F}_n = \sigma(M_k, k \leq n)$ and $\mathcal{T}_n = \sigma(M_k, k \geq n+1)$ are independent. Consequently each $\mathcal{F}_n$ is independent of $\mathcal{T} = \bigcap_n \mathcal{T}_n$ and for $A \in \mathcal{T} \subset \mathcal{F}_\infty$, we have $\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] \longrightarrow \mathbb{1}_A$ a.s. One can be puzzled by the identity $\mathbb{P}(A) = \mathbb{1}_A$ a.s. since the right-hand side is random. However this implies that either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, so $\mathbb{1}_A$ is actually constant a.s. either to 0 or to 1 respectively.

Let us end by extending the stopping theorem from Section 8.2 in the case of closed martingales. Recall that by Theorem 9.2.2 such a martingale $(M_n)_n$ converges a.s. and in $L^1$ to some $M_\infty$ $\textcircled{m}$ $\mathcal{F}_\infty$ and furthermore $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ for all $n$. For any stopping time $T$ let us set

$$M_T = \sum_{n\geq 0} M_n\, \mathbb{1}_{T=n} + M_\infty\, \mathbb{1}_{T=\infty},$$

which is $\mathcal{F}_T$-measurable according to Lemma 7.2.4.

**Theorem 9.2.5.** *Let $(M_n)_n$ be an adapted and integrable process. Then it is a closed martingale if and only if for every stopping time $T$ we have $M_T \in L^1$ and*

$$\mathbb{E}[M_T] = \mathbb{E}[M_0].$$

*Moreover in this case, for any stopping times $S \leq T$ we have:*

$$M_S = \mathbb{E}[M_T \mid \mathcal{F}_S].$$

**Remark 9.2.6.** Note that we can take $T = \infty$ in the last identity, so for any stopping time $S$, we have in the case of a closed martingale $M_S = \mathbb{E}[M_\infty \mid \mathcal{F}_S]$ and further $\mathbb{E}[M_0] = \mathbb{E}[M_S] = \mathbb{E}[M_\infty]$.

*Proof.* Let us first prove the direct implication. Since $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$, then

$$\begin{aligned}
\mathbb{E}[|M_T|] &= \sum_{n\geq 0} \mathbb{E}[|M_n|\, \mathbb{1}_{T=n}] + \mathbb{E}[|M_\infty|\, \mathbb{1}_{T=\infty}] \\
&= \sum_{n\geq 0} \mathbb{E}[|\mathbb{E}[M_\infty \mid \mathcal{F}_n]|\, \mathbb{1}_{T=n}] + \mathbb{E}[|M_\infty|\, \mathbb{1}_{T=\infty}] \\
&\leq \sum_{n\geq 0} \mathbb{E}[\mathbb{E}[|M_\infty| \mid \mathcal{F}_n]\, \mathbb{1}_{T=n}] + \mathbb{E}[|M_\infty|\, \mathbb{1}_{T=\infty}] \\
&= \sum_{n\geq 0} \mathbb{E}[|M_\infty|\, \mathbb{1}_{T=n}] + \mathbb{E}[|M_\infty|\, \mathbb{1}_{T=\infty}] \\
&= \mathbb{E}[|M_\infty|] < \infty.
\end{aligned}$$

So indeed $M_T \in L^1$. Let next $A \in \mathcal{F}_T$, i.e. $A \in \mathcal{F}$ and $A \cap \{T = n\} \in \mathcal{F}_n$ for all $n$. Then for every $n \geq 0$,

$$\mathbb{E}[M_T\, \mathbb{1}_{A\cap\{T=n\}}] = \mathbb{E}[M_n\, \mathbb{1}_{A\cap\{T=n\}}] = \mathbb{E}[M_\infty\, \mathbb{1}_{A\cap\{T=n\}}].$$

By summing over $n$ and using Fubini's Theorem (recall $M_T, M_\infty \in L^1$) we obtain

$$\mathbb{E}[M_T\, \mathbb{1}_A] = \mathbb{E}[M_\infty\, \mathbb{1}_A] \quad \text{for all } A \in \mathcal{F}_T \text{ and so} \quad M_T = \mathbb{E}[M_\infty \mid \mathcal{F}_T].$$

Now if $S \leq T$ is another stopping time, then $\mathcal{F}_S \subset \mathcal{F}_T$ (Lemma 7.2.2) so by the tower property (Lemma 6.5.1),

$$\mathbb{E}[M_T \mid \mathcal{F}_S] = \mathbb{E}[\mathbb{E}[M_\infty \mid \mathcal{F}_T] \mid \mathcal{F}_S] = \mathbb{E}[M_\infty \mid \mathcal{F}_S] = M_S.$$

Let us next prove the converse implication, so suppose that $\mathbb{E}[M_T] = \mathbb{E}[M_0]$ for all stopping times $T$. By Remark 8.2.4 we know (using only bounded stoping times) that $(M_n)_n$ is a martingale. Taking $T = \infty$, we know that $M_\infty$ is integrable and we can adapt the proof of Corollary 8.2.3 to show that $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ for all $n$. Indeed, fix $n \geq 0$ and $A \in \mathcal{F}_n$ and define the stopping time

$$T = n\, \mathbb{1}_A + \infty\, \mathbb{1}_{A^c}.$$

Then

$$\mathbb{E}[M_\infty] = \mathbb{E}[M_0] = \mathbb{E}[M_T] = \mathbb{E}[M_n\, \mathbb{1}_A] + \mathbb{E}[M_\infty\, \mathbb{1}_{A^c}],$$

and thus $\mathbb{E}[M_\infty\, \mathbb{1}_A] = \mathbb{E}[M_n\, \mathbb{1}_A]$. Since this holds for all $A \in \mathcal{F}_n$, then $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$. □

## 9.3 Uniformly integrable martingales (⋆)

We can push further the previous subsection with the notion of uniform integrability from Section 2.3.2. Recall from Theorem 2.3.14 that it is the optimal assumption to improve convergence in probability to $L^1$ or $L^p$ convergence. Recall also that boundedness in $L^p$ with some $p > 1$ implies uniform integrability.

**Lemma 9.3.1.** *Let $M$ be an integrable random variable then the family $\{\mathbb{E}[M \mid \mathcal{G}]; \mathcal{G} \subset \mathcal{F}\}$ of conditional expectations with respect to each sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{F}$ is uniformly integrable.*

*Proof.* Fix a sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{F}$ and let $Y = \mathbb{E}[M \mid \mathcal{G}]$ and $Z = \mathbb{E}[|M| \mid \mathcal{G}]$. Fix $K > 0$, recall that $|Y| \leq Z$ and that $\mathbb{1}_{Z>K}$ (m) $\mathcal{G}$ so

$$\mathbb{E}[|Y|\, \mathbb{1}_{|Y|>K}] \leq \mathbb{E}[Z\, \mathbb{1}_{Z>K}] = \mathbb{E}[|M|\, \mathbb{1}_{Z>K}].$$

Further, we have

$$|M|\, \mathbb{1}_{Z>K} = |M|\, \mathbb{1}_{|M|\leq\sqrt{K}, Z>K} + |M|\, \mathbb{1}_{|M|>\sqrt{K}, Z>K} \leq \sqrt{K}\, \mathbb{1}_{Z>K} + |M|\, \mathbb{1}_{|M|>\sqrt{K}}.$$

Taking the expectation and using the Markov inequality, we infer that

$$\mathbb{E}[|Y|\,\mathbb{1}_{|Y|>K}] \le \sqrt{K}\,\mathbb{P}(Z > K) + \mathbb{E}[|M|\,\mathbb{1}_{|M|>\sqrt{K}}] \le \frac{1}{\sqrt{K}}\,\mathbb{E}[Z] + \mathbb{E}[|M|\,\mathbb{1}_{|M|>\sqrt{K}}].$$

Recall that $Z = \mathbb{E}[|M| \mid \mathcal{G}]$ so in particular $\mathbb{E}[Z] = \mathbb{E}[|M|]$, so for any sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{F}$, we have

$$\mathbb{E}[|\mathbb{E}[M \mid \mathcal{G}]|\,\mathbb{1}_{|\mathbb{E}[M|\mathcal{G}]|>K}] \le \frac{1}{\sqrt{K}}\,\mathbb{E}[|M|] + \mathbb{E}[|M|\,\mathbb{1}_{|M|>\sqrt{K}}],$$

and the right-hand side tends to 0 as $K \to \infty$. □

**Remark 9.3.2.** Combined with Remark 9.2.6 we find that if $(M_n)_n$ is a closed martingale, then the collection $\{M_T, T \text{ stopping time}\}$ is uniformly integrable.

Let us next complete Theorem 9.2.2.

**Theorem 9.3.3.** *Let $(M_n)_n$ be a martingale. The following assertions are equivalent:*

*(i) It is uniformly integrable.*

*(ii) It is closed.*

*(iii) It converges almost surely and in $L^1$ to some $M_\infty$.*

*(iv) It converges in $L^1$ to some $M_\infty$.*

*Moreover, when this holds we have $M_\infty = \mathbb{E}[\xi \mid \mathcal{F}_\infty]$ and so $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$.*

*Proof.* Lemma 9.3.1 shows that (ii) $\Longrightarrow$ (i). Next, if $(M_n)_n$ is a uniformly integrable martingale, then it is bounded in $L^1$ so it converges a.s. by Theorem 9.1.1 and further in $L^1$ by Theorem 2.3.14 so (i) $\Longrightarrow$ (iv). The rest was proved in Theorem 9.2.2. □

## 9.4 $L^p$ convergence

Fix $p > 1$ and suppose that each $M_n \in L^p$. We wonder wether we can extend the previous $L^1$ convergence to an $L^p$ convergence. We shall rely on Doob's inequalities, which are powerful tools that allow to control the maximum of the whole trajectory of a process up to a given time $n$ simply by looking at its value at time $n$ and that are of independent interest. As an introduction, recall the Markov inequality: for any nonnegative real-valued random variable $M$ and any constant $c > 0$, we have:

$$c\,\mathbb{P}(M \ge c) = \mathbb{E}[c\,\mathbb{1}_{M\ge c}] \le \mathbb{E}[M\,\mathbb{1}_{M\ge c}] \le \mathbb{E}[M].$$

Below, we improve this bound for a submartingale by controlling the maximum up to time $n$ in terms of the value at time $n$.

