Graduate Texts in Mathematics

GTM

Jean-François Le Gall

Measure Theory, Probability, and Stochastic Processes

Springer

Graduate Texts in Mathematics 295

Graduate Texts in Mathematics

Graduate Texts in Mathematics bridge the gap between passive study and creative understanding, offering graduate-level introductions to advanced topics in mathematics. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Although these books are frequently used as textbooks in graduate courses, they are also suitable for individual study.

Jean-François Le Gall

Measure Theory, Probability, and Stochastic Processes

Springer

Jean-François Le Gall
Institut de Mathématique d'Orsay
University of Paris-Saclay
Orsay, France

ISSN 0072-5285 ISSN 2197-5612 (electronic)
Graduate Texts in Mathematics
ISBN 978-3-031-14204-8 ISBN 978-3-031-14205-5 (eBook)
https://doi.org/10.1007/978-3-031-14205-5

Mathematics Subject Classification: 60-01, 60G07, 28-01

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is based on lecture notes for a series of lectures given at the Ecole normale supérieure de Paris. The goal of these lectures was first to provide a concise but comprehensive presentation of measure theory, then to introduce the fundamental notions of modern probability theory, and finally to illustrate these concepts by the study of some important classes of stochastic processes. This text therefore consists of three parts of approximately the same size. In the first part, we present measure theory with a view toward the subsequent applications to probability theory. After introducing the basic concepts of a measure on a measurable space and of a measurable function, we construct the integral of a real function with respect to a measure, and we establish the main convergence theorems of the theory, including the monotone convergence theorem and the Lebesgue dominated convergence theorem. In the subsequent chapter, we use the notion of an outer measure to give a short and efficient construction of Lebesgue measure. We then discuss L^p spaces, with an application to the important Radon-Nikodym theorem, before turning to measures on product spaces and to the celebrated Fubini theorem. We next introduce signed measures and prove the classical Jordan decomposition theorem, and we also give an application to the classical L^p–L^q duality theorem. We conclude this part with a chapter on the change of variables formula, which is a key tool for many concrete calculations. In view of our applications to probability theory, we have chosen to present the "abstract" approach to measure theory, in contrast with the functional analytic approach. Concerning the latter approach, we only state without proofs two versions of the Riesz-Markov-Kakutani representation theorem, which is not used elsewhere in the book.

The second part of the book is devoted to the basic notions of probability theory. We start by introducing the concepts of a random variable defined on a probability space, and of the mathematical expectation of a random variable. Although these concepts are just special cases of corresponding notions in measure theory, we explain how the point of view of probability theory is different. In particular the notion of the pushforward of a measure leads to the fundamental definition of the law or distribution of a random variable. We then provide a thorough discussion of the notion of independence, and of its relations with measures on product spaces.

Although we have chosen not to develop the theory of measures on infinite product spaces, we briefly explain how Lebesgue measure makes it possible to construct infinite sequences of independent real random variables, as this is sufficient for all subsequent applications including the construction of Brownian motion. We then study the different types of convergence of random variables and give proofs of the law of large numbers and the central limit theorem, which are the most famous limit theorems of probability theory. The last chapter of this part is devoted to the definition and properties of the conditional expectation of a real random variable given some partial information represented by a sub-σ-field of the underlying probability space. Conditional expectations are the key ingredient needed for the definition and study of the most important classes of stochastic processes.

Finally, the third part of the book discusses three fundamental types of stochastic processes. We start with discrete-time martingales, which may be viewed as providing models for the evolution of the fortune of a player in a fair game. Martingales are ubiquitous in modern probability theory, and one could say that (almost) any probability question can be solved by finding the right martingale. We prove the basic convergence theorems of martingale theory and the optional stopping theorem, which, roughly speaking, says that, independently of the player's strategy, the mean value of the fortune at the end of the game will coincide with the initial one. We give several applications including a short proof of the strong law of large numbers. The next chapter is devoted to Markov chains with values in a countable state space. The key concept underlying the notion of a Markov chain is the Markov property, which asserts that the past does not give more information than the present, if one wants to predict the future evolution of the random process in consideration. The Markov property allows one to characterize the evolution of a Markov chain in terms of the so-called transition matrix, and to derive many remarkable asymptotic properties. We give specific applications to important classes of Markov chains such as random walks and branching processes. In the last chapter, we study Brownian motion, which is now a random process indexed by a continuous time parameter. We give a complete construction of (d-dimensional) Brownian motion and motivate this construction by the study of random walks over a long time interval. We investigate remarkable properties of the sample paths of Brownian motion, as well as certain explicit related distributions. We conclude this chapter with a thorough discussion of the relations between Brownian motion and harmonic functions, which is perhaps the most beautiful connection between probability theory and analysis.

The first two parts of the book should be read linearly, even though the chapter on signed measures may be skipped at first reading. A good understanding of the notions presented in the four chapters of Part II is definitely required for anybody who aims to study more advanced topics of probability theory. In contrast with the other two parts, the three chapters of Part III can be read (almost) independently, but we have tried to emphasize the relevance of certain fundamental concepts in different settings. In particular, martingales appear in the Markov chain theory in connection with discrete harmonic functions, and the same connection (involving continuous-time martingales) occurs in the study of Brownian motion. Similarly, the strong Markov property is a fundamental tool in the study of both Markov chains and Brownian motion.

The prerequisite for this book is a good knowledge of advanced calculus (set-theoretic manipulations, real analysis, metric spaces). For the reader's convenience we have recalled the basic notions of the theory of Banach spaces, including elementary Hilbert space theory, in an appendix. Apart from these prerequisites, the book is essentially self-contained and appropriate for self-study. Detailed proofs are given for all statements, with a couple of exceptions (the two versions of the Riesz-Markov-Kakutani theorem, the existence of conditional distributions) which are not used elsewhere.

A number of exercises are listed at the end of every chapter, and the reader is strongly advised to try at least some of them. Some of these exercises are straightforward applications of the main theorems, but some of them are more involved and often lead to statements that are of interest though they could not be included in the text. Most of these exercises were taken from exercise sessions at the Ecole normale supérieure, and I am indebted to Nicolas Curien, Grégory Miermont, Gilles Stoltz, and Mathilde Weill for making these problems available to me. It is a pleasure to thank Mireille Chaleyat-Maurel for her careful reading of the manuscript. And finally, last but not least, I am indebted to several anonymous reviewers whose numerous remarks helped me to improve the text.

Orsay, France Jean-François Le Gall
May 2022

Contents

List of Symbols

$\mathbb{N}$	$= \{1, 2, 3, \ldots\}$
$\mathbb{Z}$	$= \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Z}_+$	$= \{0, 1, 2, \ldots\}$
$\mathbb{R}$	the set of all real numbers
$\mathbb{R}_+$	the set of all nonnegative real numbers
$\overline{\mathbb{R}}$	$= \mathbb{R} \cup \{-\infty + \infty\} = [-\infty, +\infty]$
$\mathbb{Q}$	the set of all rational numbers
$\mathbb{C}$	the set of all complex numbers
$x \wedge y$	$= \min\{x, y\}$ (for $x, y \in \mathbb{R}$)
$x \vee y$	$= \max\{x, y\}$ (for $x, y \in \mathbb{R}$)
x^+	$= x \vee 0$ (for $x \in \mathbb{R}$)
x^-	$= (-x) \vee 0$ (for $x \in \mathbb{R}$)
$x \cdot y$	Euclidean scalar product of $x, y \in \mathbb{R}^d$
$\lfloor x \rfloor$	integer part of $x \in \mathbb{R}$ (the largest integer $n \leq x$)
$\mathbb{B}_d$	closed unit ball of $\mathbb{R}^d$
$\mathbb{S}^{d-1}$	unit sphere of $\mathbb{R}^d$
$B(x, r)$	open ball of radius r centered at $x \in \mathbb{R}^d$
$\bar{B}(x, r)$	closed ball of radius r centered at $x \in \mathbb{R}^d$
$\Gamma(\cdot)$	the Gamma function (Chapter 7)
J_φ	Jacobian of a continuously differentiable function φ (Chapter 7)
$\mathcal{P}(E)$	set of all subsets of the set E
A^c	complement of the subset A of E
$A \backslash B$	$= A \cap B^c$
$\mathbf{1}_A$	indicator function of the set A
$\mathrm{card}(A)$	cardinality of the set A
$\mathcal{B}(E)$	Borel σ-field of a metric (or topological) space E (Chapter 1)
$\sigma(\mathcal{C})$	σ-field generated by a class $\mathcal{C}$ of subsets (Chapter 1)
$\sigma(X)$	σ-field generated by a random variable X (Chapter 8)
$\mathcal{F} \vee \mathcal{G}$	$= \sigma(\mathcal{F} \cup \mathcal{G})$ (for $\mathcal{F}, \mathcal{G}$ σ-fields of E)

$\liminf A_n$	$= \bigcup_{n=1}^{\infty} \Big(\bigcap_{k=n}^{\infty} A_k \Big)$ (for $A_n \subset E$, $n \in \mathbb{N}$)
$\limsup A_n$	$= \bigcap_{n=1}^{\infty} \Big(\bigcup_{k=n}^{\infty} A_k \Big)$ (for $A_n \subset E$, $n \in \mathbb{N}$)
δ_x	Dirac measure at x (Chapter 1)
$\mathcal{F} \otimes \mathcal{G}$	product σ-field of $\mathcal{F}$ and $\mathcal{G}$ (Chapters 1, 5)
$\mu \otimes \nu$	product measure of the measures μ and ν (Chapter 5)
$\nu \ll \mu$	ν is absolutely continuous with respect to μ (Chapter 4)
$\nu \perp \mu$	ν is singular with respect to μ (Chapter 4)
$\widehat{f}, \widehat{\mu}$	Fourier transform of the function f, resp. the measure μ (Chapter 2)
$g \cdot \mu$	measure with density g with respect to the measure μ (Chapter 2)
$\lambda(\cdot)$, $\lambda_d(\cdot)$	Lebesgue measure on $\mathbb{R}^d$ (Chapter 3)
$\int f \, \mathrm{d}\mu$	$= \int f(x)\, \mu(\mathrm{d}x)$: integral of f with respect to μ (Chapter 2)
$L^p(E, \mathcal{A}, \mu)$	Lebesgue space of measurable functions for which the p-th power of the absolute value is integrable (Chapter 4)
$\\|f\\|_p$	L^p-norm of the function f (Chapter 4)
$\mathrm{dist}(x, F)$	$= \inf\{\|x - y\| : y \in F\}$, for $x \in \mathbb{R}^d$, $F \subset \mathbb{R}^d$
$\mathrm{dist}(F, F')$	$= \inf\{\|y - y'\| : y \in F, y' \in F\}$, for $F, F' \subset \mathbb{R}^d$
$(\Omega, \mathcal{A}, \mathbb{P})$	underlying probability space (Chapter 8)
$\mathbb{E}[X]$	$= \int_\Omega X(\omega)\, \mathbb{P}(\mathrm{d}\omega)$: expected value of X (Chapter 8)
$\mathbb{P}_X$	law or distribution of the random variable X (Chapter 8)
Φ_X	characteristic function of the random variable X (Chapter 8)
F_X	distribution function of the random variable X (Chapter 8)
$\mathrm{var}(X)$	variance of the random variable X (Chapter 8)
K_X	covariance matrix of the random variable X (Chapter 8)
$\mathbb{E}[X \mid \mathcal{B}]$	conditional expectation of X given the σ-field $\mathcal{B}$ (Chapter 11)
$C_c(E)$	space of all real-valued continuous functions with compact support on E
$C_b(E)$	space of all real-valued bounded continuous functions on E
$C_0(E)$	space of all real-valued continuous functions on E that tend to 0 at infinity
$C(\mathbb{R}_+, \mathbb{R}^d)$	space of all continuous functions from $\mathbb{R}_+$ into $\mathbb{R}^d$
$M_1(\mathbb{R}^d)$	space of all probability measures on $\mathbb{R}^d$
$(\mathbf{\Omega}, \mathfrak{F}, \mathbf{P}_x)$	canonical probability space for the construction of a Markov chain (Chapter 13), or of Brownian motion (Chapter 14), with initial point x

Part I
Measure Theory

Chapter 1
Measurable Spaces

The basic idea of measure theory is to assign a nonnegative real number to every subset of a given set. This number is called the measure of the subset and is required to satisfy a certain additivity property (informally, the measure of a disjoint union should be the sum of the measures of the sets in this union). This additivity property is natural if one thinks of a measure as an abstract generalization of the familiar notions of length, area or volume. For deep mathematical reasons, it is not possible in general to define the measure of every subset, and one has to restrict to a certain class (σ-field) of subsets, which is called the class of measurable sets. A set given with a σ-field is called a measurable space.

This chapter introduces the fundamental notions of a σ-field, of a measure on a measurable space, and of a measurable function, which will be used in Chapter 2 when we construct the Lebesgue integral. In view of our forthcoming applications to probability, it is important that we develop an "abstract" measure theory, where no additional structure is required on the underlying space. The last section of the chapter states the so-called monotone class theorem, which plays an important role both in measure theory and in probability theory. Roughly speaking, the monotone class theorem allows one to show that a property known to hold for certain special sets holds in fact for all measurable sets.

1.1 Measurable Sets

Measurable sets are the "regular" sets of measure theory. We introduce them in an abstract setting.

Let E be a set. The set of all subsets of E is denoted by $\mathcal{P}(E)$. We use the notation A^c for the complement of a subset A of E. If A and B are two subsets of E, we write $A\backslash B = A \cap B^c$.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_1

Definition 1.1 A σ-field $\mathcal{A}$ on E is a collection of subsets of E which satisfies the following properties:

(i) $E \in \mathcal{A}$;
(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$;
(iii) If $A_n \in \mathcal{A}$ for every $n \in \mathbb{N}$, then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

The elements of $\mathcal{A}$ are called measurable sets, or sometimes $\mathcal{A}$-measurable sets if there is ambiguity, and we say that $(E, \mathcal{A})$ is a measurable space.

Let us give a few easy consequences of the definition:

(1) $\varnothing \in \mathcal{A}$.
(2) If $A_n \in \mathcal{A}$ for every $n \in \mathbb{N}$, one has also $\bigcap_{n \in \mathbb{N}} A_n \in \mathcal{A}$ (use (ii) and (iii)).
(3) Since we may take $A_n = \varnothing$ for all $n > n_0$, property (iii) implies that $\mathcal{A}$ is closed under finite unions, meaning that the union of finitely many elements of $\mathcal{A}$ is in $\mathcal{A}$, and then $\mathcal{A}$ is also closed under finite intersections using (ii).

Examples

- $\mathcal{A} = \mathcal{P}(E)$. This is the σ-field that we (almost) always consider when E is finite or countable, but not in other cases.
- $\mathcal{A} = \{\varnothing, E\}$ is the trivial σ-field.
- The collection of all subsets of E that are at most countable, or whose complement is at most countable, forms a σ-field on E.

In view of producing more interesting examples, we note that the intersection of an arbitrary collection of σ-fields on E is a σ-field. This leads to the following definition.

Definition 1.2 Let $\mathcal{C}$ be a subset of $\mathcal{P}(E)$. The smallest σ-field on E containing $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the σ-field obtained as the intersection of all σ-fields on E that contain $\mathcal{C}$. We call $\sigma(\mathcal{C})$ the σ-field generated by $\mathcal{C}$.

If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are σ-fields on E, we will use the notation $\mathcal{A}_1 \vee \cdots \vee \mathcal{A}_k = \sigma(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_k)$ for the smallest σ-field that contains $\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_k$.

The Borel σ-Field To illustrate the interest of the notion of the σ-field generated by a class of sets, let us consider the case where E is a topological space.

Definition 1.3 Suppose that E is a metric space, or more generally a topological space, and let $\mathcal{O}$ be the class of all open subsets of E. The σ-field $\sigma(\mathcal{O})$ is called the Borel σ-field on E and is denoted by $\mathcal{B}(E)$.

The Borel σ-field on E is thus the smallest σ-field on E that contains all open sets. The elements of $\mathcal{B}(E)$ are called Borel subsets of E. Clearly, closed subsets of E are Borel sets. The σ-field $\mathcal{B}(\mathbb{R})$ is also generated by the class of all intervals

(a, b) for $a, b \in \mathbb{R}$, $a < b$, or even by the class of all $(-\infty, a)$ for $a \in \mathbb{R}$, see Exercise 1.1.

In what follows, whenever we consider a topological space, for instance $\mathbb{R}$ or $\mathbb{R}^d$, we will assume unless otherwise indicated that it is equipped with its Borel σ-field.

The Product σ-Field Another very important example of the notion of the σ-field generated by a class of sets is the product σ-field.

Definition 1.4 Let $(E_1, \mathcal{A}_1)$ and $(E_2, \mathcal{A}_2)$ be two measurable spaces. The product σ-field $\mathcal{A}_1 \otimes \mathcal{A}_2$ is the σ-field on $E_1 \times E_2$ defined by

$$\mathcal{A}_1 \otimes \mathcal{A}_2 := \sigma(A_1 \times A_2\,;\ A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2).$$

In other words, $\mathcal{A}_1 \otimes \mathcal{A}_2$ is the smallest σ-field that contains all "rectangles" $A_1 \times A_2$ for $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.

Let us state a useful technical lemma. Recall that a metric space is said to be separable if it contains a dense sequence.

Lemma 1.5 *Suppose that E and F are separable metric spaces, and equip the product $E \times F$ with the product topology. Then $\mathcal{B}(E \times F) = \mathcal{B}(E) \otimes \mathcal{B}(F)$.*

Proof The property $\mathcal{B}(E) \otimes \mathcal{B}(F) \subset \mathcal{B}(E \times F)$ holds without the separability assumption. To verify this property, first fix an open subset B of F. Then the class of all sets $A \in \mathcal{B}(E)$ such that $A \times B \in \mathcal{B}(E \times F)$ contains all open subsets of E, and is easily checked to be a σ-field. It follows that this class must contain $\mathcal{B}(E)$ and thus be equal to $\mathcal{B}(E)$. So we have proved that $A \times B \in \mathcal{B}(E \times F)$ if $A \in \mathcal{B}(E)$ and B is open. We then fix $A \in \mathcal{B}(E)$ and observe similarly that the class of all $B \in \mathcal{B}(F)$ such that $A \times B \in \mathcal{B}(E \times F)$ contains all open subsets of F and is a σ-field, so that this class must be equal to $\mathcal{B}(F)$. Finally we have proved that $A \times B \in \mathcal{B}(E \times F)$ if $A \in \mathcal{B}(E)$ and $B \in \mathcal{B}(F)$, which gives the inclusion $\mathcal{B}(E) \otimes \mathcal{B}(F) \subset \mathcal{B}(E \times F)$.

Conversely, we observe that we can find a countable collection $\mathcal{U} = \{U_n : n \in \mathbb{N}\}$ of open subsets of E, such that every open subset of E is the union of a subcollection of $\mathcal{U}$ (if $(x_k)_{k \in \mathbb{N}}$ is a dense sequence in E, we may define $\mathcal{U}$ as the collection of all open balls of rational radius centered at one of the points x_k). Let $\mathcal{V} = \{V_n : n \in \mathbb{N}\}$ be a similar collection for F. For every open set O of $E \times F$ and every $z = (x, y) \in O$, we know that O contains an open set of the form $U \times V$, where U, resp. V, is an open subset of E, resp. of F, containing x, resp. y. It follows that O can be written as the union of (at most countably many) sets of the form $U_n \times V_m$, where $n, m \in \mathbb{N}$. Hence, every open subset of $E \times F$ belongs to $\mathcal{B}(E) \otimes \mathcal{B}(F)$, and the definition of the Borel σ-field then implies $\mathcal{B}(E \times F) \subset \mathcal{B}(E) \otimes \mathcal{B}(F)$. □

In particular, we have $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$.

1.2 Positive Measures

Let $(E, \mathcal{A})$ be a measurable space.

Definition 1.6 A positive measure on $(E, \mathcal{A})$ is a mapping $\mu : \mathcal{A} \longrightarrow [0, \infty]$ which satisfies the following properties:

(i) $\mu(\varnothing) = 0$;
(ii) (*σ-additivity*) For every sequence $(A_n)_{n\in\mathbb{N}}$ of disjoint measurable sets,

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

The triple $(E, \mathcal{A}, \mu)$ is called a measure space.
The quantity $\mu(E)$ is called the total mass of the measure μ.

In what follows, we often say measure instead of positive measure. It is important that we allow the value $+\infty$: the sum $\sum_{n\in\mathbb{N}} \mu(A_n)$ always makes sense as a number in $[0, \infty]$. Similarly, if $(a_n)_{n\in\mathbb{N}}$ is an increasing sequence (resp. a decreasing sequence) in $[-\infty, \infty]$, the increasing limit (resp. the decreasing limit) of $(a_n)_{n\in\mathbb{N}}$ exists in $[-\infty, \infty]$ and will be denoted by $\lim_{n\to\infty} \uparrow a_n$ (resp. $\lim_{n\to\infty} \downarrow a_n$).

Property (ii) is called σ-additivity. It is crucial that we restrict our attention to countable collections $(A_n)_{n\in\mathbb{N}}$ in this property. Of course, property (ii) includes the case where all A_n's are empty for $n > n_0$, which corresponds to the property of finite additivity.

Properties

(1) If $A \subset B$, then $\mu(A) \leq \mu(B)$. If in addition $\mu(A) < \infty$, then

$$\mu(B\backslash A) = \mu(B) - \mu(A)\,.$$

(2) If $A, B \in \mathcal{A}$,

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

(3) (*Continuity from below*) If $A_n \in \mathcal{A}$ and $A_n \subset A_{n+1}$, for every $n \in \mathbb{N}$,

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \lim_{n\to\infty} \uparrow \mu(A_n)\,.$$

(4) (*Continuity from above*) If $B_n \in \mathcal{A}$ and $B_{n+1} \subset B_n$ for every $n \in \mathbb{N}$, and if $\mu(B_1) < \infty$, then

$$\mu\Big(\bigcap_{n\in\mathbb{N}} B_n\Big) = \lim_{n\to\infty} \downarrow \mu(B_n)\,.$$

(5) If $A_n \in \mathcal{A}$, for every $n \in \mathbb{N}$, then

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \leq \sum_{n\in\mathbb{N}} \mu(A_n)\,.$$

Properties (1) and (2) are easy and we omit the proof. Let us prove (3), (4) and (5). For (3), we set $C_1 = A_1$ and, for every $n \geq 2$,

$$C_n = A_n \backslash A_{n-1}$$

in such a way that $\bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} C_n$. Since the C_n's are disjoint,

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \mu\Big(\bigcup_{n\in\mathbb{N}} C_n\Big) = \sum_{n\in\mathbb{N}} \mu(C_n) = \lim_{N\to\infty} \uparrow \sum_{n=1}^{N} \mu(C_n) = \lim_{N\to\infty} \uparrow \mu(A_N).$$

For (4), we set $A_n = B_1 \backslash B_n$ for every n, so that the sequence $(A_n)_{n\in\mathbb{N}}$ is increasing. Then,

$$\begin{aligned}
\mu(B_1) - \mu\Big(\bigcap_{n\in\mathbb{N}} B_n\Big) = \mu\Big(B_1 \backslash \bigcap_{n\in\mathbb{N}} B_n\Big) &= \mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \\
&= \lim_{n\to\infty} \uparrow \mu(A_n) \\
&= \lim_{n\to\infty} \uparrow (\mu(B_1) - \mu(B_n)) \\
&= \mu(B_1) - \lim_{n\to\infty} \downarrow \mu(B_n).
\end{aligned}$$

The condition $\mu(B_1) < \infty$ is used (in particular) to obtain that $\mu(A_n) = \mu(B_1) - \mu(B_n)$.

Finally, for (5), we set $C_1 = A_1$, and then, for every $n \geq 2$,

$$C_n = A_n \backslash \bigcup_{k=1}^{n-1} A_k.$$

The sets C_n are disjoint, and thus

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \mu\Big(\bigcup_{n\in\mathbb{N}} C_n\Big) = \sum_{n\in\mathbb{N}} \mu(C_n) \leq \sum_{n\in\mathbb{N}} \mu(A_n).$$

Examples

(a) Let $E = \mathbb{N}$, and $\mathcal{A} = \mathcal{P}(\mathbb{N})$. The counting measure is defined by

$$\mu(A) := \text{card}(A).$$

(This definition of the counting measure makes sense on the measurable space $(E, \mathcal{P}(E))$ for any set E.) This example shows that the condition $\mu(B_1) < \infty$ in property (4) above cannot be omitted. Indeed, if we take

$$B_n = \{n, n+1, n+2, \ldots\}$$

for every $n \in \mathbb{N}$, we have $\mu(B_n) = \infty$ whereas $\bigcap_{n\in\mathbb{N}} B_n = \varnothing$ and thus $\mu(\bigcap_{n\in\mathbb{N}} B_n) = 0$.

(b) If A is a subset of E, the indicator function of A is the function $\mathbf{1}_A$ defined on E and taking values in $\{0, 1\}$, such that, for $x \in E$,

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

Let $(E, \mathcal{A})$ be a measurable space, and fix $x \in E$. Setting $\delta_x(A) = \mathbf{1}_A(x)$ for every $A \in \mathcal{A}$ defines a probability measure on E, which is called the Dirac measure (or sometimes the Dirac mass) at x. More generally, if $(x_n)_{n\in\mathbb{N}}$ is a sequence of points of E and $(\alpha_n)_{n\in\mathbb{N}}$ is a sequence in $[0, \infty]$, we can consider the measure $\sum_{n\in\mathbb{N}} \alpha_n \delta_{x_n}$ defined by

$$\Big(\sum_{n\in\mathbb{N}} \alpha_n \delta_{x_n}\Big)(A) := \sum_{n\in\mathbb{N}} \alpha_n \delta_{x_n}(A) = \sum_{n\in\mathbb{N}} \alpha_n \mathbf{1}_A(x_n).$$

This is called a *point measure* on E.

(c) *Lebesgue measure*. There exists a unique positive measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that, for every compact interval $[a, b]$ of $\mathbb{R}$, we have $\lambda\big([a, b]\big) = b - a$. The existence of this measure will be established in Chapter 3, and the fact that there is at most one such measure is discussed at the end of the present chapter.

Restriction of a Measure If μ is a positive measure on $(E, \mathcal{A})$, and $C \in \mathcal{A}$, the restriction of μ to C is the measure ν on $(E, \mathcal{A})$ defined by

$$\nu(A) := \mu(A \cap C) \;,\; \forall A \in \mathcal{A}.$$

Definitions

- The measure μ is said to be *finite* if $\mu(E) < \infty$.
- The measure μ is a *probability measure* if $\mu(E) = 1$, and we then say that $(E, \mathcal{A}, \mu)$ is a probability space.
- The measure μ is called *σ-finite* if there exists a sequence $(E_n)_{n\in\mathbb{N}}$ of measurable sets such that $E = \bigcup_{n\in\mathbb{N}} E_n$ and $\mu(E_n) < \infty$ for every n. (Clearly the sequence $(E_n)_{n\in\mathbb{N}}$ can be taken to be increasing or the E_n's can be assumed to be disjoint.)
- We say that $x \in E$ is an *atom* of the measure μ if $\mu(\{x\}) > 0$ (here we implicitly assume that singletons belong to the σ-field $\mathcal{A}$, which will always be the case in our applications).
- The measure μ is called *diffuse* if it has no atoms.

We conclude this section with a useful lemma concerning the limsup and the liminf of a sequence of measurable sets. If $(A_n)_{n\in\mathbb{N}}$ is a sequence of measurable subsets of E, we set

$$\limsup A_n = \bigcap_{n=1}^{\infty}\left(\bigcup_{k=n}^{\infty} A_k\right), \quad \liminf A_n = \bigcup_{n=1}^{\infty}\left(\bigcap_{k=n}^{\infty} A_k\right).$$

It is immediate that $\liminf A_n \subset \limsup A_n$. Also recall that, if $(a_n)_{n\in\mathbb{N}}$ is a sequence of elements of $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, we define

$$\limsup a_n := \lim_{n\to\infty} \downarrow \left(\sup_{k\geq n} a_k\right), \ \liminf a_n := \lim_{n\to\infty} \uparrow \left(\inf_{k\geq n} a_k\right),$$

where both limits exist in $\bar{\mathbb{R}}$ ($\limsup a_n$ and $\liminf a_n$ are respectively the greatest and the smallest accumulation points of the sequence $(a_n)_{n\in\mathbb{N}}$ in $\overline{\mathbb{R}}$).

Lemma 1.7 *Let μ be a measure on $(E, \mathcal{A})$. Then,*

$$\mu(\liminf A_n) \leq \liminf \mu(A_n).$$

If μ is finite, or more generally if $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) < \infty$, we have also

$$\mu(\limsup A_n) \geq \limsup \mu(A_n).$$

Proof For every $n \in \mathbb{N}$,

$$\mu\left(\bigcap_{k=n}^{\infty} A_k\right) \leq \inf_{k\geq n} \mu(A_k),$$

and thus, using the continuity from below (property (3) above),

$$\mu(\liminf A_n) \leq \lim_{n\to\infty} \inf_{k\geq n} \mu(A_k) = \liminf \mu(A_n).$$

The proof of the second part of the lemma is similar, but we now rely on the continuity from above (property (4)), and use the finiteness assumption to justify the application of this property. □

Remark Example (a) above shows that the finiteness assumption cannot be omitted in the second part of the lemma.

1.3 Measurable Functions

We now turn to measurable functions, which are the functions of interest in measure theory.

Definition 1.8 Let $(E, \mathcal{A})$ and $(F, \mathcal{B})$ be two measurable spaces. A function $f : E \longrightarrow F$ is said to be measurable if

$$\forall B \in \mathcal{B}\,,\; f^{-1}(B) \in \mathcal{A}.$$

When E and F are two topological spaces equipped with their respective Borel σ-fields, we also say that f is Borel measurable (or is a Borel function).

Remark To emphasize the dependence on the σ-field $\mathcal{A}$, we often say that f is $\mathcal{A}$-measurable. Usually, the choice of $\mathcal{B}$ will be clear from the context, but this will not necessarily be the case for the σ-field $\mathcal{A}$ (especially in probability theory).

Proposition 1.9 *The composition of two measurable functions is measurable.*

This is immediate by writing $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$.

The following criterion of measurability is often useful.

Proposition 1.10 *Let $(E, \mathcal{A})$ and $(F, \mathcal{B})$ be two measurable spaces, and consider a function $f : E \longrightarrow F$. In order for f to be measurable, it suffices that there exists a subclass $\mathcal{C}$ of $\mathcal{B}$ such that $\sigma(\mathcal{C}) = \mathcal{B}$ and the property $f^{-1}(B) \in \mathcal{A}$ holds for every $B \in \mathcal{C}$.*

Proof Set

$$\mathcal{G} = \{B \in \mathcal{B} : f^{-1}(B) \in \mathcal{A}\}.$$

Then it is straightforward to verify that $\mathcal{G}$ is a σ-field. By our assumption, $\mathcal{C} \subset \mathcal{G}$. It follows that $\mathcal{G}$ contains $\sigma(\mathcal{C}) = \mathcal{B}$, which was the desired result. □

Example Suppose that $(F, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. To show that f is measurable, it suffices to prove that the sets $f^{-1}((a, b))$ are measurable, for every $a < b$. It is even enough to consider the sets $f^{-1}((-\infty, a))$ for $a \in \mathbb{R}$.

Corollary 1.11 *Suppose that E and F are two topological spaces equipped with their respective Borel σ-fields. Then any continuous function $f : E \longrightarrow F$ is Borel measurable*

Proof We apply Proposition 1.10 to the class $\mathcal{C}$ of open subsets of F. □

Operations on Measurable Functions

Lemma 1.12 *Let $(E, \mathcal{A})$, $(F_1, \mathcal{B}_1)$ and $(F_2, \mathcal{B}_2)$ be measurable spaces, and equip the product $F_1 \times F_2$ with the product σ-field $\mathcal{B}_1 \otimes \mathcal{B}_2$. Let $f_1 : E \longrightarrow F_1$ and $f_2 : E \longrightarrow F_2$ be two functions and define $f : E \longrightarrow F_1 \times F_2$ by setting $f(x) = (f_1(x), f_2(x))$ for every $x \in E$. Then f is measurable if and only if both f_1 and f_2 are measurable.*

Proof The "only if" part is very easy and left to the reader. For the "if" part, we apply Proposition 1.10 to the class

$$\mathcal{C} = \{B_1 \times B_2 \, ; \; B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}.$$

Since $f^{-1}(B_1 \times B_2) = f_1^{-1}(B_1) \cap f_2^{-1}(B_2) \in \mathcal{A}$ if $B_1 \in \mathcal{B}_1$ and $B_2 \in \mathcal{B}_2$, we immediately get the desired result. □

Corollary 1.13 *Let $(E, \mathcal{A})$ be a measurable space, and let f and g be measurable functions from E into $\mathbb{R}$. Then the functions $f + g$, fg, $\inf(f, g)$, $\sup(f, g)$, $f^+ = \sup(f, 0)$, $f^- = \sup(-f, 0)$ are measurable.*

The proof is easy. For instance, $f + g$ is the composition of the two functions $x \longrightarrow (f(x), g(x))$ and $(a, b) \longrightarrow a + b$ which are both measurable (the second one because it is continuous, using also the equality $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$). The other cases are left to the reader.

Recall the notation $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ for the extended real line, which is equipped with its usual topology. Similarly as in the case of $\mathbb{R}$, the Borel σ-field of $\overline{\mathbb{R}}$ is generated by the intervals $[-\infty, a)$ for $a \in \mathbb{R}$.

Proposition 1.14 *Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of measurable functions from E into $\bar{\mathbb{R}}$. Then,*

$$\sup_{n \in \mathbb{N}} f_n \, , \; \inf_{n \in \mathbb{N}} f_n \, , \; \limsup_{n \to \infty} f_n \, , \; \liminf_{n \to \infty} f_n$$

are also measurable functions. In particular, if the sequence (f_n) *converges pointwise, its limit* $\lim f_n$ *is a measurable function.*

Proof Set $f(x) = \inf_{n\in\mathbb{N}} f_n(x)$. To prove that f is measurable, it enough to verify that $f^{-1}([-\infty, a)) \in \mathcal{A}$ for every $a \in \mathbb{R}$ (use Proposition 1.10). However,

$$f^{-1}([-\infty, a)) = \{x : \inf f_n(x) < a\} = \bigcup_{n\in\mathbb{N}} \{x : f_n(x) < a\}$$

giving the desired property. The case of $\sup_{n\in\mathbb{N}} f_n$ is treated similarly.

Then, we immediately get that

$$\liminf_{n\to\infty} f_n = \sup_{n\geq 0} \left(\inf_{k\geq n} f_k \right)$$

is measurable. □

The following technical lemma is often useful.

Lemma 1.15 *Let* $(f_n)_{n\in\mathbb{N}}$ *be a sequence of measurable functions from* E *into* $\mathbb{R}$. *The set* A *of all* $x \in E$ *such that* $f_n(x)$ *converges in* $\mathbb{R}$ *as* $n \to \infty$ *is measurable. Furthermore, the function* $h : E \longrightarrow \mathbb{R}$ *defined by*

$$h(x) := \begin{cases} \lim_{n\to\infty} f_n(x) & \text{if } x \in A, \\ 0 & \text{if } x \notin A, \end{cases}$$

is measurable.

Proof For the first assertion, let G be the measurable function from E into $\overline{\mathbb{R}}^2$ defined by $G(x) := (\liminf f_n(x), \limsup f_n(x))$, and set $\Delta = \{(x, x) : x \in \mathbb{R}\}$. Then

$$A = \{x \in E : -\infty < \liminf f_n(x) = \limsup f_n(x) < \infty\} = G^{-1}(\Delta).$$

The measurability of A follows since Δ is a measurable subset of $\overline{\mathbb{R}}^2$ (note that $\{(x, x) : x \in \overline{\mathbb{R}}\}$ is measurable as a closed subset of $\overline{\mathbb{R}}^2$).

Then let $F \in \mathcal{B}(\mathbb{R})$ with $0 \notin F$. We observe that

$$h^{-1}(F) = A \cap \{x \in E : \limsup f_n(x) \in F\}$$

and use the previous proposition to get that $h^{-1}(F)$ is measurable. If $0 \in F$ we just write $h^{-1}(F)^c = h^{-1}(F^c)$. □

Pushforward of a Measure The following definition is especially important in probability theory as it will lead to the fundamental notion of the law of a random variable (Definition 8.2).

Definition 1.16 Let $(E, \mathcal{A})$ and $(F, \mathcal{B})$ be two measurable spaces and let $\varphi : E \longrightarrow F$ be a measurable function. Let μ be a positive measure on $(E, \mathcal{A})$. The formula

$$\nu(B) = \mu(\varphi^{-1}(B)), \ \forall B \in \mathcal{B}$$

defines a positive measure ν on $(F, \mathcal{B})$, which is called the pushforward of μ under φ and denoted by $\varphi(\mu)$, or sometimes by $\varphi_* \mu$.

The fact that $\nu(B) = \mu(\varphi^{-1}(B))$ defines a positive measure on $(F, \mathcal{B})$ is easy and left to the reader. The measures μ and $\varphi(\mu)$ have the same total mass, but it may happen that μ is σ-finite while $\varphi(\mu)$ is not (for instance, if μ is σ-finite and not finite, and φ is a constant function).

1.4 Monotone Class

In this section, we state and prove the monotone class theorem, which is a fundamental tool in measure theory and even more in probability theory.

Definition 1.17 A subset $\mathcal{M}$ of $\mathcal{P}(E)$ is called a monotone class if

(i) $E \in \mathcal{M}$.
(ii) If $A, B \in \mathcal{M}$ and $A \subset B$, then $B \backslash A \in \mathcal{M}$.
(iii) If $A_n \in \mathcal{M}$ and $A_n \subset A_{n+1}$ for every $n \in \mathbb{N}$, then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{M}$.

Any σ-field is also a monotone class. Conversely, a monotone class $\mathcal{M}$ is a σ-field if and only if it is closed under finite intersections. Indeed, if $\mathcal{M}$ is closed under finite intersections, then by considering the complementary sets we get that $\mathcal{M}$ is closed under finite unions, and using property (iii) that $\mathcal{M}$ is also closed under countable unions.

As in the case of σ-fields, it is clear that any intersection of monotone classes is a monotone class. If $\mathcal{C}$ is any subset of $\mathcal{P}(E)$, we can thus define the monotone class $\mathcal{M}(\mathcal{C})$ generated by $\mathcal{C}$ as the intersection of all monotone classes of E that contain $\mathcal{C}$.

Theorem 1.18 (Monotone Class Theorem) *Let $\mathcal{C} \subset \mathcal{P}(E)$ be closed under finite intersections. Then $\mathcal{M}(\mathcal{C}) = \sigma(\mathcal{C})$. Consequently, if $\mathcal{M}$ is any monotone class such that $\mathcal{C} \subset \mathcal{M}$, we have also $\sigma(\mathcal{C}) \subset \mathcal{M}$.*

Remark on the Terminology Monotone classes defined in Definition 1.17 are often called λ-systems, and classes closed under finite intersections are called π-systems. The preceding theorem is then known as the $\pi - \lambda$ theorem.

Proof It is enough to prove the first assertion. Since any σ-field is a monotone class, it is clear that $\mathcal{M}(\mathcal{C}) \subset \sigma(\mathcal{C})$. To prove the reverse inclusion, it is enough to verify that $\mathcal{M}(\mathcal{C})$ is a σ-field. As explained above, it suffices to verify that $\mathcal{M}(\mathcal{C})$ is closed under finite intersections.

For every $A \in \mathcal{P}(E)$, set

$$\mathcal{M}_A = \{B \in \mathcal{M}(\mathcal{C}) : A \cap B \in \mathcal{M}(\mathcal{C})\}.$$

Fix $A \in \mathcal{C}$. Since $\mathcal{C}$ is closed under finite intersections, we have $\mathcal{C} \subset \mathcal{M}_A$. Let us verify that $\mathcal{M}_A$ is a monotone class:

- $E \in \mathcal{M}_A$ is trivial.
- If $B, B' \in \mathcal{M}_A$ and $B \subset B'$, then $A \cap (B' \backslash B) = (A \cap B') \backslash (A \cap B) \in \mathcal{M}(\mathcal{C})$ and thus $B' \backslash B \in \mathcal{M}_A$.
- If $B_n \in \mathcal{M}_A$ for every $n \geq 0$ and the sequence $(B_n)_{n \geq 0}$ is increasing, we have $A \cap (\cup_{n \in \mathbb{N}} B_n) = \cup_{n \in \mathbb{N}} (A \cap B_n) \in \mathcal{M}(\mathcal{C})$ and therefore $\cup_{n \in \mathbb{N}} B_n \in \mathcal{M}_A$.

Since $\mathcal{M}_A$ is a monotone class that contains $\mathcal{C}$, $\mathcal{M}_A$ also contains $\mathcal{M}(\mathcal{C})$. We have thus obtained that

$$\forall A \in \mathcal{C},\ \forall B \in \mathcal{M}(\mathcal{C}),\ A \cap B \in \mathcal{M}(\mathcal{C}).$$

This is not yet the desired result, but we can use the same idea one more time. Precisely, we now fix $A \in \mathcal{M}(\mathcal{C})$. According to the first part of the proof, $\mathcal{C} \subset \mathcal{M}_A$. By exactly the same arguments as in the first part of the proof, we get that $\mathcal{M}_A$ is a monotone class. It follows that $\mathcal{M}(\mathcal{C}) \subset \mathcal{M}_A$, which shows that $\mathcal{M}(\mathcal{C})$ is closed under finite intersections, and completes the proof. □

Corollary 1.19 *Let μ and ν be two measures on $(E, \mathcal{A})$. Suppose that there exists a class $\mathcal{C} \subset \mathcal{A}$, which is closed under finite intersections, such that $\sigma(\mathcal{C}) = \mathcal{A}$ and $\mu(A) = \nu(A)$ for every $A \in \mathcal{C}$.*

(1) *If $\mu(E) = \nu(E) < \infty$, then we have $\mu = \nu$.*
(2) *If there exists an increasing sequence $(E_n)_{n \in \mathbb{N}}$ of elements of $\mathcal{C}$ such that $E = \bigcup_{n \in \mathbb{N}} E_n$ and $\mu(E_n) = \nu(E_n) < \infty$ for every $n \in \mathbb{N}$, then $\mu = \nu$.*

Proof

(1) Let $\mathcal{G} = \{A \in \mathcal{A} : \mu(A) = \nu(A)\}$. By assumption, $\mathcal{C} \subset \mathcal{G}$. On the other hand, it is easy to verify that $\mathcal{G}$ is a monotone class. For instance, if $A, B \in \mathcal{G}$ and $A \subset B$, we have $\mu(B \backslash A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \backslash A)$, and hence $B \backslash A \in E$ (note that we here use the fact that both μ and ν are finite).

    If follows that $\mathcal{G}$ contains $\mathcal{M}(\mathcal{C})$, which is equal to $\sigma(\mathcal{C})$ by the monotone class theorem. By our assumption $\sigma(\mathcal{C}) = \mathcal{A}$, we get $\mathcal{G} = \mathcal{A}$, which means that $\mu = \nu$.

(2) For every $n \in \mathbb{N}$, let μ_n and ν_n be the respective restrictions of μ and ν to E_n. Part (1) applies to μ_n and ν_n, and we get $\mu_n = \nu_n$. Finally, for every $A \in \mathcal{A}$,

$$\mu(A) = \lim \uparrow \mu(A \cap E_n) = \lim \uparrow \nu(A \cap E_n) = \nu(A). \qquad \square$$

Consequences *Uniqueness of Lebesgue measure.* If there exists a positive measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda\big([a, b]\big) = b - a$ for every reals $a \leq b$, then this measure is necessarily unique. Indeed, supposing that λ' is another measure satisfying the same property as λ, we can apply part (2) of the corollary to $\mu = \lambda$ and $\nu = \lambda'$, taking for $\mathcal{C}$ the class of all compact intervals (which generates the Borel σ-field) and $E_n = [-n, n]$ for every $n \in \mathbb{N}$.

The same argument shows that, for every $r > 0$, the pushforward of Lebesgue measure under the mapping $x \mapsto rx$ is $r^{-1}\lambda$.

In a similar manner, one deduces from the previous corollary that a finite measure μ on $\mathbb{R}$ is characterized by the values of $\mu((-\infty, a])$ for every $a \in \mathbb{R}$.

1.5 Exercises

Exercise 1.1 Check that the σ-field $\mathcal{B}(\mathbb{R})$ is also generated by the class of all intervals (a, b), $a, b \in \mathbb{R}$, $a < b$, or by the class of all $(-\infty, a)$, $a \in \mathbb{R}$, or even by the intervals $(-\infty, a)$, $a \in \mathbb{Q}$ (one can also replace open intervals by closed intervals).

Exercise 1.2 For every integer $n \geq 1$, let $\mathcal{F}_n$ be the σ-field on $\mathbb{N}$ defined by $\mathcal{F}_n := \sigma(\{1\}, \{2\}, \ldots, \{n\})$. Show that $(\mathcal{F}_n)_{n\geq 1}$ is an increasing sequence of σ-fields of $\mathbb{N}$, but $\cup_{n\geq 1}\mathcal{F}_n$ is not a σ-field.

Exercise 1.3 Let $C([0, 1], \mathbb{R}^d)$ be the space of all continuous functions from $[0, 1]$ into $\mathbb{R}^d$, which is equipped with the topology induced by the sup norm. Let $\mathcal{C}_1$ be the Borel σ-field on $C([0, 1], \mathbb{R}^d)$, and let $\mathcal{C}_2$ be the smallest σ-field on $C([0, 1], \mathbb{R}^d)$ such that all functions $f \mapsto f(t)$, for $t \in [0, 1]$, are measurable from $(C([0, 1], \mathbb{R}^d), \mathcal{C}_2)$ into $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (justify the existence of this smallest σ-field).

(1) Show that $\mathcal{C}_2 \subset \mathcal{C}_1$.
(2) Show that any open ball in $C([0, 1], \mathbb{R}^d)$ is in $\mathcal{C}_2$.
(3) Conclude that $\mathcal{C}_2 = \mathcal{C}_1$.

Exercise 1.4 Let $(E, \mathcal{A}, \mu)$ be a measure space, with $\mu(E) > 0$, and let $f : E \longrightarrow \mathbb{R}$ be a measurable function. Show that, for every $\varepsilon > 0$, there exists a measurable set $A \in \mathcal{A}$ such that $\mu(A) > 0$ and $|f(x) - f(y)| < \varepsilon$ for every $x, y \in A$.

Exercise 1.5 Let E be an arbitrary set, and let $\mathcal{A}$ be a σ-field on E. Show that $\mathcal{A}$ cannot be countably infinite. (*Hint:* By contradiction, suppose that $\mathcal{A}$ is countably infinite and introduce the *atoms* of $\mathcal{A}$ defined for every $x \in E$ by

$$E_x := \bigcap_{A \in \mathcal{A}, x \in A} A.$$

Then show that $E_x = E_y$ defines an equivalence relation on E, and that the equivalence classes for this relation form a partition of E by measurable sets.)

Exercise 1.6 (Egoroff's Theorem) Let $(E, \mathcal{A})$ be a measurable space and let μ be a finite measure on E. Let $(f_n)_{n \in \mathbb{R}}$ be a sequence of measurable functions from E into $\mathbb{R}$, and assume that $f_n(x) \longrightarrow f(x)$ for every $x \in E$. Show that, for every $\varepsilon > 0$, there exists $A \in \mathcal{A}$ such that $\mu(E \backslash A) < \varepsilon$ and the convergence of the sequence f_n to f holds uniformly on A. *Hint:* Consider, for every integers $k, n \geq 1$, the set

$$E_{k,n} = \bigcap_{j=n}^{\infty} \left\{ x \in E : |f_j(x) - f(x)| \leq \frac{1}{k} \right\}.$$

Exercise 1.7 Let $f : [0, 1] \longrightarrow \mathbb{R}$ be a continuous function. For every $y \in \mathbb{R}$, let $N(y) \in \{0, 1, 2, \ldots\} \cup \{\infty\}$ be the number of solutions of the equation $f(x) = y$. Show that $y \mapsto N(y)$ is a measurable function.

Exercise 1.8 This exercise uses the existence of Lebesgue measure on $\mathbb{R}$, which is the unique measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda([a, b]) = b - a$ for every $a \leq b$ (cf. Chapter 3).

(1) Let $\varepsilon > 0$. Construct a dense open subset O_ε of $\mathbb{R}$ such that $\lambda(O_\varepsilon) < \varepsilon$.
(2) Infer that there exists a closed subset F_ε of $\mathbb{R}$ with empty interior such that $\lambda(A \cap F_\varepsilon) \geq \lambda(A) - \varepsilon$ for every $A \in \mathcal{B}(\mathbb{R})$.

Chapter 2
Integration of Measurable Functions

In this chapter, we construct the Lebesgue integral of real-valued measurable functions with respect to a positive measure. The definition of the Lebesgue integral is very natural and easy for (nonnegative) simple functions, which take only finitely many values. Then the integral of a nonnegative measurable function f is obtained as the supremum of integrals of simple functions g that are bounded above by f. For functions taking both positive and negative values, the definition proceeds by linearity, but one must restrict to integrable functions f, which are such that the integral of $|f|$ is finite.

After constructing the integral of measurable functions, we establish the three main convergence theorems, namely the monotone convergence theorem, Fatou's lemma and the dominated convergence theorem. These remarkably simple statements are of constant use in applications of the Lebesgue integral, both in measure theory and in probability theory. Roughly speaking, they provide conditions that ensure that the integral of the limit of a sequence of measurable functions is equal to the limit of the integrals of the functions in the sequence. The last section gives typical applications to the continuity and differentiability of integrals of functions depending on a parameter. Important special cases of these applications are the Fourier transform and the convolution of functions.

2.1 Integration of Nonnegative Functions

Throughout this chapter, we consider a measurable space $(E, \mathcal{A})$ equipped with a positive measure μ.

Simple Functions A measurable function $f : E \longrightarrow \mathbb{R}$ is called a simple function if it takes only finitely many values. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the distinct values taken

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_2

by f. We may assume that these values are ranked in increasing order, $\alpha_1 < \alpha_2 < \cdots < \alpha_n$, and write

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

where, for every $i \in \{1, \ldots, n\}$, we have set $A_i = f^{-1}(\{\alpha_i\}) \in \mathcal{A}$, and we recall that $\mathbf{1}_A$ stands for the indicator function of the set A. We note that E is the disjoint union of $A_1, \ldots, A_n$. The formula $f = \sum_{i=1}^{n} \alpha_i \, \mathbf{1}_{A_i}$ will be called the canonical representation of f.

Definition 2.1 Let f be a simple function taking values in $\mathbb{R}_+$ with canonical representation $f = \sum_{i=1}^{n} \alpha_i \, \mathbf{1}_{A_i}$. The integral of f with respect to μ is defined by

$$\int f \, \mathrm{d}\mu := \sum_{i=1}^{n} \alpha_i \, \mu(A_i)$$

with the particular convention $0 \times \infty = 0$ in the case where $\alpha_i = 0$ and $\mu(A_i) = \infty$. In particular, for every $A \in \mathcal{A}$,

$$\int \mathbf{1}_A \mathrm{d}\mu = \mu(A).$$

The sum $\sum_{i=1}^{n} \alpha_i \, \mu(A_i)$ makes sense as an element of $[0, \infty]$, and the value ∞ may occur if $\mu(A_i) = \infty$ for some i. The fact that we consider only *nonnegative* simple functions avoids having to consider expressions such has $\infty - \infty$. The convention $0 \times \infty = 0$ will be in force throughout this book.

Suppose that we have another expression of the simple function f in the form

$$f = \sum_{j=1}^{m} \beta_j \, \mathbf{1}_{B_j}$$

where the measurable sets B_j still form a partition of E but the reals β_j are no longer necessarily distinct. Then it is easy to verify that we have also

$$\int f \, \mathrm{d}\mu = \sum_{j=1}^{m} \beta_j \, \mu(B_j).$$

Indeed, for every $i \in \{1, \dots, n\}$, A_i must be the disjoint union of the sets B_j for indices j such that $\beta_j = \alpha_i$. By the additivity property of the measure μ, we have thus

$$\mu(A_i) = \sum_{\{j:\beta_j=\alpha_i\}} \mu(B_j)$$

and this leads to the desired result.

Properties Let f and g be nonnegative simple functions on E.

(1) For every $a, b \in \mathbb{R}_+$,

$$\int (af + bg)\mathrm{d}\mu = a \int f\,\mathrm{d}\mu + b \int g\,\mathrm{d}\mu.$$

(2) If $f \leq g$, meaning that $f(x) \leq g(x)$ for every $x \in E$,

$$\int f\,\mathrm{d}\mu \leq \int g\,\mathrm{d}\mu.$$

Proof

(1) Let

$$f = \sum_{i=1}^{n} \alpha_i\, \mathbf{1}_{A_i}\ ,\ g = \sum_{k=1}^{m} \alpha'_k\, \mathbf{1}_{A'_k}$$

be the canonical representations of f and g. By writing each set A_i as the disjoint union of the sets $A_i \cap A'_k$, $k \in \{1, \dots, m\}$, and similarly each set A'_k as the disjoint union of the sets $A_i \cap A'_k$, $i \in \{1, \dots, n\}$, we see that we can write

$$f = \sum_{j=1}^{p} \beta_j\, \mathbf{1}_{B_j}\ ,\ g = \sum_{j=1}^{p} \gamma_j\, \mathbf{1}_{B_j}$$

with the *same* partition $B_1, \dots, B_p$ of E (but the numbers β_j, resp. the numbers γ_j, are no longer necessarily distinct). By the remark following the definition of $\int f\,\mathrm{d}\mu$, we have

$$\int f\,\mathrm{d}\mu = \sum_{j=1}^{p} \beta_j\, \mu(B_j)\ ,\ \int g\,\mathrm{d}\mu = \sum_{j=1}^{p} \gamma_j\, \mu(B_j).$$

and similarly $\int (af + bg)\mathrm{d}\mu = \sum_{j=1}^{p} (a\beta_j + b\gamma_j)\, \mu(B_j)$. The desired result follows.

(2) By part (1), we have

$$\int g\, \mathrm{d}\mu = \int f\, \mathrm{d}\mu + \int (g - f)\mathrm{d}\mu \geq \int f\, \mathrm{d}\mu. \qquad \square$$

We let $\mathcal{E}_+$ stand for the space of all nonnegative simple functions.

Definition 2.2 Let $f : E \longrightarrow [0, \infty]$ be a measurable function. The integral of f with respect to μ is defined by

$$\int f\, \mathrm{d}\mu := \sup_{h \in \mathcal{E}_+, h \leq f} \int h\, \mathrm{d}\mu\, .$$

Property (2) above shows that this definition is consistent with the preceding one when f a simple function. We note that we allow f to take the value $+\infty$. This turns out to be very useful in practice.

The integral of f with respect to μ is often written in different ways. The expressions

$$\int f\, \mathrm{d}\mu, \int f(x)\, \mathrm{d}\mu(x), \int f(x)\mu(\mathrm{d}x), \int \mu(\mathrm{d}x) f(x)$$

all refer to the same quantity. Also, if A is a measurable subset of E, it will be convenient to write

$$\int_A f\, \mathrm{d}\mu = \int \mathbf{1}_A f\, \mathrm{d}\mu.$$

Important Convention In the remaining part of this chapter, "nonnegative measurable function" means "measurable function with values in $[0, \infty]$" (note however that simple functions take finite values by definition).

Proposition 2.3 *Let f and g be nonnegative measurable functions on E.*

(1) *If $f \leq g$, $\int f\, \mathrm{d}\mu \leq \int g\, \mathrm{d}\mu$.*
(2) *If $\mu(\{x \in E : f(x) > 0\}) = 0$, then $\int f\, \mathrm{d}\mu = 0$.*

Property (1) is obvious from the definition. As for property (2), it is enough to verify it when f is a simple function (if $h \in \mathcal{E}_+$ and $h \leq f$, $\{x \in E : h(x) > 0\} \subset \{x \in E : f(x) > 0\}$) and then this is also obvious from the definition.

Theorem 2.4 (Monotone Convergence Theorem) *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative measurable functions on E such that $f_n \le f_{n+1}$ for every $n \in \mathbb{N}$. For every $x \in E$, set $f(x) = \lim \uparrow f_n(x)$. Then,*

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} \uparrow \int f_n \,\mathrm{d}\mu.$$

Proof The function f is measurable by Proposition 1.14. By property (1) above, we have

$$\int f \,\mathrm{d}\mu \ge \lim_{n\to\infty} \uparrow \int f_n \,\mathrm{d}\mu$$

and it thus suffices to establish the reverse inequality. To this end, let $h = \sum_{i=1}^{k} \alpha_i \mathbf{1}_{A_i}$ be a nonnegative simple function such that $h \le f$. Let $a \in [0, 1)$, and, for every $n \in \mathbb{N}$, set

$$E_n = \{x \in E : ah(x) \le f_n(x)\}.$$

Then E_n is measurable. Furthermore, using the fact that f_n increases to f, and the condition $a < 1$, we obtain that E is the increasing union of the sequence E_n.

Then, the definition of E_n readily gives the inequality $f_n \ge a\mathbf{1}_{E_n} h$, hence

$$\int f_n \,\mathrm{d}\mu \ge \int a\mathbf{1}_{E_n} h \,\mathrm{d}\mu = a \sum_{i=1}^{k} \alpha_i \,\mu(A_i \cap E_n).$$

Since E is the increasing union of the sequence $(E_n)_{n\in\mathbb{N}}$, we get that A_i is also the increasing union of the sequence $(A_i \cap E_n)_{n\in\mathbb{N}}$, for every $i \in \{1, \dots, k\}$, and consequently $\mu(A_i \cap E_n) \uparrow \mu(A_i)$ as $n \to \infty$. It follows that

$$\lim_{n\to\infty} \uparrow \int f_n \,\mathrm{d}\mu \ge a \sum_{i=1}^{k} \alpha_i \,\mu(A_i) = a \int h \,\mathrm{d}\mu.$$

By letting a tend to 1, we get

$$\lim_{n\to\infty} \uparrow \int f_n \,\mathrm{d}\mu \ge \int h \,\mathrm{d}\mu.$$

Since $\int f \,\mathrm{d}\mu$ is the supremum of the quantities in the right-hand side when h varies among $\{g \in \mathcal{E}_+ : g \le f\}$, we obtain the desired inequality. □

Proposition 2.5

(1) *Let f be a nonnegative measurable function on E. There exists an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of nonnegative simple functions such that $f_n \longrightarrow f$*

pointwise as $n \to \infty$*. If* f *is also bounded, the sequence* $(f_n)_{n\in\mathbb{N}}$ *can be chosen to converge uniformly to* f*.*

(2) *Let* f *and* g *be two nonnegative measurable functions on* E *and* $a, b \in \mathbb{R}_+$*. Then,*

$$\int (af + bg)\mathrm{d}\mu = a \int f \,\mathrm{d}\mu + b \int g \,\mathrm{d}\mu.$$

(3) *Let* $(f_n)_{n\in\mathbb{N}}$ *be a sequence of nonnegative measurable functions on* E*. Then*

$$\int \Big(\sum_{n\in\mathbb{N}} f_n\Big)\mathrm{d}\mu = \sum_{n\in\mathbb{N}} \int f_n \,\mathrm{d}\mu.$$

Remark $\sum_{n\in\mathbb{N}} f_n$ is an increasing limit of nonnegative measurable functions and thus also measurable.

Proof

(1) For every $n \in \mathbb{N}$, for every $i \in \{0, 1, \dots, n2^n - 1\}$, set

$$A_n = \{x \in E : f(x) \geq n\}$$
$$B_{n,i} = \{x \in E : i2^{-n} \leq f(x) < (i+1)2^{-n}\}.$$

Consider then, for every $n \in \mathbb{N}$, the simple function

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

One easily verifies that $(f_n)_{n\in\mathbb{N}}$ is an increasing sequence that converges to f. If f is bounded, we have $0 \leq f - f_n \leq 2^{-n}$ as soon as n is large enough.

(2) By (1), we can find two increasing sequences $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ of nonnegative simple functions that converge to f and g respectively. Then $af + bg$ is also the increasing limit of $af_n + bg_n$ as $n \to \infty$. Hence, by using the monotone convergence theorem and the properties of integrals of simple functions, we have

$$\int (af + bg)\mathrm{d}\mu = \lim_{n\to\infty} \uparrow \int (af_n + bg_n)\mathrm{d}\mu = \lim_{n\to\infty} \uparrow \Big(a \int f_n \,\mathrm{d}\mu + b \int g_n \,\mathrm{d}\mu\Big)$$
$$= a \int f \,\mathrm{d}\mu + b \int g \,\mathrm{d}\mu.$$

(3) For every $m \geq 1$, part (2) gives

$$\int \Big(\sum_{n=1}^{m} f_n\Big) \mathrm{d}\mu = \sum_{n=1}^{m} \int f_n \, \mathrm{d}\mu.$$

and it then suffices to let $m \uparrow \infty$ using the monotone convergence theorem. □

Remark Consider the special case where $E = \mathbb{N}$ and μ is the counting measure. Then, for any nonnegative function f on E, we get

$$\int f \, \mathrm{d}\mu = \sum_{n \in \mathbb{N}} f(n),$$

by applying (3) to the functions $f_n(k) = f(n)\, \mathbf{1}_{\{n\}}(k)$. Then, considering the functions $f_n(k) = a_{n,k}$, we also get the well-known property

$$\sum_{k \in \mathbb{N}} \Big(\sum_{n \in \mathbb{N}} a_{n,k}\Big) = \sum_{n \in \mathbb{N}} \Big(\sum_{k \in \mathbb{N}} a_{n,k}\Big),$$

for any double sequence $(a_{n,k})_{n,k \in \mathbb{N}}$ of nonnegative real numbers.

Corollary 2.6 *Let g be a nonnegative measurable function, and, for every $A \in \mathcal{A}$, set*

$$\nu(A) = \int_A g \, \mathrm{d}\mu = \int \mathbf{1}_A g \, \mathrm{d}\mu.$$

Then ν is a positive measure on E, which is called the measure of density g with respect to μ, and denoted by $\nu = g \cdot \mu$. Moreover, for every nonnegative measurable function f,

$$\int f \, \mathrm{d}\nu = \int fg \, \mathrm{d}\mu. \tag{2.1}$$

Instead of $\nu = g \cdot \mu$, one often writes $\nu(\mathrm{d}x) = g(x)\, \mu(\mathrm{d}x)$, or $g = \dfrac{\mathrm{d}\nu}{\mathrm{d}\mu}$. If we have both $\nu(\mathrm{d}x) = g(x)\, \mu(\mathrm{d}x)$ and $\theta(\mathrm{d}x) = h(x)\, \nu(\mathrm{d}x)$, then we have also $\theta(\mathrm{d}x) = h(x)g(x)\, \mu(\mathrm{d}x)$. This "associativity" property immediately follows from (2.1).

Proof It is obvious that $\nu(\varnothing) = 0$. On the other hand, if $(A_n)_{n\in\mathbb{N}}$ is a sequence of disjoint measurable sets, we have

$$\nu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \int \sum_{n\in\mathbb{N}} \mathbf{1}_{A_n} g \, \mathrm{d}\mu = \sum_{n\in\mathbb{N}} \int \mathbf{1}_{A_n} g \, \mathrm{d}\mu = \sum_{n\in\mathbb{N}} \nu(A_n),$$

using Proposition 2.5 (3). We have proved that ν is a measure.

Formula (2.1) holds when f is an indicator function by the definition, it then extends to the case when f is a nonnegative simple function by linearity (using Proposition 2.5 (2)). Finally, in the general case, we use the fact that f is the increasing limit of a sequence of nonnegative simple functions (Proposition 2.5 (1)) together with the monotone convergence theorem. □

Remark For every $A \in \mathcal{A}$, the condition $\mu(A) = 0$ implies $\nu(A) = 0$ (by Proposition 2.3 (2)). We will see later, under suitable assumptions, that conversely a measure ν that satisfies this property must be of the form $f \cdot \mu$.

We say that a property depending on $x \in E$ holds $\mu(\mathrm{d}x)$ almost everywhere, or μ a.e. (or even a.e. if there is no risk of confusion), if it holds for every $x \in E$ except possibly for x belonging to a set of μ-measure zero. For instance, if f and g are two measurable real functions on E, $f = g$, μ a.e., means

$$\mu(\{x \in E : f(x) \neq g(x)\}) = 0.$$

Proposition 2.7 *Let f be a nonnegative measurable function.*

(1) *For every $a \in (0, \infty)$,*

$$\mu(\{x \in E : f(x) \geq a\}) \leq \frac{1}{a} \int f \, \mathrm{d}\mu.$$

(2) *We have*

$$\int f \, \mathrm{d}\mu < \infty \Rightarrow f < \infty\,, \ \mu \ a.e.$$

(3) *We have*

$$\int f \, \mathrm{d}\mu = 0 \Leftrightarrow f = 0\,, \ \mu \ a.e.$$

(4) *If g is another nonnegative measurable function,*

$$f = g\,, \ \mu \ a.e. \Rightarrow \int f \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu.$$

Proof

(1) Set $A_a = \{x \in E : f(x) \geq a\}$. Then $f \geq a\mathbf{1}_{A_a}$ and thus

$$\int f \, d\mu \geq \int a\mathbf{1}_{A_a} \, d\mu = a\mu(A_a).$$

(2) For every $n \geq 1$, set $A_n = \{x \in E : f(x) \geq n\}$ and $A_\infty = \{x \in E : f(x) = \infty\}$. By (1) we have $\mu(A_n) \leq n^{-1} \int f \, d\mu < \infty$ for every n, and then

$$\mu(A_\infty) = \mu\Big(\bigcap_{n\geq 1} A_n\Big) = \lim_{n\to\infty} \downarrow \mu(A_n) \leq \lim_{n\to\infty} \frac{1}{n} \int f \, d\mu = 0.$$

(3) We already saw in Proposition 2.3 (2) that $f = 0$ a.e. implies $\int f \, d\mu = 0$. Conversely, assume that $\int f \, d\mu = 0$, and for every $n \geq 1$, set $B_n = \{x \in E : f(x) \geq n^{-1}\}$. By (1),

$$\mu(B_n) \leq n \int f \, d\mu = 0$$

and thus $\mu(B_n) = 0$, which implies $\mu(\{x : f(x) > 0\}) = \mu\Big(\bigcup_{n\geq 1} B_n\Big) = 0$.

(4) We use the notation $f \vee g = \max(f, g)$ and $f \wedge g = \min(f, g)$. Then $f = g$ a.e. implies $f \vee g = f \wedge g$ a.e., and thus

$$\int (f \vee g) d\mu = \int (f \wedge g) d\mu + \int (f \vee g - f \wedge g) d\mu = \int (f \wedge g) d\mu,$$

where we applied (3) to $f \vee g - f \wedge g$. Since $\int f \wedge g \, d\mu \leq \int f \, d\mu \leq \int f \vee g \, d\mu$, and similarly for $\int g \, d\mu$, we conclude that

$$\int f \, d\mu = \int (f \vee g) d\mu = \int g \, d\mu.$$

□

Theorem 2.8 (Fatou's Lemma) *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative measurable functions. Then,*

$$\int (\liminf f_n) \, d\mu \leq \liminf \int f_n \, d\mu.$$

Proof We have

$$\liminf f_n = \lim_{k\to\infty} \uparrow \Big(\inf_{n\geq k} f_n\Big)$$

and thus, by the monotone convergence theorem,

$$\int (\liminf f_n) \mathrm{d}\mu = \lim_{k\to\infty} \uparrow \int \Big(\inf_{n\geq k} f_n\Big) \mathrm{d}\mu.$$

On the other hand, for every integer $p \geq k$,

$$\inf_{n\geq k} f_n \leq f_p$$

which implies

$$\int \Big(\inf_{n\geq k} f_n\Big) \mathrm{d}\mu \leq \inf_{p\geq k} \int f_p \, \mathrm{d}\mu.$$

By taking a monotone limit as $k \uparrow \infty$, we get

$$\begin{aligned} \int (\liminf f_n) \mathrm{d}\mu = \lim_{k\to\infty} \uparrow \int \Big(\inf_{n\geq k} f_n\Big) \mathrm{d}\mu &\leq \lim_{k\to\infty} \uparrow \inf_{p\geq k} \int f_p \, \mathrm{d}\mu \\ &= \liminf \int f_n \, \mathrm{d}\mu, \end{aligned}$$

which completes the proof. □

One might naively think that an analog of Fatou's lemma holds if the liminf is replaced by the limsup, in the form

$$\int (\limsup f_n) \, \mathrm{d}\mu \geq \limsup \int f_n \, \mathrm{d}\mu.$$

It is very easy to give examples showing that this does not hold: for instance, if $E = \mathbb{N}$ and μ is the counting measure, we can take $f_n(k) = 1$ if $k > n$ and $f_n(k) = 0$ if $k \leq n$, in such a way that $\limsup f_n$ is the function identically equal to 0, but $\int f_n \, \mathrm{d}\mu = \infty$ for every n.

We conclude this section with a simple observation, which is however of constant use in probability theory. Recall from Definition 1.16 the notion of the pushforward of a measure.

Proposition 2.9 *Let $(F, \mathcal{B})$ be a measurable space, and let $\varphi : E \longrightarrow F$ be a measurable function. Let ν be the pushforward of μ under φ. Then, for any nonnegative measurable function h on F, we have*

$$\int_E h(\varphi(x)) \, \mu(\mathrm{d}x) = \int_F h(y) \, \nu(\mathrm{d}y). \tag{2.2}$$

Proof If $h = \mathbf{1}_B$ with $B \in \mathcal{B}$, formula (2.2) holds by the definition of the pushforward of a measure. The formula also holds when h is a nonnegative simple function by linearity, using Proposition 2.5 (2). Then we just have to write h as the increasing limit of a sequence of nonnegative simple functions (Proposition 2.5 (1)) and to use the monotone convergence theorem twice. □

2.2 Integrable Functions

We will now discuss the integral of measurable functions of arbitrary sign. In contrast with the case of nonnegative functions (Definition 2.2), we need to restrict ourselves to a particular class, which will be the class of integrable functions.