**Theorem 9.4.1** (Doob's maximal inequalities)**.** *Let $(M_n)_n$ be a nonnegative submartingale and for $n \ge 0$ define*

$$\overline{M}_n = \sup_{k\le n} M_k.$$

*Then the following assertions hold for every $n \ge 0$.*

*(i) For any $c > 0$ it holds:*

$$c\,\mathbb{P}(\overline{M}_n \ge c) \le \mathbb{E}[M_n\,\mathbb{1}_{\overline{M}_n\ge c}] \le \mathbb{E}[M_n].$$

*(ii) For any $p > 1$, it holds:*

$$\|\overline{M}_n\|_p \le \frac{p}{p-1}\,\|M_n\|_p.$$

By Lemma 8.1.4, this theorem applies to $|M_n|$ when $(M_n)_n$ is a martingale and to $M_n^+$ if $(M_n)_n$ is a submartingale.

*Proof.* (i) The second inequality is immediate, let us prove the first one. Define the stopping time $T_c = \inf\{n \geq 0 : M_n \geq c\}$ and notice that $\{\overline{M}_n \geq c\} = \{T_c \leq n\} = \bigcup_{k=0}^n \{T_c = k\}$. Since $\{T_c = k\} \in \mathcal{F}_k$ and $\mathbb{E}[M_n \mid \mathcal{F}_k] \geq M_k \geq c$ on this event, then by the characterisation of the conditional expectation for the last equality, we have:

$$c\,\mathbb{P}(T_c = k) = \mathbb{E}[c\,\mathbb{1}_{T_c=k}] \leq \mathbb{E}[M_k\,\mathbb{1}_{T_c=k}] \leq \mathbb{E}[\mathbb{E}[M_n \mid \mathcal{F}_k]\,\mathbb{1}_{T_c=k}] = \mathbb{E}[M_n\,\mathbb{1}_{T_c=k}].$$

The claim follows by summing over $k$, recalling that $\{\overline{M}_n \geq c\} = \bigcup_{k=0}^n \{T_c = k\}$.

(ii) Fix $p > 1$ and $n \geq 1$ and assume $\mathbb{E}[M_n^p] < \infty$ as otherwise it is immediate. According to Lemma 8.1.4 the process $(M_n^p)_n$ is a submartingale, and in particular $\mathbb{E}[M_k^p] \leq \mathbb{E}[M_n^p]$ for any $k \leq n$. Hence

$$\mathbb{E}[(\overline{M}_n)^p] = \mathbb{E}\left[\sup_{k\leq n} M_k^p\right] \leq \mathbb{E}\left[\sum_{k=0}^n M_k^p\right] = \sum_{k\leq n} \mathbb{E}[M_k^p] \leq n \sup_{k\leq n} \mathbb{E}[M_k^p] = n\,\mathbb{E}[M_n^p] < \infty.$$

Next by Fubini's theorem, the first part, and Fubini's theorem again,

$$\begin{aligned}
\mathbb{E}[(\overline{M}_n)^p] &= \mathbb{E}\left[\int_0^\infty p x^{p-1}\,\mathbb{1}_{x\leq \overline{M}_n}\,\mathrm{d}x\right] \\
&= \int_0^\infty p x^{p-1}\,\mathbb{P}(\overline{M}_n \geq x)\,\mathrm{d}x \\
&\leq \int_0^\infty p x^{p-2}\,\mathbb{E}[M_n\,\mathbb{1}_{\overline{M}_n \geq x}]\,\mathrm{d}x \\
&= \mathbb{E}\left[M_n \int_0^\infty p x^{p-2}\,\mathbb{1}_{x\leq \overline{M}_n}\,\mathrm{d}x\right] \\
&= \frac{p}{p-1}\,\mathbb{E}[M_n(\overline{M}_n)^{p-1}].
\end{aligned}$$

Let $q = p/(p-1)$ be such that $1/p + 1/q = 1$, then by Hölder's inequality,

$$\mathbb{E}[(\overline{M}_n)^p] \leq q\,\mathbb{E}[M_n(\overline{M}_n)^{p-1}] \leq q\,\mathbb{E}[M_n^p]^{1/p}\,\mathbb{E}[(\overline{M}_n)^{q(p-1)}]^{1/q} = q\,\mathbb{E}[M_n^p]^{1/p}\,\mathbb{E}[(\overline{M}_n)^p]^{1-1/p}.$$

Since $\mathbb{E}[(\overline{M}_n)^p] < \infty$, then we may divide both sides by $\mathbb{E}[(\overline{M}_n)^p]^{1-1/p} > 0$ to obtain the inequality

$$\|\overline{M}_n\|_p \leq \frac{p}{p-1}\,\|M_n\|_p$$

as wanted. □

Recall from Corollary 2.3.15 and the remark below it that in general, boundedness in $L^p$ and convergence in probability imply convergence in $L^q$ for $q < p$ but not in $L^p$. Here the martingale structure allows us to get enough control in $L^p$ via Theorem 9.4.1.

**Theorem 9.4.2.** *Let $p > 1$ and let $(M_n)_n$ be a martingale bounded in $L^p$ in the sense that*

$$\sup_{n\geq 0} \mathbb{E}[|M_n|^p] < \infty.$$

*Then $M$ is closed in that there exists an integrable random variable $M_\infty$ which satisfies*

$$\mathbb{E}[M_\infty \mid \mathcal{F}_n] = M_n \qquad \text{for all } n.$$

*Moreover $M_\infty \in L^p$ and more precisely, letting $\overline{M}_\infty = \sup_{n\geq 0} |M_n|$, we have*

$$\frac{p-1}{p}\,\|\overline{M}_\infty\|_p \leq \|M_\infty\|_p \leq \sup_{n\geq 0} \|M_n\|_p \qquad \text{and finally} \qquad M_n \underset{n\to\infty}{\longrightarrow} M_\infty \quad \text{a.s. and in } L^p.$$

*Proof.* Since $(M_n)_n$ is bounded in $L^p$, then it is also bounded in $L^1$ so it converges a.s. by Theorem 9.1.1 to $M_\infty \in L^1$ which satisfies $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ for all $n$. Let $\overline{M}_n = \sup_{k\le n} |M_k|$, then by Theorem 9.4.1 applied to the nonnegative submartingale $(|M_n|)_n$, we have:

$$\mathbb{E}[\overline{M}_n^p] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}[|M_n|^p] \le \left(\frac{p}{p-1}\right)^p \sup_{n\ge 0} \mathbb{E}[|M_n|^p] < \infty.$$

By monotone convergence, the left-hand side converges to $\mathbb{E}[\overline{M}_\infty^p]$ which therefore has finite expectation. The random variables $|M_n|^p$ are thus dominated by the integrable random variable $\overline{M}_\infty^p$, so by dominated convergence $\mathbb{E}[|M_n|^p] \longrightarrow \mathbb{E}[|M_\infty|^p]$ and we infer by letting $n \longrightarrow \infty$ in the previous display that

$$\left(\frac{p-1}{p}\right)^p \mathbb{E}[\overline{M}_\infty^p] \le \mathbb{E}[|M_\infty|^p] \le \sup_{n\ge 0} \mathbb{E}[|M_n|^p].$$

Dominated convergence applied to $|M_n - M_\infty|^p$ also implies the $L^p$ convergence $M_n \longrightarrow M_\infty$. □

## 9.5 The case of bounded increments (⋆)

A martingale with bounded increments has a simple destiny: it either converges to a finite limit or oscillates between $\pm\infty$; in particular it cannot tend to infinity!

**Theorem 9.5.1** (Destiny of a martingale)**.** *Let $(X_n)_{n\ge 0}$ be a martingale with bounded increments, and let*

$$A_{conv} = \{X_n \text{ converges to a finite limit}\}$$

*and*

$$A_{osc} = \{X \text{ oscillates}\} = \{\liminf_{n\to\infty} X_n = -\infty\} \cap \{\limsup_{n\to\infty} X_n = \infty\}.$$

*Then*

$$\mathbb{P}(A_{conv} \cup A_{osc}) = 1.$$

*Proof.* A union bound shows that:

$$\mathbb{P}(A_{\text{conv}}^c \cap A_{\text{osc}}^c) \le \mathbb{P}(A_{\text{conv}}^c \cap \{\liminf X_n > -\infty\}) + \mathbb{P}(A_{\text{conv}}^c \cap \{\limsup X_n < \infty\}).$$

Let us prove that the first probability on the right vanishes. Then so does the second one by replacing $X_n$ by $-X_n$. For $K \ge 1$, we let $T_K = \inf\{n \ge 0 : X_n < -K\}$ and observe that $\{\liminf_n X_n > -\infty\} = \bigcup_K \{T_K = \infty\}$, so

$$\mathbb{P}(A_{\text{conv}}^c \cap \{\liminf X_n > -\infty\}) = \mathbb{P}\Big(\bigcup_{K\ge 1} (A_{\text{conv}}^c \cap \{T_K = \infty\})\Big) \le \sum_{K\ge 1} \mathbb{P}(A_{\text{conv}}^c \cap \{T_K = \infty\}).$$

It now suffices to prove that for any $K$ fixed, each probability on the right vanishes.