If $f : E \longrightarrow \mathbb{R}$ is a measurable function, we denote the positive part and the negative part of f by $f^+ = \sup(f, 0)$ and $f^- = \sup(-f, 0)$ respectively. Note that both f^+ and f^- are nonnegative measurable functions, and $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

Definition 2.10 Let $f : E \longrightarrow \mathbb{R}$ be a measurable function. We say that f is integrable with respect to μ (or μ-integrable, or simply integrable if there is no ambiguity) if

$$\int |f| \, \mathrm{d}\mu < \infty.$$

In that case, we set

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

If $A \in \mathcal{A}$, we write

$$\int_A f \, \mathrm{d}\mu = \int \mathbf{1}_A f \, \mathrm{d}\mu.$$

Remarks

(1) If $\int |f| \, \mathrm{d}\mu < \infty$, we have $\int f^+ \mathrm{d}\mu \leq \int |f| \mathrm{d}\mu < \infty$ and similarly $\int f^- \mathrm{d}\mu < \infty$, which shows that the definition $\int f \, \mathrm{d}\mu$ makes sense (without the condition $\int |f| \, \mathrm{d}\mu < \infty$, the definition might lead to $\int f \, \mathrm{d}\mu = \infty - \infty$!). When f is nonnegative, the definition of $\int f \, \mathrm{d}\mu$ is of course consistent with the previous section.

(2) The reader may observe that we can give a sense to $\int f \, \mathrm{d}\mu$ if (at least) *one of the two integrals* $\int f^+ \mathrm{d}\mu$ and $\int f^- \mathrm{d}\mu$ is finite. For instance, we can define $\int f \, \mathrm{d}\mu = -\infty$ if $\int f^+ \mathrm{d}\mu < \infty$ and $\int f^- \mathrm{d}\mu = \infty$. In this book, we will not use this convention, and whenever we consider the integral of functions of arbitrary sign we will assume that they are integrable.

We let $\mathcal{L}^1(E, \mathcal{A}, \mu)$ be the space of all μ-integrable functions $f : E \longrightarrow \mathbb{R}$. We will also use the notation $\mathcal{L}^1_+(E, \mathcal{A}, \mu)$ for the space of all nonnegative μ-integrable functions.

Properties

(a) For every $f \in \mathcal{L}^1(E, \mathcal{A}, \mu)$, $|\int f \,\mathrm{d}\mu| \le \int |f| \,\mathrm{d}\mu$.
(b) $\mathcal{L}^1(E, \mathcal{A}, \mu)$ is a linear space and $f \mapsto \int f \,\mathrm{d}\mu$ is a linear form on this space.
(c) If $f, g \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ and $f \le g$, then $\int f \,\mathrm{d}\mu \le \int g \,\mathrm{d}\mu$.
(d) If $f \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ and g is a measurable function such that $f = g$, μ a.e., then $g \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ and $\int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$.
(e) Under the assumptions of Proposition 2.9, let $h : F \longrightarrow \mathbb{R}$ be measurable. Then h is ν-integrable if and only if $h \circ \varphi$ is μ-integrable, and then

$$\int_E h(\varphi(x))\, \mu(\mathrm{d}x) = \int_F h(y)\, \nu(\mathrm{d}y).$$

Remark It is also true that, if $f \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ and g is nonnegative and measurable, the property $f \le g$ implies $\int f \,\mathrm{d}\mu \le \int g \,\mathrm{d}\mu$. Indeed, if $\int g \,\mathrm{d}\mu < \infty$ we apply (c), and if $\int g \,\mathrm{d}\mu = \infty$ the result is trivial.

Proof

(a) We write

$$\Big|\int f \,\mathrm{d}\mu\Big| = \Big|\int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu\Big| \le \int f^+ \mathrm{d}\mu + \int f^- \mathrm{d}\mu = \int |f| \mathrm{d}\mu.$$

(b) Let $f \in \mathcal{L}^1(E, \mathcal{A}, \mu)$. For $a \in \mathbb{R}$,

$$\int |af| \mathrm{d}\mu = |a| \int |f| \mathrm{d}\mu < \infty.$$

If $a \ge 0$,

$$\int (af) \mathrm{d}\mu = \int (af)^+ \mathrm{d}\mu - \int (af)^- \mathrm{d}\mu = a \int f^+ \mathrm{d}\mu - a \int f^- \mathrm{d}\mu = a \int f \mathrm{d}\mu$$

and, if $a < 0$,

$$\int (af) \mathrm{d}\mu = \int (af)^+ \mathrm{d}\mu - \int (af)^- \mathrm{d}\mu = (-a) \int f^- \mathrm{d}\mu + a \int f^+ \mathrm{d}\mu = a \int f \mathrm{d}\mu.$$

If $f, g \in \mathcal{L}^1(E, \mathcal{A}, \mu)$, the inequality $|f + g| \le |f| + |g|$ implies that $f + g \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ and moreover the fact that

$$(f + g)^+ - (f + g)^- = f + g = f^+ - f^- + g^+ - g^-$$

implies

$$(f+g)^+ + f^- + g^- = (f+g)^- + f^+ + g^+.$$

Using the additivity property of the integral of nonnegative measurable functions (Proposition 2.5 (2)), we get

$$\int (f+g)^+ \mathrm{d}\mu + \int f^- \mathrm{d}\mu + \int g^- \mathrm{d}\mu = \int (f+g)^- \mathrm{d}\mu + \int f^+ \mathrm{d}\mu + \int g^+ \mathrm{d}\mu,$$

and therefore, since all integrals are finite,

$$\int (f+g)^+ \mathrm{d}\mu - \int (f+g)^- \mathrm{d}\mu = \int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu + \int g^+ \mathrm{d}\mu - \int g^- \mathrm{d}\mu,$$

which gives $\int (f+g)\mathrm{d}\mu = \int f\, \mathrm{d}\mu + \int g\, \mathrm{d}\mu$.

(c) We just write $\int g\, \mathrm{d}\mu = \int f\, \mathrm{d}\mu + \int (g-f)\mathrm{d}\mu$.

(d) The equality $f = g$ a.e. implies $f^+ = g^+$ and $f^- = g^-$ a.e. It then suffices to use Proposition 2.7 (4).

(e) This immediately follows from formula (2.2) applied to h^+ and h^-. □

Remark By combining (c) and (d), we also get that, if $f, g \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ and $f \le g$ a.e., then $\int f\, \mathrm{d}\mu \le \int f \vee g\, \mathrm{d}\mu = \int g\, \mathrm{d}\mu$.

Extension to Complex-Valued Functions Let $f : E \longrightarrow \mathbb{C}$ be a measurable function (this is equivalent to saying that both the real part $\mathrm{Re}(f)$ of f and the imaginary part $\mathrm{Im}(f)$ of f are measurable). We say that f is integrable (with respect to μ) if

$$\int |f| \mathrm{d}\mu < \infty.$$

In that case, we set

$$\int f\, \mathrm{d}\mu = \int \mathrm{Re}(f)\mathrm{d}\mu + i \int \mathrm{Im}(f)\mathrm{d}\mu.$$

We denote the set of all complex-valued integrable functions by $\mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$. Properties (a), (b), (d) above are still valid if $\mathcal{L}^1(E, \mathcal{A}, \mu)$ is replaced by $\mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$ (the easiest way to get (a) is to observe that

$$\Big| \int f\, \mathrm{d}\mu \Big| = \sup_{a \in \mathbb{C}, |a|=1} a \cdot \int f\, \mathrm{d}\mu = \sup_{a \in \mathbb{C}, |a|=1} \int a \cdot f\, \mathrm{d}\mu,$$

where $a \cdot z$ denotes the Euclidean scalar product on $\mathbb{C}$ identified with $\mathbb{R}^2$). Notice that, in the complex analog of (b), $\mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$ is viewed as a complex vector space.

Theorem 2.11 (Dominated Convergence Theorem) *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{L}^1(E, \mathcal{A}, \mu)$ (resp. in $\mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$). We assume that:*

(1) *There exists a measurable function f with values in $\mathbb{R}$ (resp. with values in $\mathbb{C}$) such that*

$$f_n(x) \underset{n\to\infty}{\longrightarrow} f(x) \qquad \mu \text{ a.e.}$$

(2) *There exists a nonnegative measurable function g such that $\int g \, \mathrm{d}\mu < \infty$ and for every $n \in \mathbb{N}$,*

$$|f_n(x)| \le g(x) \qquad \mu \text{ a.e.}$$

Then $f \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ (resp. $f \in \mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$), and we have

$$\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu$$

and

$$\lim_{n\to\infty} \int |f_n - f| \mathrm{d}\mu = 0.$$

Proof We first assume that the following stronger assumptions hold:

(1)' For every $x \in E$,

$$f_n(x) \longrightarrow f(x).$$

(2)' There exists a nonnegative measurable function g such that $\int g \, \mathrm{d}\mu < \infty$ and, for every $n \in \mathbb{N}$, for every $x \in E$,

$$|f_n(x)| \le g(x).$$

The property $f \in \mathcal{L}^1(E, \mathcal{A}, \mu)$ (resp. $f \in \mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$) is then immediate since $|f| \le g$ and $\int g \, \mathrm{d}\mu < \infty$. Then, since $|f - f_n| \le 2g$ and $|f - f_n| \longrightarrow 0$ as $n \to \infty$, we can apply Fatou's lemma to get

$$\liminf \int (2g - |f - f_n|) \, \mathrm{d}\mu \ge \int \liminf (2g - |f - f_n|) \, \mathrm{d}\mu = 2 \int g \, \mathrm{d}\mu.$$

By the linearity of the integral of integrable functions, it follows that

$$2\int g\,\mathrm{d}\mu - \limsup \int |f - f_n|\mathrm{d}\mu \geq 2\int g\,\mathrm{d}\mu,$$

and therefore

$$\limsup \int |f - f_n|\mathrm{d}\mu = 0,$$

so that $\int |f - f_n|\mathrm{d}\mu \longrightarrow 0$ as $n \to \infty$. Finally, we have

$$\Big|\int f\,\mathrm{d}\mu - \int f_n\,\mathrm{d}\mu\Big| \leq \int |f - f_n|\mathrm{d}\mu.$$

In the general case where we only assume (1) and (2), we set

$$A = \{x \in E : f_n(x) \underset{n\to\infty}{\longrightarrow} f(x) \text{ and } |f_n(x)| \leq g(x) \text{ for every } n \in \mathbb{N}\}.$$

We note that A is a measurable set (use Lemma 1.15) and $\mu(A^c) = 0$. We can then apply the first part of the proof to the functions

$$\tilde{f}_n(x) = \mathbf{1}_A(x) f_n(x)\,, \quad \tilde{f}(x) = \mathbf{1}_A(x) f(x).$$

We have $f = \tilde{f}$ a.e., $f_n = \tilde{f}_n$ a.e. for every $n \in \mathbb{N}$, and thus $\int f_n \mathrm{d}\mu = \int \tilde{f}_n \mathrm{d}\mu$, $\int f\,\mathrm{d}\mu = \int \tilde{f}\,\mathrm{d}\mu$ and $\int |f_n - f|\mathrm{d}\mu = \int |\tilde{f}_n - \tilde{f}|\mathrm{d}\mu$. Furthermore properties (1)' and (2)' are satisfied by the functions $\tilde{f}_n$ and $\tilde{f}$. The desired results then follow from the first part of the proof. □

2.3 Integrals Depending on a Parameter

We start with a continuity theorem for the integral of a function depending on a parameter. We consider a metric space (U, d), which will be the parameter set.

Theorem 2.12 (Continuity of Integrals Depending on a Parameter) *Let* $f : U \times E \longrightarrow \mathbb{R}$ *(or* $\mathbb{C}$*), and let* $u_0 \in E$*. Assume that*

(i) *for every* $u \in U$*, the function* $x \mapsto f(u, x)$ *is measurable;*
(ii) $\mu(\mathrm{d}x)$ *a.e., the function* $u \mapsto f(u, x)$ *is continuous at* u_0*;*
(iii) *there exists a function* $g \in \mathcal{L}^1_+(E, \mathcal{A}, \mu)$ *such that, for every* $u \in U$*,*

$$|f(u, x)| \leq g(x) \qquad \mu(\mathrm{d}x) \text{ a.e.}$$

Then the function $F(u) = \int f(u,x)\mu(\mathrm{d}x)$ is well defined for every $u \in U$ and is continuous at u_0.

Proof Assumption (iii) implies that the function $x \mapsto f(u,x)$ is integrable, for every $u \in U$, so that $F(u)$ is well defined. Then let $(u_n)_{n\geq 1}$ be a sequence in U that converges to u_0. Assumption (ii) ensures that

$$f(u_n, x) \underset{n\to\infty}{\longrightarrow} f(u_0, x)\;, \quad \mu(\mathrm{d}x) \text{ a.e.}$$

Thanks to (iii), we can then apply the dominated convergence theorem, which gives

$$\lim_{n\to\infty} \int f(u_n, x)\,\mu(\mathrm{d}x) = \int f(u_0, x)\,\mu(\mathrm{d}x).$$

□

Examples

(a) Let μ be a diffuse measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\varphi \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$. Then the function $F : \mathbb{R} \to \mathbb{R}$ defined by

$$F(u) := \int_{(-\infty,u]} \varphi(x)\,\mu(\mathrm{d}x) = \int \mathbf{1}_{(-\infty,u]}(x)\varphi(x)\,\mu(\mathrm{d}x)$$

is continuous. Indeed, we can apply the theorem to $f(u,x) = \mathbf{1}_{(-\infty,u]}(x)\varphi(x)$, taking $g = |\varphi|$ and observing that, for every fixed $u_0 \in \mathbb{R}$ and for every $x \in \mathbb{R}\backslash\{u_0\}$, the function $u \mapsto f(u,x)$ is continuous at u_0 (and $\mu(\{u_0\}) = 0$ since μ is diffuse).

(b) *Fourier transform.* Let λ denote Lebesgue measure on $\mathbb{R}$, whose existence will be established in the next chapter. If $\varphi \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, the function $\widehat{\varphi} : \mathbb{R} \longrightarrow \mathbb{C}$ defined by

$$\widehat{\varphi}(u) := \int e^{\mathrm{i}ux}\,\varphi(x)\,\lambda(\mathrm{d}x)$$

is continuous on $\mathbb{R}$. The function $\widehat{\varphi}$ is called the Fourier transform of φ. In probability theory, we will also need to consider the Fourier transform of finite measures. If μ is a finite measure on $\mathbb{R}$, its Fourier transform is defined by

$$\widehat{\mu}(u) := \int e^{\mathrm{i}ux}\,\mu(\mathrm{d}x), \quad u \in \mathbb{R}.$$

Again the dominated convergence theorem implies that $\widehat{\mu}$ is continuous (and bounded) on $\mathbb{R}$.

(c) *Convolution*. Let $\varphi \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, and let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be a bounded continuous function. Then the function $h * \varphi$ defined on $\mathbb{R}$ by

$$h * \varphi(u) := \int h(u - x)\, \varphi(x)\, \lambda(\mathrm{d}x)$$

is continuous (and bounded).

We will now state a theorem of differentiability of integrals depending on a parameter. We let I be an open interval of $\mathbb{R}$.

Theorem 2.13 (Differentiation Under the Integral Sign) *We consider a function $f : I \times E \longrightarrow \mathbb{R}$. Let $u_0 \in I$. Suppose that*

(i) *for every $u \in I$, the function $x \mapsto f(u, x)$ belongs to $\mathcal{L}^1(E, \mathcal{A}, \mu)$;*
(ii) *$\mu(\mathrm{d}x)$ a.e., the function $u \mapsto f(u, x)$ is differentiable at u_0 and its derivative is denoted by*

$$\frac{\partial f}{\partial u}(u_0, x)\ ;$$

(iii) *there exists a function $g \in \mathcal{L}^1_+(E, \mathcal{A}, \mu)$ such that, for every $u \in I$,*

$$|f(u, x) - f(u_0, x)| \leq g(x)|u - u_0|\,, \qquad \mu(\mathrm{d}x) \text{ a.e.}$$

Then the function $F(u) = \int f(u, x)\mu(\mathrm{d}x)$ is differentiable at u_0, and its derivative is

$$F'(u_0) = \int \frac{\partial f}{\partial u}(u_0, x)\, \mu(\mathrm{d}x).$$

Remarks

(1) The derivative $\frac{\partial f}{\partial u}(u_0, x)$ is defined in (ii) only for x belonging to the complement of a measurable set H of μ-measure zero. We can extend its definition to every $x \in E$ by assigning the value 0 when x belongs to H. The function $x \mapsto \frac{\partial f}{\partial u}(u_0, x)$ is then measurable (on the complement of H it is the pointwise limit of the functions φ_n introduced in the proof below, so that we can use Lemma 1.15) and even integrable thanks to (iii). This shows that the formula for $F'(u_0)$ makes sense.
(2) It is easy to write an analog of Theorem 2.13 for complex-valued functions: just deal separately with the real and the imaginary part.

Proof Let $(u_n)_{n\geq 1}$ be a sequence in $I\backslash\{u_0\}$ that converges to u_0, and let

$$\varphi_n(x) = \frac{f(u_n, x) - f(u_0, x)}{u_n - u_0}$$

for every $x \in E$. Thanks to (ii), $\varphi_n(x)$ converges to $\frac{\partial f}{\partial u}(u_0, x)$, $\mu(\mathrm{d}x)$ a.e. Furthermore (iii) allows us to apply the dominated convergence theorem and to get

$$\lim_{n\to\infty} \frac{F(u_n) - F(u_0)}{u_n - u_0} = \lim_{n\to\infty} \int \varphi_n(x)\,\mu(\mathrm{d}x) = \int \frac{\partial f}{\partial u}(u_0, x)\,\mu(\mathrm{d}x).$$

□

In many applications, assumptions (ii) and (iii) hold in the stronger form:

(ii)' $\mu(\mathrm{d}x)$ a.e. the function $u \longrightarrow f(u, x)$ is differentiable on I;
(iii)' there exists a function $g \in \mathcal{L}^1_+(E, \mathcal{A}, \mu)$ such that, $\mu(\mathrm{d}x)$ a.e.,

$$\forall u \in I\ , \quad \left|\frac{\partial f}{\partial u}(u, x)\right| \leq g(x).$$

(Notice that (iii)'⇒(iii) thanks to the mean value theorem.) Under the assumptions (i),(ii)',(iii)', F is differentiable on I. Exercise 2.14 below shows that the more general statement of the theorem is sometimes necessary.

Examples

(a) Let $\varphi \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ be such that

$$\int |x\varphi(x)|\,\lambda(\mathrm{d}x) < \infty.$$

Then the Fourier transform $\widehat{\varphi}(u)$ is differentiable on $\mathbb{R}$, and

$$\widehat{\varphi}'(u) = \mathrm{i}\int x\, e^{\mathrm{i}ux}\,\varphi(x)\,\lambda(\mathrm{d}x).$$

(b) Let $\varphi \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, and let $h : \mathbb{R} \longrightarrow \mathbb{R}$ be continuously differentiable, and assume that h and h' are both bounded. Then the convolution $h * \varphi$ is differentiable on $\mathbb{R}$, and

$$(h * \varphi)' = h' * \varphi.$$

This argument can be iterated. For instance, if h is infinitely differentiable with compact support, $h * \varphi$ is also infinitely differentiable.

2.4 Exercises

Exercise 2.1 Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of integrable functions on E. Assume that

$$\sum_{n=1}^{\infty} \int |f_n| \, \mathrm{d}\mu < \infty.$$

Verify that the series $\sum_{n=1}^{\infty} f_n(x)$ is absolutely convergent for μ a.e. $x \in E$. If $F(x)$ denotes the sum of this series (and $F(x) = 0$ if the series is not absolutely convergent), prove that the function F is integrable and

$$\int F \, \mathrm{d}\mu = \sum_{n=1}^{\infty} \int f_n \, \mathrm{d}\mu.$$

Exercise 2.2 Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $(A_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{A}$. Also let $f : E \longrightarrow \mathbb{R}$ be an integrable function. Assume that

$$\lim_{n\to\infty} \int |\mathbf{1}_{A_n} - f| \, \mathrm{d}\mu = 0.$$

Prove that there exists $A \in \mathcal{A}$ such that $f = \mathbf{1}_A$, μ a.e.

Exercise 2.3 Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $(A_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{A}$. Show that the condition

$$\sum_{n=1}^{\infty} \mu(A_n) < \infty.$$

implies $\mu(\limsup A_n) = 0$.

Application Let λ denote Lebesgue measure on $\mathbb{R}$. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers, and let $(\alpha_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers. Show that the condition

$$\sum_{n\in\mathbb{N}} \sqrt{\alpha_n} < \infty$$

implies

$$\sum_{n\in\mathbb{N}} \frac{\alpha_n}{|x-a_n|} < \infty\,, \quad \lambda(\mathrm{d}x) \text{ a.e.}$$

Hint: Consider the sets $A_n = \{x \in \mathbb{R} : |x - a_n| \le \sqrt{\alpha_n}\}$.

Exercise 2.4 Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $f : E \longrightarrow \mathbb{R}$ be a measurable function. Show that

$$\int |f|\,\mathrm{d}\mu < \infty \iff \sum_{n\in\mathbb{Z}} 2^n \mu(\{x \in E : 2^n \le |f(x)| < 2^{n+1}\}) < \infty.$$

Assume that μ is finite. Show that

$$\int |f|\,\mathrm{d}\mu < \infty \iff \sum_{n=1}^{\infty} \mu(\{x \in E : |f(x)| \ge n\}) < \infty.$$

Exercise 2.5 Let μ be a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$. Assume that $\int_{[a,b]} f\,\mathrm{d}\mu = 0$ for every reals $a \le b$. Show that $f = 0$, μ a.e. (*Hint:* Use the monotone class theorem to verify that $\int_B f\,\mathrm{d}\mu = 0$ for every $B \in \mathcal{B}(\mathbb{R})$.)

Exercise 2.6 Compute

$$\lim_{n\to\infty} \int_0^n \left(1 + \frac{x}{n}\right)^n e^{-2x}\,\mathrm{d}x.$$

Let $\alpha \in \mathbb{R}$. Prove that the limit

$$\lim_{n\to\infty} \int_0^n \left(1 - \frac{x}{n}\right)^n x^{\alpha-1}\,\mathrm{d}x$$

exists in $[0, \infty]$ and is finite if and only if $\alpha > 0$.

Exercise 2.7 Let $(E, \mathcal{A}, \mu)$ be a measure space. Prove that μ is σ-finite if and only if there exists a measurable function f on E such that $\int f\,\mathrm{d}\mu < \infty$ and $f(x) > 0$ for every $x \in E$.

Exercise 2.8 (Scheffé's Lemma) Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $(f_n)_{n\in\mathbb{N}}$ and f be nonnegative measurable functions on E. Assume that $\int f\mathrm{d}\mu < \infty$, and that $f_n(x) \longrightarrow f(x)$ as $n \to \infty$, μ a.e. Show that the condition

$$\int f_n \mathrm{d}\mu \underset{n\to\infty}{\longrightarrow} \int f\mathrm{d}\mu$$

implies that

$$\int |f_n - f| \mathrm{d}\mu \underset{n\to\infty}{\longrightarrow} 0.$$

Exercise 2.9 Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $f : E \longrightarrow \mathbb{R}$ be an integrable function.

(1) Show that

$$\lim_{n\to\infty} \int |f| \, \mathbf{1}_{\{|f|\geq n\}} \, \mathrm{d}\mu = 0,$$

where we write $\{|f| \geq n\} = \{x \in E : |f(x)| \geq n\}$.

(2) Show that, for every $\varepsilon > 0$, there exists $\delta > 0$ such that, for every $A \in \mathcal{A}$,

$$\mu(A) < \delta \Rightarrow \int_A |f| \, \mathrm{d}\mu < \varepsilon.$$

(3) Suppose that $(E, \mathcal{A}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, where λ is Lebesgue measure. Show that the function F defined by

$$F(x) := \begin{cases} \int_{[0,x]} f \, \mathrm{d}\lambda & \text{if } x \geq 0, \\ -\int_{[x,0]} f \, \mathrm{d}\lambda & \text{if } x < 0, \end{cases}$$

is uniformly continuous on $\mathbb{R}$.

Exercise 2.10 Let $(E, \mathcal{A}, \mu)$ be a measure space, and let $f \in \mathcal{L}^1_{\mathbb{C}}(E, \mathcal{A}, \mu)$. Show that, if $|\int f \mathrm{d}\mu| = \int |f| \mathrm{d}\mu$, there exists a complex number $a \in \mathbb{C}$ such that $|a| = 1$ and $f = a|f|$, μ a.e.

Exercise 2.11 Consider the measure space $(\mathbb{R}, \mathcal{B}(R), \lambda)$, where λ is Lebesgue measure. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be an integrable function. Show that, for every $\alpha > 0$,

$$\lim_{n\to\infty} n^{-\alpha} f(nx) = 0\,, \quad \lambda(\mathrm{d}x) \text{ a.e.}$$

(*Hint:* Consider, for every $\eta > 0$ and $n \geq 1$, the set

$$A_{\eta,n} = \{x \in \mathbb{R} : n^{-\alpha} |f(nx)| > \eta\},$$

and use Exercise 2.3.)

Exercise 2.12 (Convergence in Measure) Let $(E, \mathcal{A})$ be a measurable space, and let μ be a finite measure on E. If $(f_n)_{n\in\mathbb{N}}$ and f are real measurable functions on E, we say that f_n converges in μ-measure to f if, for every $\varepsilon > 0$,

$$\mu(\{x \in E : |f_n(x) - f(x)| > \varepsilon\}) \underset{n\to\infty}{\longrightarrow} 0.$$

(1) Show that if $f_n(x) \longrightarrow f(x)$, $\mu(\mathrm{d}x)$ a.e., then f_n converges in μ-measure to $f(x)$. Give an example (using Lebesgue measure) showing that the converse is not true in general.
(2) Using Exercise 2.3, show that, if the sequence f_n converges in μ-measure to f, we can find a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that $f_{n_k}(x) \longrightarrow f(x)$ as $k \to \infty$, $\mu(\mathrm{d}x)$ a.e.
(3) Show that (still under our assumption that μ is finite) the conclusion of the dominated convergence theorem remains valid if we replace the condition

$$f_n \underset{n\to\infty}{\longrightarrow} f(x)\,, \quad \mu(\mathrm{d}x) \text{ a.e.}$$

by the weaker assumption f_n converges in μ-measure to f.

Exercise 2.13 Let $\varphi : [0, 1] \longrightarrow \mathbb{R}$ be integrable with respect to Lebesgue measure.

(1) For every $t \geq 0$, set

$$F(t) = \int_0^1 \sqrt{\varphi(x)^2 + t}\, \mathrm{d}x.$$

Prove that the function F is continuous on $[0, \infty)$ and differentiable on $(0, \infty)$. Give a necessary and sufficient condition (on φ) for F to be differentiable at 0.
(2) For every $t \in \mathbb{R}$, set

$$G(t) = \int_0^1 |\varphi(x) - t|\, \mathrm{d}x.$$

Prove that G is continuous on $\mathbb{R}$. For a fixed $t_0 \in \mathbb{R}$, give a necessary and sufficient condition for G to be differentiable at t_0.

Exercise 2.14 Let μ be a diffuse measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and let $\varphi \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$. Assume that

$$\int |x\varphi(x)|\, \mu(\mathrm{d}x) < \infty,$$

and, for every $u \in \mathbb{R}$, set

$$F(u) = \int_{\mathbb{R}} (u - x)^+ \varphi(x)\, \mu(\mathrm{d}x).$$

Show that F is differentiable on $\mathbb{R}$, and its derivative is

$$F'(u) = \int_{(-\infty,u]} \varphi(x)\, \mu(\mathrm{d}x).$$

Chapter 3
Construction of Measures

While the preceding chapter dealt with the Lebesgue integral with respect to a given measure μ on a measurable space, we are now interested in constructing certain measures of particular interest. Our main goal is to prove the existence of Lebesgue measure on $\mathbb{R}$ or on $\mathbb{R}^d$, but we provide tools that could be used as well to construct more general measures.

The first section introduces the notion of an outer measure. An outer measure satisfies weaker properties than a measure, but it turns out that it is possible from an outer measure to construct a measure defined on an appropriate σ-field. This approach leads to a relatively simple construction of Lebesgue measure on $\mathbb{R}$ or on $\mathbb{R}^d$. We discuss several properties of Lebesgue measure, as well as its connections with the Riemann integral. We also provide an example of a non-measurable subset of $\mathbb{R}$, which illustrates the fact that Lebesgue measure could not be defined on *all* subsets of $\mathbb{R}$. Another application of outer measures is the construction (and characterization) of all finite measures on $\mathbb{R}$, leading to the so-called Stieltjes integral. Although this chapter focusses on measures on Euclidean spaces, we give several statements that are valid in a more general setting. In the last section, we state without proof the Riesz-Markov-Kakutani representation theorem, which is a cornerstone of the functional-analytic approach to measure theory.

3.1 Outer Measures

Recall our notation $\mathcal{P}(E)$ for the set of all subsets of a set E.

Definition 3.1 Let E be a set. A mapping $\mu^* : \mathcal{P}(E) \longrightarrow [0, \infty]$ is called an outer measure (or exterior measure) if

(i) $\mu^*(\varnothing) = 0$;
(ii) μ^* is increasing: $A \subset B \;\Rightarrow\; \mu^*(A) \le \mu^*(B)$;

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_3

(iii) μ^* is σ-subadditive, meaning that, for every sequence $(A_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(E)$,

$$\mu^*\Big(\bigcup_{k\in\mathbb{N}} A_k\Big) \le \sum_{k\in\mathbb{N}} \mu^*(A_k).$$

The properties of an outer measure are less stringent than those of a measure (σ-additivity is replaced by σ-subadditivity). On the other hand, an outer measure is defined on every subset of E, whereas a measure is defined only on elements of a σ-field on E.

We will later provide examples of interesting outer measures. Our goal in this section is to show how, starting from an outer measure μ^*, one can construct a measure on a certain σ-field $\mathcal{M}(\mu^*)$ which depends on μ^*. In the remaining part of this section, we fix an outer measure μ^*.

Definition 3.2 A subset B of E is said to be μ^*-measurable if, for every subset A of E, we have

$$\mu^*(A) = \mu^*(A \cap B) + \mu^*(A \cap B^c).$$

We let $\mathcal{M}(\mu^*) \subset \mathcal{P}(E)$ denote the set of all μ^*-measurable sets.

Remark The inequality $\mu^*(A) \le \mu^*(A \cap B) + \mu^*(A \cap B^c)$ always holds by σ-subadditivity. To verify that a set B is μ^*-measurable, we have to check the reverse inequality.

Theorem 3.3

(i) *$\mathcal{M}(\mu^*)$ is a σ-field, which contains all subsets B of E such that $\mu^*(B) = 0$.*
(ii) *The restriction of μ^* to $\mathcal{M}(\mu^*)$ is a measure.*

Proof

(i) Write $\mathcal{M} = \mathcal{M}(\mu^*)$ to simplify notation. If $\mu^*(B) = 0$, the inequality

$$\mu^*(A) \ge \mu^*(A \cap B^c) = \mu^*(A \cap B) + \mu^*(A \cap B^c)$$

immediately shows that $B \in \mathcal{M}$.

Then it is obvious that $\varnothing \in \mathcal{M}$ and that $B \in \mathcal{M}$ implies $B^c \in \mathcal{M}$. To complete the proof of (i), it remains to verify that $\mathcal{M}$ is closed under countable unions. We start by proving that $\mathcal{M}$ is closed under finite unions. Let $B_1, B_2 \in \mathcal{M}$. Then, for every $A \in \mathcal{P}(E)$, the fact that $B_1 \in \mathcal{M}$ implies that

$$\begin{aligned}\mu^*(A \cap (B_1 \cup B_2)) &= \mu^*(A \cap (B_1 \cup B_2) \cap B_1) + \mu^*(A \cap (B_1 \cup B_2) \cap B_1^c)\\ &= \mu^*(A \cap B_1) + \mu^*(A \cap B_2 \cap B_1^c).\end{aligned}$$

Hence, using successively the properties $B_2 \in \mathcal{M}$ and $B_1 \in \mathcal{M}$,

$$\begin{aligned}\mu^*(A \cap (B_1 \cup B_2)) + \mu^*(A \cap (B_1 \cup B_2)^c) \\ &= \mu^*(A \cap B_1) + \mu^*(A \cap B_1^c \cap B_2) + \mu^*(A \cap B_1^c \cap B_2^c) \\ &= \mu^*(A \cap B_1) + \mu^*(A \cap B_1^c) \\ &= \mu^*(A),\end{aligned}$$

which shows that $B_1 \cup B_2 \in \mathcal{M}$. Since $\mathcal{M}$ is closed under finite unions and under the replacement of a set by its complement, $\mathcal{M}$ is closed under finite intersections. Consequently, if $B, B' \in \mathcal{M}$, then $B' \backslash B = B' \cap B^c \in \mathcal{M}$.

Thanks to the last remark, the proof of (i) will be complete if we can prove that, for any sequence $(B_k)_{k \in \mathbb{N}}$ of *disjoint* sets belonging to $\mathcal{M}$, we have $\bigcup_{k \in \mathbb{N}} B_k \in \mathcal{M}$. To this end, we show by induction that, for every integer $m \in \mathbb{N}$ and every subset A of E,

$$\mu^*(A) = \sum_{k=1}^{m} \mu^*(A \cap B_k) + \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{m} B_k^c\Big)\Big). \tag{3.1}$$

When $m = 1$, this is just the property $B_1 \in \mathcal{M}$. To go from step m to step $m + 1$, we use the property $B_{m+1} \in \mathcal{M}$ and then the fact that the sets B_k are disjoint (so that $B_{m+1} \subset B_k^c$ for $1 \le k \le m$), to get

$$\begin{aligned}\mu^*\Big(A \cap \Big(\bigcap_{k=1}^{m} B_k^c\Big)\Big) &= \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{m} B_k^c\Big) \cap B_{m+1}\Big) + \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{m+1} B_k^c\Big)\Big) \\ &= \mu^*(A \cap B_{m+1}) + \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{m+1} B_k^c\Big)\Big)\end{aligned}$$

and combining this equality with the induction hypothesis we get the desired result at step $m + 1$. This completes the proof of (3.1).