The stopped process $(X_{n\wedge T_K})_n$ is a martingale and moreover, because the increments are bounded, say $|X_{n+1} - X_n| \le M$, then $X_{n\wedge T_K} \ge -M - K$ for every $n$. We infer from Corollary 9.1.4 that it converges a.s. to a finite (even integrable) limit. Since the martingale is unstopped when $T_K = \infty$, then indeed:

$$\mathbb{P}(A_{\text{conv}}^c \cap \{T_K = \infty\}) \le \mathbb{P}(X_{n\wedge T_K} \text{ does not converge to a finite limit}) = 0,$$

and the proof is complete. □

As a corollary, we can extend the Borel–Cantelli lemma from Lemma 2.1.15.

**Corollary 9.5.2.** *Fix a sequence of events $A_n \in \mathcal{F}_n$ for every $n \ge 1$ and define two nondecreasing processes by:*

$$Y_n = \sum_{k=1}^n \mathbb{1}_{A_k} \qquad \text{and} \qquad Z_n = \sum_{k=1}^n \mathbb{E}[\mathbb{1}_{A_k} \mid \mathcal{F}_{k-1}].$$

*Then a.s. we have $Z_\infty < \infty \iff Y_\infty < \infty$.*

*Proof.* Let $X_0 = 0$ and $X_n = Y_n - Z_n$ for every $n \geq 1$. The sequence $(X_n)_n$ has bounded increments (by 1) and $X_n$ is clearly $\mathcal{F}_n$-measurable; finally one easily gets $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ so it is a martingale with bounded increments. By the previous theorem, a.s. it either converges to a finite limit or oscillates between $\pm\infty$. If $Y_\infty = \infty$ and $Z_\infty < \infty$ or $Y_\infty < \infty$ and $Z_\infty = \infty$, then $X_n \longrightarrow \infty$ and $X_n \longrightarrow -\infty$ respectively, so this occurs with probability 0 and the claim follows. □

Note that $Y_\infty < \infty$ iff $A_k$ occurs for only finitely many indices $k$. On the one hand if $\sum_k \mathbb{P}(A_k) < \infty$, then $\mathbb{E}[Y_\infty] < \infty$ so $Y_\infty < \infty$ a.s. On the other hand, if the events are independent, then for $\mathcal{F}_n = \sigma(A_k, k \leq n)$ we have that $\mathbb{E}[\mathbb{1}_{A_k} \mid \mathcal{F}_{k-1}] = \mathbb{P}(A_k)$. Therefore if $\sum_k \mathbb{P}(A_k) = \infty$, then $Y_\infty$ a.s. We thus indeed recover the Borel–Cantelli lemma.

## 9.6 Law of Large Numbers

Martingales were originally introduce to generalise cumulative sums of independent and centred sequences and extend the of strong Law of Large Numbers & Central Limit Theorem. We focus here on the former and differ the latter to the next section. For the rest of this section, we let $(M_n)_n$ be a martingale. Henceforth we shift the notation to see it as the sum of its increments and we shall write:

$$M_n = \sum_{i=0}^{n} X_i \qquad \text{so} \qquad X_n = M_n - M_{n-1} \qquad \text{and} \qquad X_0 = M_0.$$

Then the process $(M_n)_n$ is a martingale if and only if for every $n \geq 0$, it holds:

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = 0.$$

When working with non i.i.d. random variables, we need more than just integrable random variables. We shall work with martingales in $L^2$, i.e. such that $\mathbb{E}[X_n^2] < \infty$ for every $n$. In this case, the increments satisfy an orthogonality property, namely for every $n \geq 0$ and $k \geq 1$ we have:

$$\mathbb{E}[X_{n+k}X_n] = \mathbb{E}[\mathbb{E}[X_{n+k}X_n \mid \mathcal{F}_{n+k-1}]] = \mathbb{E}[\mathbb{E}[X_{n+k} \mid \mathcal{F}_{n+k-1}]X_n] = 0. \tag{9.1}$$

This property will be crucial in this section.

### 9.6.1 A first easy strong law

First, observe that boundedness in $L^2$ can be checked by considering a deterministic series.

**Lemma 9.6.1.** *A martingale is bounded in $L^2$ in that $\sup_n \mathbb{E}[M_n^2] < \infty$ if and only if the series $\sum_n \mathbb{E}[X_n^2]$ converges. When this holds $M_n$ converges a.s. and in $L^2$ to some limit $M_\infty$ which satisfies $\mathbb{E}[M_\infty \mid \mathcal{F}_n] = M_n$ for all $n$.*

*Proof.* The claim relies on the orthogonality of the increments (9.1). Indeed, this implies that all the crossed products in the expansion of $\mathbb{E}[M_n^2] = \mathbb{E}[(\sum_{i=0}^n X_i)^2]$ vanish and we obtain:

$$\mathbb{E}[M_n^2] = \mathbb{E}\left[\left(\sum_{i=0}^{n} X_i\right)^2\right] = \sum_{i=0}^{n} \mathbb{E}[X_i^2].$$

Thus the claimed equivalence. The rest follows from Theorem 9.4.2 for $p = 2$. □

Let us derive an easy LLN for martingales in $L^2$. We shall rely on Kronecker's lemma which is very useful when it comes to proving a LLN.

**Lemma 9.6.2** (Kronecker). *Let $(x_n)_n$ be real numbers and let $(a_n)_n$ be positive numbers with $a_n \uparrow \infty$.*

$$\textit{If the series} \quad \sum_n \frac{x_n}{a_n} \quad \textit{converges, then} \quad \frac{1}{a_n}\sum_{k=0}^{n} x_k \underset{n\to\infty}{\longrightarrow} 0.$$

*Proof.* Let us write $y_k = \sum_{i=1}^k x_i/a_i$, then

$$\sum_{k=1}^n x_k = \sum_{k=1}^n a_k(y_k - y_{k-1}) = \sum_{k=1}^n a_k y_k - \sum_{k=1}^n a_{k-1}y_{k-1} - \sum_{k=1}^n (a_k - a_{k-1})y_{k-1} = a_n y_n - \sum_{k=1}^n (a_k - a_{k-1})y_{k-1}.$$

Under our assumption the sequence $(y_n)_n$ has a finite limit, say $y_\infty$. Consequently, for any $\varepsilon > 0$, there exists $k_0$ such that $|y_{k-1} - y_\infty| < \varepsilon$ for every $k > k_0$. Since $a_n \longrightarrow \infty$, then we infer that:

$$\limsup_{n\to\infty} \frac{1}{a_n}\sum_{k=1}^n (a_k - a_{k-1})y_{k-1} = \limsup_{n\to\infty} \frac{1}{a_n}\sum_{k=k_0}^n (a_k - a_{k-1})y_{k-1} \le \limsup_{n\to\infty} \frac{a_n - a_{k_0-1}}{a_n}(y_\infty + \varepsilon) = y_\infty + \varepsilon.$$

Similarly the liminf is lower bounded by $y_\infty - \varepsilon$. Since $\varepsilon$ is arbitrary, then

$$\frac{1}{a_n}\sum_{k=1}^n (a_k - a_{k-1})y_{k-1} \underset{n\to\infty}{\longrightarrow} y_\infty,$$

and the claim follows. □

Here is a first LNN for martingales. Note that if the increments are i.i.d. then the assumption is clearly satisfied.

**Proposition 9.6.3.** *Let $(M_n)_n$ be an $L^2$-martingale.*

$$\text{If} \quad \sum_{n\ge 1} \frac{\mathbb{E}[X_n^2]}{n^2} < \infty, \quad \text{then} \quad \frac{M_n}{n} \underset{n\to\infty}{\longrightarrow} 0 \ \text{ a.s. and in } L^2.$$

*Proof.* First note that the $L^2$ convergence follows from Lemma 9.6.1 and Kronecker's lemma 9.6.2 applied to $x_k = \mathbb{E}[X_k^2]$ and $a_n = n^2$. Indeed, we have

$$\mathbb{E}[(n^{-1}M_n)^2] = \frac{1}{n^2}\sum_{k=0}^n \mathbb{E}[X_k^2] \underset{n\to\infty}{\longrightarrow} 0 \qquad \text{since} \qquad \sum_{n\ge 1} \frac{\mathbb{E}[X_n^2]}{n^2} < \infty.$$

Let us prove the almost sure convergence by applying Kronecker's lemma 9.6.2 to $x_k = X_k$ and $a_n = n$. Indeed, define $\widetilde{X}_k = k^{-1}X_k$ and then $\widetilde{M}_n = \sum_{k=1}^n \widetilde{X}_k$. Then $\mathbb{E}[\widetilde{X}_{n+1} \mid \mathcal{F}_k] = (n+1)^{-1}\,\mathbb{E}[X_{n+1} \mid \mathcal{F}_k] = 0$ so $(\widetilde{M}_n)_n$ inherits the martingale property from $(M_n)_n$. Further, our assumption reads $\sum_n \mathbb{E}[\widetilde{X}_n^2] < \infty$, so Lemma 9.6.1 implies that $\widetilde{M}_n$ converges a.s. (and in $L^2$) to a finite limit $\widetilde{M}_\infty$. This means formally that the event

$$A = \{\omega \in \Omega : \text{ the series } \sum_k \frac{X_k(\omega)}{k} \text{ converges}\}$$

has probability 1. As alluded, we then apply Kronecker's lemma 9.6.2 to $x_k = X_k(\omega)$ and $a_k = k$ to infer that for every $\omega \in A$, it holds $n^{-1}M_n = n^{-1}\sum_{k=1}^n X_k(\omega) \longrightarrow 0$. Since $\mathbb{P}(A) = 1$, this means that $n^{-1}M_n \longrightarrow 0$ almost surely as wanted. □

**Remark 9.6.4.** Proposition 9.6.3 can be used to prove the strong LLN for i.i.d. integrable random variables. Indeed one of the key points in the proof of Theorem 2.4.2 was to show with the notation there that $n^{-1}\sum_{k\le n} Y_k \longrightarrow 0$ and one can check that this sum defines an $L^2$ martingale which satisfies the assumptions of Proposition 9.6.3.

### 9.6.2 The bracket process & further strong laws

In the LLN and CLT for martingales, the normalising factor is not always of order $n$ and $\sqrt{n}$ respectively. It involves more generally the so-called *bracket process*.

**Definition 9.6.5.** Let $M_n = \sum_{k\le n} X_k$ be an $L^2$-martingale. We define a nondecreasing process $(\langle M\rangle_n)_{n\ge 0}$ by $\langle M\rangle_0 = 0$ and for $n \ge 1$,

$$\langle M\rangle_n = \sum_{k=1}^{n} \mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] = \sum_{k=1}^{n} \mathbb{E}[M_k^2 - M_{k-1}^2 \mid \mathcal{F}_{k-1}].$$

Let also $\langle M\rangle_\infty = \uparrow\lim_n \langle M\rangle_n \in [0,\infty]$.

**Remark 9.6.6.** The analogous bracket process, also called *quadratic variation* plays an important role in the study of continuous-time processes.

The next lemma characterises the bracket process. It is actually a particular case of Lemma 8.3.1 applied to the submartingale $(M_n^2)$ and we refer the interested reader to this more general result.

**Lemma 9.6.7.** *The process $(\langle M\rangle_n)_{n\ge 0}$ is the almost surely unique predictable process such that the difference $M_n^2 - M_0^2 - \langle M\rangle_n$ defines a martingale null at 0.*

*Proof.* First it is clear from the definition that $\langle M\rangle_n \,\textcircled{m}\, \mathcal{F}_{n-1}$. In addition, by construction, for every $n \ge 0$, we get by expanding $M_{n+1}^2 = M_n^2 + X_{n+1}^2 + 2M_nX_{n+1}$:

$$\begin{aligned}\mathbb{E}[M_{n+1}^2 - M_0^2 - \langle M\rangle_{n+1} \mid \mathcal{F}_n] &= M_n^2 + \mathbb{E}[X_{n+1}^2 \mid \mathcal{F}_n] - M_0^2 - \langle M\rangle_{n+1}\\ &= M_n^2 - M_0^2 - \langle M\rangle_n,\end{aligned}$$

hence the martingale property. Suppose next that $(A_n)_{n\ge 0}$ is a predictable process such that $\widetilde{M}_n = M_n^2 - M_0^2 - A_n$ defines a martingale null at 0, then:

$$A_{n+1} - A_n = (M_{n+1}^2 - M_n^2) - (\widetilde{M}_{n+1} - \widetilde{M}_n).$$

Since $A_{n+1} - A_n \,\textcircled{m}\, \mathcal{F}_n$ and $(\widetilde{M}_n)_n$ is martingale, then we infer that

$$A_{n+1} - A_n = \mathbb{E}[A_{n+1} - A_n \mid \mathcal{F}_n] = \mathbb{E}[M_{n+1}^2 - M_n^2 \mid \mathcal{F}_n] = \langle M\rangle_{n+1} - \langle M\rangle_n.$$

Since $A_0 = 0 = \langle M\rangle_0$, then we infer that the two sequences $(A_n)_n$ and $(\langle M\rangle_n)_n$ coincide. □

Let $T$ be a stopping time and recall from Lemma 8.2.1 that the stopped process defined by $M_n^T = M_{n\wedge T}$ for every $n \ge 0$ remains a martingale in $L^2$. We can therefore define its bracket process $(\langle M^T\rangle_n)_{n\ge 0}$. The next lemma shows that the stopped bracket process is the bracket process of the stopped process.

**Lemma 9.6.8.** *For any stopping time $T$, we have $(\langle M\rangle_{n\wedge T})_n = (\langle M^T\rangle_n)_n$.*

*Proof.* According to Lemma 9.6.7, the process $(\langle M^T\rangle_n)_n$ is the unique predictable process $(A_n)_n$ such that $(M_{n\wedge T}^2 - A_n)_n$ is a martingale started at $M_0^2$. Let us prove that $(\langle M\rangle_{n\wedge T})_n$ also satisfies these two properties. First, by decomposing according to the value of $T$, we have

$$\langle M\rangle_{n\wedge T} = \sum_{j=0}^{n-1} \mathbb{1}_{T=j}\langle M\rangle_j + \mathbb{1}_{T>n-1}\langle M\rangle_n,$$

and each random variable on the right is $\mathcal{F}_{n-1}$-measurable, hence $(\langle M\rangle_{n\wedge T})_n$ is indeed predictable. Next simply observe that since $(M_n^2 - \langle M\rangle_n)_n$ is a martingale, then so is the stopped process $(M^2 - \langle M\rangle)_{n\wedge T} = M_{n\wedge T}^2 - \langle M\rangle_{n\wedge T}$. The claim follows then by uniqueness of $(\langle M^T\rangle_n)_n$. □

Let us observe that $\mathbb{E}[M_n^2] = \mathbb{E}[\langle M\rangle_n]$ so by monotone convergence $(M_n)_n$ is bounded in $L^2$ if and only if $\mathbb{E}[\langle M\rangle_\infty] < \infty$. In this case, Theorem 9.4.2 applies and the martingale converges almost surely an in $L^2$. In the next result, we remove the assumption $\mathbb{E}[\langle M\rangle_\infty] < \infty$ and prove almost sure convergences; note however that we may not have convergence in $L^2$.

**Theorem 9.6.9.** *Let $(M_n)_n$ be an $L^2$-martingale.*

*(i) On the event that $\langle M\rangle_\infty < \infty$ we have $M_n \longrightarrow M_\infty$ a.s where $\mathbb{E}[M_\infty^2] < \infty$. Conversely, if $(M_n)_n$ has bounded increments, then on the event that it converges a.s. we have $\langle M\rangle_\infty < \infty$ a.s.*

*(ii) On the event that $\langle M\rangle_\infty = \infty$ we have $M_n/\langle M\rangle_n \longrightarrow 0$ a.s.*

*Proof.* Fix $k \geq 0$ and let $T_k = \inf\{n \geq 0 : \langle M\rangle_{n+1} > k\}$, which is a stopping time since $(\langle M\rangle_{n+1})_n$ is adapted. Then by Lemma 9.6.8 the bracket of the stopped process $(M_{n\wedge T_k})_n$ is $(\langle M\rangle_{n\wedge T_k})_n$. The latter is bounded by $k$, so $(M_{n\wedge T_k})_n$ is bounded in $L^2$ by the previous remark and Lemma 9.6.1 shows that it converges a.s. The limit is square-integrable by Fatou's lemma. On the event $\{\langle M\rangle_\infty < \infty\} = \bigcup_k \{T_k = \infty\}$, this implies that $(M_n)_n$ converges a.s.

Conversely, assume that $M_0 = 0$, or otherwise subtract it, and suppose that there exists $K > 0$ such that $|\Delta M_n| \leq K$ for all $n$ a.s. Then by Lemma 9.6.8, for any $n, k \geq 0$ by the martingale property we have $\mathbb{E}[M_{n\wedge T}^2 - \langle M\rangle_{n\wedge T}] = 0$ for any stopping time $T$. Fix $k > 0$ and let $T^k = \inf\{n \geq 0 : |M_n| > k\}$, then $M_{n\wedge T^k} \leq k + K$ and so $\mathbb{E}[\langle M\rangle_{n\wedge T^k}] \leq (k+K)^2$ for all $n$, hence $\uparrow\lim_n \langle M\rangle_{n\wedge T^k} < \infty$ a.s. On the event that $M$ converges we have $\{\sup_n |M_n| < \infty\} = \bigcup_k \{T^k = \infty\}$ and so $\langle M\rangle_\infty < \infty$ a.s.

Let us prove the last claim. The process $H = 1/(1 + \langle M\rangle)$ is bounded by 1 and predictable so $H \bullet M$ is a martingale in $L^2$. Moreover,

$$(H \bullet M)_k - (H \bullet M)_{k-1} = H_k(M_k - M_{k-1}) = \frac{X_k}{1 + \langle M\rangle_k}$$

and since the denominator is $\mathcal{F}_{k-1}$-measurable, then

$$\begin{aligned}
\langle H \bullet M\rangle_n &= \sum_{k=1}^n \mathbb{E}[((H \bullet M)_k - (H \bullet M)_{k-1})^2 \mid \mathcal{F}_{k-1}] \\
&= \sum_{k=1}^n \frac{\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]}{(1 + \langle M\rangle_k)^2} \\
&= \sum_{k=1}^n \frac{\langle M\rangle_k - \langle M\rangle_{k-1}}{(1 + \langle M\rangle_k)^2} \\
&\leq \sum_{k=1}^n \Big( \frac{1}{1 + \langle M\rangle_{k-1}} - \frac{1}{1 + \langle M\rangle_k} \Big) \\
&= 1 - \frac{1}{1 + \langle M\rangle_n},
\end{aligned}$$

where the inequality follows from the fact that $\langle M\rangle$ is nondecreasing. In particular, we see that $\langle H \bullet M\rangle_\infty \leq 1$ so by the first part the series $H \bullet M = \sum_k X_k/(1 + \langle M\rangle_k)$ converges a.s. Kronecker's lemma 9.6.2 applied to $a_n = 1 + \langle M\rangle_n$ and $x_n = X_n$ finally shows that

$$\frac{M_n}{1 + \langle M\rangle_n} \xrightarrow[n\to\infty]{a.s.} 0$$

on the event $\langle M\rangle_n \uparrow \langle M\rangle_\infty = \infty$. □

This theorem applied to sums of independent random variables reads as follows.