It follows from (3.1) that

$$\mu^*(A) \ge \sum_{k=1}^{m} \mu^*(A \cap B_k) + \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{\infty} B_k^c\Big)\Big)$$

and, letting m tend to ∞,

$$\mu^*(A) \geq \sum_{k=1}^{\infty} \mu^*(A \cap B_k) + \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{\infty} B_k^c\Big)\Big) \tag{3.2}$$
$$\geq \mu^*\Big(A \cap \Big(\bigcup_{k=1}^{\infty} B_k\Big)\Big) + \mu^*\Big(A \cap \Big(\bigcap_{k=1}^{\infty} B_k^c\Big)\Big),$$

by σ-subadditivity. Recalling the remark after Definition 3.2, we conclude that $\bigcup_{k=1}^{\infty} B_k \in \mathcal{M}$, and this completes the proof of (i).

(ii) Let μ denote the restriction of μ^* to $\mathcal{M}$. We already know that $\mu(\varnothing) = 0$. Let $(B_k)_{k\in\mathbb{N}}$ be a sequence of disjoint elements of $\mathcal{M}$. We can apply (3.2) with $A = \bigcup_{k=1}^{\infty} B_k$. It follows that

$$\mu^*\Big(\bigcup_{k=1}^{\infty} B_k\Big) \geq \sum_{k=1}^{\infty} \mu^*(B_k).$$

Since the reverse inequality holds by σ-subadditivity, we have obtained the σ-additivity of the restriction of μ^* to $\mathcal{M}$. This completes the proof. □

3.2 Lebesgue Measure

For every subset A of $\mathbb{R}$, we set

$$\lambda^*(A) = \inf\Big\{\sum_{i\in\mathbb{N}} (b_i - a_i) : A \subset \bigcup_{i\in\mathbb{N}} (a_i, b_i)\Big\}.$$

The infimum is over all countable covers of A by open intervals (a_i, b_i), $i \in \mathbb{N}$ (it is trivial that such covers exist). Note that the infimum makes sense as a number in $[0, \infty]$: the value ∞ may occur if A is unbounded.

Theorem 3.4

(i) *λ^* is an outer measure on $\mathbb{R}$.*
(ii) *The σ-field $\mathcal{M}(\lambda^*)$ contains $\mathcal{B}(\mathbb{R})$.*
(iii) *For every reals $a \leq b$, $\lambda^*([a, b]) = \lambda^*((a, b)) = b - a$.*

The restriction of λ^* to $\mathcal{B}(\mathbb{R})$ is a positive measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ called **Lebesgue measure** on $\mathbb{R}$. It will be denoted by λ. As a consequence of the monotone class theorem (see the end of Chapter 1), this is the unique measure on $\mathcal{B}(\mathbb{R})$ that satisfies the property $\lambda([a, b]) = b - a$ for every compact interval $[a, b]$.

Proof

(i) It is immediate that $\lambda^*(\varnothing) = 0$ and $\lambda^*(A) \leq \lambda^*(B)$ if $A \subset B$. It remains to establish subadditivity. To this end, consider a sequence $(A_n)_{n\in\mathbb{N}}$ of subsets of $\mathbb{R}$. We can assume that $\lambda^*(A_n) < \infty$ for every n (otherwise there is nothing to prove). Let $\varepsilon > 0$. For every fixed $n \in \mathbb{N}$, we can find a sequence of open intervals $(a_i^{(n)}, b_i^{(n)})$, $i \in \mathbb{N}$, such that

$$A_n \subset \bigcup_{i\in\mathbb{N}} (a_i^{(n)}, b_i^{(n)})$$

and

$$\sum_{i\in\mathbb{N}} (b_i^{(n)} - a_i^{(n)}) \leq \lambda^*(A_n) + \frac{\varepsilon}{2^n}.$$

Then we just have to notice that the collection of all intervals $(a_i^{(n)}, b_i^{(n)})$ for $n \in \mathbb{N}$ and $i \in \mathbb{N}$ forms a countable cover of $\bigcup_{n\in\mathbb{N}} A_n$, and thus

$$\lambda^*\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \leq \sum_{n\in\mathbb{N}} \sum_{i\in\mathbb{N}} (b_i^{(n)} - a_i^{(n)}) \leq \sum_{n\in\mathbb{N}} \lambda^*(A_n) + \varepsilon.$$

The desired result follows since ε was arbitrary.

(ii) Since $\mathcal{M}(\lambda^*)$ is a σ-field, it is enough to verify that it contains a class that generates the Borel σ-field, and, for instance, we can consider the class of all intervals $(-\infty, \alpha]$, for $\alpha \in \mathbb{R}$. So fix $\alpha \in \mathbb{R}$ and set $B = (-\infty, \alpha]$. We have to check that, for every subset A of $\mathbb{R}$,

$$\lambda^*(A) \geq \lambda^*(A \cap B) + \lambda^*(A \cap B^c).$$

Let $((a_i, b_i))_{i\in\mathbb{N}}$ be a cover of A, and $\varepsilon > 0$. The intervals $(a_i \wedge \alpha, (b_i \wedge \alpha) + \varepsilon 2^{-i})$ cover $A \cap B$, and the intervals $(a_i \vee \alpha, b_i \vee \alpha)$ cover $A \cap B^c$. Hence

$$\lambda^*(A \cap B) \leq \sum_{i\in\mathbb{N}} ((b_i \wedge \alpha) - (a_i \wedge \alpha)) + \varepsilon,$$

$$\lambda^*(A \cap B^c) \leq \sum_{i\in\mathbb{N}} ((b_i \vee \alpha) - (a_i \vee \alpha)).$$

By summing these two inequalities, we get

$$\lambda^*(A \cap B) + \lambda^*(A \cap B^c) \leq \sum_{i \in \mathbb{N}} (b_i - a_i) + \varepsilon.$$

Since ε was arbitrary, we get

$$\lambda^*(A \cap B) + \lambda^*(A \cap B^c) \leq \sum_{i \in \mathbb{N}} (b_i - a_i),$$

and since $\lambda^*(A)$ is defined as the infimum of the sums in the right-hand side over all covers of A, we obtain the desired inequality.

(iii) From the definition, it is immediate that

$$\lambda^*([a, b]) \leq b - a.$$

To get the reverse inequality, assume that

$$[a, b] \subset \bigcup_{i \in \mathbb{N}} (a_i, b_i).$$

By compactness, we can find a sufficiently large integer N such that

$$[a, b] \subset \bigcup_{i=1}^{N} (a_i, b_i).$$

An elementary argument left to the reader then shows that this implies

$$b - a \leq \sum_{i=1}^{N} (b_i - a_i) \leq \sum_{i=1}^{\infty} (b_i - a_i).$$

This gives the desired inequality $b - a \leq \lambda^*([a, b])$. Finally, it is easy to see that $\lambda^*((a, b)) = \lambda^*([a, b])$ (for instance by observing that $\lambda^*(\{a\}) = \lambda^*(\{b\}) = 0$). □

Extension in Higher Dimension

We now turn to the construction of Lebesgue measure on $\mathbb{R}^d$, when $d \geq 2$. An open box (resp. a closed box) is a subset P of $\mathbb{R}^d$ of the form

$$P = (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d) \quad (\text{resp. } P = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d])$$

where $a_1 \le b_1, \dots, a_d \le b_d$. The volume of P is by definition

$$\mathrm{vol}(P) = \prod_{j=1}^{d} (b_j - a_j).$$

We then define, for every subset A of $\mathbb{R}^d$,

$$\lambda^*(A) := \inf \Big\{ \sum_{i \in \mathbb{N}} \mathrm{vol}(P_i) : A \subset \bigcup_{i \in \mathbb{N}} P_i \Big\}.$$

where the infimum is over all covers of A by a countable collection $(P_i)_{i \in \mathbb{N}}$ of open boxes.

We have the following generalization of Theorem 3.4.

Theorem 3.5

(i) *λ^* is an outer measure on $\mathbb{R}^d$.*
(ii) *The σ-field $\mathcal{M}(\lambda^*)$ contains $\mathcal{B}(\mathbb{R}^d)$.*
(iii) *For every (open or closed) box P, $\lambda^*(P) = \mathrm{vol}(P)$.*

The restriction of λ^* to $\mathcal{B}(\mathbb{R}^d)$ is **Lebesgue measure** on $\mathbb{R}^d$, and will again be denoted by λ, or sometimes by λ_d if there is a risk of confusion.

Proof The proof of (i) is exactly the same as in the case $d = 1$. For (ii), it is enough to prove that, if $A \subset \mathbb{R}^d$ is a set of the form

$$A = \mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, a] \times \mathbb{R} \times \cdots \times \mathbb{R},$$

with $a \in \mathbb{R}$, then $A \in \mathcal{M}(\lambda^*)$ (it is easy to verify that sets of this form generate $\mathcal{B}(\mathbb{R}^d)$). The proof is then very similar to the case $d = 1$, and we omit the details. Finally, for (iii), the point is to verify that, if P is a closed box and if

$$P \subset \bigcup_{i=1}^{n} P_i$$

where the P_i's are open boxes, then

$$\mathrm{vol}(P) \le \sum_{i=1}^{n} \mathrm{vol}(P_i).$$

We leave this as an exercise for the reader (*Hint:* The volume of a box can be obtained as the limit as $n \to \infty$ of 2^{-dn} times the number of cubes of the form $[k_1 2^{-n}, (k_1+1)2^{-n}] \times \cdots \times [k_d 2^{-n}, (k_d+1)2^{-n}]$, $k_i \in \mathbb{Z}$, that intersect this box). $\square$

Remark Our study of product measures in Chapter 5 below will give another way of constructing Lebesgue measure on $\mathbb{R}^d$, from the special case of Lebesgue measure on $\mathbb{R}$.

Notation We (almost always) write

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} f(x)\,\lambda(\mathrm{d}x)$$

for the integral of f with respect to Lebesgue measure (provided that this integral is well defined). When $d = 1$, for $a \le b$, we also write

$$\int_a^b f(x)\,\mathrm{d}x = \int_{[a,b]} f(x)\,\lambda(\mathrm{d}x).$$

A natural question is whether $\mathcal{M}(\lambda^*)$ is much larger than the σ-field $\mathcal{B}(\mathbb{R}^d)$. We will see that in a certain sense these two σ-fields are not much different. We start with a preliminary proposition, which we state in a general setting since the proof is not more difficult that in the case of $\mathbb{R}$.

Proposition 3.6 *Let $(E, \mathcal{A}, \mu)$ be a measure space. The class of μ-negligible sets is defined as*

$$\mathcal{N} := \{A \in \mathcal{P}(E) : \exists B \in \mathcal{A},\ A \subset B \text{ and } \mu(B) = 0\}.$$

The completed σ-field of $\mathcal{A}$ with respect to μ is defined as the smallest σ-field that contains both $\mathcal{A}$ and $\mathcal{N}$. This σ-field is denoted by $\bar{\mathcal{A}}$. Then there exists a unique measure on $(E, \bar{\mathcal{A}})$ whose restriction to $\mathcal{A}$ is μ.

Proof We first observe that a more explicit description of $\bar{\mathcal{A}}$ can be given by saying that $\bar{\mathcal{A}} = \mathcal{B}$, where

$$\mathcal{B} = \{A \in \mathcal{P}(E) : \exists B, B' \in \mathcal{A},\ B \subset A \subset B' \text{ and } \mu(B'\backslash B) = 0\}.$$

In fact, it is straightforward to verify that $\mathcal{B}$ is a σ-field (we omit the details). It is then clear that $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{N} \subset \mathcal{B}$, which implies that $\bar{\mathcal{A}} \subset \mathcal{B}$. Finally, if $A \in \mathcal{B}$, we choose B and B' as in the definition of $\mathcal{B}$, and we observe that $A = B \cup (A\backslash B)$, with $B \in \mathcal{A}$ and $A\backslash B \in \mathcal{N}$ (because $A\backslash B \subset B'\backslash B$). The property $\mathcal{B} \subset \bar{\mathcal{A}}$ follows.

Once we have checked that $\bar{\mathcal{A}} = \mathcal{B}$, we extend μ to $\bar{\mathcal{A}}$ in the following way. If $A \in \bar{\mathcal{A}} = \mathcal{B}$, and if B and B' are as in the definition of $\mathcal{B}$, we set $\mu(A) = \mu(B) = \mu(B')$. This does not depend on the choice of B and B': if $\tilde{B}$ and $\tilde{B}'$ provide another possible choice, we have both $\mu(\tilde{B}) \le \mu(B')$ and $\mu(\tilde{B}') \ge \mu(B)$ which forces $\mu(B) = \mu(B') = \mu(\tilde{B}) = \mu(\tilde{B}')$. Finally it is easy to check that the extension of μ to $\bar{\mathcal{A}}$ is a measure. Indeed, if we consider a sequence $(A_n)_{n\in\mathbb{N}}$ of disjoint sets in $\bar{\mathcal{A}}$, then, for every $n \in \mathbb{N}$, we can choose $B_n \in \mathcal{A}$, such that $B_n \subset A_n$ and $A_n\backslash B_n$ is

μ-negligible, and we have

$$\sum_n \mu(A_n) = \sum_n \mu(B_n) = \mu\Big(\bigcup_{n\in\mathbb{N}} B_n\Big) = \mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big),$$

where the last equality holds because $\bigcup_{n\in\mathbb{N}} A_n \backslash \bigcup_{n\in\mathbb{N}} B_n \subset \bigcup_{n\in\mathbb{N}}(A_n\backslash B_n)$ is μ-negligible. □

Remark Let f and g be two real functions defined on E. Assume that g is $\mathcal{A}$-measurable and that $f = g$, μ a.e. (by definition, this only means that $\{x \in E : f(x) \neq g(x)\}$ is μ-negligible, which makes sense even if we do not assume that f is measurable). Then f is $\bar{\mathcal{A}}$-measurable. Indeed, we know that there exists a set $C \in \mathcal{A}$ such that $\mu(C) = 0$ and $f(x) = g(x)$ if $x \notin C$. Then, for any Borel subset H of $\mathbb{R}$, we have $g^{-1}(H)\backslash C \subset f^{-1}(H) \subset g^{-1}(H) \cup C$, and this implies $f^{-1}(H) \in \bar{\mathcal{A}}$, as in the preceding proof.

Proposition 3.7 *The σ-field $\mathcal{M}(\lambda^*)$ is equal to the completed σ-field $\overline{\mathcal{B}}(\mathbb{R}^d)$ of $\mathcal{B}(\mathbb{R}^d)$ with respect to Lebesgue measure on $\mathbb{R}^d$.*

Proof The fact that $\overline{\mathcal{B}}(\mathbb{R}^d) \subset \mathcal{M}(\lambda^*)$ is immediate. Indeed, if A is a λ-negligible subset of $\mathbb{R}^d$, there exists $B \in \mathcal{B}(\mathbb{R}^d)$ such that $A \subset B$ and $\lambda(B) = 0$. Then $\lambda^*(A) \leq \lambda^*(B) = \lambda(B) = 0$, and by Theorem 3.3 (i), this implies that $A \in \mathcal{M}(\lambda^*)$.

Conversely, let $A \in \mathcal{M}(\lambda^*)$. We aim to show that $A \in \overline{\mathcal{B}}(\mathbb{R}^d)$. Without loss of generality, we may assume that $A \subset (-K, K)^d$ for some $K > 0$ (otherwise, we write A as the increasing union of the sets $A \cap (-n, n)^d$). Then $\lambda^*(A) < \infty$, and thus, for every $n \geq 1$, we can find a countable collection $(P_i^n)_{i\in\mathbb{N}}$ of open boxes such that

$$A \subset \bigcup_{i\in\mathbb{N}} P_i^n\,, \qquad \sum_{i\in\mathbb{N}} \mathrm{vol}(P_i^n) \leq \lambda^*(A) + \frac{1}{n}.$$

We may assume that the boxes P_i^n are contained in $(-K, K)^d$ (the intersection of an open box with $(-K, K)^d$ is again an open box). Set

$$B_n = \bigcup_{i\in\mathbb{N}} P_i^n, \qquad B = \bigcap_{n\in\mathbb{N}} B_n.$$

Then $B \in \mathcal{B}(\mathbb{R}^d)$ and $A \subset B$, and furthermore, for every n,

$$\lambda(B) \leq \lambda(B_n) \leq \sum_i \mathrm{vol}(P_i^n) \leq \lambda^*(A) + \frac{1}{n}$$

which implies $\lambda(B) \leq \lambda^*(A)$ and then $\lambda(B) = \lambda^*(A)$ since we have also $\lambda(B) = \lambda^*(B) \geq \lambda^*(A)$. If we replace A by $(-K, K)^d\backslash A$, the same argument gives $\widetilde{B} \in \mathcal{B}(\mathbb{R}^d)$, with $\widetilde{B} \subset (-K, K)^d$, such that $(-K, K)^d \backslash A \subset \widetilde{B}$ and $\lambda^*((-K, K)^d \backslash A) = \lambda(\widetilde{B})$. If $B' = (-K, K)^d \backslash \widetilde{B}$, we have then $B' \subset A$ and $\lambda(B') = \lambda^*(A) = \lambda(B)$. Finally we have found two Borel sets B and B' such that $B' \subset A \subset B$ and $\lambda(B \backslash B') = 0$, so that we get $A \in \overline{\mathcal{B}}(\mathbb{R}^d)$. □

Theorem 3.8 *Lebesgue measure on $\mathbb{R}^d$ is invariant under the translations: for every $A \in \mathcal{B}(\mathbb{R}^d)$ and every $x \in \mathbb{R}^d$, we have $\lambda(x + A) = \lambda(A)$.*

Conversely, if μ is a measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which is finite on bounded sets and invariant under translations, there exists a constant $c \geq 0$ such that $\mu = c\lambda$.

Proof For $x \in \mathbb{R}^d$, write σ_x for the translation $\sigma_x(y) = y - x$ for $y \in \mathbb{R}^d$. The pushforward $\sigma_x(\lambda)$ satisfies

$$\forall A \in \mathcal{B}(\mathbb{R}^d),\ \sigma_x(\lambda)(A) = \lambda(\sigma_x^{-1}(A)) = \lambda(x + A).$$

The equality $\sigma_x(\lambda)(A) = \lambda(A)$ is trivial for any box A (since A and $x + A$ are boxes with the same volume). Thanks to an application of Corollary 1.19 to the class of open boxes, it follows that $\sigma_x(\lambda) = \lambda$.

Conversely, let μ be a measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which is invariant under the translations and takes finite values on boxes. Set $c = \mu([0, 1)^d)$. Since $[0, 1)^d$ is the disjoint union of n^d boxes which are translations of $[0, \frac{1}{n})^d$, it follows that, for every $n \geq 1$,

$$\mu([0, \frac{1}{n})^d) = \frac{c}{n^d}.$$

Then, let $a_1, \ldots, a_d \geq 0$, and denote the integer part of a real x by $\lfloor x \rfloor$. Writing

$$\prod_{j=1}^{d} [0, \frac{\lfloor n a_j \rfloor}{n}) \subset \prod_{j=1}^{d} [0, a_j) \subset \prod_{j=1}^{d} [0, \frac{\lfloor n a_j \rfloor + 1}{n})$$

we get

$$(\prod_{j=1}^{d} \lfloor n a_j \rfloor) \frac{c}{n^d} = \mu\Big(\prod_{j=1}^{d} [0, \frac{\lfloor n a_j \rfloor}{n}) \Big) \leq \mu\Big(\prod_{j=1}^{d} [0, a_j) \Big)$$

$$\leq \mu\Big(\prod_{j=1}^{d} [0, \frac{\lfloor n a_j \rfloor + 1}{n}) \Big) = (\prod_{j=1}^{d} \lfloor n a_j \rfloor + 1) \frac{c}{n^d}.$$

Letting n tend to ∞, it follows that

$$\mu\Big(\prod_{j=1}^{d} [0, a_j) \Big) = c \prod_{j=1}^{n} a_j = c\lambda\Big(\prod_{j=1}^{d} [0, a_j) \Big)$$

and, using the invariance of μ under translations, we find that μ and $c\,\lambda$ coincide on sets of the form

$$\prod_{j=1}^{d}[a_j, b_j).$$

Again Corollary 1.19 allows us to conclude that $\mu = c\,\lambda$. □

Remark For every $a \in \mathbb{R}\backslash\{0\}$, the pushforward of Lebesgue measure λ on $\mathbb{R}^d$ under the mapping $x \mapsto a\,x$ is $|a|^{-d}\lambda$. This is an immediate application of Theorem 3.8 (or of Corollary 1.19).

Proposition 3.9 *Lebesgue measure on $\mathbb{R}^d$ is regular, in the sense that, for every $A \in \overline{\mathcal{B}}(\mathbb{R}^d)$, we have*

$$\begin{aligned}\lambda(A) &= \inf\{\lambda(U) : U \textit{ open set},\ A \subset U\}\\ &= \sup\{\lambda(F) : F \textit{ compact set},\ F \subset A\}.\end{aligned}$$

Proof The quantity $\inf\{\lambda(U) : U \text{ open set},\ A \subset U\}$ is always greater than or equal to $\lambda(A)$. To get the reverse inequality, we may assume that $\lambda(A) < \infty$. Then, by the definition of $\lambda(A) = \lambda^*(A)$, we can, for every $\varepsilon > 0$, find a cover of A by open boxes P_i, $i \in \mathbb{N}$, such that $\sum_{i\in\mathbb{N}} \lambda(P_i) \le \lambda(A) + \varepsilon$. The open set $U = \bigcup_{i\in\mathbb{N}} P_i$ contains A, and $\lambda(U) \le \sum_{i\in\mathbb{N}} \lambda(P_i) \le \lambda(A)+\varepsilon$, which gives the desired inequality since ε was arbitrary.

To get the second equality of the proposition, we may assume that A is contained in a compact set C (otherwise, write $\lambda(A) = \lim \uparrow \lambda(A \cap [-n, n]^d)$). For every $\varepsilon > 0$, the first part of the proof allows us to find an open set U containing $C\backslash A$ and such that $\lambda(U) < \lambda(C\backslash A) + \varepsilon$. But then $F = C\backslash U$ is a compact set contained in A, and

$$\lambda(F) \ge \lambda(C) - \lambda(U) \ge \lambda(C) - (\lambda(C\backslash A) + \varepsilon) = \lambda(A) - \varepsilon,$$

which gives the second equality. □

The preceding proposition is by no means special to Lebesgue measure, and in fact it holds in much greater generality for measures on metric spaces. We content ourselves with a statement about finite measures.

Proposition 3.10 *Let (E, d) be a metric space, and let μ be a finite measure on $(E, \mathcal{B}(E))$. Then, for every $A \in \mathcal{B}(E)$,*

$$\begin{aligned}\mu(A) &= \inf\{\mu(U) : U \textit{ open set},\ A \subset U\}\\ &= \sup\{\mu(F) : F \textit{ closed set},\ F \subset A\}.\end{aligned}$$

Proof Let $\mathcal{O}$ be the class of all open subsets of E, and let $\mathcal{C}$ be the class of all sets $A \in \mathcal{B}(E)$ that satisfy the property of the proposition. Since the Borel σ-field is (by definition) generated by $\mathcal{O}$, it suffices to prove that $\mathcal{O} \subset \mathcal{C}$ and that $\mathcal{C}$ is a σ-field.

So suppose that $A \in \mathcal{O}$. Then the first equality in the proposition is trivial. As for the second one, we observe that, for every $n \geq 1$, the set

$$F_n = \{x \in E : d(x, A^c) \geq \frac{1}{n}\}$$

is contained in A and is closed (because the function $x \mapsto d(x, A^c) = \inf\{d(x, y) : y \in A^c\}$ is continuous). On the other hand, A is the increasing union of the sequence $(F_n)_{n\geq 1}$, which implies

$$\mu(A) = \lim \uparrow \mu(F_n).$$

This gives the second equality. We have thus proved that $\mathcal{O} \subset \mathcal{C}$.

It remains to show that $\mathcal{C}$ is a σ-field. Plainly, $\varnothing \in \mathcal{C}$ and, because complements of open sets are closed, one immediately sees that $A \in \mathcal{C}$ implies $A^c \in \mathcal{C}$ (here we also use the fact that μ is finite). We still have to show that $\mathcal{C}$ is closed under countable unions. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{C}$. We need to verify that $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{C}$.

Let $\varepsilon > 0$. For every n, we can find an open set U_n containing A_n such that $\mu(U_n) \leq \mu(A_n) + \varepsilon 2^{-n}$, and therefore

$$\mu\Big(\bigcup_{n\in\mathbb{N}} U_n \backslash \bigcup_{n\in\mathbb{N}} A_n\Big) \leq \sum_{n\in\mathbb{N}} \mu(U_n \backslash A_n) \leq \varepsilon.$$

Since $\bigcup_{n\in\mathbb{N}} U_n$ is open, this gives the first of the two desired equalities for $\bigcup_{n\in\mathbb{N}} A_n$.

Then, let N be an integer large enough so that

$$\mu\Big(\bigcup_{n=1}^{N} A_n\Big) \geq \mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) - \varepsilon.$$

For every $n \in \{1, \ldots, N\}$, we can find a closed set $F_n \subset A_n$ such that $\mu(A_n \backslash F_n) \leq \varepsilon 2^{-n}$. Hence

$$F = \bigcup_{n=1}^{N} F_n$$

is closed and

$$\mu\Big(\Big(\bigcup_{n=1}^{N} A_n\Big)\backslash F\Big) \leq \mu\Big(\bigcup_{n=1}^{N}(A_n \backslash F_n)\Big) \leq \sum_{n=1}^{N} \mu(A_n \backslash F_n) \leq \varepsilon.$$

It follows that

$$\mu\Big(\Big(\bigcup_{n\in\mathbb{N}} A_n\Big)\backslash F\Big) = \mu\Big(\Big(\bigcup_{n\in\mathbb{N}} A_n\Big)\backslash\Big(\bigcup_{n=1}^{N} A_n\Big)\Big) + \mu\Big(\Big(\bigcup_{n=1}^{N} A_n\Big)\backslash F\Big) \leq 2\varepsilon,$$

and we conclude that $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{C}$, which completes the proof. □

3.3 Relation with Riemann Integrals

In this section, we briefly discuss the relation between the Lebesgue integral that we constructed in Chapter 2 and the (older) Riemann integral. Roughly speaking, the Riemann integral on the real line involves approximating the function in consideration by *step functions*, which are constant on intervals, whereas we saw in the previous chapter that the definition of the Lebesgue integral involves the more general *simple functions*. The Lebesgue integral is in fact much more powerful than the Riemann integral, and, in particular, the convergence theorems that we obtained in Chapter 2 would be difficult to derive, even for continuous functions defined on a compact interval of the real line, if one relied only on the theory of the Riemann integral.

Let us briefly present the construction of the Riemann integral. Throughout this section, we consider functions defined on a fixed interval $[a, b]$ of $\mathbb{R}$, with $a < b$. A real function h defined on $[a, b]$ is called a step function if there exist a subdivision $a = x_0 < x_1 < \cdots < x_N = b$ of the interval $[a, b]$, and reals $y_1, \ldots, y_N$ such that, for every $i \in \{1, \ldots, N\}$, we have $h(x) = y_i$ for every $x \in (y_{i-1}, y_i)$. We then set

$$I(h) = \sum_{i=1}^{N} y_i(x_i - x_{i-1}).$$

Clearly any step function h is Borel measurable and integrable with respect to Lebesgue measure, and $I(h) = \int_{[a,b]} h(x)\,\mathrm{d}x$ (in particular, $I(h)$ does not depend on the subdivision chosen to represent h). If h and h' are two step functions and $h \leq h'$, one easily verifies that $I(h) \leq I(h')$.

Let Step($[a, b]$) denotes the set of all step functions on $[a, b]$. A bounded function $f : [a, b] \longrightarrow \mathbb{R}$ is called Riemann-integrable if

$$\sup_{h\in \mathrm{Step}([a,b]),\, h\leq f} I(h) = \inf_{h\in \mathrm{Step}([a,b]),\, h\geq f} I(h)$$

and then this value is called the Riemann integral of f and denoted by $I(f)$ (if f is a step function, this is consistent with the preceding definition).

In the next statement, $\overline{\mathcal{B}}([a,b])$ stands for the completed σ-field of $\mathcal{B}([a,b])$ with respect to Lebesgue measure (cf. Proposition 3.6).

Proposition 3.11 *Let f be a Riemann-integrable function on $[a,b]$. Then f is measurable from $([a,b],\overline{\mathcal{B}}([a,b]))$ into $(\mathbb{R},\mathcal{B}(\mathbb{R}))$, and*

$$I(f) = \int_{[a,b]} f(x)\,\mathrm{d}x.$$

Proof By definition, there exists a sequence $(h_n)_{n\in\mathbb{N}}$ of step functions on $[a,b]$ such that $h_n \geq f$ and $I(h_n) \downarrow I(f)$ as $n \to \infty$. Up to replacing h_n par $h_1 \wedge h_2 \wedge \cdots \wedge h_n$, we can assume that the sequence $(h_n)_{n\in\mathbb{N}}$ is decreasing. We then set

$$h_\infty = \lim \downarrow h_n \geq f.$$

Similarly, we can find an increasing sequence $(\widetilde{h}_n)$ of step functions on $[a,b]$ such that $\widetilde{h}_n \leq f$ and $I(\widetilde{h}_n) \uparrow I(f)$, and we set

$$\widetilde{h}_\infty = \lim \uparrow \widetilde{h}_n \leq f.$$

The functions h_∞ and $\widetilde{h}_\infty$ are bounded and Borel measurable. By the dominated convergence theorem, we have

$$\int_{[a,b]} h_\infty(x)\mathrm{d}x = \lim \downarrow \int_{[a,b]} h_n(x)\mathrm{d}x = \lim \downarrow I(h_n) = I(f),$$
$$\int_{[a,b]} \widetilde{h}_\infty(x)\mathrm{d}x = \lim \uparrow \int_{[a,b]} \widetilde{h}_n(x)\mathrm{d}x = \lim \uparrow I(\widetilde{h}_n) = I(f).$$

Hence

$$\int_{[a,b]} (h_\infty(x) - \widetilde{h}_\infty(x))\mathrm{d}x = 0.$$

Since $h_\infty \geq \widetilde{h}_\infty$, this implies $h_\infty = \widetilde{h}_\infty$, λ a.e. (Proposition 2.7 (3)). Since $h_\infty \geq f \geq \widetilde{h}_\infty$, it follows that $f = h_\infty$, λ a.e., and from the remark after Proposition 3.6 we get that f is $\overline{\mathcal{B}}([a,b])$-measurable. Finally, since $f = h_\infty$, λ a.e., we have $\int_{[a,b]} f(x)\,\mathrm{d}x = \int_{[a,b]} h_\infty(x)\,\mathrm{d}x = I(f)$. □

Remark It is worth emphasizing that, even for functions defined on a compact interval of $\mathbb{R}$, the Lebesgue integral is much more general than the Riemann integral (see Exercise 3.5 below for a characterization of Riemann-integrable functions). A simple example is the indicator function $\mathbf{1}_{\mathbb{Q}\cap[0,1]}$, which is *not* Riemann-integrable but is obviously integrable with respect to Lebesgue measure. Furthermore, as we already mentioned, the convergence theorems proved in Chapter 2 are very powerful tools that are not available in the theory of the Riemann integral.

3.4 A Subset of $\mathbb{R}$ Which Is Not Measurable

It is not so easy to give examples of subsets of $\mathbb{R}$ that are not Borel measurable. In this section, we provide such an example, whose construction strongly relies on the axiom of choice. Consider the quotient space $\mathbb{R}/\mathbb{Q}$ of equivalence classes of real numbers modulo rationals (in other words, equivalence classes for the equivalence relation on $\mathbb{R}$ defined by setting $x \sim y$ if and only if $x - y \in \mathbb{Q}$). For each $a \in \mathbb{R}/\mathbb{Q}$, choose a representative x_a of a (here we rely on the axiom of choice, which allows us to make such choices). Clearly we can assume that $x_a \in [0, 1]$ (replace x_a by $x_a - \lfloor x_a \rfloor$). Set

$$F = \{x_a \,;\ a \in \mathbb{R}/\mathbb{Q}\} \subset [0, 1].$$

Then F is not a Borel set, and is not even measurable with respect to the completed σ-field $\overline{\mathcal{B}}(\mathbb{R})$.

To verify this, let us suppose that F is Borel measurable (the argument also works if we assume that F is $\overline{\mathcal{B}}(\mathbb{R})$-measurable) and show that this leads to a contradiction. We have

$$\mathbb{R} \subset \bigcup_{q \in \mathbb{Q}} (q + F)$$

because, for any real x, F contains a representative y of the equivalence class of x in $\mathbb{R}/\mathbb{Q}$, which means that $x = q + y$ for some rational q. It follows that $\lambda(F) > 0$, because otherwise $\mathbb{R}$ would be contained in a countable union of sets of zero Lebesgue measure.

However, the sets $q + F$, $q \in \mathbb{Q}$ are disjoint (if, given two rationals q, q', we have $q + x_a = q' + x_{a'}$ for some $a, a' \in \mathbb{R}/\mathbb{Q}$, it follows that $x_a - x_{a'} = q' - q \in \mathbb{Q}$ and therefore $a = a'$ and $q = q'$). From the fact that

$$\bigcup_{q \in \mathbb{Q} \cap [0,1]} (q + F) \subset [0, 2]$$

we deduce that

$$\sum_{q \in \mathbb{Q} \cap [0,1]} \lambda(q + F) \leq 2$$

and necessarily $\lambda(F) = 0$, which gives the announced contradiction.

Remark Other ways of constructing non-measurable sets are presented in Exercise 3.6. All these examples involve making uncountably many choices (in the above example, choosing a representative in each equivalence class of $\mathbb{R}/\mathbb{Q}$). As an informal general principle, all subsets of $\mathbb{R}$ or of $\mathbb{R}^d$ that are obtained by a "constructive" method will be measurable.

3.5 Finite Measures on $\mathbb{R}$ and the Stieltjes Integral

The next theorem provides a description of all finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The statement could be extended to measures that are only finite on compact subsets of $\mathbb{R}$ (see the discussion at the end of this section) but for the sake of simplicity we restrict our attention to finite measures.

Theorem 3.12

(i) *Let μ be a finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For every $x \in \mathbb{R}$, let*

$$F_\mu(x) = \mu((-\infty, x]).$$

The function F_μ is increasing, bounded, right-continuous and such that $F_\mu(-\infty) = 0$ (where $F_\mu(-\infty)$ obviously stands for the limit of $F_\mu(x)$ as $x \to -\infty$).

(ii) *Conversely, consider a function $F : \mathbb{R} \longrightarrow \mathbb{R}_+$ which is increasing, bounded, right-continuous and such that $F(-\infty) = 0$. Then, there exists a unique finite measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F = F_\mu$.*

Proof

(i) It is straightforward to verify that F_μ satisfies the stated properties. For instance, if $(x_n)_{n\in\mathbb{N}}$ is a real sequence decreasing to x, $(-\infty, x]$ is the decreasing intersection of the sequence $(-\infty, x_n]$, and thus

$$F_\mu(x_n) = \mu((-\infty, x_n]) \underset{n\to\infty}{\longrightarrow} \mu((-\infty, x]) = F_\mu(x).$$

Similarly, if $x_n \downarrow -\infty$ as $n \to \infty$, the decreasing intersection of the sequence $(-\infty, x_n]$ is $\varnothing$, which implies $F_\mu(x_n) \longrightarrow 0$.