**Corollary 9.6.10.** *Let $(X_k)_{k\geq 1}$ be independent random variables with $\mathbb{E}[X_k] = 0$ and $\mathrm{Var}(X_k) = \sigma_k^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$ and $V_n = \sigma_1^2 + \cdots + \sigma_n^2 \uparrow V_\infty \in [0, \infty]$.*

*(i) If $V_\infty < \infty$, then $S_n$ converges a.s. to a finite limit. The converse holds if in addition $\sup_k |X_k| < K$ a.s. for some constant $K < \infty$.*

*(ii) If $V_\infty = \infty$, then $S_n/V_n \longrightarrow 0$ a.s.*

As another corollary, we can extend the Borel–Cantelli lemma further than Corollary 9.5.2.

**Theorem 9.6.11.** *Fix a sequence of events $A_n \in \mathcal{F}_n$ for every $n \geq 1$ and define two nondecreasing processes by:*

$$Y_n = \sum_{k=1}^{n} \mathbb{1}_{A_k} \qquad \text{and} \qquad Z_n = \sum_{k=1}^{n} \mathbb{E}[\mathbb{1}_{A_k} \mid \mathcal{F}_{k-1}].$$

*Then a.s. we have:*

$$(Z_\infty < \infty \implies Y_\infty < \infty) \qquad \text{and} \qquad (Z_\infty = \infty \implies Y_n/Z_n \to 1).$$

*Proof.* By construction, $Z$ is predictable and nondecreasing and $M = Y - Z$ is a martingale with bounded increments, so in $L^2$, hence $Y = M + Z$ is the decomposition of a submartingale. Note that

$$\begin{aligned}
\langle M \rangle_n &= \sum_{k=1}^{n} \mathbb{E}[(M_n - M_{-1})^2 \mid \mathcal{F}_{n-1}] \\
&= \sum_{k=1}^{n} \mathbb{E}[(\mathbb{1}_{A_n} - \mathbb{E}[\mathbb{1}_{A_n} \mid \mathcal{F}_{n-1}])^2 \mid \mathcal{F}_{n-1}] \\
&= \sum_{k=1}^{n} \mathbb{E}[\mathbb{1}_{A_n} \mid \mathcal{F}_{n-1}] - \mathbb{E}[\mathbb{1}_{A_n} \mid \mathcal{F}_{n-1}]^2 \\
&\leq \sum_{k=1}^{n} \mathbb{E}[\mathbb{1}_{A_n} \mid \mathcal{F}_{n-1}] = Z_n.
\end{aligned}$$

Consider now the three cases (we drop the "a.s." everywhere):

- If $Z_\infty < \infty$, then $\langle M \rangle_\infty < \infty$, so $M$ converges by Theorem 9.6.9 and so $Y_\infty \leq \sup_n |M_n| + Z_\infty < \infty$.
- If $Z_\infty = \infty$ and $\langle M \rangle_\infty < \infty$, then $M$ converges again and so $Y_n/Z_n = 1 + M_n/Z_n \to 1$.
- If $Z_\infty = \infty$ and $\langle M \rangle_\infty = \infty$, then $|M_n|/Z_n \leq |M_n|/\langle M \rangle_n \to 0$ and again $Y_n/Z_n = 1 + M_n/Z_n \to 1$. □

Note that when $Z_\infty = \infty$, we have $Y_\infty = \infty$. Therefore we arrive at the dichotomy: a.s. we have,

- either $Z_\infty < \infty$ and then $Y_\infty < \infty$, that is $A_k$ occurs for only finitely many indices $k$,
- or $Z_\infty = \infty$ and then $Y_\infty = \infty$, that is $A_k$ occurs for infinitely many indices $k$.

On the one hand $\mathbb{E}[\mathbb{E}[\mathbb{1}_{A_k} \mid \mathcal{F}_{k-1}]] = \mathbb{P}(A_k)$, so if $\sum_k \mathbb{P}(A_k) < \infty$, then $\mathbb{E}[Z_\infty] < \infty$ so $Y_\infty < \infty$ a.s. On the other hand, if the events are independent, then for $\mathcal{F}_n = \sigma(A_k, k \leq n)$ we have that $\mathbb{E}[\mathbb{1}_{A_k} \mid \mathcal{F}_{k-1}] = \mathbb{P}(A_k)$. Therefore if $\sum_k \mathbb{P}(A_k) = \infty$, then $Y_\infty$ a.s. We thus indeed recover the Borel–Cantelli lemma.

## 9.7 Central Limit Theorems

Let us continue with a Central Limit Theorem for $L^2$ martingales, which generalises the case where the increments are independent in Theorem 2.7.2. We then apply this result to prove the CLT for finite-state Markov chains stated in Theorem 5.1.4.

### 9.7.1 Martingale Central Limit Theorem

As for the case of independent increments, rather than a fixed martingale $(M_n)_{n\geq 0}$, one can consider but a triangular array $(M_{n,k})_{n\geq k\geq 0}$ of martingales, that is for every $n \geq 0$, the path $(M_{n,k})_{k\geq 0}$ is a martingale. We shall not emphasise this in order to keep light notation, but the reader may keep it in mind.

**Theorem 9.7.1** (Martingale Lindeberg's CLT). *Suppose that there exist a positive sequence $a_n \longrightarrow \infty$ and a constant $\sigma^2 > 0$ such that*

$$\frac{\langle M\rangle_n}{a_n} \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2.$$

*Second, assume the Lindeberg condition: for any $\varepsilon > 0$, one has*

$$\frac{1}{a_n}\sum_{k=1}^{n} \mathbb{E}\left[|X_k|^2\, \mathbb{1}_{|\Delta X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\right] \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{9.2}$$

*Then*

$$\frac{M_n}{\sqrt{a_n}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,\sigma^2) \qquad \text{and} \qquad \frac{M_n}{\sqrt{\langle M\rangle_n}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

*If instead $\sigma^2 = 0$, then $a_n^{-1/2} M_n \longrightarrow 0$ in probability.*

As for independent random variables a stronger condition but often simpler to verify is the so-called Lyapunov condition (9.3), which is often checked with $2+\delta = 3$ or 4 in practice (provided such a moment exists). This recovers similarly Theorem 2.7.3 in the case of independent increments.

**Proposition 9.7.2** (Lyapunov's CLT). *Suppose that there exists $\delta > 0$ such that*

$$\frac{1}{a_n^{1+\delta/2}}\sum_{k=1}^{n} \mathbb{E}\left[|X_k|^{2+\delta} \mid \mathcal{F}_{k-1}\right] \xrightarrow[n\to\infty]{} 0. \tag{9.3}$$

*Then* (9.2) *holds.*

*Proof.* We use the conditional Hölder inequality and then the conditional Markov inequality. Indeed, with $p = (2+\delta)/2$ and $q = (2+\delta)/\delta$ so $1/p + 1/q = 1$, we have for every $\varepsilon > 0$:

$$\begin{aligned} a_n^{-1}\,\mathbb{E}\left[|X_k|^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\right] &\le a_n^{-1}\,\mathbb{E}\left[|X_k|^{2+\delta} \mid \mathcal{F}_{k-1}\right]^{1/p} \mathbb{P}(|X_k| > \varepsilon\sqrt{a_n} \mid \mathcal{F}_{k-1})^{1/q} \\ &\le a_n^{-1}\,\mathbb{E}\left[|X_k|^{2+\delta} \mid \mathcal{F}_{k-1}\right]^{1/p} \left((\varepsilon\sqrt{a_n})^{-(2+\delta)}\,\mathbb{E}[|X_k|^{2+\delta} \mid \mathcal{F}_{k-1}]\right)^{1/q} \\ &= \varepsilon^{-\delta} a_n^{-(1+\delta/2)}\,\mathbb{E}[|X_k|^{2+\delta} \mid \mathcal{F}_{k-1}], \end{aligned}$$

and the claim follows. □

The philosophy of the proof of Theorem 9.7.1 is to follow as closely as possible that of Theorem 2.7.2 and we invite the reader to have a look at the latter first. The lack of independence causes several issues but we can deal with them with some tricks.

*Proof of Theorem 9.7.1.* Recall that a random variable $Z$ with the Gaussian law $\mathcal{N}(0,\sigma^2)$ is characterised by its characteristic function $\mathbb{E}[\exp(itZ)] = \exp(-t^2\sigma^2/2)$ for all $t \in \mathbb{R}$. Moreover the pointwise convergence of characteristic functions is equivalent to the convergence in distribution so our claim follows if we prove that for each $t \in \mathbb{R}$ fixed, we have:

$$\mathbb{E}\left[\exp\left(it\frac{M_n}{\sqrt{a_n}}\right)\right] \xrightarrow[n\to\infty]{} \exp\left(-\frac{t^2\sigma^2}{2}\right).$$

Step 1: Reduction to an additional assumption. Fix a constant $C > \sigma^2$. Since $\langle M\rangle_n/a_n \longrightarrow \sigma^2$ in probability then in particular with a probability tending to 1 we have $\langle M\rangle_1 \le \cdots \le \langle M\rangle_n \le Ca_n$. This bound will help us at several occasions so we shall replace $M_n = \sum_{k=1}^n X_k$ by

$$\widetilde{M}_n = \sum_{k=1}^{n} X_k\, \mathbb{1}_{\langle M\rangle_k \le Ca_n} = M_{n\wedge T_n}, \qquad \text{where} \qquad T_n = \inf\{k \ge 0 :\ \langle M\rangle_{k+1} > Ca_n\},$$

is a stopping time. Then $(M_{k\wedge T_n})_{n\geq k\geq 0}$ is an array of martingales, and since $|X_k\,\mathbb{1}_{\langle M\rangle_k\leq Ca_n}|\leq |X_k|$ then it also satisfies the Lindeberg condition (9.2). In addition, on the event $\{\langle M\rangle_n\leq Ca_n\}$ each indicator equals 1 so we have $\widetilde{M}_n=M_n$ and $\langle\widetilde{M}\rangle_n=\langle M\rangle_n$. Hence for any $\varepsilon>0$, it holds:

$$\mathbb{P}(|\langle\widetilde{M}\rangle_n/a_n-\sigma^2|>\varepsilon)\leq\mathbb{P}(|\langle M\rangle_n/a_n-\sigma^2|>\varepsilon)+\mathbb{P}(\langle M\rangle_n/a_n>C)\underset{n\to\infty}{\longrightarrow}0.$$