(ii) The uniqueness of μ is a consequence of Corollary 1.19, since the class $\mathcal{C} = \{(-\infty, x]; x \in \mathbb{R}\}$ is closed under finite intersections and generates $\mathcal{B}(\mathbb{R})$.

To show existence, we set, for every $A \subset \mathbb{R}$,

$$\mu^*(A) = \inf\Big\{ \sum_{i\in\mathbb{N}} (F(b_i) - F(a_i)) : A \subset \bigcup_{i\in\mathbb{N}} (a_i, b_i] \Big\}.$$

(Here it is more convenient to cover A by left-open and right-closed intervals, in contrast with the open intervals we used for Lebesgue measure.) Arguments very similar to those we used in the proof of Theorem 3.4 in the case of Lebesgue measure show that μ^* is an outer measure and the intervals $(-\infty, \alpha]$ are in $\mathcal{M}(\mu^*)$ (the latter property is even easier here). It follows that the σ-field $\mathcal{M}(\mu^*)$ contains $\mathcal{B}(\mathbb{R})$, and the restriction of μ^* to $\mathcal{B}(\mathbb{R})$ is a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which we denote by μ. Clearly, μ is a finite measure.

To complete the proof, we have to verify that $\mu((-\infty, x]) = F(x)$ for every $x \in \mathbb{R}$. To this end, it suffices to check that $\mu((a, b]) = F(b) - F(a)$ for every $a < b$ (then let $a \to -\infty$). The bound

$$\mu((a, b]) \leq F(b) - F(a)$$

is immediate from the construction of μ^*.

To get the reverse inequality, let $((x_i, y_i])_{i\in\mathbb{N}}$ be a countable cover of $(a, b]$, and let $\varepsilon \in (0, b - a)$. For every $i \in \mathbb{N}$, we can find $y_i' > y_i$ such that $F(y_i') \leq F(y_i) + \varepsilon 2^{-i}$. By compactness, we can find N_ε large enough so that $[a + \varepsilon, b]$ is covered by the finite collection $((x_i, y_i'))_{i\in\{1,\dots,N_\varepsilon\}}$. An elementary argument then shows that

$$\begin{aligned} F(b) - F(a+\varepsilon) \leq \sum_{i=1}^{N_\varepsilon}(F(y_i') - F(x_i)) &\leq \sum_{i=1}^{\infty}(F(y_i') - F(x_i)) \\ &\leq \sum_{i=1}^{\infty}(F(y_i) - F(x_i)) + \varepsilon. \end{aligned}$$

Since ε was arbitrary and $F(a + \varepsilon) \longrightarrow F(a)$ as $\varepsilon \to 0$, we get

$$F(b) - F(a) \leq \sum_{i=1}^{\infty}(F(y_i) - F(x_i))$$

which gives the lower bound $\mu((a, b]) \geq F(b) - F(a)$ and completes the proof. □

Let F be as in part (ii) of the proposition, and let μ be the finite measure such that $F = F_\mu$. Then we may define, for every bounded Borel measurable function $f : \mathbb{R} \longrightarrow \mathbb{R}$,

$$\int f(x)\,\mathrm{d}F(x) := \int f(x)\,\mu(\mathrm{d}x),$$

and $\int f(x)\, dF(x)$ is called the Stieltjes integral of f with respect to F. The notation is motivated by the equality, for every $a \le b$,

$$\int \mathbf{1}_{(a,b]}(x)\, \mathrm{d}F(x) = F(b) - F(a).$$

We have also, for instance,

$$\int \mathbf{1}_{[a,b]}(x)\, dF(x) = \lim_{n\to\infty} \int \mathbf{1}_{(a-\frac{1}{n},b]}(x)\, dF(x) = F(b) - F(a-),$$

where $F(a-)$ denotes the left limit of F at a. In particular (taking $b = a$), the function F is continuous if and only if μ is diffuse.

Extension to Measures that Are Finite on Compact Sets of $\mathbb{R}$ The formula

$$F(x) = \begin{cases} \mu((0,x]) & \text{if } x \ge 0, \\ -\mu((x,0]) & \text{if } x < 0, \end{cases}$$

gives a one-to-one correspondence between measures μ on $\mathbb{R}$ that take finite values on compact sets (these measures are called Radon measures on $\mathbb{R}$) and functions $F : \mathbb{R} \longrightarrow \mathbb{R}$ that are increasing and right-continuous and vanish at 0. This is an easy extension of Proposition 3.12 and we omit the details. The equality $\mu((a,b]) = F(b) - F(a)$ for every $a < b$ still holds, and the Stieltjes integral $\int f(x)\, \mathrm{d}F(x)$ makes sense, say when f is bounded with compact support. Note that, in the special case $F(x) = x$, μ is Lebesgue measure.

3.6 The Riesz-Markov-Kakutani Representation Theorem

In this section, we briefly discuss a famous representation theorem initially due to Riesz in the case of functions defined on the unit interval. This theorem is a cornerstone of the functional-analytic approach to measure theory. We say little about this approach in the present book, because the "abstract" approach that we have presented is much more adapted to applications to probability theory.

Recall that a topological space X is called locally compact if every point of X has a compact neighborhood. Let us consider a locally compact metric space X. The vector space of all real continuous functions with compact support on X is denoted by $C_c(X)$. A linear form J on $C_c(X)$ is a linear mapping from $C_c(X)$ into $\mathbb{R}$. It is said to be positive if $J(f) \ge 0$ whenever $f \ge 0$.

A *Radon measure* on X is a measure μ on $(X, \mathcal{B}(X))$ such that $\mu(K) < \infty$ for every compact subset K of X. If μ is a Radon measure on X, the formula

$$J(f) = \int f\, \mathrm{d}\mu$$

defines a linear form J on $C_c(X)$. Notice that the integral is well defined since there exist a constant C and a compact subset K of X such that $|f| \le C\,\mathbf{1}_K$, and the property $\mu(K) < \infty$ ensures that f is integrable. Moreover the linear form J is positive.

The Riesz-Markov-Kakutani representation theorem provides a converse to the preceding observations. Under suitable assumptions, any positive linear form on $C_c(X)$ is of the previous type.

Theorem 3.13 *Let X be a separable locally compact metric space, and let J be a positive linear form on $C_c(X)$. Then there exists a unique Radon measure μ on $(X, \mathcal{B}(X))$ such that*

$$\forall f \in C_c(X),\ J(f) = \int f \, \mathrm{d}\mu.$$

The measure μ is regular in the sense that, for every $A \in \mathcal{B}(X)$,

$$\begin{aligned}\mu(A) &= \inf\{\mu(U) : U \text{ open set},\ A \subset U\}\\ &= \sup\{\mu(K) : K \text{ compact set},\ K \subset A\}.\end{aligned}$$

Moreover, for every open subset U of X,

$$\mu(U) = \sup\{J(f) : f \in C_c(X),\ 0 \le f \le \mathbf{1}_U\}.$$

Example If $X = \mathbb{R}$, we may take $J(f) = I(f)$, where $I(f)$ is as above the Riemann integral of the function f (it is easy to verify that any function in $C_c(\mathbb{R})$ is Riemann integrable). One verifies that J is a positive linear form on $C_c(\mathbb{R})$, and the associated measure is (of course) Lebesgue measure. The Riesz-Markov-Kakutani representation theorem thus provides another construction of Lebesgue measure.

We omit the proof of Theorem 3.13, as this theorem is not used elsewhere in the present book. We refer to Chapter 2 of Rudin [22], which gives a slightly more precise statement.

3.7 Exercises

Exercise 3.1 For any function $g : \mathbb{R} \longrightarrow \mathbb{R}$, show that g is $\overline{\mathcal{B}}(\mathbb{R})$-measurable if and only if there exist two Borel measurable functions f and h such that $f \le g \le h$ and $f = h$, λ a.e.

Exercise 3.2 Let E be a compact metric space, and let μ be a probability measure on $(E, \mathcal{B}(E))$. Prove that there exists a unique compact subset K of E such that $\mu(K) = 1$ and $\mu(H) < 1$ for every compact subset H of K such that $H \ne K$. (*Hint:*

Define K as the intersection of all compact subsets K' of E such that $\mu(K') = 1$, and use Proposition 3.10 to verify that $\mu(K) = 1$.)

Exercise 3.3 Let $F : \mathbb{R} \longrightarrow \mathbb{R}$ be a bounded continuous increasing function such that $F(-\infty) = 0$, and let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuously differentiable function.

(1) Show that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \Big|\varphi(F(\frac{i}{n})) - \varphi(F(\frac{i-1}{n})) - \varphi'(F(\frac{i-1}{n}))\Big(F(\frac{i}{n}) - F(\frac{i-1}{n})\Big)\Big| = 0.$$

(2) Prove that $\varphi(F(1)) = \varphi(F(0)) + \int_0^1 \varphi'(F(s))\,\mathrm{d}F(s)$.

Exercise 3.4 (Hausdorff Dimension) If A is a subset of $\mathbb{R}^d$, we denote the diameter of A by $\Delta(A) = \sup\{|x - y| : x, y \in A\} \in [0, \infty]$. For every $\alpha > 0$ and every $\varepsilon > 0$, we set, for every subset A of $\mathbb{R}^d$,

$$\mu_{\alpha,\varepsilon}(A) = \inf_{(U_k)_{k\in\mathbb{N}} \in \mathcal{R}_\varepsilon(A)} \sum_{k\in\mathbb{N}} \Delta(U_k)^\alpha,$$

where $\mathcal{R}_\varepsilon(A)$ is the set of all countable coverings of A by open sets of diameter smaller than ε.

(1) Observe that $\mu_{\alpha,\varepsilon}(A) \geq \mu_{\alpha,\varepsilon'}(A)$ if $\varepsilon < \varepsilon'$, and thus one can define a mapping $A \mapsto \mu_\alpha(A)$ by

$$\mu_\alpha(A) := \lim_{\varepsilon\downarrow 0} \mu_{\alpha,\varepsilon}(A) \in [0, \infty].$$

(2) Prove that μ_α is an outer measure on $\mathbb{R}^d$.
(3) Verify that, for every subset A of $\mathbb{R}^d$, there exists a (unique) real number $\dim(A) \in [0, d]$, called the Hausdorff dimension of A, such that, for every $\alpha > 0$,

$$\mu_\alpha(A) = \begin{cases} 0 & \text{if } \alpha > \dim(A) \\ \infty & \text{if } \alpha < \dim(A). \end{cases}$$

(4) Let $\alpha > 0$ and let A be a Borel subset of $\mathbb{R}^d$. Assume that there exists a measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $\mu(A) > 0$ and $\mu(B) \leq r^\alpha$ for every open ball B of radius $r > 0$ centered at a point of A. Prove that $\dim(A) \geq \alpha$.

Exercise 3.5 Let $f : [0, 1] \longrightarrow \mathbb{R}$ be a bounded function. For every $x \in [0, 1]$, we define the oscillation of f at x by

$$\omega(f, x) := \lim_{h\to 0, h>0} \sup\{|f(u) - f(v)| : u, v \in [0, 1], |u - x| < h, |v - x| < h\}.$$

(1) Verify that f is continuous at x if and only if $\omega(f, x) = 0$.
(2) Let I be a compact subinterval of $[0, 1]$ and let $\varepsilon > 0$. Prove that the condition $\omega(f, x) < \varepsilon$ for every $x \in I$ implies the existence of $\alpha > 0$ such that $|f(u) - f(v)| < \varepsilon$ whenever $u, v \in I$ and $|u - v| < \alpha$.
(3) Let $\varepsilon > 0$ and $D_\varepsilon = \{x \in [0, 1] : \omega(f, x) \geq \varepsilon\}$. Prove that D_ε is closed.
(4) Assume that f is Riemann-integrable. Prove that the set of discontinuity points of f has Lebesgue measure 0. (*Hint:* If g and h are two step functions such that $g \leq f \leq h$, one has $h - g \geq \varepsilon$ at every point of D_ε, except possibly at a finite number of points.)
(5) Prove the converse to question (4): if the set of discontinuity points of f has Lebesgue measure 0, f is Riemann-integrable.

Exercise 3.6 Let $(E, \mathcal{A}, \mu)$ be a measure space, such that $0 < \mu(E) < \infty$. We suppose that there exists a bijection $T : E \longrightarrow E$, such that both T and T^{-1} are measurable, and T preserves μ in the sense that $T(\mu) = \mu$ (equivalently $\mu(T^{-1}(A)) = \mu(A)$ for every $A \in \mathcal{A}$). We also assume that T has no periodic point, so that the condition $T^n(x) = x$ for $x \in E$ and $n \in \mathbb{Z}$ can only hold if $n = 0$. For every $x \in E$, the orbit of x is defined by $\mathcal{O}(x) := \{T^n(x) : x \in \mathbb{Z}\}$. Note that orbits form a partition of E ($T^n(x) = T^m(y)$ can only hold if $\mathcal{O}(x) = \mathcal{O}(y)$). We then construct a subset A of E by choosing a single point in each orbit (*this strongly relies on the axiom of choice*).

(1) Prove that A is not $\mathcal{A}$-measurable.
(2) Verify that the assumptions of the exercise hold if $(E, \mathcal{A}) = ([0, 1), \mathcal{B}([0, 1)))$, μ is the restriction of Lebesgue measure to $[0, 1)$, and $T(x) = x + \alpha - \lfloor x + \alpha \rfloor$ where $\alpha \in [0, 1)$ is irrational.

Chapter 4
L^p Spaces

This chapter is mainly devoted to the study of the Lebesgue space L^p of measurable real functions whose p-th power of the absolute value is integrable on a given measure space. The fundamental inequalities of Hölder, Minkowski and Jensen provide essential ingredients for this study. We investigate the Banach space structure of L^p, and in the special case $p = 2$, the Hilbert space structure of L^2, which has important applications in probability theory.

Density theorems showing that any function of L^p can be well approximated by "smoother" functions play an important role in many analytic developments. As an application of the Hilbert space structure of L^2, we establish the Radon-Nikodym theorem, which allows one to decompose an arbitrary measure as the sum of a measure having a density with respect to the reference measure and a singular measure. The Radon-Nikodym theorem will be a crucial ingredient of the theory of conditional expectations in Chapter 11.

4.1 Definitions and the Hölder Inequality

Throughout this chapter, we consider a measure space $(E, \mathcal{A}, \mu)$. For a real $p \geq 1$, we let $\mathcal{L}^p(E, \mathcal{A}, \mu)$ denote the space of all measurable functions $f : E \longrightarrow \mathbb{R}$ such that

$$\int |f|^p \mathrm{d}\mu < \infty.$$

Note that the case $p = 1$ was already considered in Chapter 2. We also introduce the space $\mathcal{L}^\infty(E, \mathcal{A}, \mu)$ of all measurable functions $f : E \longrightarrow \mathbb{R}$ such that there exists a constant $C \in \mathbb{R}_+$ with

$$|f| \leq C, \ \ \mu \text{ a.e.}$$

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_4

We can also introduce the similar spaces $\mathcal{L}^p_{\mathbb{C}}(E, \mathcal{A}, \mu)$ and $\mathcal{L}^\infty_{\mathbb{C}}(E, \mathcal{A}, \mu)$ obtained by considering complex-valued functions instead of real functions. However, in this chapter, we will focus on the real case.

For every $p \in [1, \infty]$, we define an equivalence relation on $\mathcal{L}^p$ by setting

$$f \sim g \quad \text{if and only if} \;\; f = g, \;\; \mu \text{ a.e.}$$

We then consider the quotient space

$$L^p(E, \mathcal{A}, \mu) = \mathcal{L}^p(E, \mathcal{A}, \mu)/\sim \;.$$

An element of $L^p(E, \mathcal{A}, \mu)$ is thus an equivalence class of measurable functions that are equal μ a.e. In what follows we will almost always abuse notation and identify an element of $L^p(E, \mathcal{A}, \mu)$ with one of its representatives in $\mathcal{L}^p(E, \mathcal{A}, \mu)$: when we speak about a function in $L^p(E, \mathcal{A}, \mu)$, we mean that we have chosen a particular representative in $\mathcal{L}^p(E, \mathcal{A}, \mu)$, and the subsequent considerations will not depend on the choice of this representative.

We will often write $L^p(\mu)$, or even simply L^p, instead of $L^p(E, \mathcal{A}, \mu)$ if there is no risk of confusion. Notice that the space L^1 is just the space of all (equivalence classes of) integrable functions.

For every measurable function $f : E \longrightarrow \mathbb{R}$, and every $p \in [1, \infty)$, we set

$$\|f\|_p = \Big(\int |f|^p \mathrm{d}\mu\Big)^{1/p}$$

with the convention $\infty^{1/p} = \infty$, and

$$\|f\|_\infty = \inf\{C \in [0, \infty] : |f| \le C, \;\; \mu \text{ a.e.}\}$$

in such a way that $|f| \le \|f\|_\infty$, μ a.e. and $\|f\|_\infty$ is the smallest number in $[0, \infty]$ with this property. We observe that, if f and g are two measurable functions such that $f = g$, μ a.e., we have $\|f\|_p = \|g\|_p$ and thus we can define $\|f\|_p$ for $f \in L^p(E, \mathcal{A}, \mu)$.

Let $p, q \in [1, \infty]$. We say that p and q are *conjugate exponents* if

$$\frac{1}{p} + \frac{1}{q} = 1.$$

In particular, $p = 1$ and $q = \infty$ are conjugate.

Theorem 4.1 (Hölder Inequality) *Let p and q be conjugate exponents. Then, if f and g are two measurable functions from E into $\mathbb{R}$,*

$$\int |fg| \, \mathrm{d}\mu \leq \|f\|_p \|g\|_q .$$

In particular, $fg \in L^1(E, \mathcal{A}, \mu)$ if $f \in L^p(E, \mathcal{A}, \mu)$ and $g \in L^q(E, \mathcal{A}, \mu)$.

Remark In the last assertion, we implicitly use the fact that, if f and g are defined up to a set of zero μ-measure, the product fg (as well as the sum $f + g$) is also well defined up to a set of zero μ-measure.

Proof If $\|f\|_p = 0$, we have $f = 0$, μ a.e., which implies $\int |fg| \mathrm{d}\mu = 0$, and the inequality is trivial. We can thus assume that $\|f\|_p > 0$ and $\|g\|_q > 0$. Without loss of generality, we can also assume that $f \in L^p(E, \mathcal{A}, \mu)$ and $g \in L^q(E, \mathcal{A}, \mu)$ (otherwise $\|f\|_p \|g\|_q = \infty$ and there is nothing to prove).

The case $p = 1$ and $q = \infty$ is very easy, since we have $|fg| \leq \|g\|_\infty |f|$, μ a.e., which implies

$$\int |fg| \, \mathrm{d}\mu \leq \|g\|_\infty \int |f| \mathrm{d}\mu = \|g\|_\infty \|f\|_1.$$

In what follows we therefore assume that $1 < p < \infty$ (and thus $1 < q < \infty$).

Let $\alpha \in (0, 1)$. Then, for every $x \in \mathbb{R}_+$

$$x^\alpha - \alpha x \leq 1 - \alpha.$$

Indeed, define $\varphi_\alpha(x) = x^\alpha - \alpha x$ for $x \geq 0$. Then, for $x > 0$, we have $\varphi'_\alpha(x) = \alpha(x^{\alpha-1} - 1)$, and thus $\varphi'(x) > 0$ if $x \in (0, 1)$ and $\varphi'(x) < 0$ if $x \in (1, \infty)$. Hence φ_α attains its maximum at $x = 1$, which gives the desired inequality. By applying this inequality to $x = \frac{u}{v}$, where $u \geq 0$ and $v > 0$, we get

$$u^\alpha v^{1-\alpha} \leq \alpha u + (1 - \alpha) v,$$

and this inequality still holds if $v = 0$. We then take $\alpha = \frac{1}{p}$ (so that $1 - \alpha = \frac{1}{q}$) and

$$u = \frac{|f(x)|^p}{\|f\|_p^p} \qquad , \qquad v = \frac{|g(x)|^q}{\|g\|_q^q}$$

to arrive at

$$\frac{|f(x)g(x)|}{\|f\|_p \|g\|_q} \leq \frac{1}{p} \frac{|f(x)|^p}{\|f\|_p^p} + \frac{1}{q} \frac{|g(x)|^q}{\|g\|_q^q}.$$

By integrating the latter inequality with respect to $\mu(dx)$, we get

$$\frac{1}{\|f\|_p\|g\|_q}\int |fg|d\mu \leq \frac{1}{p}+\frac{1}{q}=1,$$

which completes the proof. □

The special case $p = q = 2$ of the Hölder inequality is of particular importance.

Cauchy-Schwarz Inequality If f and g are two measurable functions from E into $\mathbb{R}$,

$$\int |fg|\, d\mu \leq \Big(\int |f|^2 d\mu\Big)^{1/2}\Big(\int |g|^2 d\mu\Big)^{1/2} = \|f\|_2\|g\|_2.$$

Let us give some easy but important consequences of the Hölder inequality.

Corollary 4.2 *Suppose that μ is a finite measure. Then, if p and q are conjugate exponents, with $p > 1$, we have, for any real measurable function f on E,*

$$\|f\|_1 \leq \mu(E)^{1/q}\, \|f\|_p$$

and consequently $L^p \subset L^1$ for every $p \in (1,\infty]$. More generally, for any $1 \leq r < r' < \infty$,

$$\|f\|_r \leq \mu(E)^{\frac{1}{r}-\frac{1}{r'}}\, \|f\|_{r'},$$

and thus $L^q \subset L^p$ when $1 \leq p < q \leq \infty$.

Proof The bound $\|f\|_1 \leq \mu(E)^{1/q}\, \|f\|_p$ follows from the Hölder inequality by taking $g = 1$. For the second bound of the corollary, just replace f by $|f|^r$ and take $p = r'/r$. □

When μ is a probability measure, the corollary gives $\|f\|_p \leq \|f\|_q$ for every $1 \leq p < q \leq \infty$ (the case $q = \infty$ is immediate). This last inequality can also be derived from Jensen's inequality stated in the next theorem.

Theorem 4.3 (Jensen's Inequality) *Suppose that μ is a probability measure, and let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}_+$ be a convex function. Then, for every $f \in L^1(E, \mathcal{A}, \mu)$,*

$$\int \varphi \circ f\, d\mu \geq \varphi\Big(\int f\, d\mu\Big).$$

Remark The integral $\int \varphi \circ f\, d\mu$ is well defined as the integral of a nonnegative measurable function.

Proof Set

$$\mathcal{E}_\varphi = \{(a,b) \in \mathbb{R}^2 : \forall x \in \mathbb{R},\ \varphi(x) \geq ax + b\}.$$

Then elementary properties of convex functions show that

$$\forall x \in \mathbb{R}\,, \quad \varphi(x) = \sup_{(a,b)\in\mathcal{E}_\varphi} (ax+b).$$

Since $\varphi \circ f \geq af + b$ for every $(a,b) \in \mathcal{E}_\varphi$, we get

$$\int \varphi \circ f \, \mathrm{d}\mu \geq \sup_{(a,b)\in\mathcal{E}_\varphi} \int (af+b)\mathrm{d}\mu = \sup_{(a,b)\in\mathcal{E}_\varphi} \Big(a \int f \mathrm{d}\mu + b\Big) = \varphi\Big(\int f \, \mathrm{d}\mu\Big). \quad \square$$

4.2 The Banach Space $L^p(E, \mathcal{A}, \mu)$

We now state another fundamental inequality of the theory of L^p spaces.

Theorem 4.4 (Minkowski's Inequality) *Let $p \in [1,\infty]$, $f, g \in L^p(E, \mathcal{A}, \mu)$. Then, $f+g \in L^p(E, \mathcal{A}, \mu)$ and*

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p\,.$$

Proof The cases $p = 1$ and $p = \infty$ are very easy, just by using the inequality $|f+g| \leq |f| + |g|$. So we assume that $1 < p < \infty$. Writing

$$|f+g|^p \leq (|f|+|g|)^p \leq \big(2\max(|f|,|g|)\big)^p \leq 2^p(|f|^p + |g|^p)$$

we see that $\int |f+g|^p \mathrm{d}\mu < \infty$ and thus $f+g \in L^p$. Then, by integrating the inequality

$$|f+g|^p = |f+g| \times |f+g|^{p-1} \leq |f|\,|f+g|^{p-1} + |g|\,|f+g|^{p-1}$$

with respect to μ, we get

$$\int |f+g|^p \mathrm{d}\mu \leq \int |f|\,|f+g|^{p-1}\mathrm{d}\mu + \int |g|\,|f+g|^{p-1}\mathrm{d}\mu.$$

By applying the Hölder inequality to the conjugate exponents p and $q = p/(p-1)$, we get from the preceding display that

$$\int |f+g|^p \mathrm{d}\mu \leq \|f\|_p \Big(\int |f+g|^p \mathrm{d}\mu\Big)^{\frac{p-1}{p}} + \|g\|_p \Big(\int |f+g|^p \mathrm{d}\mu\Big)^{\frac{p-1}{p}}.$$

If $\int |f+g|^p d\mu = 0$, the inequality in the theorem is trivial. Otherwise, we can divide each side of the preceding inequality by $(\int |f+g|^p d\mu)^{(p-1)/p}$ and we get the desired result. □

Theorem 4.5 (Riesz) *For every $p \in [1, \infty]$, the space $L^p(E, \mathcal{A}, \mu)$ equipped with the norm $f \mapsto \|f\|_p$ is a Banach space (that is, a complete normed linear space, see the Appendix below).*

Proof Consider first the case $1 \le p < \infty$. Let us verify that $f \mapsto \|f\|_p$ defines a norm on L^p. We have

$$\|f\|_p = 0 \Rightarrow \int |f|^p d\mu = 0 \Rightarrow f = 0,\ \mu \text{ a.e.}$$

which means that $f = 0$ in L^p (f belongs to the equivalence class of the function 0; the point of replacing $\mathcal{L}^p(E, \mathcal{A}, \mu)$ by $L^p(E, \mathcal{A}, \mu)$ is indeed to get the property $\|f\|_p = 0 \Rightarrow f = 0$). The property $\|\lambda f\|_p = |\lambda|\|f\|_p$ for $\lambda \in \mathbb{R}$ is immediate and Minkowski's inequality gives the triangle inequality.

We then have to show that L^p equipped with the norm $\|\cdot\|_p$ is complete. Let $(f_n)_{n\ge1}$ be a Cauchy sequence in $(L^p, \|\cdot\|_p)$. We may find a strictly increasing sequence $(k_n)_{n\in\mathbb{N}}$ of positive integers such that, for every $n \ge 1$,

$$\|f_{k_{n+1}} - f_{k_n}\|_p \le 2^{-n}.$$

Set $g_n = f_{k_n}$, so that $\|g_{n+1} - g_n\|_p \le 2^{-n}$ for every $n \in \mathbb{N}$. Using the monotone convergence theorem and then Minkowski's inequality, we have

$$\begin{aligned}\int \Big(\sum_{n=1}^{\infty} |g_{n+1} - g_n|\Big)^p d\mu &= \lim_{N\uparrow\infty} \uparrow \int \Big(\sum_{n=1}^{N} |g_{n+1} - g_n|\Big)^p d\mu \\ &\le \lim_{N\uparrow\infty} \uparrow \Big(\sum_{n=1}^{N} \|g_{n+1} - g_n\|_p\Big)^p \\ &= \Big(\sum_{n=1}^{\infty} \|g_{n+1} - g_n\|_p\Big)^p \\ &< \infty.\end{aligned}$$

We have thus, by Proposition 2.7 (2),

$$\sum_{n=1}^{\infty} |g_{n+1} - g_n| < \infty\,, \quad \mu \text{ a.e.}$$

For every x such that $\sum_{n=1}^{\infty} |g_{n+1}(x) - g_n(x)| < \infty$, we can therefore set

$$h(x) = g_1(x) + \sum_{n=1}^{\infty}(g_{n+1}(x) - g_n(x)) = \lim_{n\to\infty} g_n(x)$$

where the series is absolutely convergent. On the (measurable) set N of zero μ-measure on which $\sum_{n=1}^{\infty} |g_{n+1}(x) - g_n(x)| = \infty$, we take $h(x) = 0$. The function h is then measurable by Lemma 1.15. Since g_n converges to h, μ a.e., we have $|h| = \liminf |g_n|$, μ a.e., and Fatou's lemma immediately gives

$$\int |h|^p \mathrm{d}\mu \leq \liminf \int |g_n|^p \mathrm{d}\mu \leq \sup_{n\geq 1} \int |g_n|^p \mathrm{d}\mu < \infty,$$

because the Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ is bounded in L^p. Hence, $h \in L^p$. Then, thanks again to Fatou's lemma, we have, for every $n \geq 1$,

$$\|h - g_n\|_p^p = \int |h - g_n|^p \mathrm{d}\mu \leq \liminf_{m\to\infty} \int |g_m - g_n|^p \mathrm{d}\mu = \liminf_{m\to\infty} \|g_m - g_n\|_p^p \leq (2^{-n+1})^p$$

where, for $m > n$, we have used the bound

$$\|g_m - g_n\|_p \leq \|g_{n+1} - g_n\|_p + \cdots + \|g_m - g_{m-1}\|_p \leq 2^{-n+1}.$$

The preceding bound shows that the sequence $(g_n)_{n\in\mathbb{N}}$ converges to h in L^p. Because a Cauchy sequence having a convergent subsequence must also converge, we have obtained that the sequence $(f_n)_{n\in\mathbb{N}}$ converges to h, which completes the proof in the case $1 \leq p < \infty$.

Let us turn to the (easier) case $p = \infty$. The fact that $f \mapsto \|f\|_\infty$ defines a norm on L^∞ is proved in the same way as in the case $p < \infty$. Then, let $(f_n)_{n\geq 1}$ be a Cauchy sequence in L^∞. By the definition of the L^∞-norm, for every $m > n \geq 1$, there is a measurable set $N_{m,n}$ of zero μ-measure such that we have $|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty$ for every $x \in E\backslash N_{n,m}$. Let N be the (countable) union of all sets $N_{m,n}$ for $m > n \geq 1$, so that we have again $\mu(N) = 0$, and $|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty$ for every $m > n \geq 1$ and every $x \in E\backslash N$. Hence, for every $x \in E\backslash N$, the sequence $(f_n(x))_{n\geq 1}$ is Cauchy in $\mathbb{R}$ and converges to a limit, which we denote by $g(x)$. We also take $g(x) = 0$ if $x \in N$. The function g is measurable, and by passing to the limit $m \to \infty$ in the preceding bound for $|f_n(x) - f_m(x)|$ we get

$$\sup_{x\in E\backslash N} |f_n(x) - g(x)| \leq \sup_{m\in\{n+1,n+2,\ldots\}} \|f_n - f_m\|_\infty.$$

The right-hand side tends to 0 as $n \to \infty$ and it follows that $\|f_n - g\|_\infty \longrightarrow 0$ as $n \to \infty$. Consequently $g \in L^\infty$ and $f_n \longrightarrow g$ in L^∞. This completes the proof. □

Example If $E = \mathbb{N}$ and μ is the counting measure, then, for every $p \in [1, \infty)$, the space $L^p(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$ is the space of all real sequences $a = (a_n)_{n\in\mathbb{N}}$ such that

$$\sum_{n=1}^{\infty} |a_n|^p < \infty$$

equipped with the norm

$$\|a\|_p = \Big(\sum_{n=1}^{\infty} |a_n|^p\Big)^{1/p}.$$

The space L^∞ is simply the space of all bounded real sequences $(a_n)_{n\in\mathbb{N}}$ with the norm $\|a\|_\infty = \sup_{n\in\mathbb{N}} |a_n|$. Note that in this case there are no (nonempty) sets of zero measure, and thus L^p coincides with $\mathcal{L}^p$. This space is usually denoted by $\ell^p = \ell^p(\mathbb{N})$. It plays an important role in the theory of Banach spaces.

In the proof of Theorem 4.5, we have derived an intermediate result which deserves to be stated formally.

Proposition 4.6 *Let $p \in [1, \infty)$ and let $(f_n)_{n\in\mathbb{N}}$ be a convergent sequence in $L^p(E, \mathcal{A}, \mu)$ with limit f. Then there is a subsequence $(f_{k_n})_{n\in\mathbb{N}}$ that converges pointwise to f except on a measurable set of zero μ-measure.*

Remark The result also holds for $p = \infty$, but in that case there is no need to extract a subsequence, as the convergence in L^∞ is equivalent to uniform convergence except on a set of zero measure.

Let us mention a useful by-product of Lemma 4.6. If $(f_n)_{n\in\mathbb{N}}$ is a convergent sequence in $L^p(E, \mathcal{A}, \mu)$ with limit f, and if we also know that $f_n(x) \longrightarrow g(x)$, $\mu(\mathrm{d}x)$ a.e., then $f = g$, μ a.e.

For $p \in [1, \infty)$, one may ask whether conversely a sequence $(f_n)_{n\in\mathbb{N}}$ in $L^p(E, \mathcal{A}, \mu)$ that converges μ a.e. also converges in the Banach space $L^p(E, \mathcal{A}, \mu)$. This is not true in general, but the dominated convergence theorem implies that, if the following two conditions hold,

(i) $f_n \longrightarrow f$, μ a.e.,
(ii) there exists a nonnegative measurable function g such that $\int g^p \mathrm{d}\mu < \infty$ and $|f_n| \le g$, μ a.e., for every $n \in \mathbb{N}$,

then the functions f_n, f are in L^p and $f_n \longrightarrow f$ in L^p.

See Exercise 4.3 for another simple criterion that allows one to derive L^p convergence from almost everywhere convergence.

The case $p = 2$ of Theorem 4.5 is especially important since the space L^2 has a structure of Hilbert space (see the Appendix below for basic facts about Hilbert spaces).

Theorem 4.7 *The space $L^2(E, \mathcal{A}, \mu)$ equipped with the scalar product*

$$\langle f, g \rangle = \int fg \, \mathrm{d}\mu$$

is a real Hilbert space.

Proof The Cauchy-Schwarz inequality shows that, if $f, g \in L^2$, fg is integrable and thus $\langle f, g \rangle$ is well defined. Then it is clear that $(f, g) \mapsto \langle f, g \rangle$ defines a scalar product on L^2, and $\langle f, f \rangle = \|f\|_2^2$. The fact that $(L^2, \|\cdot\|_2)$ is complete is a special case of Theorem 4.5. □

Classical results of the theory of Hilbert spaces can thus be applied to $L^2(E, \mathcal{A}, \mu)$. If particular, if $\Phi : L^2(E, \mathcal{A}, \mu) \longrightarrow \mathbb{R}$ is a continuous linear form, there exists a unique element g of L^2 such that $\Phi(f) = \langle f, g \rangle$ for every $f \in L^2$ (Theorem A.2 in the Appendix below). This result will be useful in what follows.