The stopped martingales thus satisfy the assumptions of the theorem, but also the extra condition $\langle\widetilde{M}\rangle_n\leq Ca_n$ for all $n\geq 1$ by construction. Finally, since $M_n$ and $\widetilde{M}_n$ can only differ when $\langle M\rangle_n>Ca_n$, then we conclude that

$$\begin{aligned}\left|\mathbb{E}\Big[\exp\Big(it\frac{M_n}{\sqrt{a_n}}\Big)\Big]-\mathbb{E}\Big[\exp\Big(it\frac{\widetilde{M}_n}{\sqrt{a_n}}\Big)\Big]\right|&\leq\mathbb{E}\Big[\Big|\exp\Big(it\frac{M_n}{\sqrt{a_n}}\Big)\Big]-\mathbb{E}\Big[\exp\Big(it\frac{\widetilde{M}_n}{\sqrt{a_n}}\Big)\Big|\,\mathbb{1}_{\langle M\rangle_n>Ca_n}\Big]\\&\leq 2\,\mathbb{P}(\langle M\rangle_n>Ca_n),\end{aligned}$$

which converges to 0. Therefore it is sufficient to prove the theorem for $\widetilde{M}_n$, or equivalently, we may assume that almost surely, we have

$$\frac{\langle M\rangle_n}{a_n}\leq C\qquad\text{for all } n\geq 1,\tag{9.4}$$

which we do for the rest of the proof.

Step 2: Proof under the additional assumption. We aim at showing that

$$\mathbb{E}\Big[\exp\Big(it\frac{M_n}{\sqrt{a_n}}+\frac{t^2\sigma^2}{2}\Big)\Big]\underset{n\to\infty}{\longrightarrow}1$$

under the extra assumption (9.4). Since $\langle M\rangle_n/a_n\longrightarrow\sigma^2$ in probability, then it is tempting to replace the latter by the former in the previous expectation and indeed:

$$\left|\mathbb{E}\Big[\exp\Big(it\frac{M_n}{\sqrt{a_n}}+\frac{t^2\sigma^2}{2}\Big)\Big]-\mathbb{E}\Big[\exp\Big(it\frac{M_n}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_n}{2a_n}\Big)\Big]\right|\leq\mathbb{E}\Big[\Big|\exp\Big(\frac{t^2\sigma^2}{2}\Big)-\exp\Big(\frac{t^2\langle M\rangle_n}{2a_n}\Big)\Big|\Big].$$

The term inside the expectation on the right tends to 0 in probability, and the extra assumption (9.4) provides enough domination to conclude that the expectation tends to 0. It thus remains to prove that

$$\mathbb{E}\Big[\exp\Big(it\frac{M_n}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_n}{2a_n}\Big)\Big]\underset{n\to\infty}{\longrightarrow}1,$$

which we shall do using (9.4) again. Indeed, notice that we have a telescopic sum:

$$\begin{aligned}\exp\Big(it\frac{M_n}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_n}{2a_n}\Big)-1&=\sum_{k=1}^n\Big(\exp\Big(it\frac{M_k}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_k}{2a_n}\Big)-\exp\Big(it\frac{M_{k-1}}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_{k-1}}{2a_n}\Big)\Big)\\&=\sum_{k=1}^n\exp\Big(it\frac{M_{k-1}}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_k}{2a_n}\Big)\Big(\exp\Big(it\frac{X_k}{\sqrt{a_n}}\Big)-\exp\Big(-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)\Big).\end{aligned}$$

Notice that in the last sum, each random variable is $\mathcal{F}_{k-1}$-measurable, except $X_k$. Then by conditioning with respect to $\mathcal{F}_{k-1}$, we have:

$$\begin{aligned}&\left|\mathbb{E}\Big[\exp\Big(it\frac{M_{k-1}}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_k}{2a_n}\Big)\Big(\exp\Big(it\frac{X_k}{\sqrt{a_n}}\Big)-\exp\Big(-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)\Big)\Big]\right|\\&=\left|\mathbb{E}\Big[\exp\Big(it\frac{M_{k-1}}{\sqrt{a_n}}+\frac{t^2\langle M\rangle_k}{2a_n}\Big)\Big(\mathbb{E}\Big[\exp\Big(it\frac{X_k}{\sqrt{a_n}}\Big)\,\Big|\,\mathcal{F}_{k-1}\Big]-\exp\Big(-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)\Big)\Big]\right|\\&\leq\exp\Big(\frac{t^2C}{2}\Big)\,\mathbb{E}\Big[\Big|\mathbb{E}\Big[\exp\Big(it\frac{X_k}{\sqrt{a_n}}\Big)\,\Big|\,\mathcal{F}_{k-1}\Big]-\exp\Big(-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)\Big|\Big].\end{aligned}$$

Using the triangle inequality, let us further upper bound the term in last expectation by

$$\left|\mathbb{E}\Big[\exp\Big(it\frac{X_k}{\sqrt{a_n}}\Big)\,\Big|\,\mathcal{F}_{k-1}\Big]-\Big(1-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)\right|+\left|\mathbb{E}\Big[\exp\Big(-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)-\Big(1-\frac{t^2\,\mathbb{E}[X_k^2\mid\mathcal{F}_{k-1}]}{2a_n}\Big)\right|.$$

Applying Lemma 2.7.4 with $n = 2$ (removing the useless factor 1/6) and taking the conditional expectation, we infer that for every $\varepsilon > 0$ and every $u \in \mathbb{R}$ (which plays the role of $t/\sqrt{a_n}$ to lighten the notation), we have since $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$:

$$\begin{aligned}\Big|\mathbb{E}[e^{iuX_k} \mid \mathcal{F}_{k-1}] - \Big(1 - \frac{u^2\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]}{2}\Big)\Big| &= \Big|\mathbb{E}\Big[e^{iuX_k} - \Big(1 + iuX_k - u^2\frac{X_k^2}{2}\Big) \,\Big|\, \mathcal{F}_{k-1}\Big]\Big| \\ &\le \mathbb{E}\big[\min(u^2X_k^2, |u|^3\,|X_k|^3) \mid \mathcal{F}_{k-1}\big] \\ &\le u^2\,\mathbb{E}\big[X_k^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big] + |u|^3\,\mathbb{E}\big[|X_k|^3\,\mathbb{1}_{|X_k|\le\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big] \\ &\le u^2\,\mathbb{E}\big[X_k^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big] + \varepsilon\sqrt{a_n}\,|u|^3\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}].\end{aligned}$$

On the other hand, it is straightforward to show that for $x \ge 0$ it holds $1 - x \le e^{-x} \le 1 - x + x^2/2$, and thus, for every $u \in \mathbb{R}$, we have:

$$\begin{aligned}\Big|\exp\Big(-\frac{u^2\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]}{2}\Big) - \Big(1 - \frac{u^2\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]}{2}\Big)\Big| &\le \frac{1}{2}\Big(\frac{u^2\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]}{2}\Big)^2 \\ &\le \frac{u^4}{8}\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] \sup_{k\le n} \mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}].\end{aligned}$$

Let us note that:

$$\begin{aligned}\sup_{k\le n} \mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] &\le \sup_{k\le n} \mathbb{E}\big[X_k^2\,\mathbb{1}_{|X_k|\le\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big] + \sup_{k\le n} \mathbb{E}\big[X_k^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big] \\ &\le \varepsilon^2 a_n + \sum_{k=1}^n \mathbb{E}\big[X_k^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big].\end{aligned}$$

Let $L(n,\varepsilon) = \sum_{k=1}^n \mathbb{E}[X_k^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}]$, then combining all the previous bounds, with $u = t/\sqrt{a_n}$, we obtain:

$$\begin{aligned}&\Big|\mathbb{E}\Big[\exp\Big(it\frac{M_n}{\sqrt{a_n}} + \frac{t^2\,\langle M\rangle_n}{2a_n}\Big) - 1\Big]\Big| \\ &\le e^{t^2C/2} \sum_{k=1}^n \mathbb{E}\Big[\frac{t^2}{a_n}\,\mathbb{E}\big[X_k^2\,\mathbb{1}_{|X_k|>\varepsilon\sqrt{a_n}} \mid \mathcal{F}_{k-1}\big] + \varepsilon\frac{|t|^3}{a_n}\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}] + \frac{t^4}{8a_n^2}\,\mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]\big(\varepsilon^2 a_n + L(n,\varepsilon)\big)\Big] \\ &\le e^{t^2C/2}\,\mathbb{E}\Big[\frac{t^2}{a_n}L(n,\varepsilon) + \varepsilon|t|^3\frac{\langle M\rangle_n}{a_n} + \frac{t^4}{8}\frac{\langle M\rangle_n}{a_n}(\varepsilon^2 + a_n^{-1}L(n,\varepsilon))\Big].\end{aligned}$$

Now recall that $a_n^{-1}\langle M\rangle_n \longrightarrow \sigma^2$ in probability and $a_n^{-1}L(n,\varepsilon) \longrightarrow 0$ in probability by the Lindeberg condition (9.2). Using the extra assumption (9.4) that $L(n,\varepsilon) \le \langle M\rangle_n \le Ca_n$ almost surely we can apply the dominated convergence theorem once again to infer that the expectation above tends to $\varepsilon|t|^3\sigma^2 + \varepsilon^2 t^2\sigma^2/8$. By letting further $\varepsilon \longrightarrow 0$, we may now conclude. □

### 9.7.2 Markov chain Central Limit Theorem

Let us prove Theorem 5.1.4 by an application of Theorem 9.7.1. Recall the former result: let $(X_n)_{n\ge0}$ be an irreducible Markov chain on a finite set $\mathbb{X}$ with stationary probability $\pi$. For any function $f : \mathbb{X} \longrightarrow \mathbb{R}$, we have the convergence in distribution:

$$\frac{1}{\sqrt{n}}\sum_{k=0}^n \big(f(X_k) - \pi(f)\big) \quad \xrightarrow[n\to\infty]{(d)} \quad \mathcal{N}(0,\sigma^2),$$

where $\sigma^2$ is a constant that we discussed after the statement and that we shall see appearing in the proof. Replacing $f$ by $f - \pi(f)$ if necessary, it is sufficient to consider $f$ such that $\pi(f) = 0$.

The proof uses the solution of the so-called Poisson equation that we put in a separate lemma.

**Lemma 9.7.3.** *Let $P$ be an irreducible transition matrix on a finite state space $\mathbb{X}$, with stationary probability $\pi$. Let $f : \mathbb{X} \to \mathbb{R}$ be such that $\pi(f) = 0$. Then the equation*

$$u - Pu = f \tag{9.5}$$

*has a solution, given explicitly by $u(x) = \sum_{k\geq 0} P^k f(x)$ for every $x \in \mathbb{X}$.*

*Proof.* Let us first prove that $u$ is well-defined. Indeed, recall that in a finite state-space, the Döblin condition is always satisfied, so by Theorem 5.2.13 there exist $k \geq 1$ and $\delta > 0$ such that for any $n \geq 1$ and any initial position $x$, we have:

$$\sum_{y\in\mathbb{X}} |P^n(x,y) - \pi(y)| \leq 2(1-\delta)^{\lfloor n/k \rfloor}.$$

Consequently,

$$\left|\sum_{y\in\mathbb{X}} P^n(x,y)f(y) - \sum_{y\in\mathbb{X}} \pi(y)f(y)\right| \leq (\max f) \sum_{y\in\mathbb{X}} |P^n(x,y) - \pi(y)| \leq 2(\max f)(1-\delta)^{\lfloor n/k \rfloor}.$$

Recall that we assume that $\sum_{y\in\mathbb{X}} \pi(y)f(y) = \pi(f) = 0$, hence $\sum_n |P^n f(x)| < \infty$ and $u(x)$ is well-defined as an absolutely convergent series. The Poisson equation (9.5) then follows easily from the explicit expression:

$$Pu = P\sum_{k\geq 0} P^k f = \sum_{k\geq 0} P^{k+1} f = \sum_{k\geq 1} P^k f = \sum_{k\geq 0} P^k f + f = u - f,$$

which is equivalent to (9.5). □

We shall prove Theorem 5.1.4 by constructing a martingale using the Poisson equation (9.5) and applying Theorem 9.7.1 to the latter.

*Proof of Theorem 5.1.4.* Recall that we assume $\pi(f) = 0$, otherwise simply replace $f$ by $f - \pi(f)$. Let $u$ be the solution to the Poisson equation (9.5) and define for all $n, k \geq 1$:

$$Y_k = u(X_k) - Pu(X_{k-1}) \qquad \text{and then} \qquad M_n = \sum_{k=1}^{n} Y_k.$$

Then observe that:

$$\begin{aligned}
\sum_{k=0}^{n} f(X_k) &= \sum_{k=0}^{n} (u(X_k) - Pu(X_k)) \\
&= \sum_{k=1}^{n} (u(X_k) - Pu(X_{k-1})) - \sum_{k=1}^{n} (Pu(X_k) - Pu(X_{k-1})) \\
&= M_n - (Pu(X_n) - u(X_0)).
\end{aligned}$$

Since $\mathbb{X}$ is a finite state, then $Pu(X_n) - u(X_0)$ is uniformly bounded in $n$, so it vanishes once divided by $\sqrt{n}$ and our claim is thus equivalent to

$$\frac{M_n}{\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2). \tag{9.6}$$

We claim that $(M_n)_n$ is a martingale for the filtration $\mathcal{F}_n = \sigma(X_k, k \leq n)$. Indeed:

$$\begin{aligned}
\mathbb{E}[Y_k \mid \mathcal{F}_{k-1}] &= \mathbb{E}[u(X_k) \mid \mathcal{F}_{k-1}] - Pu(X_{k-1}) \\
&= \mathbb{E}[u(X_k) \mid X_{k-1}] - Pu(X_{k-1}) \qquad \text{(Markov property)} \\
&= Pu(X_{k-1}) - Pu(X_{k-1}) \\
&= 0.
\end{aligned}$$

See the proof of Theorem 8.4.1 for more details on the Markov property in this context. According to Theorem 9.7.1, the convergence (9.6) then holds as soon as $n^{-1}\langle M\rangle_n \longrightarrow \sigma^2$ in probability and that the

Lindeberg condition (9.2) is satisfied with $a_n = n$. This last condition is trivial here since the increments $Y_k$ are uniformly bounded (because $\mathbb{X}$ is finite), so each conditional expectation in (9.2) vanishes when $n$ is large enough. Let us focus on the bracket process $\langle M \rangle_n = \sum_{k=1}^n \mathbb{E}[Y_k^2 \mid \mathcal{F}_{k-1}]$. We have by the Markov property at the third line:

$$\begin{aligned}
\mathbb{E}[Y_k^2 \mid \mathcal{F}_{k-1}] &= \mathbb{E}[(u(X_k) - Pu(X_{k-1}))^2 \mid \mathcal{F}_{k-1}] \\
&= \mathbb{E}[u(X_k)^2 \mid \mathcal{F}_{k-1}] + (Pu(X_{k-1}))^2 - 2Pu(X_{k-1})\,\mathbb{E}[u(X_k) \mid \mathcal{F}_{k-1}] \\
&= \mathbb{E}[u(X_k)^2 \mid X_{k-1}] + (Pu(X_{k-1}))^2 - 2Pu(X_{k-1})\,\mathbb{E}[u(X_k) \mid X_{k-1}] \\
&= Pu(X_{k-1})^2 - (Pu(X_{k-1}))^2 \\
&= \Psi(X_{k-1}),
\end{aligned}$$

where $\Psi = Pu^2 - (Pu)^2$. Applying Corollary 5.1.2 to this function $\Psi$, we obtain:

$$\frac{\langle M \rangle_n}{n} = \frac{1}{n} \sum_{k=1}^n \Psi(X_{k-1}) \xrightarrow[n\to\infty]{a.s.} \pi(\Psi) = \sigma^2.$$

Theorem 9.7.1 then shows that the convergence (9.6) holds and this completes the proof. □

The reader can see that the proof is rather robust and indeed CLT's hold also in infinite state spaces $\mathbb{X}$ but one has to be more careful. For example, the solution to the Poisson equation $u - Pu = f$ may not exist in general. Thus different versions of CLT exist under good conditions.

## 9.8 Stochastic Gradient Descent & Robbins–Monro Algorithm

In many applications, one aims at minimising or maximising a real-valued function which depends on many parameters, and often in a not so explicit way. Such a function could for example quantify some cost in economy, the energy efficiency in some chemical reaction, optimise a transport system, compute a maximum likelihood estimator, etc. Let us see an application of martingales theory to an algorithm solving numerically this deterministic problem. In particular, Corollary 9.1.4 will ensure that our algorithm converges. Let us start with the deterministic setting first. We shall not try to provide the minimal assumption here and satisfy ourselves with easy proofs in simple cases.

### 9.8.1 The Gradient Descent algorithm

The context is the following. Suppose that a parameter $\theta \in \mathbb{R}^d$ produces an effect that we quantify by $f(\theta) \in \mathbb{R}$ and suppose that $f$ has a unique minimiser, i.e. a solution $\theta^* \in \mathbb{R}^d$ of:

$$f(\theta^*) = \min_{\theta \in \mathbb{R}^d} f(\theta).$$

The question is: If $f$ is not explicit, how can we find $\theta^*$ in practice? Not that we do not care about the minimum value $f(\theta^*)$, we only want to be able to choose the optimal parameter $\theta^*$. A well-used numerical scheme is called the *gradient descent*. Recall that the gradient (if it exists) is the vector of partial derivatives: $\nabla f(\theta) = (\partial_1 f(\theta), \dots, \partial_d f(\theta))$. We shall suppose that $\nabla f$ satisfies the so-called separating condition:

$$\langle \theta - \theta^*, \nabla f(\theta) \rangle > 0 \qquad \text{for every } \ \theta \in \mathbb{R}^d \setminus \{\theta^*\}, \tag{9.7}$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^d$. In dimension $d = 1$, this simply means that $f'$ is strictly negative before $\theta^*$ and is strictly positive after $\theta^*$. This condition ensures that $f$ has a unique local minimum, at $\theta^*$, which is the global minimum. Recall in Section 5.3.3 an idea to deal with functions that admit several local minima.