Remark Like preceding results, Theorem 4.7 extends to the case of complex functions. The space $L^2_{\mathbb{C}}(E, \mathcal{A}, \mu)$ (defined from $\mathcal{L}^2_{\mathbb{C}}(E, \mathcal{A}, \mu)$ in the same way as we defined $L^2(E, \mathcal{A}, \mu)$ from $\mathcal{L}^2(E, \mathcal{A}, \mu)$) is a complex Hilbert space for the scalar product

$$\langle f, g \rangle = \int f\bar{g} \, \mathrm{d}\mu,$$

where as usual $\bar{z}$ denotes the complex conjugate of $z \in \mathbb{C}$.

4.3 Density Theorems in L^p Spaces

Let (E, d) be a metric space. Recall that a function $f : E \longrightarrow \mathbb{R}$ is said to be Lipschitz if there exists a constant $K \geq 0$ such that

$$\forall x, y \in E \,, \quad |f(x) - f(y)| \leq K \, d(x, y).$$

A measure μ on $(E, \mathcal{B}(E))$ is said to be outer regular if, for every $A \in \mathcal{B}(E)$,

$$\mu(A) = \inf\{\mu(U) : U \text{ open set}, \;\; A \subset U\}.$$

This property holds as soon as μ is finite (Proposition 3.10). Theorem 3.13 (which we stated without proof) also shows that it holds for a Radon measure on a separable locally compact metric space, and in fact we will recover that result in the proof of the next statement.

Considering the general case of a measurable space $(E, \mathcal{A})$, we have introduced the notion of a simple function in Chapter 2. If μ is a measure on $(E, \mathcal{A})$, it is immediate that a simple function that is integrable with respect to μ is also in $L^p(\mu)$ for every $p \in [1, \infty)$.

Theorem 4.8 *Let $p \in [1, \infty)$.*

(1) *If $(E, \mathcal{A}, \mu)$ is a measure space, the set of all integrable simple functions is dense in $L^p(E, \mathcal{A}, \mu)$.*
(2) *If (E, d) is a metric space, and μ is an outer regular measure on $(E, \mathcal{B}(E))$, then the set of all bounded Lipschitz functions that belong to $L^p(E, \mathcal{B}(E), \mu)$ is dense in $L^p(E, \mathcal{B}(E), \mu)$.*
(3) *If (E, d) is a separable locally compact metric space, and μ is a Radon measure on E, then the set of all Lipschitz functions with compact support is dense in $L^p(E, \mathcal{B}(E), \mu)$.*

Proof

(1) Since we can decompose $f = f^+ - f^-$, it is enough to prove that, if f is a nonnegative function in L^p, then f is the limit in L^p of a sequence of simple functions. By Proposition 2.5 we can write

$$f = \lim_{n\to\infty} \uparrow \varphi_n$$

where, for every $n \in \mathbb{N}$, φ_n is a simple function and $0 \le \varphi_n \le f$. Then $\int |\varphi_n|^p d\mu \le \int |f|^p d\mu < \infty$ and thus $\varphi_n \in L^p$ (which is equivalent to $\varphi_n \in L^1$ for a simple function). Since $|f - \varphi_n|^p \le f^p$, the dominated convergence theorem gives

$$\lim_{n\to\infty} \int |f - \varphi_n|^p d\mu = 0.$$

(2) Thanks to part (1), it is enough to prove that any integrable simple function can be approximated in L^p by a bounded Lipschitz function. Clearly, it is enough to treat the case where $f = \mathbf{1}_A$, with $A \in \mathcal{B}(E)$ and $\mu(A) < \infty$. Then let $\varepsilon > 0$. Since μ is outer regular, we can find an open set O containing A and such that $\mu(O\backslash A) < (\varepsilon/2)^p$ (in particular $\mu(O) < \infty$). It follows that

$$\|\mathbf{1}_O - \mathbf{1}_A\|_p < \frac{\varepsilon}{2}.$$

Then, for every $k \ge 1$, set $\varphi_k(x) = (k\, d(x, O^c)) \wedge 1$ for every $x \in E$. The function φ_k is Lipschitz and bounded, and $\varphi_k \uparrow \mathbf{1}_O$ pointwise as $k \to \infty$. By dominated convergence, we have $\int |\mathbf{1}_O - \varphi_k|^p \mathrm{d}\mu \longrightarrow 0$ as $k \to \infty$, and thus we may find k large enough such that

$$\|\mathbf{1}_O - \varphi_k\|_p < \frac{\varepsilon}{2}.$$

For this value of k, we have $\|\mathbf{1}_A - \varphi_k\|_p \leq \|\mathbf{1}_A - \mathbf{1}_O\|_p + \|\mathbf{1}_O - \varphi_k\|_p < \varepsilon$.

(3) We use the following topological lemma, whose proof is postponed after the proof of Theorem 4.8. If A is a subset of E, A° denotes the interior of A, and $\bar{A}$ denotes the closure of A.

Lemma 4.9 *Let E be a separable locally compact metric space. There exists an increasing sequence $(L_n)_{n\in\mathbb{N}}$ of compact subsets of E such that, for every $n \in \mathbb{N}$, $L_n \subset L_{n+1}^\circ$ and*

$$E = \bigcup_{n\geq 1} L_n = \bigcup_{n\geq 1} L_n^\circ.$$

The lemma easily implies that any Radon measure μ on E is outer regular (this was already stated without proof in Theorem 3.13). Indeed, let A be a Borel subset of E, and let $\varepsilon > 0$. For every $n \in \mathbb{N}$, we may apply Proposition 3.10 to the restriction of μ to L_n° (which is a finite measure) and find an open subset O_n of E such that $A \cap L_n^\circ \subset O_n$ and

$$\mu((O_n \cap L_n^\circ)\backslash(A \cap L_n^\circ)) \leq \varepsilon\, 2^{-n}.$$

Up to replacing O_n by $O_n \cap L_n^\circ$, we can assume that $O_n \subset L_n^\circ$. Then the union $O := \bigcup_{n\in\mathbb{N}} O_n$ is an open subset of E and

$$\mu(O\backslash A) \leq \sum_{n\geq 1} \mu(O_n\backslash(A \cap L_n^\circ)) \leq \varepsilon.$$

This proves that μ is outer regular.

We can then apply part (2) of the theorem, and we see that it is enough to prove the following claim: if f is a bounded Lipschitz function on E such that $\int |f|^p \mathrm{d}\mu < \infty$, then f is the limit in L^p of a sequence of Lipschitz functions with compact support (such functions are automatically in L^p). In order to prove our claim, we start by using dominated convergence to get

$$\lim_{n\to\infty} \int_{(L_n^\circ)^c} |f|^p \mathrm{d}\mu = 0,$$

and thus $\|f - f\mathbf{1}_{L_n^\circ}\|_p \longrightarrow 0$ as $n \to \infty$. On the other hand, for every fixed $n \in \mathbb{N}$ and every $k \in \mathbb{N}$, if

$$\varphi_{n,k}(x) := (k\, d(x, (L_n^\circ)^c)) \wedge 1\ , \quad x \in E,$$

we have $\varphi_{n,k} \in L^p$ (since $\varphi_{n,k} \le \mathbf{1}_{L_n^\circ}$) and moreover, using again dominated convergence, $\varphi_{n,k}$ converges to $\mathbf{1}_{L_n^\circ}$ in L^p as $k \to \infty$. Finally, writing

$$\begin{aligned}\|f - f\varphi_{n,k}\|_p &\le \|f - f\mathbf{1}_{L_n^\circ}\|_p + \|f\mathbf{1}_{L_n^\circ} - f\varphi_{n,k}\|_p\\ &\le \|f - f\mathbf{1}_{L_n^\circ}\|_p + \|f\|_\infty \|\mathbf{1}_{L_n^\circ} - \varphi_{n,k}\|_p,\end{aligned}$$

we see that, for every $\varepsilon > 0$, we can choose n and then k large enough so that $\|f - f\varphi_{n,k}\|_p < \varepsilon$. This gives our claim, since the function $f\varphi_{n,k}$ is Lipschitz with compact support. □

Proof of Lemma 4.9 We first show that E is the union of an increasing sequence $(K_n)_{n\ge 1}$ of compact sets. To this end, let $(x_p)_{p\in\mathbb{N}}$ be a dense sequence in E. Let I be the set of all pairs (p, k) of positive integers such that the ball $\bar{B}_E(x_p, 2^{-k})$ is compact (we use the notation $\bar{B}_E(x, r)$ for the closed ball of radius r centered at x in E). For every $x \in E$, we can find an integer $k \in \mathbb{N}$ such that $\bar{B}_E(x, 2^{-k})$ is compact, and then an integer $p \in \mathbb{N}$ such that $d(x, x_p) < 2^{-k-1}$, which implies that the ball $\bar{B}_E(x_p, 2^{-k-1})$ contains x and is compact, and in particular $(p, k+1) \in I$. It follows that

$$E = \bigcup_{(p,k)\in I} \bar{B}_E(x_p, 2^{-k}).$$

On the other hand, since I is countable, we can find an increasing sequence $(I_n)_{n\in\mathbb{N}}$ of finite subsets of I such that I is the union of the sets I_n. Then, if we set

$$K_n = \bigcup_{(p,k)\in I_n} \bar{B}_E(x_p, 2^{-k})$$

for every $n \in \mathbb{N}$, the sets K_n are compact and their union is E.

We then construct the sequence $(L_n)_{n\in\mathbb{N}}$ by induction on n. We take $L_1 = K_1$. If L_n has been constructed, we find a cover of the compact set $K_{n+1} \cup L_n$ by a finite union $V_1 \cup V_2 \cup \ldots \cup V_p$ of open balls with compact closure whose centers belong to $K_{n+1} \cup L_n$, and we take $L_{n+1} = \bar{V}_1 \cup \bar{V}_2 \cup \ldots \cup \bar{V}_p$. The sequence $(L_n)_{n\in\mathbb{N}}$ satisfies the desired properties. □

Consequences For every $p \in [1, \infty)$:

(i) The space $C_c(\mathbb{R}^d)$ of all continuous functions with compact support on $\mathbb{R}^d$ is dense in $L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$. The measure λ can be replaced by any Radon measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

(ii) A step function on $\mathbb{R}$ is a real function defined on $\mathbb{R}$ that vanishes outside some compact interval $[a, b]$ and whose restriction to $[a, b]$ satisfies the properties stated at the beginning of Section 3.3. Then the set of all step functions on $\mathbb{R}$ is dense in $L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. Indeed, it is enough to verify that any function $f \in C_c(\mathbb{R})$ is the limit in L^p of a sequence of step functions. This follows from the fact that f can be written as the pointwise limit

$$f = \lim_{n\to\infty} \Big(\sum_{k\in\mathbb{Z}} f(\frac{k}{n})\, \mathbf{1}_{[\frac{k}{n}, \frac{k+1}{n})} \Big),$$

and dominated convergence shows that this limit also holds in L^p.

Application (Riemann-Lebesgue Lemma) Recall from Section 2.3 the definition of the Fourier transform $\widehat{f}$ of a function $f \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. We claim that

$$\widehat{f}(\xi) \underset{|\xi|\to\infty}{\longrightarrow} 0.$$

We first observe that it is enough to prove this when f is a step function on $\mathbb{R}$. Indeed, by property (ii) above we can find a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of step functions with compact support such that $\|\varphi_n - f\|_1$ tends to 0 as $n \to \infty$. Then, we just have to observe that

$$\sup_{\xi\in\mathbb{R}} |\widehat{f}(\xi) - \widehat{\varphi}_n(\xi)| = \sup_{\xi\in\mathbb{R}} \Big| \int f(x) e^{\mathrm{i}x\xi} dx - \int \varphi_n(x) e^{\mathrm{i}x\xi} dx \Big| \le \|f - \varphi_n\|_1.$$

Finally, if φ is a step function on $\mathbb{R}$, we can write

$$\varphi(x) = \sum_{j=1}^{p} \alpha_j\, \mathbf{1}_{(x_j, x_{j+1})}(x), \quad \lambda(\mathrm{d}x) \text{ a.e.}$$

where $\alpha_1, \ldots, \alpha_p \in \mathbb{R}$ and $x_1 \le x_2 \le \cdots \le x_{p+1}$, and we have

$$\widehat{\varphi}(\xi) = \sum_{j=1}^{p} \alpha_j \Big(\frac{e^{\mathrm{i}\xi x_{j+1}} - e^{\mathrm{i}\xi x_j}}{\mathrm{i}\xi} \Big) \underset{|\xi|\to\infty}{\longrightarrow} 0.$$

4.4 The Radon-Nikodym Theorem

Definition 4.10 Let μ and ν be two measures on $(E, \mathcal{A})$. We say that:

(i) ν is absolutely continuous with respect to μ (notation $\nu \ll \mu$) if

$$\forall A \in \mathcal{A}, \quad \mu(A) = 0 \Rightarrow \nu(A) = 0.$$

(ii) ν is singular with respect to μ (notation $\nu \perp \mu$) if there exists $N \in \mathcal{A}$ such that $\mu(N) = 0$ and $\nu(N^c) = 0$.

Example Let f be a nonnegative measurable function on E. The measure $\nu = f \cdot \mu$ defined in Corollary 2.6 by

$$\nu(A) := \int_A f \,\mathrm{d}\mu$$

is absolutely continuous with respect to μ.

Recall that the measure μ is σ-finite if E is a countable union of sets of finite μ-measure.

Theorem 4.11 (Radon-Nikodym) *Suppose that the measure μ is σ-finite, and let ν be another σ-finite measure on $(E, \mathcal{A})$. Then, there exists a unique pair (ν_a, ν_s) of σ-finite measures on $(E, \mathcal{A})$ such that*

(1) $\nu = \nu_a + \nu_s$.
(2) $\nu_a \ll \mu$ *and* $\nu_s \perp \mu$.

Moreover, there exists a nonnegative measurable function $g : E \longrightarrow \mathbb{R}_+$ such that $\nu_a = g \cdot \mu$, meaning that

$$\forall A \in \mathcal{A}, \quad \nu_a(A) = \int_A g \,\mathrm{d}\mu,$$

and the function g is unique in the sense that, if $\tilde{g}$ is another function satisfying the same properties, we have $\tilde{g} = g$, μ a.e.

The first part of the theorem, namely the decomposition of ν as the sum of an absolutely continuous part and a singular part, is known as *Lebesgue decomposition*. If $\nu \ll \mu$, the theorem gives the existence of a nonnegative measurable function g such that $\nu = g \cdot \mu$. In that case, the function g is often called the *Radon-Nikodym derivative* of μ with respect to ν.

Proof We provide details in the case when both μ and ν are finite, and, at the end of the proof, we explain the (straightforward) extension to the σ-finite case.
Step 1: μ and ν are finite and $\nu \leq \mu$. We assume that $\nu \leq \mu$, meaning that $\nu(A) \leq \mu(A)$ for every $A \in \mathcal{A}$ (and in particular $\nu \ll \mu$), which easily implies that $\int g \, d\nu \leq \int g \,\mathrm{d}\mu$ for every nonnegative measurable function g. Consider the linear mapping $\Phi : L^2(E, \mathcal{A}, \mu) \longrightarrow \mathbb{R}$ defined by

$$\Phi(f) := \int f \, d\nu.$$

Note that the integral is well defined since $\int |f| d\nu \leq \int |f| \mathrm{d}\mu$ and we know that $L^2(\mu) \subset L^1(\mu)$ when μ is finite. Furthermore the mapping Φ makes sense since $\Phi(f)$ does not depend on the representative of f chosen to compute $\int f d\nu$:

$$f = \tilde{f},\ \mu \text{ a.e.} \Rightarrow f = \tilde{f},\ \nu \text{ a.e.} \Rightarrow \int f d\nu = \int \tilde{f} d\nu.$$

The Cauchy-Schwarz inequality then shows that

$$|\Phi(f)| \leq \Big(\int f^2 d\nu\Big)^{1/2} \nu(E)^{1/2} \leq \Big(\int f^2 \mathrm{d}\mu\Big)^{1/2} \nu(E)^{1/2} = \nu(E)^{1/2} \|f\|_{L^2(\mu)},$$

where we write $\|f\|_{L^2(\mu)}$ (instead of $\|f\|_2$) for the norm on $L^2(E, \mathcal{A}, \mu)$, to avoid confusion with the norm on $L^2(E, \mathcal{A}, \nu)$. It follows that Φ is a continuous linear form on $L^2(E, \mathcal{A}, \mu)$ and, thanks to the classical property of Hilbert spaces stated in Theorem A.2, we know that there exists a function $g \in L^2(E, \mathcal{A}, \mu)$ such that

$$\forall f \in L^2(E, \mathcal{A}, \mu), \quad \Phi(f) = \langle f, g \rangle_{L^2(E,\mathcal{A},\mu)} = \int fg \, \mathrm{d}\mu.$$

In particular, taking $f = \mathbf{1}_A$, we have

$$\forall A \in \mathcal{A}, \quad \nu(A) = \int_A g \, \mathrm{d}\mu.$$

We can also observe that $0 \leq g \leq 1$, μ a.e. In fact, for every $\varepsilon > 0$,

$$\begin{aligned}
\mu(\{x \in E : g(x) \geq 1 + \varepsilon\}) &\geq \nu(\{x \in E : g(x) \geq 1 + \varepsilon\}) \\
&= \int_{\{x : g(x) \geq 1+\varepsilon\}} g \, \mathrm{d}\mu \\
&\geq (1 + \varepsilon)\mu(\{x \in E : g(x) \geq 1 + \varepsilon\})
\end{aligned}$$

which implies that $\mu(\{x \in E : g(x) \geq 1 + \varepsilon\}) = 0$, and then, by taking a sequence of values of ε decreasing to 0, that $\mu(\{x \in E : g(x) > 1\} = 0$. A similar argument left to the reader shows that $g \geq 0$, μ a.e. Up to replacing g by the function $(g \vee 0) \wedge 1$ (which is equal to g, μ a.e.), we can assume that $0 \leq g(x) \leq 1$ for every $x \in E$. In particular, we have obtained the properties stated in the theorem with $\nu_a = g \cdot \mu$ and $\nu_s = 0$. The uniqueness of g will be derived in the second step under more general assumptions.

Step 2: μ and ν are finite. We apply the first part of the proof after replacing μ by $\mu + \nu$. It follows that there exists a measurable function h such that $0 \leq h \leq 1$ and, for every function $f \in L^2(\mu + \nu)$,

$$\int f \, \mathrm{d}\nu = \int fh \, \mathrm{d}(\mu + \nu).$$

In particular, we get for every bounded measurable function f,

$$\int f\,\mathrm{d}\nu = \int fh\,\mathrm{d}\mu + \int fh\,\mathrm{d}\nu$$

and it follows that

$$\int f(1-h)\,\mathrm{d}\nu = \int fh\,\mathrm{d}\mu. \tag{4.1}$$

Using the monotone convergence theorem, we get that this last equality holds for any nonnegative measurable function f.

Set $N = \{x \in E : h(x) = 1\}$. Then, by taking $f = \mathbf{1}_N$ in (4.1), we obtain that $\mu(N) = 0$. The measure

$$\nu_s = \mathbf{1}_N \cdot \nu \qquad (\text{i.e. } \forall A \in \mathcal{A},\ \nu_s(A) = \nu(A \cap N))$$

is thus singular with respect to μ. On the other hand, if in (4.1) we replace f by $\mathbf{1}_{N^c}(1-h)^{-1} f$, we obtain that, for every nonnegative measurable function f,

$$\int_{N^c} f\,d\nu = \int_{N^c} f\,\frac{h}{1-h}\,\mathrm{d}\mu = \int fg\,\mathrm{d}\mu,$$

where $g = \mathbf{1}_{N^c}\frac{h}{1-h}$. Setting

$$\nu_a = \mathbf{1}_{N^c} \cdot \nu = g \cdot \mu$$

we get that properties (1) and (2) of the theorem both hold, and ν_a is the measure of density g with respect to μ. Note that $\int g\,\mathrm{d}\mu = \nu_a(E) < \infty$.

Let us verify the uniqueness of the pair (ν_a, ν_s). If $(\tilde{\nu}_a, \tilde{\nu}_s)$ is another pair satisfying the same properties (1) and (2), we have

$$\forall A \in \mathcal{A}, \quad \nu_s(A) - \tilde{\nu}_s(A) = \tilde{\nu}_a(A) - \nu_a(A). \tag{4.2}$$

Thanks to the properties $\nu_s \perp \mu$ and $\tilde{\nu}_s \perp \mu$, we can find two measurable sets N and $\tilde{N}$ of zero μ-measure such that $\nu_s(N^c) = 0$ and $\tilde{\nu}_s(\tilde{N}^c) = 0$, and then, for every $A \in \mathcal{A}$,

$$\begin{aligned}\nu_s(A) - \tilde{\nu}_s(A) &= \nu_s(A \cap (N \cup \tilde{N})) - \tilde{\nu}_s(A \cap (N \cup \tilde{N}))\\ &= \tilde{\nu}_a(A \cap (N \cup \tilde{N})) - \nu_a(A \cap (N \cup \tilde{N})) = 0\end{aligned}$$

using the fact that $\mu(N \cup \tilde{N}) = 0$ and the properties $\nu_a \ll \mu$ and $\tilde{\nu}_a \ll \mu$. It follows that $\nu_s = \tilde{\nu}_s$, and then $\nu_a = \tilde{\nu}_a$ using (4.2).

Finally, to get the uniqueness of g, suppose that there is another function $\tilde{g}$ such that $\nu_a = \tilde{g} \cdot \mu$. Then, writing $\{\tilde{g} > g\}$ for the set $\{x \in E : \tilde{g}(x) > g(x)\}$,

$$\int_{\{\tilde{g}>g\}} \tilde{g}\, \mathrm{d}\mu = \nu_a(\{\tilde{g} > g\}) = \int_{\{\tilde{g}>g\}} g\, \mathrm{d}\mu,$$

which implies

$$\int_{\{\tilde{g}>g\}} (\tilde{g} - g)\, \mathrm{d}\mu = 0$$

and Proposition 2.7 gives $\mathbf{1}_{\{\tilde{g}>g\}}(\tilde{g} - g) = 0$, μ a.e., which implies $\tilde{g} \leq g$, μ a.e. By interchanging g and $\tilde{g}$, we also get $g \leq \tilde{g}$, μ a.e., and we conclude that $g = \tilde{g}$, μ a.e. as desired.

Step 3: General case. If μ and ν are only supposed to be σ-finite, we may construct a sequence $(E_n)_{n\in\mathbb{N}}$ of disjoint measurable subsets of E such that $\mu(E_n) < \infty$ and $\nu(E_n) < \infty$ for every $n \in \mathbb{N}$, and $E = \bigcup_{n\in\mathbb{N}} E_n$ (we first find a countable partition $(E'_n)_{n\in\mathbb{N}}$ of E such that $\mu(E'_n) < \infty$ for every n, and another partition $(E''_n)_{m\in\mathbb{N}}$ of E such that $\nu(E''_m) < \infty$ for every m, and we order the collection $(E'_n \cap E''_m)_{(n,m)\in\mathbb{N}^2}$ in a sequence). Let μ_n be the restriction of μ to E_n and let ν_n be the restriction of ν to E_n. By step 2, we can write, for every $n \in \mathbb{N}$,

$$\nu_n = \nu_a^n + \nu_s^n$$

where $\nu_s^n \perp \mu_n$, and $\nu_a^n = g_n \cdot \mu_n$, and we can assume that the measurable function g_n vanishes on E_n^c (since $\mu_n(E_n^c) = 0$, we can always impose the latter condition by replacing g_n by $\mathbf{1}_{E_n} g_n$). The desired result follows by setting

$$\nu_a = \sum_{n\in\mathbb{N}} \nu_a^n\,, \quad \nu_s = \sum_{n\in\mathbb{N}} \nu_s^n\,, \quad g = \sum_{n\in\mathbb{N}} g_n,$$

noting that, for every $x \in E$, there is most one value of n for which $g_n(x) > 0$ (because the sets E_n are disjoint). The uniqueness of the triple (ν_a, ν_s, g) is obtained by very similar arguments as in the case of finite measures. In the analog of (4.2), one should suppose that $A \subset E_n$ for some $n \in \mathbb{N}$, and when proving the uniqueness of g, one should replace $\{\tilde{g} > g\}$ by $\{\tilde{g} > g\} \cap E_n$, but otherwise the proof goes through without change. □

Remark One may wonder whether the σ-finiteness assumption is really necessary in Theorem 4.11. To give a simple example, suppose that μ is the counting measure on $([0, 1], \mathcal{B}([0, 1]))$ and ν is Lebesgue measure on $([0, 1], \mathcal{B}([0, 1]))$. Then it is trivial that $\nu \ll \mu$ (the only set of zero μ-measure is $\varnothing$), but the reader will easily convince himself that one cannot write ν in the form $g \cdot \mu$ with some nonnegative measurable function g. This is not a contradiction since μ is not σ-finite.

Examples

(1) Take $(E, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and suppose that $\mu = f \cdot \lambda$, where λ is Lebesgue measure. Assume that the function f is positive on $(0, \infty)$ and vanishes on $(-\infty, 0]$. Then, if $\nu = g \cdot \lambda$ is another measure absolutely continuous with respect to λ, the Lebesgue decomposition of ν with respect to μ reads

$$\nu = h \cdot \mu + \theta$$

where $h(x) = \mathbf{1}_{(0,\infty)}(x)\, g(x)/f(x)$ and $\theta(\mathrm{d}x) = \mathbf{1}_{(-\infty,0]}(x)\, \nu(\mathrm{d}x)$.

(2) Take $E = [0, 1)$ and $\mathcal{A} = \mathcal{B}([0, 1))$, and also consider, for every $n \in \mathbb{N}$, the σ-field

$$\mathcal{F}_n = \sigma([\frac{i-1}{2^n}, \frac{i}{2^n}); i \in \{1, 2, \dots, 2^n\}),$$

so that any $A \in \mathcal{F}_n$ is a finite union of intervals $[(i-1)2^{-n}, i2^{-n})$. Write λ for Lebesgue measure on $[0, 1)$ and let ν be a finite measure on $\mathcal{B}([0, 1))$. By restricting λ and ν to sets in $\mathcal{F}_n$, we can view both λ and ν as measures on $([0, 1), \mathcal{F}_n)$, and we observe that ν is then absolutely continuous with respect to λ (just note that any nonempty set of $\mathcal{F}_n$ has positive λ-measure). Furthermore, an immediate verification shows that the Radon-Nikodym derivative of ν with respect to λ, when both ν and λ are viewed as measures on $([0, 1), \mathcal{F}_n)$, is given by

$$f_n(x) = \sum_{i=1}^{2^n} \frac{\nu([(i-1)2^{-n}, i2^{-n}))}{2^{-n}}\, \mathbf{1}_{[(i-1)2^{-n}, i2^{-n})}(x).$$

At the end of Section 12.3 below, we use martingale theory to prove that there is a Borel measurable function f such that $f_n(x) \longrightarrow f(x)$, λ a.e., and that the absolutely continuous part of the Lebesgue decomposition of ν with respect to λ (considering now ν and λ as measures on the σ-field $\mathcal{A} = \mathcal{B}([0, 1))$) is $f \cdot \lambda$.

Proposition 4.12 *Let μ be a σ-finite measure on $(E, \mathcal{A})$, and let ν be a finite measure on $(E, \mathcal{A})$. The following two conditions are equivalent:*

(i) $\nu \ll \mu$;
(ii) *For every $\varepsilon > 0$, there exists $\delta > 0$ such that, for every $A \in \mathcal{A}$, $\mu(A) < \delta$ implies $\nu(A) < \varepsilon$.*

Proof The fact that (ii) $\Rightarrow$ (i) is immediate. Conversely, if $\nu \ll \mu$, Theorem 4.11 allows us to find a nonnegative measurable function g on E such that $\nu = g \cdot \mu$. Note that $\int g\, \mathrm{d}\mu = \nu(E) < \infty$, and thus the dominated convergence theorem shows that

$$\int \mathbf{1}_{\{g>n\}} g\, \mathrm{d}\mu \underset{n\to\infty}{\longrightarrow} 0.$$

Then, given $\varepsilon > 0$, we can choose $M \in \mathbb{N}$ large enough so that

$$\int \mathbf{1}_{\{g>M\}} g \, \mathrm{d}\mu < \frac{\varepsilon}{2}.$$

If we fix $\delta = \varepsilon/(2M)$, we get for every $A \in \mathcal{A}$ such that $\mu(A) < \delta$,

$$\nu(A) = \int_A g \, \mathrm{d}\mu \leq \int \mathbf{1}_{\{g>M\}} g \, \mathrm{d}\mu + \int \mathbf{1}_{A\cap\{g\leq M\}} \, g \, \mathrm{d}\mu < \frac{\varepsilon}{2} + M\mu(A) \leq \varepsilon.$$

□

4.5 Exercises

Exercise 4.1 When $1 < p < \infty$, show that equality in the Hölder inequality $\int |fg| \, \mathrm{d}\mu \leq \|f\|_p \|g\|_q$ holds if and only if there exist two nonnegative reals α, β not both zero, such that $\alpha|f|^p = \beta|g|^q$, μ a.e.

Exercise 4.2

(1) Show that, if $\mu(E) < \infty$, one has $\|f\|_\infty = \lim_{p\to\infty} \|f\|_p$ for any measurable function $f : E \longrightarrow \mathbb{R}$.
(2) Show that the result of question (1) still holds when $\mu(E) = \infty$ if we assume that $\|f\|_p < \infty$ for some $p \in [1, \infty)$.

Exercise 4.3 We assume that $\mu(E) < \infty$. Let $(f_n)_{n\in\mathbb{N}}$ and f be real measurable functions on E, and let $p \in [1, \infty)$. Show that the conditions

(i) $f_n \longrightarrow f$, μ a.e.,
(ii) there exists a real $r > p$ such that $\sup_{n\in\mathbb{N}} \int |f_n|^r \mathrm{d}\mu < \infty$,

imply that $f_n \longrightarrow f$ in L^p.

Exercise 4.4 Let $p \in [1, \infty)$ and let $(f_n)_{n\in\mathbb{N}}$ and f be functions in $L^p(E, \mathcal{A}, \mu)$. Assume that $f_n \longrightarrow f$, μ a.e., and $\|f_n\|_p \longrightarrow \|f\|_p$ as $n \to \infty$. Show that $f_n \longrightarrow f$ in L^p. This extends Exercise 2.8. (*Hint:* Argue as in the proof of the dominated convergence theorem.)

Exercise 4.5 Let $p \in [1, \infty)$ and let $(f_n)_{n\in\mathbb{N}}$ and f be nonnegative functions in $L^p(E, \mathcal{A}, \mu)$. Show that, if $f_n \longrightarrow f$ in L^p, then, for every $r \in [1, p]$, f_n^r converges to f^r in $L^{p/r}$. (*Hint:* If $x, y \geq 0$ and $r \geq 1$, check that $|x^r - y^r| \leq r|x - y|(x^{r-1} + y^{r-1})$.)

Exercise 4.6 Let f be a real measurable function on $(E, \mathcal{A}, \mu)$ such that $\mu(\{x \in E : f(x) \neq 0\}) > 0$. For every real $p > 0$, set $\varphi(p) = \int |f|^p d\mu$.

(1) Prove that the set $I = \{p > 0 : \varphi(p) < \infty\}$ is an interval (possibly a singleton).
(2) Prove that the function $t \mapsto \log \varphi(t)$ is convex on I and that φ is continuous on I.
(3) Let $p, q, r \in I$ with $p \leq r \leq q$. Prove that $\varphi(r)^{1/r} \leq \max\{\varphi(p)^{1/p}, \varphi(q)^{1/q}\}$.

Exercise 4.7 (Hardy's Inequality) Let $p \in (1, \infty)$.

(1) Let $f : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ be a nonnegative continuous function with compact support contained in $(0, \infty)$. Set $F(0) = 0$ and, for every $x > 0$,

$$F(x) = \frac{1}{x} \int_0^x f(t)\, dt.$$

Prove that

$$\int_0^\infty F(x)^p\, dx = \frac{p}{p-1} \int_0^\infty f(x) F(x)^{p-1}\, dx.$$

(*Hint:* Use the fact that $F(x)^p = F(x)F(x)^{p-1} = (f(x) - xF'(x))F(x)^{p-1}$ and an integration by parts.) Infer that $F \in L^p(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \lambda)$, and

$$\|F\|_p \leq \frac{p}{p-1} \|f\|_p. \tag{4.3}$$

(2) We now assume only that $f \in L^p(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \lambda)$. Verify that the definition of F still makes sense and the bound (4.3) still holds.

Exercise 4.8

(1) Let f and g are two nonnegative measurable functions on $(E, \mathcal{A}, \mu)$ such that $fg \geq 1$. Prove that

$$\int_E f\, d\mu \int_E g\, d\mu \geq \mu(E)^2.$$

(2) Characterize the measure spaces $(E, \mathcal{A}, \mu)$ on which there exists a measurable function $f > 0$ such that both f and $1/f$ are integrable.

Exercise 4.9

(1) Let $p \in [1, \infty)$, let q be the conjugate exponent of p, and let $f \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. For every $x \in \mathbb{R}$, set $F(x) = \int_0^x f(t)\mathrm{d}t$ (where as usual $\int_0^x f(t)\mathrm{d}t = -\int_x^0 f(t)\mathrm{d}t$ if $x < 0$). Verify that F is well-defined, and prove that

$$\lim_{h \to 0, h > 0} \frac{\sup_{x \in \mathbb{R}} |F(x+h) - F(x)|}{|h|^{1/q}} = 0,$$

where $|h|^{1/q} = 1$ if $p = 1$. (*Hint:* Use the last question of Exercise 2.9.)

(2) Let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuously differentiable function, such that $g \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. Assume that $g' \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ for some $p \in [1, \infty)$. Prove that $g(x) \longrightarrow 0$ as $|x| \to \infty$.

Exercise 4.10 Give a proof of Proposition 4.12 that does not use Theorem 4.11. (*Hint:* Use Lemma 1.7 and Exercise 2.3.)

Chapter 5
Product Measures

Given two measurable spaces E and F and two measures μ and ν defined on E and F respectively, one can construct a measure on the Cartesian product $E \times F$ equipped with the product σ-field. This measure is called the product measure of μ and ν and denoted by $\mu \otimes \nu$. The integral of a real function $f(x, y)$ with respect to $\mu \otimes \nu$ can be evaluated by computing first the integral of the function $x \mapsto f(x, y)$ with respect to the measure $\mu(\mathrm{d}x)$ and then integrating the value of this integral with respect to $\nu(\mathrm{d}y)$, or conversely. This is the famous Fubini theorem, which holds under appropriate conditions on the function f, and in particular if f is nonnegative and measurable. Beyond its important applications in analysis (integration by parts, convolution, etc.) or in probability theory, the Fubini theorem is also an essential tool for the effective calculation of integrals. As a typical application, we compute the volume of the unit ball in $\mathbb{R}^d$.

5.1 Product σ-Fields

Let $(E, \mathcal{A})$ and $(F, \mathcal{B})$ be two measurable spaces. The product σ-field $\mathcal{A} \otimes \mathcal{B}$ was introduced in Chapter 1 as the σ-field on $E \times F$ defined by

$$\mathcal{A} \otimes \mathcal{B} := \sigma(A \times B \,:\, A \in \mathcal{A}, B \in \mathcal{B}).$$

Sets of the form $A \times B$, for $A \in \mathcal{A}$ and $B \in \mathcal{B}$, will be called (measurable) rectangles. It is easy to verify that $\mathcal{A} \otimes \mathcal{B}$ is also the smallest σ-field on $E \times F$ for which the two canonical projections $\pi_1 : E \times F \longrightarrow E$ and $\pi_2 : E \times F \longrightarrow F$ are measurable.

Let $(G, \mathcal{C})$ be another measurable space, and consider a function $f : G \longrightarrow E \times F$, which we write $f(x) = (f_1(x), f_2(x))$ for $x \in G$. By Lemma 1.12, f is measurable (when $E \times F$ is equipped with the product σ-field) if and only if both f_1 and f_2 are measurable.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_5

Remark The definition of the product σ-field is easily extended to a finite number of measurable spaces, say $(E_1, \mathcal{A}_1), \ldots, (E_n, \mathcal{A}_n)$:

$$\mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \cdots \otimes \mathcal{A}_n = \sigma(A_1 \times \cdots \times A_n \,:\, A_i \in \mathcal{A}_i \text{ for every } i \in \{1, \ldots, n\})$$

and the expected "associativity" properties hold. For instance, if $n = 3$,

$$(\mathcal{A}_1 \otimes \mathcal{A}_2) \otimes \mathcal{A}_3 = \mathcal{A}_1 \otimes (\mathcal{A}_2 \otimes \mathcal{A}_3) = \mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \mathcal{A}_3 \,.$$

Consider again the two measurable spaces $(E, \mathcal{A})$ and $(F, \mathcal{B})$. We introduce some notation in view of the next proposition. If $C \subset E \times F$, and $x \in E$, we let C_x be the subset of F defined by

$$C_x := \{y \in F : (x, y) \in C\}.$$

Similarly, if $y \in F$, we let C^y be the subset of E defined by

$$C^y := \{x \in E : (x, y) \in C\}.$$

If f is a function on $E \times F$, and $x \in E$, we let f_x be the function on F defined by $f_x(y) = f(x, y)$. Similarly, if $y \in F$, f^y is the function on E defined by $f^y(x) = f(x, y)$.

Proposition 5.1

(i) *Let $C \in \mathcal{A} \otimes \mathcal{B}$. Then, for every $x \in E$, $C_x \in \mathcal{B}$ and, for every $y \in F$, $C^y \in \mathcal{A}$.*
(ii) *Let $(G, \mathcal{G})$ be a measurable space and let $f : E \times F \longrightarrow G$ be measurable with respect to $\mathcal{A} \otimes \mathcal{B}$. Then, for every $x \in E$, f_x is $\mathcal{B}$-measurable, and, for every $y \in F$, f^y is $\mathcal{A}$-measurable.*

Proof

(i) Fix $x \in E$ and set

$$\mathcal{C} = \{C \in \mathcal{A} \otimes \mathcal{B} : C_x \in \mathcal{B}\}.$$

Then $\mathcal{C}$ contains all measurable rectangles (if $C = A \times B$, we have $C_x = B$ or $C_x = \varnothing$ according as $x \in A$ or $x \notin A$). On the other hand, it is easy to verify that $\mathcal{C}$ is a σ-field, and it follows that $\mathcal{C} = \mathcal{A} \otimes \mathcal{B}$. This gives the property $C_x \in \mathcal{B}$ for every $C \in \mathcal{A} \otimes \mathcal{B}$. The property $C^y \in \mathcal{A}$ is derived in a similar manner.

(ii) For every measurable subset D of G,

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

which belongs to $\mathcal{B}$ by part (i). □

5.2 Product Measures

We start by constructing the product measure of two σ-finite measures.

Theorem 5.2 *Let μ and ν be two σ-finite measures on $(E, \mathcal{A})$ and on $(F, \mathcal{B})$ respectively.*

(i) *There exists a unique measure m on $(E \times F, \mathcal{A} \otimes \mathcal{B})$ such that, for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$,*

$$m(A \times B) = \mu(A)\nu(B) \tag{5.1}$$

with the usual convention $0 \times \infty = 0$. The measure m is σ-finite, and is denoted by $m = \mu \otimes \nu$.

(ii) *For every $C \in \mathcal{A} \otimes \mathcal{B}$, the function $x \mapsto \nu(C_x)$ is $\mathcal{A}$-measurable on E, and the function $y \mapsto \mu(C^y)$ is $\mathcal{B}$-measurable on F, and we have*

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

Proof *Uniqueness* We verify that there is at most one measure m on $E \times F$ such that (5.1) holds. We first observe that we can find an increasing sequence $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$, and an increasing sequence $(B_n)_{n\in\mathbb{N}}$ in $\mathcal{B}$, such that $\mu(A_n) < \infty$ and $\nu(B_n) < \infty$, for every $n \in \mathbb{N}$, and $E = \bigcup_{n\in\mathbb{N}} A_n$, $F = \bigcup_{n\in\mathbb{N}} B_n$. Then, if $G_n = A_n \times B_n$, we have also

$$E \times F = \bigcup_{n\in\mathbb{N}} G_n.$$

Let m and m' be two measures on $\mathcal{A} \otimes \mathcal{B}$ that satisfy the property (5.1). Then,

- we have $m(C) = m'(C)$ for every measurable rectangle C, and the class $\mathcal{R}$ of all such rectangles is closed under finite intersections and generates the σ-field $\mathcal{A} \otimes \mathcal{B}$, by the definition of this σ-field;
- for every $n \in \mathbb{N}$, $G_n \in \mathcal{R}$ and $m(G_n) = \mu(A_n)\nu(B_n) = m'(G_n) < \infty$.

By Corollary 1.19, this suffices to give $m = m'$.

We also note that, once we have established the existence of m satisfying (5.1), the fact that it is σ-finite will be immediate since we will have $m(G_n) = \mu(A_n)\nu(B_n) < \infty$ for every $n \in \mathbb{N}$.

Existence We define the measure m via the first equality in (ii),

$$m(C) := \int_E \nu(C_x)\, \mu(\mathrm{d}x), \tag{5.2}$$

for every $C \in \mathcal{A}\otimes\mathcal{B}$. We first need to explain why the right-hand side of (5.2) makes sense. We note that $\nu(C_x)$ is well defined for every $x \in E$ thanks to Proposition 5.1. To verify that the integral in (5.2) makes sense, we also need to check that the function $x \mapsto \nu(C_x)$ is $\mathcal{A}$-measurable.

Suppose first that ν is finite and let $\mathcal{G}$ be the class of all sets $C \in \mathcal{A}\otimes\mathcal{B}$ such that the function $x \mapsto \nu(C_x)$ is $\mathcal{A}$-measurable. Then,

- $\mathcal{G}$ contains all measurable rectangles: if $C = A \times B$, $\nu(C_x) = \mathbf{1}_A(x)\nu(B)$.
- $\mathcal{G}$ is a monotone class: if $C, C' \in \mathcal{G}$ and $C \subset C'$, we have $\nu((C\backslash C')_x) = \nu(C_x) - \nu(C'_x)$ (here we use the finiteness of ν), and it follows that $C\backslash C' \in \mathcal{G}$. Similarly, if $(C_n)_{n\in\mathbb{N}}$ is an increasing sequence in $\mathcal{G}$, we have $\nu((\bigcup_{n\in\mathbb{N}} C_n)_x) = \lim \uparrow \nu((C_n)_x)$ for every $x \in E$, and therefore $\bigcup_{n\in\mathbb{N}} C_n \in \mathcal{G}$.

By the monotone class theorem (Theorem 1.18), $\mathcal{G}$ contains the σ-field generated by the measurable rectangles, and so we must have $\mathcal{G} = \mathcal{A}\otimes\mathcal{B}$. This gives the desired measurability property for the mapping $x \mapsto \nu(C_x)$.

In the general case where ν is only σ-finite, we consider the sequence $(B_n)_{n\in\mathbb{N}}$ introduced in the uniqueness part of the proof. For every $n \in \mathbb{N}$, we may replace ν by its restriction to B_n, which we denote by ν_n, and the finite case shows that $x \mapsto \nu_n(C_x)$ is measurable. Finally, we just have to write $\nu(C_x) = \lim \uparrow \nu_n(C_x)$ for every $x \in E$.

The preceding discussion shows that the definition (5.2) of $m(C)$ makes sense for every $C \in \mathcal{A}\otimes\mathcal{B}$. It is then easy to verify that m is a measure on $\mathcal{A}\otimes\mathcal{B}$: if $(C_n)_{n\in\mathbb{N}}$ is a sequence of disjoint elements of $\mathcal{A}\otimes\mathcal{B}$, the sets $(C_n)_x$, $n \in \mathbb{N}$ are also disjoint, for every $x \in E$, and thus

$$\begin{aligned} m\Big(\bigcup_{n\in\mathbb{N}} C_n\Big) &= \int_E \nu\Big(\bigcup_{n\in\mathbb{N}} (C_n)_x\Big)\,\mu(\mathrm{d}x) \\ &= \int_E \sum_{n\in\mathbb{N}} \nu((C_n)_x)\,\mu(\mathrm{d}x) \\ &= \sum_{n\in\mathbb{N}} \int_E \nu((C_n)_x)\,\mu(\mathrm{d}x) \\ &= \sum_{n\in\mathbb{N}} m(C_n), \end{aligned}$$

where, in the third equality, we use Proposition 2.5 to interchange integral and sum.

It is immediate that m verifies the property $m(A \times B) = \mu(A)\nu(B)$ for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$. So we have proved the existence of a measure m satisfying the properties in (i) and the first equality in (ii).

Furthermore, we can use exactly the same arguments to verify that the function $y \mapsto \mu(C^y)$ is $\mathcal{B}$-measurable, for every $C \in \mathcal{A} \otimes \mathcal{B}$, and that we can define a measure m' on $E \times F$ by the formula

$$m'(C) := \int_F \mu(C^y)\, \nu(\mathrm{d}y).$$

Since m' satisfies the same property $m'(A \times B) = \mu(A)\nu(B)$, the uniqueness part forces $m = m'$. It follows that m also satisfies the second equality in (ii) and the proof of the theorem is complete. □

Remarks

(i) The assumption that μ and ν are σ-finite is necessary at least for part (ii) of the theorem. Indeed, take $(E, \mathcal{A}) = (F, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $\mu = \lambda$ be Lebesgue measure and let ν be the counting measure. Then if $C = \{(x, x) : x \in \mathbb{R}\}$, we have $C_x = C^x = \{x\}$ for every $x \in \mathbb{R}$, and thus

$$\infty = \int \nu(C_x)\, \lambda(\mathrm{d}x) \neq \int \lambda(C^y)\, \nu(\mathrm{d}y) = 0.$$

(ii) Suppose now that we consider an arbitrary finite number of σ-finite measures $\mu_1, \ldots, \mu_n$, defined on $(E_1, \mathcal{A}_1), \ldots, (E_n, \mathcal{A}_n)$ respectively. We can then define the product measure $\mu_1 \otimes \cdots \otimes \mu_n$ on $(E_1 \times \cdots \times E_n, \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n)$ by setting

$$\mu_1 \otimes \cdots \otimes \mu_n = \mu_1 \otimes (\mu_2 \otimes (\cdots \otimes \mu_n)).$$

The way we insert parentheses is in fact unimportant since $\mu_1 \otimes \cdots \otimes \mu_n$ is characterized by its values on (measurable) rectangles,

$$\mu_1 \otimes \cdots \otimes \mu_n(A_1 \times \cdots \times A_n) = \mu_1(A_1) \ldots \mu_n(A_n).$$

Example If $(E, \mathcal{A}) = (F, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $\mu = \nu = \lambda$, one easily checks that $\lambda \otimes \lambda$ is Lebesgue measure on $\mathbb{R}^2$ (just observe that Lebesgue on $\mathbb{R}^2$ is characterized by its values on rectangles of the form $[a, b] \times [c, d]$, again by a straightforward application of Corollary 1.19). This fact is easily generalized to higher dimensions, and we will use it without further comment in what follows. We may also observe that it would have been enough to construct Lebesgue measure on the real line in Chapter 3.

5.3 The Fubini Theorems

We start with the case of nonnegative functions. As in the preceding section, we consider two measurable spaces $(E, \mathcal{A})$ and $(F, \mathcal{B})$, and the product $E \times F$ is equipped with the σ-field $\mathcal{A} \otimes \mathcal{B}$.

Theorem 5.3 (Fubini-Tonelli) *Let μ and ν be two σ-finite measures on $(E, \mathcal{A})$ and on $(F, \mathcal{B})$ respectively, and let $f : E \times F \longrightarrow [0, \infty]$ be a measurable function.*

(i) *The functions*

$$E \ni x \mapsto \int_F f(x, y)\, \nu(\mathrm{d}y),$$

$$F \ni y \mapsto \int_E f(x, y)\, \mu(\mathrm{d}x),$$

with values in $[0, \infty]$, are respectively $\mathcal{A}$-measurable and $\mathcal{B}$-measurable.

(ii) *We have*

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

Remark The existence of the integrals $\int_E f(x, y)\, \mu(\mathrm{d}x)$ and $\int_F f(x, y)\, \nu(\mathrm{d}y)$ is justified by Proposition 5.1 (ii).

Proof

(i) Let $C \in \mathcal{A} \otimes \mathcal{B}$. If $f = \mathbf{1}_C$, we know from Theorem 5.2 that the function $x \mapsto \int_F f(x, y)\nu(\mathrm{d}y) = \nu(C_x)$ is $\mathcal{A}$-measurable, and similarly $y \mapsto \int_E f(x, y)\mu(\mathrm{d}x) = \mu(C^y)$ is $\mathcal{B}$-measurable. By a linearity argument we obtain that (i) holds for every (nonnegative) simple function. Finally, for a general nonnegative measurable function, f, Proposition 2.5 allows us to write $f = \lim \uparrow f_n$, where $(f_n)_{n\in\mathbb{N}}$ is an increasing sequence of nonnegative simple functions. It follows that, for every $x \in E$, the function $f_x(y) = f(x, y)$ is also the increasing limit of the simple functions $(f_n)_x(y) = f_n(x, y)$, and the monotone convergence theorem gives

$$\int_F f(x, y)\, \nu(\mathrm{d}y) = \lim \uparrow \int_F f_n(x, y)\, \nu(\mathrm{d}y).$$

Hence $x \mapsto \int_F f(x, y)\, \nu(\mathrm{d}y)$ is a pointwise limit of measurable functions and is therefore also measurable. The argument is the same for the function $y \mapsto \int_E f(x, y)\, \mu(\mathrm{d}x)$.

(ii) If $f = \mathbf{1}_C$, the desired formula reduces to

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

which holds by Theorem 5.2. By linearity, we get that the formula also holds when f is a nonnegative simple function. In the general case, we use Proposition 2.5 (as in part (i) of the proof) to write $f = \lim f_n$, where $(f_n)_{n\in\mathbb{N}}$ is an increasing sequence of nonnegative simple functions, and we observe that we have

$$\int_E \Big(\int_F f(x,y)\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x) = \lim_{n\to\infty} \uparrow \int_E \Big(\int_F f_n(x,y)\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x)$$

by two successive applications of the monotone convergence theorem. Since we have also $\int_{E\times F} f\,\mathrm{d}\mu\otimes\nu = \lim \uparrow \int_{E\times F} f_n\,\mathrm{d}\mu\otimes\nu$, again by monotone convergence, the desired formula follows from the case of simple functions. □

We now consider functions of arbitrary sign and derive another form of the Fubini theorem. We let μ and ν be as in Theorem 5.3.

Theorem 5.4 (Fubini-Lebesgue) *Let $f \in \mathcal{L}^1(E\times F, \mathcal{A}\otimes\mathcal{B}, \mu\otimes\nu)$. Then:*

(a) $\mu(\mathrm{d}x)$ *a.e., the function $y \mapsto f(x,y)$ belongs to $\mathcal{L}^1(F,\mathcal{B},\nu)$,*
$\nu(\mathrm{d}y)$ *a.e., the function $x \mapsto f(x,y)$ belongs to $\mathcal{L}^1(E,\mathcal{A},\mu)$.*
(b) *The functions $x \mapsto \int_F f(x,y)\,\nu(\mathrm{d}y)$ and $y \mapsto \int_E f(x,y)\,\mu(\mathrm{d}x)$ belong to $\mathcal{L}^1(E,\mathcal{A},\mu)$ and $\mathcal{L}^1(F,\mathcal{B},\nu)$ respectively.*
(c) *We have*

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

Remarks

(i) In (b), the function $x \mapsto \int_F f(x,y)\,\nu(\mathrm{d}y)$ is only defined (thanks to (a)) outside a measurable set of values of x of zero μ-measure. We can always agree that this function is equal to 0 on the latter set of zero measure, and similarly for the function $y \mapsto \int_E f(x,y)\,\mu(\mathrm{d}x)$.
(ii) The same statement holds for a function $f \in \mathcal{L}^1_{\mathbb{C}}(E\times F, \mathcal{A}\otimes\mathcal{B}, \mu\otimes\nu)$ with the obvious adaptations (in fact we just have to deal separately with the real and the imaginary part of f).

Proof

(a) By Theorem 5.3 applied to $|f|$, we have

$$\int_E \Big(\int_F |f(x,y)|\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x) = \int |f|\,\mathrm{d}\mu\otimes\nu < \infty.$$

Thanks to Proposition 2.7 (2), this implies that $\int_F |f(x,y)|\,\nu(\mathrm{d}y) < \infty$, $\mu(\mathrm{d}x)$ a.e. Thus the function $y \mapsto f(x,y)$ (which is measurable by Theorem 5.1) belongs to $\mathcal{L}^1(F,\mathcal{B},\nu)$, except for x belonging to the $\mathcal{A}$-measurable set

$$N := \Big\{x \in E : \int_F |f(x,y)|\nu(\mathrm{d}y) = \infty\Big\},$$

which has zero μ-measure. The same argument applies to the function $x \mapsto f(x,y)$.

(b) If $x \in N^c$, $\int_F f(x,y)\nu(\mathrm{d}y)$ is well defined since the the function $y \mapsto f(x,y)$ is ν-integrable. As already mentioned in the remark following the theorem, we agree that $\int_F f(x,y)\nu(\mathrm{d}y) = 0$ if $x \in N$. Then the function

$$x \mapsto \int f(x,y)\,\nu(\mathrm{d}y) = \mathbf{1}_{N^c}(x)\int f^+(x,y)\,\nu(\mathrm{d}y) - \mathbf{1}_{N^c}(x)\int f^-(x,y)\,\nu(\mathrm{d}y)$$

is $\mathcal{A}$-measurable by Theorem 5.3 (i), and we have also, thanks to part (ii) of the same theorem,

$$\int_E \Big|\int_F f(x,y)\,\nu(\mathrm{d}y)\Big|\mu(\mathrm{d}x) \le \int_E \Big(\int_F |f(x,y)|\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x) = \int |f|\,\mathrm{d}\mu\otimes\nu < \infty.$$

This proves that the function $x \mapsto \int f(x,y)\,\nu(\mathrm{d}y)$ is in $\mathcal{L}^1(E,\mathcal{A},\mu)$. The same argument applies to the function $y \mapsto \int_E f(x,y)\,\mu(\mathrm{d}x)$.

(c) Theorem 5.3 gives the two equalities

$$\int_E \Big(\int_F f^+(x,y)\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x) = \int_{E\times F} f^+\,\mathrm{d}\mu\otimes\nu,$$
$$\int_E \Big(\int_F f^-(x,y)\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x) = \int_{E\times F} f^-\,\mathrm{d}\mu\otimes\nu,$$

and we can replace the integral over E by an integral over N^c in the left-hand sides of these formulas. Then we just have to subtract the second equality from the first one (using part (b)), in order to arrive at the first equality in (c), and the second equality is derived in a similar manner. □

In what follows, we will refer to either Theorem 5.3 or Theorem 5.4 as "the Fubini theorem". It will be clear which of the two statements is considered.

Remark The assumption $f \in \mathcal{L}^1(\mu \otimes \nu)$ in Theorem 5.4 is crucial. It may indeed happen that both (a) and (b) hold, so that the quantities

$$\int_E \Big(\int_F f(x,y)\,\nu(\mathrm{d}y)\Big)\mu(\mathrm{d}x) \text{ and } \int_F \Big(\int_E f(x,y)\,\mu(\mathrm{d}x)\Big)\nu(\mathrm{d}y)$$

are well defined, but the formula in (c) does not hold! To give an example, consider the function

$$f(x,y) = 2e^{-2xy} - e^{-xy}$$

defined for $(x,y) \in (0,\infty) \times (0,1]$. Then, for every $y \in (0,1]$,

$$\int_{(0,\infty)} f(x,y)\,\mathrm{d}x = 2\int_0^\infty e^{-2xy}\mathrm{d}x - \int_0^\infty e^{-xy}\mathrm{d}x = 0$$

and for every $x \in (0,\infty)$,

$$\int_{(0,1]} f(x,y)\mathrm{d}y = 2\int_0^1 e^{-2xy}\mathrm{d}y - \int_0^1 e^{-xy}\mathrm{d}y = \frac{e^{-x} - e^{-2x}}{x}.$$

We then see that

$$\int_{(0,1]} \Big(\int_{(0,\infty)} f(x,y)\,\mathrm{d}x\Big)\mathrm{d}y = 0$$

whereas

$$\int_{(0,\infty)} \Big(\int_{(0,1]} f(x,y)\,\mathrm{d}y\Big)\mathrm{d}x = \int_0^\infty \frac{e^{-x} - e^{-2x}}{x}\mathrm{d}x > 0.$$

This is not a contradiction with Theorem 5.4, because

$$\int_{(0,\infty)\times(0,1)} |f(x,y)|\,\mathrm{d}x\mathrm{d}y = \infty$$

(to see this, note that there exists a constant $c > 0$ such that $f(x,y) \geq c$ if $x \leq 1/(8y)$ and use $\int_0^1 y^{-1}\mathrm{d}y = \infty$).

In practice, one should remember that the application of the Fubini theorem is always valid for nonnegative functions (Theorem 5.3) but that, in the case of functions of arbitrary sign, it is necessary to verify that

$$\int |f|\,\mathrm{d}\mu \otimes \nu < \infty.$$

Most of the time this verification is done by using the case of nonnegative functions.

Notation When the application of the Fubini theorem is valid (and only in this case), we often remove parentheses and write

$$\int f \, \mathrm{d}\mu \otimes \nu = \int_E \int_F f(x, y)\, \mu(\mathrm{d}x)\nu(\mathrm{d}y).$$

5.4 Applications

5.4.1 Integration by Parts

In this section, we extend the classical integration by parts formula for the Riemann integral. Let f and g be two measurable functions from $\mathbb{R}$ into $\mathbb{R}$. Suppose that f and g are locally integrable, meaning that the restriction of f (or of g) to any compact interval of $\mathbb{R}$ is integrable with respect to Lebesgue measure on this interval. Then set, for every $x \in \mathbb{R}$,

$$F(x) = \int_0^x f(t)\,\mathrm{d}t = \begin{cases} \int_{[0,x]} f(t)\,\mathrm{d}t & \text{if } x \geq 0 \\ -\int_{[x,0]} f(t)\,\mathrm{d}t & \text{if } x < 0 \end{cases}$$

$$G(x) = \int_0^x g(t)\,\mathrm{d}t.$$

With this notation, we have $F(b) - F(a) = \int_a^b f(t)\,\mathrm{d}t$ for every $a, b \in \mathbb{R}$ with $a \leq b$, and the similar formula for $G(b) - G(a)$. The integration by parts formula states that

$$F(b)G(b) = F(a)G(a) + \int_a^b f(t)G(t)\mathrm{d}t + \int_a^b F(t)g(t)\mathrm{d}t.$$

Let us justify this equality, which is easily seen to be equivalent to

$$\int_a^b f(t)(G(t) - G(a))\,\mathrm{d}t = \int_a^b (F(b) - F(t))g(t)\,\mathrm{d}t.$$

To establish the latter identity, we write

$$\begin{aligned}\int_a^b f(t)(G(t) - G(a))\,\mathrm{d}t &= \int_a^b f(t)\Big(\int_a^t g(s)\mathrm{d}s\Big)\mathrm{d}t \\ &= \int_a^b \Big(\int_a^b \mathbf{1}_{\{s \leq t\}} f(t)g(s)\mathrm{d}s\Big)\mathrm{d}t\end{aligned}$$

$$= \int_a^b \Big(\int_a^b \mathbf{1}_{\{s \le t\}} f(t) g(s) \mathrm{d}t \Big) \mathrm{d}s$$

$$= \int_a^b g(s) \Big(\int_s^b f(t) \mathrm{d}t \Big) \mathrm{d}s$$

$$= \int_a^b g(s)(F(b) - F(s)) \mathrm{d}s.$$

In the third equality, we have applied Theorem 5.4 to the function

$$\varphi(s,t) = \mathbf{1}_{\{s \le t\}} f(t) g(s)$$

noting that, thanks to Theorem 5.3, we have

$$\int_{[a,b]^2} |\varphi(s,t)| \mathrm{d}s \mathrm{d}t \le \int_{[a,b]^2} |f(t)||g(s)| \mathrm{d}s \mathrm{d}t = \Big(\int_{[a,b]} |f(t)| \mathrm{d}t \Big) \Big(\int_{[a,b]} |g(s)| \mathrm{d}s \Big) < \infty.$$

5.4.2 Convolution

Throughout this section, we consider the measure space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$. Let f and g be two real measurable functions on $\mathbb{R}^d$. For $x \in \mathbb{R}^d$, the convolution

$$f * g(x) = \int_{\mathbb{R}^d} f(x-y) g(y) \,\mathrm{d}y$$

is well defined provided

$$\int_{\mathbb{R}^d} |f(x-y) g(y)| \,\mathrm{d}y < \infty.$$

In that case, the fact that Lebesgue measure on $\mathbb{R}^d$ is invariant under translations and under the symmetry $y \to -y$ shows that $g * f(x)$ is also well defined and $g * f(x) = f * g(x)$.

Proposition 5.5 *Let $f \in L^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$ and $g \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$, where $p \in [1, \infty]$. Then, for λ a.e. $x \in \mathbb{R}^d$, the convolution $f * g(x)$ is well defined. Moreover $f * g \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$ and $\|f * g\|_p \le \|f\|_1 \|g\|_p$.*

Remark Again, we may agree that, on the set of zero Lebesgue measure where $f * g$ is not well defined, we take $f * g(x) = 0$.

Proof We leave aside the case $p = \infty$, which is easier (and where stronger results are proved in Proposition 5.6 below). We assume that $\|f\|_1 > 0$, since otherwise the results are trivial. Since the measure with density $\|f\|_1^{-1} |f(x)|$ with respect to

λ is a probability measure, Jensen's inequality gives, for every $x \in \mathbb{R}^d$,

$$\Big(\int_{\mathbb{R}^d} |g(x-y)|\,|f(y)|\,\mathrm{d}y\Big)^p = \|f\|_1^p\Big(\int_{\mathbb{R}^d} |g(x-y)|\,\frac{|f(y)|}{\|f\|_1}\,\mathrm{d}y\Big)^p$$
$$\le \|f\|_1^{p-1}\int_{\mathbb{R}^d} |g(x-y)|^p\,|f(y)|\,\mathrm{d}y.$$

Using Theorem 5.3, we have then

$$\begin{aligned}\int_{\mathbb{R}^d}\Big(\int_{\mathbb{R}^d} |g(x-y)|\,|f(y)|\,\mathrm{d}y\Big)^p\mathrm{d}x &\le \|f\|_1^{p-1}\int_{\mathbb{R}^d}\Big(\int_{\mathbb{R}^d} |g(x-y)|^p\,|f(y)|\,\mathrm{d}y\Big)\mathrm{d}x\\ &= \|f\|_1^{p-1}\int_{\mathbb{R}^d}\Big(\int_{\mathbb{R}^d} |g(y)|^p\,|f(x-y)|\,\mathrm{d}y\Big)\mathrm{d}x\\ &= \|f\|_1^{p-1}\int_{\mathbb{R}^d} |g(y)|^p\Big(\int_{\mathbb{R}^d} |f(x-y)|\mathrm{d}x\Big)\mathrm{d}y\\ &= \|f\|_1^p\,\|g\|_p^p\\ &< \infty,\end{aligned}$$

which shows that

$$\int_{\mathbb{R}^d} |f(x-t)|\,|g(t)|\mathrm{d}t < \infty, \qquad \lambda(\mathrm{d}x) \text{ a.e.}$$

and gives the first assertion. The second one also follows since the previous calculation gives

$$\int_{\mathbb{R}^d} |f * g(x)|^p\mathrm{d}x \le \int_{\mathbb{R}^d}\Big(\int_{\mathbb{R}^d} |g(x-y)|\,|f(y)|\,\mathrm{d}y\Big)^p\mathrm{d}x \le \|f\|_1^p\,\|g\|_p^p. \qquad \square$$

The next proposition gives another important case where the convolution $f * g$ is well defined and enjoys nice continuity properties.

Proposition 5.6 *Let $p \in [1,\infty)$ and $q \in (1,\infty]$ be conjugate exponents ($\frac{1}{p}+\frac{1}{q}=1$). Let $f \in L^p(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d),\lambda)$ and $g \in L^q(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d),\lambda)$. Then, the convolution $f*g(x)$ is well defined and $|f*g(x)| \le \|f\|_p\|g\|_q$, for every $x \in \mathbb{R}^d$. Furthermore, the function $f * g$ is uniformly continuous on $\mathbb{R}^d$.*

Proof By the Hölder inequality,

$$\int_{\mathbb{R}^d} |f(x-y)g(y)|\,\mathrm{d}y \le \Big(\int |f(x-y)|^p\mathrm{d}y\Big)^{1/p}\|g\|_q = \|f\|_p\|g\|_q.$$

This gives the first assertion and also proves that $f * g$ is bounded above by $\|f\|_p\|g\|_q$. To get the property of uniform continuity, we use the next lemma.

Lemma 5.7 *Set $\sigma_x(y) = y - x$ for every $x, y \in \mathbb{R}^d$. Let $p \in [1, \infty)$ and $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$, Then the mapping $x \mapsto f \circ \sigma_x$ is uniformly continuous from $\mathbb{R}^d$ into $L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$.*

Given this lemma, it is easy to complete the proof of the proposition. Indeed, for every $x, x' \in \mathbb{R}^d$,

$$
\begin{aligned}
|f * g(x) - f * g(x')| &\le \int |f(x-y) - f(x'-y)|\,|g(y)|\,\mathrm{d}y \\
&\le \|g\|_q \Big(\int |f(x-y) - f(x'-y)|^p \mathrm{d}y \Big)^{1/p} \\
&= \|g\|_q \, \|f \circ \sigma_{-x} - f \circ \sigma_{-x'}\|_p
\end{aligned}
$$

and we use Lemma 5.7 to say that $\|f \circ \sigma_{-x} - f \circ \sigma_{-x'}\|_p$ tends to 0 when $x - x' \to 0$. □

Proof of Lemma 5.7 Suppose first that f belongs to the space $C_c(\mathbb{R}^d)$ of all real continuous functions with compact support on $\mathbb{R}^d$. Then,

$$
\|f \circ \sigma_x - f \circ \sigma_y\|_p^p = \int |f(z-x) - f(z-y)|^p \mathrm{d}z = \int |f(z) - f(z-(y-x))|^p \mathrm{d}z
$$

which tends to 0 when $y - x \to 0$ by an application of the dominated convergence theorem. In the general case, Theorem 4.8 (3) allows us to find a sequence $(f_n)_{n\in\mathbb{N}}$ in $C_c(\mathbb{R}^d)$ that converges to f in $L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$. Then, for every $x, y \in \mathbb{R}^d$,

$$
\begin{aligned}
&\|f \circ \sigma_x - f \circ \sigma_y\|_p \\
&\quad \le \|f \circ \sigma_x - f_n \circ \sigma_x\|_p + \|f_n \circ \sigma_x - f_n \circ \sigma_y\|_p + \|f_n \circ \sigma_y - f \circ \sigma_y\|_p \\
&\quad = 2\|f - f_n\|_p + \|f_n \circ \sigma_x - f_n \circ \sigma_y\|_p .
\end{aligned}
$$

Fix $\varepsilon > 0$ and first choose n large enough so that $\|f - f_n\|_p < \varepsilon/4$. Since $f_n \in C_c(\mathbb{R}^d)$, we can then choose $\delta > 0$ such that $\|f_n \circ \sigma_x - f_n \circ \sigma_y\|_p \le \varepsilon/2$ if $|x - y| < \delta$. The preceding display then shows that $\|f \circ \sigma_x - f \circ \sigma_y\|_p \le \varepsilon$ if $|x - y| < \delta$. This completes the proof. □

Approximations of the Dirac Measure Let us say that a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of nonnegative continuous functions on $\mathbb{R}^d$ is an approximation of the Dirac measure δ_0 if:

- For every $n \in \mathbb{N}$,

$$
\int_{\mathbb{R}^d} \varphi_n(x)\,\mathrm{d}x = 1. \tag{5.3}
$$

- For every $\delta > 0$,

$$\lim_{n\to\infty} \int_{\{|x|>\delta\}} \varphi_n(x)\,\mathrm{d}x = 0. \tag{5.4}$$

It is easy to construct approximations of δ_0. If φ is any nonnegative continuous function on $\mathbb{R}^d$ such that $\int \varphi(x)\mathrm{d}x = 1$, we can set $\varphi_n(x) = n^d\varphi(nx)$ for every $x \in \mathbb{R}^d$ and every $n \in \mathbb{N}$. Note that, if we start from a function φ with compact support, all functions φ_n will vanish outside the same closed ball.

Proposition 5.8 *Let $(\varphi_n)_{n\in\mathbb{N}}$ be an approximation of the Dirac measure δ_0.*

(i) *For any bounded continuous function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, the sequence $\varphi_n * f$ converges to f as $n \to \infty$, uniformly on every compact subset of $\mathbb{R}^d$.*
(ii) *Let $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$, where $p \in [1, \infty)$. Then, $\varphi_n * f \longrightarrow f$ in L^p.*

Proof

(i) First note that $\varphi_n * f$ is well defined since φ_n is integrable and f is bounded. Then, we have, for every $\delta > 0$,

$$\varphi_n * f(x) - f(x) \tag{5.5}$$
$$= \int_{|y|\le\delta} (f(x-y) - f(x))\varphi_n(y)\mathrm{d}y + \int_{|y|>\delta} (f(x-y) - f(x))\varphi_n(y)\mathrm{d}y.$$

Fix $\varepsilon > 0$ and a compact subset H of $\mathbb{R}^d$. Since f is uniformly continous on any compact subset of $\mathbb{R}^d$, we can choose $\delta > 0$ such that $|f(x-y)-f(x)| \le \varepsilon$ for every $x \in H$ and $y \in \mathbb{R}^d$ such that $|y| \le \delta$. Recalling (5.3), we get that the first term in the right-hand side of (5.5) is smaller than ε in absolute value, for every $x \in H$ and $n \in \mathbb{N}$. Furthermore, (5.4) and the boundedness of f show that the second term in the right-hand side of (5.5) tends to 0 when $n \to \infty$, uniformly when x varies in $\mathbb{R}^d$. It follows that we have $|\varphi_n * f(x) - f(x)| < 2\varepsilon$ for every $x \in H$, as soon as n is sufficiently large..