The gradient descent is designed as follows: Fix a sequence $(\gamma_n)_{n\geq 1}$ of real numbers such that:

$$\forall n \geq 1 : \ \gamma_n > 0, \qquad \sum_n \gamma_n = \infty, \qquad \sum_n \gamma_n^2 < \infty,$$

and fix an arbitrary initial value $\theta_0 \in \mathbb{R}^d$. Then recursively define:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f(\theta_{n-1}), \qquad n \geq 1. \tag{9.8}$$

In this scheme we correct step by step the parameter $\theta_n$ by following the slope given by the gradient. Indeed, consider the dimension $d = 1$ for simplicity. Since $\gamma_n > 0$ then either $f'(\theta_{n-1}) > 0$ and then $\theta_n < \theta_{n-1}$, or $f'(\theta_{n-1}) < 0$ and then $\theta_n > \theta_{n-1}$. By (9.7), in both cases we have $f(\theta_n) < f(\theta_{n-1})$. Roughly speaking, we want $\gamma_n \longrightarrow 0$ to ensure that the sequence converges, but $\sum_n \gamma_n = \infty$ to avoid it to converge too fast, before reaching $\theta^*$.

**Proposition 9.8.1.** *Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be continuously differentiable and suppose that $\nabla f$ is bounded and satisfies* (9.7)*. For every $\theta_0 \in \mathbb{R}^d$, the sequence $(\theta_n)_n$ defined by* (9.8) *converges to $\theta^*$.*

*Proof.* Let us write for every $k \geq 1$:

$$\begin{aligned} |\theta_k - \theta^*|^2 &= |\theta_{k-1} - \theta^*|^2 + |\theta_k - \theta_{k-1}|^2 + 2\langle \theta_{k-1} - \theta^*, \theta_k - \theta_{k-1}\rangle \\ &= |\theta_{k-1} - \theta^*|^2 + \gamma_k^2 |\nabla f(\theta_{k-1})|^2 - 2\gamma_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1})\rangle. \end{aligned}$$

Consequently,

$$2\sum_{k=1}^{n} \gamma_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1})\rangle = |\theta_0 - \theta^*|^2 - |\theta_n - \theta^*|^2 + \sum_{k=1}^{n} \gamma_k^2 |\nabla f(\theta_{k-1})|^2.$$

Under (9.7) each term on the left is positive; on the right, the sum is convergent under our assumptions that $\sum_n \gamma_n^2 < \infty$ and that $\nabla f$ is bounded. Therefore the series on the left is convergent and then since both sums converge, the sequence $|\theta_n - \theta^*|^2$ has a finite limit. Further, by (9.7) and since $\nabla f$ is continuous, then for every $\varepsilon > 0$, the quantity:

$$\inf_{\theta \,:\, |\theta - \theta^*| > \varepsilon} \langle \theta - \theta^*, \nabla f(\theta)\rangle$$

is a positive number, say $\eta > 0$. If the limit of $|\theta_n - \theta^*|^2$ is nonzero, then by choosing $\varepsilon$ small enough, we have $|\theta_{k-1} - \theta^*| > \varepsilon$ for every $k$ large enough, say $k \geq k_\varepsilon$, which leads to:

$$\infty > \sum_{k\geq 1} \gamma_k \langle \theta_{k-1} - \theta^*, \nabla f(\theta_{k-1})\rangle \geq \varepsilon \sum_{k \geq k_\varepsilon} \gamma_k = \infty$$

by our assumption. We conclude that $|\theta_n - \theta^*|^2$ converges to 0. □

### 9.8.2 Stochastic Gradient Descent

Often in practice, the effect of the parameter $\theta$ is random, say $F(X, \theta)$ for some function $F$ and some random variable $X$, and we aim at minimising $f(\theta) = \mathbb{E}[F(X, \theta)]$ as before, but we can only observe $F(X, \theta)$. Let us assume that $\nabla f(\theta) = \mathbb{E}[\nabla F(X, \theta)]$. One typical application the reader may have in mind is the response $F(X, \theta)$ of a patient to a dose of medicine $\theta$ (with possibly several components): we aim at finding the dose that provides the best average response over patients but we can only observe the effect on each patient.

A naive idea to find $\theta^*$ here is to use the same numerical scheme as in (9.8), but at every step, approximate $\nabla f(\theta_{n-1})$ using the observable $\nabla F(X, \theta_{n-1})$. Formally, let $(X_{n,i})_{n,i\geq 0}$ be i.i.d. random variables with the same law as $X$ and let

$$\widehat{\nabla f}(\theta_{n-1})_k = \frac{1}{k}\sum_{i=1}^{k} \nabla F(X_{n,i}, \theta_{n-1}),$$

which converges a.s. to $\nabla f(\theta_{n-1})$ by the Law of Large Numbers. Then we may use the scheme (9.8) with $\nabla f(\theta_{n-1})$ being replaced by $\widehat{\nabla f}(\theta_{n-1})_k$ for some large $k$. This may however not be convenient in practice for it requires a large number $n \times k$ of random variables of the form $X_{n,i}$. If we think again of the medicine tested on patients, this requires a lot of trials! Also even for computer simulations, especially when the dimension $d$ of $\theta$ is large, this may take too long computation time. This is typically the case in neural networks.

Instead, the Robbins–Monro algorithm constructs a random sequence $(\Theta_n)_n$ which in some sense allows to take $k = 1$ above. Precisely let $(X_n)_{n\geq 1}$ be i.i.d. copies of $X$, let $\mathcal{F}_n = \sigma(X_k, k \leq n)$ and construct recursively starting from $\Theta_0 \in \mathbb{R}^d$ an adapted process by:

$$\Theta_n = \Theta_{n-1} - \gamma_n \nabla F(X_n, \Theta_{n-1}), \qquad \text{where} \qquad \forall n \geq 1 : \ \gamma_n > 0, \quad \sum_n \gamma_n = \infty, \quad \sum_n \gamma_n^2 < \infty. \tag{9.9}$$

**Theorem 9.8.2.** *Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be continuously differentiable; suppose that $\nabla f(\theta) = \mathbb{E}[\nabla F(X, \theta)]$ where $\nabla F$ is bounded and suppose that (9.7) holds. For every $\Theta_0 \in \mathbb{R}^d$, the sequence $(\Theta_n)_n$ defined by (9.9) converges almost surely to $\theta^*$.*

The proof is quite close to that of Proposition 9.8.1. The almost sure convergence is provided by a supermartingale uniformly bounded below as in Corollary 9.1.4.

*Proof.* Let us write for every $k \geq 1$:

$$\begin{aligned} |\Theta_k - \theta^*|^2 &= |\Theta_{k-1} - \theta^*|^2 + |\Theta_k - \Theta_{k-1}|^2 + 2\langle \Theta_{k-1} - \theta^*, \Theta_k - \Theta_{k-1}\rangle \\ &= |\Theta_{k-1} - \theta^*|^2 + \gamma_k^2 |\nabla F(X_k, \Theta_{k-1})|^2 - 2\gamma_k \langle \Theta_{k-1} - \theta^*, \nabla F(X_k, \Theta_{k-1})\rangle. \end{aligned}$$

Consequently, the sequence

$$M_n = |\Theta_n - \theta^*|^2 - \sum_{k=1}^n \gamma_k^2 \, \mathbb{E}[|\nabla F(X_k, \Theta_{k-1})|^2 \mid \mathcal{F}_{k-1}],$$

satisfies:

$$\begin{aligned} \mathbb{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] &= \mathbb{E}[|\Theta_n - \theta^*|^2 - |\Theta_{n-1} - \theta^*|^2 - \gamma_n^2 |\nabla F(X_n, \Theta_{n-1})|^2 \mid \mathcal{F}_{n-1}] \\ &= -2\gamma_n \langle \Theta_{n-1} - \theta^*, \mathbb{E}[\nabla F(X_n, \Theta_{n-1}) \mid \mathcal{F}_{n-1}]\rangle \\ &= -2\gamma_n \langle \Theta_{n-1} - \theta^*, \nabla f(\Theta_{n-1})\rangle, \end{aligned}$$

where the last equality follows from Theorem 6.5.4. The assumption (9.7) therefore implies that $(M_n)_n$ is a supermartingale. Since we assume that $\sum_n \gamma_n^2 < \infty$ and that $\nabla F$ is bounded, then this supermartingale is bounded below by a constant and thus it converges almost surely by Corollary 9.1.4 to a finite limit. The sum in the definition of $M_n$ also has a finite limit, again because $\sum_n \gamma_n^2 < \infty$ and $\nabla F$ is bounded. Hence $|\Theta_n - \theta^*|$ converges almost surely to a finite limit. It remains to prove that this limit is 0.

By similar arguments, we have:

$$2\,\mathbb{E}\Big[\sum_{k=1}^n \gamma_k \langle \Theta_{k-1} - \theta^*, \nabla f(\Theta_{k-1})\rangle\Big] = \mathbb{E}[|\Theta_0 - \theta^*|^2 - |\Theta_n - \theta^*|^2] + \sum_{k=1}^n \gamma_k^2 \, \mathbb{E}\Big[|\nabla F(X_k, \Theta_{k-1})|^2\Big],$$

and the last sum converges. Consequently $\sum_k \gamma_k \langle \Theta_{k-1} - \theta^*, \nabla f(\Theta_{k-1})\rangle < \infty$ almost surely. Now recall that $\sum_k \gamma_k = \infty$, so we must have $\langle \Theta_{k-1} - \theta^*, \nabla f(\Theta_{k-1})\rangle \longrightarrow 0$ almost surely. On the other hand, by (9.7) and since $\nabla f$ is continuous, then for every $\varepsilon > 0$ we have:

$$\inf\{\langle \theta - \theta^*, \nabla f(\theta)\rangle : \ \theta \in \mathbb{R}^d, |\theta - \theta^*| \geq \varepsilon\} > 0.$$

So the fact that $\langle \Theta_{k-1} - \theta^*, \nabla f(\Theta_{k-1})\rangle \longrightarrow 0$ implies that $\limsup_k |\Theta_{k-1} - \theta^*| \leq \varepsilon$ for every $k$ large enough. Taking the intersection over $\varepsilon \in \mathbb{Q}$, we conclude that almost surely, we have $\limsup_k |\Theta_{k-1} - \theta^*| = 0$. □