(ii) The fact that $\varphi_n * f$ is well defined λ a.e. and belongs to L^p follows from Proposition 5.5 since φ_n is integrable. We then observe that

$$\begin{aligned}\int |\varphi_n * f(x) - f(x)|^p\mathrm{d}x &\le \int \Big(\int \varphi_n(y)|f(x-y) - f(x)|\mathrm{d}y\Big)^p \mathrm{d}x \\ &\le \int \Big(\int \varphi_n(y)|f(x-y) - f(x)|^p\mathrm{d}y\Big)\mathrm{d}x \\ &= \int \Big(\int \varphi_n(y)|f(x-y) - f(x)|^p\mathrm{d}x\Big)\mathrm{d}y \\ &= \int \varphi_n(y)\Big(\int |f(x-y) - f(x)|^p\mathrm{d}x\Big)\mathrm{d}y\end{aligned}$$

where the second inequality is a consequence of Jensen's inequality (note that $\varphi_n(y)\mathrm{d}y$ is a probability measure), and the next equality uses Theorem 5.3. For every $y \in \mathbb{R}^d$, set

$$h(y) = \int |f(x-y) - f(x)|^p \mathrm{d}x = \|f \circ \sigma_y - f\|_p^p,$$

with the notation of Lemma 5.7. The function h is bounded and $h(y) \longrightarrow 0$ as $y \to 0$ by Lemma 5.7. It then immediately follows from (5.3) and (5.4) that

$$\lim_{n\to\infty} \int \varphi_n(y)h(y)\,\mathrm{d}y = 0.$$

This completes the proof. □

Application In dimension $d = 1$, we can take

$$\varphi_n(x) = c_n(1-x^2)^n\,\mathbf{1}_{\{|x|\le 1\}}$$

where the constant c_n is chosen so that $\int \varphi_n(x)\mathrm{d}x = 1$. Then let $[a, b]$ be a compact subinterval of $(0, 1)$, and let f be a continuous function on $[a, b]$. It is easy to extend f to a continuous function on $\mathbb{R}$ that vanishes outside $[0, 1]$ (for instance, take f to be affine on both intervals $[0, a]$ and $[b, 1]$). Then the preceding proposition shows that

$$\varphi_n * f(x) = c_n \int (1-(x-y)^2)^n \mathbf{1}_{\{|x-y|\le 1\}} f(y)\mathrm{d}y \underset{n\to\infty}{\longrightarrow} f(x)$$

uniformly on $[a, b]$. If $x \in [a, b]$, we can remove the indicator function $\mathbf{1}_{\{|x-y|\le 1\}}$ in the last display, since the conditions $x \in [a, b]$ and $|x - y| > 1$ force $f(y) = 0$. It now follows that f is a uniform limit of polynomials on the interval $[a, b]$ (this is a special case of the Stone-Weierstrass theorem).

5.4.3 The Volume of the Unit Ball

In this section, it is convenient to write λ_d for Lebesgue measure on $\mathbb{R}^d$, in order to keep track of the dimension. We denote the closed unit ball of $\mathbb{R}^d$ by $\mathbb{B}_d$. Our goal is to compute the volume $\gamma_d = \lambda_d(\mathbb{B}_d)$ of $\mathbb{B}_d$. As we already observed in Chapter 3, for every $a > 0$, the pushforward of λ_d under the mapping $x \mapsto ax$ is $a^{-d}\lambda_d$. In particular, we have

$$\lambda_d(a\mathbb{B}_d) = a^d\lambda_d(\mathbb{B}_d).$$

Then, using Theorem 5.3 and assuming that $d \geq 2$, we have

$$\begin{aligned}
\gamma_d = \int_{\mathbb{R}^d} \mathbf{1}_{\mathbb{B}_d}(x)\mathrm{d}x &= \int_{\mathbb{R}^d} \mathbf{1}_{\{x_1^2+\cdots+x_d^2\leq 1\}}\mathrm{d}x_1\ldots\mathrm{d}x_d \\
&= \int_{-1}^{1}\Big(\int_{\mathbb{R}^{d-1}} \mathbf{1}_{\{x_1^2+\cdots+x_{d-1}^2\leq 1-x_d^2\}}\mathrm{d}x_1\ldots\mathrm{d}x_{d-1}\Big)\mathrm{d}x_d \\
&= \int_{-1}^{1} \lambda_{d-1}\Big(\sqrt{1-x_d^2}\,\mathbb{B}_{d-1}\Big)\,\mathrm{d}x_d \\
&= \gamma_{d-1}\int_{-1}^{1}(1-x_d^2)^{(d-1)/2}\,\mathrm{d}x_d \\
&= \gamma_{d-1} I_{d-1}
\end{aligned}$$

where we have set, for every integer $n \geq 0$,

$$I_n = \int_{-1}^{1}(1-x^2)^{n/2}\mathrm{d}x.$$

An integration by parts shows that, if $n \geq 2$,

$$I_n = \frac{n}{n+1} I_{n-2}.$$

Using the special cases $I_0 = 2$, $I_1 = \pi/2$, we get by induction that, for every integer $d \geq 2$,

$$I_{d-1}I_{d-2} = \frac{2\pi}{d}.$$

Consequently, for every $d \geq 3$,

$$\gamma_d = I_{d-1}I_{d-2}\gamma_{d-2} = \frac{2\pi}{d}\,\gamma_{d-2}.$$

From the special cases $\gamma_1 = 2$, $\gamma_2 = \gamma_1 I_1 = \pi$, we conclude that, for every integer $k \geq 1$,

$$\gamma_{2k} = \frac{\pi^k}{k!}\,, \qquad \gamma_{2k+1} = \frac{\pi^k}{(k+\frac{1}{2})(k-\frac{1}{2})\cdots\frac{3}{2}\cdot\frac{1}{2}}. \tag{5.6}$$

A reformulation of these formulas in terms of the Gamma function will be given in Section 7.2 below.

5.5 Exercises

Exercise 5.1 Let $(E, \mathcal{A})$ be a measurable space and let μ be a σ-finite measure on $(E, \mathcal{A})$. Let $f : E \longrightarrow \mathbb{R}_+$ be a nonnegative measurable function. Prove that

$$\int_E f \,\mathrm{d}\mu = \int_0^\infty \mu(\{x \in E : f(x) \geq t\}) \,\mathrm{d}t,$$

and more generally, if $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is a continuously differentiable increasing function such that $g(0) = 0$,

$$\int_E g \circ f \,\mathrm{d}\mu = \int_0^\infty \mu(\{x \in E : f(x) \geq t\})\, g'(t) \,\mathrm{d}t.$$

Exercise 5.2 Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a Borel measurable function. Prove that, for λ a.e. $y \in \mathbb{R}$, the set $\{x \in \mathbb{R} : f(x) = y\}$ has Lebesgue measure 0.

Exercise 5.3 Let f and g be two real functions defined on a compact interval I of $\mathbb{R}$. Assume that f and g are both increasing or both nondecreasing. Prove that, if μ is any probability measure on $(I, \mathcal{B}(I))$,

$$\int_I fg \,\mathrm{d}\mu \geq \int_I f \mathrm{d}\mu \times \int_I g \mathrm{d}\mu.$$

(*Hint:* Consider the function $F(x, y) = (f(x) - f(y))(g(x) - g(y))$.)

Exercise 5.4 For every $x, y \in [0, 1]$ such that $(x, y) \neq (0, 0)$, set

$$F(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2}.$$

Also set $F(0, 0) = 0$. Verify that the two quantities

$$\int_0^1 \Big(\int_0^1 F(x, y) \,\mathrm{d}y\Big)\mathrm{d}x \;, \quad \int_0^1 \Big(\int_0^1 F(x, y) \,\mathrm{d}x\Big)\mathrm{d}y$$

are well defined, and compute them explicitly. What can you infer about the value of

$$\int_{[0,1]^2} |F(x, y)| \,\mathrm{d}x\mathrm{d}y\ ?$$

Exercise 5.5 Compute in two different ways the integral

$$\int_{\mathbb{R}_+^2} \frac{1}{(1+y)(1+x^2y)}\,\mathrm{d}x\mathrm{d}y$$

and deduce that

$$\int_0^\infty \frac{\log x}{x^2-1}\,\mathrm{d}x = \frac{\pi^2}{4}.$$

Hint: Write $\dfrac{1}{(1+y)(1+x^2y)} = \dfrac{1}{(1-x^2)(1+y)} - \dfrac{x^2}{(1-x^2)(1+x^2y)}$.

Exercise 5.6 Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be two measurable spaces, and let μ and ν be two σ-finite measures on $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ respectively. Let φ be a nonnegative $\mathcal{X} \otimes \mathcal{Y}$-measurable function on $X \times Y$. For every $x \in E$, let $F(x) = \int_Y \varphi(x, y)\,\nu(\mathrm{d}y)$.

(1) Let $p \in [1, \infty)$. Verify the inequality

$$\|F\|_p \leq \int_Y \|\varphi(\cdot, y)\|_p\,\nu(\mathrm{d}y)$$

where $\varphi(\cdot, y)$ stands for the measurable function $x \mapsto \varphi(x, y)$ defined on X, and $\|\cdot\|_p$ refers to the measure μ on X. (*Hint:* Consider first the case where μ and ν are finite and φ is bounded, and write $F^p = F \times F^{p-1}$.)

(2) Suppose now that $p > 1$, and let $f \in L^p(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \lambda)$. For every $x > 0$, set

$$F(x) = \frac{1}{x}\int_0^x f(t)\,\mathrm{d}t.$$

Use the result of question (1) to give a new proof of Hardy's inequality of Exercise 4.7,

$$\|F\|_p \leq \frac{p}{p-1}\,\|f\|_p$$

(*Hint:* Consider $\varphi(x, y) = |f(xy)|$ for $y \in [0, 1]$.)

Exercise 5.7 Prove that the convolution operation acting on $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ has no neutral element (one cannot find a function $f \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ such that $f * g = g$ for every $g \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$).

Exercise 5.8 Let $p \in [1, \infty)$ and let $q \in (1, \infty]$ be the conjugate exponent of p. Let $f \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ and $g \in L^q(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$.

(1) Suppose that $p > 1$. Prove that

$$\lim_{|x|\to\infty} f * g(x) = 0.$$

Hint: Use the density of continuous functions with compact support in L^p, cf. Chapter 4.

(2) Consider the case $p = 1, q = \infty$. Show that the conclusion of question (1) remains true if we assume that $\lim_{|x|\to\infty} g(x) = 0$, but may fail without this assumption.

Exercise 5.9

(1) Let A be a Borel subset of $\mathbb{R}$ such that $0 < \lambda(A) < \infty$. Using a suitable convolution, show that the set $A - A = \{x - y : x \in A, y \in A\}$ contains a neighborhood of 0.

(2) Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a Borel measurable function such that $f(x + y) = f(x) + f(y)$ for every $x, y \in \mathbb{R}$. Using question (1), show that f is bounded on a neighborhood of 0. Infer that f is continuous at 0, and then that f must be a linear function.

Chapter 6
Signed Measures

In contrast with the preceding chapters, we now consider signed measures, which can take both positive and negative values. The main result of this chapter is the Jordan decomposition of a signed measure as the difference of two positive measures supported on disjoint sets. We also state a version of the Radon-Nikodym theorem for signed measures, and, as an application, we prove an important theorem of functional analysis stating that the space L^q is the topological dual of L^p when p and q are conjugate exponents and $p \in [1, \infty)$. We conclude this section by stating another version of the Riesz-Markov-Kakutani theorem, which (under appropriate assumptions) identifies the space of all finite measures on E with the topological dual of the space $C_0(E)$ of continuous functions that vanish at infinity.

6.1 Definition and Total Variation

Throughout this chapter, Section 6.4 excepted, $(E, \mathcal{A})$ is a measurable space. Let us start with a basic definition.

Definition 6.1 A signed measure μ on $(E, \mathcal{A})$ is a mapping $\mu : \mathcal{A} \longrightarrow \mathbb{R}$ such that $\mu(\varnothing) = 0$ and, for any sequence $(A_n)_{n\in\mathbb{N}}$ of disjoint $\mathcal{A}$-measurable sets, the series

$$\sum_{n\in\mathbb{N}} \mu(A_n)$$

is absolutely convergent and

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_6

To avoid confusion with the previous chapters, we will always include the adjective "signed" when we consider a signed measure (when we just say "measure" we mean a positive measure). Plainly, a signed measure is finitely additive since we can always take $A_n = \varnothing$ for $n \geq n_0$.

Remark A positive measure ν on $(E, \mathcal{A})$ is a signed measure only if it is finite ($\nu(E) < \infty$). So signed measures are not more general than positive measures.

Theorem 6.2 *Let μ be a signed measure on $(E, \mathcal{A})$. For every $A \in \mathcal{A}$, set*

$$|\mu|(A) = \sup \left\{ \sum_{n \in \mathbb{N}} |\mu(A_n)| : A = \bigcup_{n \in \mathbb{N}} A_n \,,\; A_n \in \mathcal{A},\; A_n \text{ disjoint} \right\}$$

where the supremum is over all representations of A as a countable disjoint union of $\mathcal{A}$-measurable sets $(A_n)_{n \in \mathbb{N}}$. Then $|\mu|$ is a finite (positive) measure on $(E, \mathcal{A})$, and $|\mu(A)| \leq |\mu|(A)$ for every $A \in \mathcal{A}$,

The measure $|\mu|$ is called the *total variation* of the signed measure μ.

Proof We first prove that $|\mu|$ is a measure. It is obvious that $|\mu|(\varnothing) = 0$. Then let $(B_i)_{i \in \mathbb{N}}$ be a sequence of disjoint $\mathcal{A}$-measurable sets and set $B = \bigcup_{i \in \mathbb{N}} B_i$. Fix $i \in \mathbb{N}$. By definition, if $t_i \in [0, |\mu|(B_i))$ (or $t_i = 0$ in case $|\mu|(B_i) = 0$), we can write $B_i = \bigcup_{n \in \mathbb{N}} A_{n,i}$ as a countable union of disjoint $\mathcal{A}$-measurable sets, in such a way that

$$\sum_{n \in \mathbb{N}} |\mu(A_{n,i})| \geq t_i \,.$$

Since the collection $(A_{n,i})_{n,i \in \mathbb{N}}$ is countable and B is the disjoint union of the sets in this collection, we have

$$|\mu|(B) \geq \sum_{i \in \mathbb{N}} \sum_{n \in \mathbb{N}} |\mu(A_{n,i})| \geq \sum_{i \in \mathbb{N}} t_i \,.$$

However each t_i can taken arbitrarily close to $|\mu|(B_i)$, and it follows that

$$|\mu|(B) \geq \sum_{i \in \mathbb{N}} |\mu|(B_i).$$

Note that the argument also works if $|\mu|(B_i) = \infty$ for some i, which we have not excluded until now.

To get the reverse inequality, let $(A_n)_{n\in\mathbb{N}}$ be a sequence of disjoint $\mathcal{A}$-measurable sets whose union is B. Then, for every $n \in \mathbb{N}$, A_n is the disjoint union of the sets $A_n \cap B_i$, $i \in \mathbb{N}$, and therefore the definition of a signed measure gives

$$\begin{aligned}\sum_{n\in\mathbb{N}} |\mu(A_n)| &= \sum_{n\in\mathbb{N}} \Big| \sum_{i\in\mathbb{N}} \mu(A_n \cap B_i) \Big| \\ &\le \sum_{n\in\mathbb{N}} \sum_{i\in\mathbb{N}} |\mu(A_n \cap B_i)| \\ &= \sum_{i\in\mathbb{N}} \sum_{n\in\mathbb{N}} |\mu(A_n \cap B_i)| \\ &\le \sum_{i\in\mathbb{N}} |\mu|(B_i),\end{aligned}$$

where the last inequality follows from the definition of $|\mu|(B_i)$ as a supremum (the sets $A_n \cap B_i$, $n \in \mathbb{N}$, are disjoint and their union is B_i). By taking the supremum over all choices of $(A_n)_{n\in\mathbb{N}}$, we get

$$|\mu|(B) \le \sum_{i\in\mathbb{N}} |\mu|(B_i)$$

which completes the proof of our claim that $|\mu|$ is a measure.

The inequality $|\mu(A)| \le |\mu|(A)$ is obvious from the definition of $|\mu|(A)$. So it only remains to verify that $|\mu|$ is a finite measure. We rely on the next lemma.

Lemma 6.3 *Suppose that there exists $A \in \mathcal{A}$ such that $|\mu|(A) = \infty$. Then one can find two disjoint $\mathcal{A}$-measurable sets B and C such that $A = B \cup C$ and $|\mu(B)| > 1$, $|\mu|(C) = \infty$.*

Proof of Lemma 6.3 Since $|\mu|(A) = \infty$, we can write $A = \bigcup_{n\in\mathbb{N}} A_n$ where the sets A_n are $\mathcal{A}$-measurable and disjoint, and

$$\sum_{n\in\mathbb{N}} |\mu(A_n)| > 2(1 + |\mu(A)|).$$

At least one of the two inequalities

$$\sum_{n\in\mathbb{N}} \mu(A_n)^+ > 1 + |\mu(A)|$$

and

$$\sum_{n\in\mathbb{N}} \mu(A_n)^- > 1 + |\mu(A)|$$

must then hold. Suppose that the first one holds (the other case is treated in a similar manner). We then set

$$B = \bigcup_{\{n:\mu(A_n)>0\}} A_n$$

in such a way that

$$\mu(B) = \sum_{n\in\mathbb{N}} \mu(A_n)^+ > 1 + |\mu(A)|.$$

Moreover, if $C = A\backslash B$,

$$|\mu(C)| = |\mu(A) - \mu(B)| \geq |\mu(B)| - |\mu(A)| > 1.$$

On the other hand, since $A = B \cup C$ and $|\mu|$ is a measure we have $|\mu|(B) = \infty$ or $|\mu|(C) = \infty$, which gives the desired result up to interchanging the roles of B and C if necessary. This completes the proof of the lemma. □

Let us now complete the proof of Theorem 6.2. We argue by contradiction and suppose that $|\mu|(E) = \infty$. From Lemma 6.3 applied to $A = E$, we can find disjoint measurable sets B_0 and C_0 such that $|\mu(B_0)| > 1$ and $|\mu|(C_0) = \infty$. But then, we can also apply Lemma 6.3 to $A = C_0$ and we find disjoint measurable sets B_1 and C_1 such that $C_0 = B_1 \cup C_1$, $|\mu(B_1)| > 1$ and $|\mu|(C_1) = \infty$. By induction, we construct a sequence of disjoint measurable sets $(B_n)_{n\in\mathbb{N}}$, such that $|\mu(B_n)| > 1$ for every n. This contradicts the fact that the series

$$\sum_{n\in\mathbb{N}} \mu(B_n)$$

has to be absolutely convergent, by the definition of a signed measure. This contradiction completes the proof. □

Example Let ν be a (positive) measure on $(E, \mathcal{A})$, and $g \in \mathcal{L}^1(E, \mathcal{A}, \nu)$.Then the formula

$$\mu(A) := \int_A g \, \mathrm{d}\nu\,,\ \ A \in \mathcal{A}\,,$$

defines a signed measure. In fact, if A is the disjoint union of a sequence $(A_n)_{n\in\mathbb{N}}$ of measurable sets, the bound $|\mu(A_n)| \leq \int_{A_n} |g|\, \mathrm{d}\nu$, together with Proposition 2.5 (3), shows that the series $\sum_{n\in\mathbb{N}} |\mu(A_n)|$ converges (its sum is bounded above by $\int_A |g| \mathrm{d}\nu$) and then the equality

$$\mu(A) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

follows by observing that

$$g\,\mathbf{1}_A = \lim_{k\to\infty} g \sum_{n\le k} \mathbf{1}_{A_n} \quad \text{in } L^1,$$

which holds by the dominated convergence theorem. We will see later that we have $|\mu| = |g| \cdot \nu$ in that case, with the notation of Theorem 4.11.

6.2 The Jordan Decomposition

It is obvious that the difference of two finite (positive) measures is a signed measure. Conversely, if μ is a signed measure on $(E, \mathcal{A})$, one immediately checks that the formulas

$$\mu^+(A) = \frac{1}{2}(\mu(A) + |\mu|(A)),$$
$$\mu^-(A) = \frac{1}{2}(|\mu|(A) - \mu(A)),$$

for every $A \in \mathcal{A}$, define two finite positive measures on $(E, \mathcal{A})$. Moreover, $\mu = \mu^+ - \mu^-$ and $|\mu| = \mu^+ + \mu^-$.

Recall the notation $g \cdot \nu$ in Theorem 4.11.

Theorem 6.4 (Jordan Decomposition) *Let μ be a signed measure on $(E, \mathcal{A})$. There exists a measurable subset B of E, which is unique up to a set of $|\mu|$-measure zero, such that $\mu^+ = \mathbf{1}_B \cdot |\mu|$ is the restriction of $|\mu|$ to B and $\mu^- = \mathbf{1}_{B^c} \cdot |\mu|$ is the restriction of $|\mu|$ to B^c. Moreover, we have, for every $A \in \mathcal{A}$,*

$$\mu^+(A) = \mu^+(A \cap B) = \mu(A \cap B)\ , \quad \mu^-(A) = \mu^-(A \cap B^c) = -\mu(A \cap B^c),$$

and consequently

$$\mu(A) = \mu^+(A \cap B) - \mu^-(A \cap B^c),$$
$$|\mu|(A) = \mu^+(A \cap B) + \mu^-(A \cap B^c).$$

Proof From the bound $|\mu(A)| \le |\mu|(A)$, we have $\mu^+(A) \le |\mu|(A)$ and $\mu^-(A) \le |\mu|(A)$ for every $A \in \mathcal{A}$. In particular μ^+ and μ^- are absolutely continuous with respect to $|\mu|$. By the Radon-Nikodym theorem, there exist two measurable functions h_1 and h_2 with values in $[0, \infty)$ such that $\mu^+ = h_1 \cdot |\mu|$ and $\mu^- = h_2 \cdot |\mu|$. Note that h_1 and h_2 are $|\mu|$-integrable.

We have then, for every $A \in \mathcal{A}$,

$$\mu(A) = \mu^+(A) - \mu^-(A) = \int_A h_1 \mathrm{d}|\mu| - \int_A h_2 \mathrm{d}|\mu| = \int_A h\, \mathrm{d}|\mu|,$$

where $h = h_1 - h_2$. Let μ' and μ'' be the (positive) measures defined by $\mu' = h^+ \cdot |\mu|$ and $\mu'' = h^- \cdot |\mu|$ and note that we have also $\mu(A) = \mu'(A) - \mu''(A)$ for every $A \in \mathcal{A}$, by the formula in the last display. Let $B = \{x \in E : h(x) > 0\}$, so that $\mu'(B^c) = 0$ and $\mu''(B) = 0$. For every $A \in \mathcal{A}$,

$$\mu'(A) = \int_A h^+ \mathrm{d}|\mu| = \int_{A\cap B} h^+ \mathrm{d}|\mu| = \int_{A\cap B} h \mathrm{d}\mu = \mu(A \cap B)$$

and similarly $\mu''(A) = -\mu(A \cap B^c)$.

We claim that we have also $|\mu|(A) = \mu'(A) + \mu''(A)$ for every $A \in \mathcal{A}$. The lower bound $|\mu|(A) \geq \mu'(A) + \mu''(A)$ is immediate since $\mu'(A) = \mu(A \cap B)$ and $\mu''(A) = -\mu(A \cap B^c)$ and we just use the definition of $|\mu|$. To get the corresponding upper bound, consider a disjoint sequence $(A_n)_{n\in\mathbb{N}}$ of measurable sets whose union is A. Then,

$$\begin{aligned}\sum_{n\in\mathbb{N}} |\mu(A_n)| &\leq \sum_{n\in\mathbb{N}} (|\mu(A_n \cap B)| + |\mu(A_n \cap B^c)|) \\ &= \sum_{n\in\mathbb{N}} (\mu'(A_n) + \mu''(A_n)) \\ &= \mu'(A) + \mu''(A).\end{aligned}$$

Since this holds for any choice of the sequence $(A_n)_{n\in\mathbb{N}}$, the definition of $|\mu|$ shows that $|\mu|(A) \leq \mu'(A) + \mu''(A)$, which completes the proof of our claim.

Next, since we have both $\mu(A) = \mu'(A) - \mu''(A)$ and $|\mu|(A) = \mu'(A) + \mu''(A)$ for every $A \in \mathcal{A}$, we immediately get that $\mu' = \mu^+$ and $\mu'' = \mu^-$. We have then, for every $A \in \mathcal{A}$,

$$|\mu|(A \cap B) = \mu^+(A \cap B) + \mu^-(A \cap B) = \mu^+(A \cap B) = \mu^+(A)$$

because $\mu^-(B) = \mu''(B) = 0$ and $\mu^+(B^c) = \mu'(B^c) = 0$. It follows that μ^+ is the restriction of $|\mu|$ to B, and we get similarly that μ^- is the restriction of $|\mu|$ to B^c. The formulas stated in the lemma also follow from our discussion. Finally, the uniqueness of B is a consequence of the uniqueness of g in Theorem 4.11. □

Remark If $\mu = \mu_1 - \mu_2$ is another decomposition of μ as the difference of two (finite) positive measures, we have necessarily $\mu_1 \geq \mu^+$ and $\mu_2 \geq \mu^-$. In fact, for every $A \in \mathcal{A}$,

$$\mu_1(A) \geq \mu_1(A \cap B) \geq \mu(A \cap B) = \mu^+(A \cap B) = \mu^+(A).$$

Integration with Respect to a Signed Measure Let μ be a signed measure and let $\mu = \mu^+ - \mu^-$ be its Jordan decomposition. If $f \in \mathcal{L}^1(E, \mathcal{A}, |\mu|)$, we set

$$\int f\,\mathrm{d}\mu := \int f \mathrm{d}\mu^+ - \int f \mathrm{d}\mu^- = \int f(\mathbf{1}_B - \mathbf{1}_{B^c})\mathrm{d}|\mu|.$$

It is then immediate that

$$\Big|\int f\,\mathrm{d}\mu\Big| \leq \int |f|\mathrm{d}|\mu|.$$

Proposition 6.5 *Let ν be a positive measure on $(E, \mathcal{A})$. Let $g \in \mathcal{L}^1(E, \mathcal{A}, \nu)$ and let $\mu = g \cdot \nu$ be the signed measure defined by*

$$\mu(A) := \int_A g\,\mathrm{d}\nu\,,\ \ A \in \mathcal{A}.$$

Then $|\mu| = |g| \cdot \nu$. Moreover, for every function $f \in \mathcal{L}^1(E, \mathcal{A}, |\mu|)$, we have $fg \in \mathcal{L}^1(E, \mathcal{A}, \nu)$), and

$$\int f\,\mathrm{d}\mu = \int fg\,\mathrm{d}\nu.$$

Proof The fact that μ is a signed measure was already explained in the example at the end of Section 6.1. Then, let B be as in Theorem 6.4. We have, for every $A \in \mathcal{A}$,

$$|\mu|(A) = \mu(A \cap B) - \mu(A \cap B^c) = \int_{A\cap B} g\,\mathrm{d}\nu - \int_{A\cap B^c} g\,\mathrm{d}\nu = \int_A gh\,\mathrm{d}\nu,$$

where we have set $h = \mathbf{1}_B - \mathbf{1}_{B^c}$. Taking $A = \{x \in E : g(x)h(x) < 0\}$, we infer from this equality that $gh \geq 0$, ν a.e. Hence $gh = |gh| = |g|$, ν a.e., so that

$$|\mu|(A) = \int_A |g|\mathrm{d}\nu,$$

and we have proved that $|\mu| = |g|\cdot\nu$. It readily follows that we have also $\mu^+ = g^+\cdot\nu$ and $\mu^- = g^- \cdot \nu$.

Then, by Proposition 2.6, if $f \in \mathcal{L}^1(E, \mathcal{A}, |\mu|)$,

$$\int |f|\,|g|\mathrm{d}\nu = \int |f|\,\mathrm{d}|\mu| < \infty$$

and thus $fg \in \mathcal{L}^1(\nu)$. The equality $\int f\mathrm{d}\mu = \int fg\,\mathrm{d}\nu$ then also follows from Proposition 2.6 by decomposing $f = f^+ - f^-$ and $\mu = \mu^+ - \mu^-$. □

The Radon-Nikodym Theorem for Signed Measures Let ν be a positive measure, and let μ be a signed measure. We say that μ is absolutely continuous with respect to ν, and we write $\mu \ll \nu$, if

$$\forall A \in \mathcal{A}, \quad \nu(A) = 0 \Rightarrow \mu(A) = 0.$$

Theorem 6.6 *Let μ be a signed measure, and let ν be a σ-finite positive measure. The following three properties are equivalent.*

(i) $\mu \ll \nu$.
(ii) *For every $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\forall A \in \mathcal{A}, \quad \nu(A) \leq \delta \Rightarrow |\mu|(A) \leq \varepsilon.$$

(iii) *There exists $g \in \mathcal{L}^1(E, \mathcal{A}, \nu)$ such that*

$$\forall A \in \mathcal{A}, \quad \mu(A) = \int_A g \, \mathrm{d}\nu.$$

Proof (ii)$\Rightarrow$(i) is immediate. Let us show that (i)$\Rightarrow$(iii). If $\mu \ll \nu$, we have also $\mu^+ \ll \nu$, since with the notation of Theorem 6.4,

$$\nu(A) = 0 \Rightarrow \nu(A \cap B) = 0 \Rightarrow \mu^+(A) = \mu(A \cap B) = 0.$$

Similarly, $\mu^- \ll \nu$. Thus the Radon-Nikodym theorem for positive measures (Theorem 4.11) allows us to write $\mu^+ = g_1 \cdot \nu$ and $\mu^- = g_2 \cdot \nu$ where g_1 and g_2 are nonnegative measurable functions such that $\int g_1 \mathrm{d}\nu = \mu^+(E) < \infty$ and $\int g_2 \mathrm{d}\nu = \mu^-(E) < \infty$. (iii) follows, with $g = g_1 - g_2$.

It remains to show that (iii)$\Rightarrow$(ii). This is very similar to the proof of Proposition 4.12. Assume that (iii) holds. By Proposition 6.5, we have $|\mu| = |g| \cdot \nu$. Moreover, the dominated convergence theorem gives

$$\lim_{n\to\infty} \int_{\{|g|\geq n\}} |g| \, \mathrm{d}\nu = 0.$$

Hence, if $\varepsilon > 0$ is fixed, we can choose N large enough so that

$$\int_{\{|g|\geq N\}} |g| \, \mathrm{d}\nu < \frac{\varepsilon}{2}.$$

Then, taking $\delta = \varepsilon/(2N)$, we have, for every $A \in \mathcal{A}$ such that $\nu(A) \leq \delta$,

$$|\mu|(A) = \int_A |g| \mathrm{d}\nu \leq \int_{\{|g|\geq N\}} |g| \, \mathrm{d}\nu + \int_{A\cap\{|g|<N\}} |g| \, \mathrm{d}\nu \leq \frac{\varepsilon}{2} + N \frac{\varepsilon}{2N} = \varepsilon.$$

This completes the proof. □

6.3 The Duality Between L^p and L^q

Let ν be a σ-finite (positive) measure on $(E, \mathcal{A})$. Let $p \in [1, \infty]$ and let q be the conjugate exponent of p. Then, if we fix $g \in L^q(E, \mathcal{A}, \nu)$, the formula

$$\Phi_g(f) := \int fg \, \mathrm{d}\nu$$

defines a continuous linear form on $L^p(E, \mathcal{A}, \nu)$. Indeed, the Hölder inequality shows on one hand that $\Phi_g(f)$ is well defined (fg is integrable with respect to ν), and on the other hand that

$$|\Phi_g(f)| \leq C_g \, \|f\|_p \, , \; \forall f \in L^p(E, \mathcal{A}, \nu),$$

where $C_g = \|g\|_q$. We also get that the operator norm of Φ_g, which is defined by

$$\|\Phi_g\| := \sup\{|\Phi_g(f)| : f \in L^p(E, \mathcal{A}, \nu), \; \|f\|_p \leq 1\}$$

satisfies the bound $\|\Phi_g\| \leq \|g\|_q$.

One may then ask whether all continuous linear forms on $L^p(E, \mathcal{A}, \nu)$ are of the preceding type. In the special case $p = q = 2$, we know from the theory of Hilbert spaces that the answer is yes. The next theorem shows more generally that the answer is yes when $p < \infty$.

Theorem 6.7 *Let ν be a σ-finite measure on $(E, \mathcal{A})$, let $p \in [1, \infty)$ and let q be the conjugate exponent of p. Then, if Φ is a continuous linear form on $L^p(E, \mathcal{A}, \nu)$, there exists a unique function $g \in L^q(E, \mathcal{A}, \nu)$ such that, for every $f \in L^p(E, \mathcal{A}, \nu)$,*

$$\Phi(f) = \int fg \, \mathrm{d}\nu.$$

Moreover, the operator norm of Φ is

$$\|\Phi\| = \|g\|_q.$$

With the notation introduced before the theorem, we thus see that the mapping $g \mapsto \Phi_g$ identifies $L^q(\nu)$ with the topological dual of $L^p(\nu)$ (that is, the linear space of all continuous linear forms on $L^p(\nu)$ equipped with the operator norm, see the Appendix below). When $p = \infty$, the mapping $g \mapsto \Phi_g$ still makes sense from $L^1(\nu)$ into the dual of $L^\infty(\nu)$, but, as we will explain later on an example, there may exist continuous linear forms on $L^\infty(\nu)$ that cannot be written as Φ_g for some $g \in L^1(\nu)$.

Proof Suppose first that $\nu(E) < \infty$. Then, for every $A \in \mathcal{A}$, set

$$\mu(A) = \Phi(\mathbf{1}_A),$$

which makes sense since indicator functions belong to $L^p(\nu)$ when ν is finite. The first step of the proof is to verify that $A \mapsto \mu(A)$ is a signed measure on $(E, \mathcal{A})$. It is trivial that $\mu(\varnothing) = 0$. Then let $(A_n)_{n\in\mathbb{N}}$ be a sequence of disjoint measurable sets, and let A be the union of the sets A_n, $n \in \mathbb{N}$. Then,

$$\mathbf{1}_A = \lim_{k\to\infty} \sum_{n=1}^{k} \mathbf{1}_{A_n}$$

and convergence holds in $L^p(\nu)$ by dominated convergence. Using the fact that Φ is continuous, we get

$$\mu(A) = \lim_{k\to\infty} \Phi\Big(\sum_{n=1}^{k} \mathbf{1}_{A_n}\Big) = \lim_{k\to\infty} \sum_{n=1}^{k} \mu(A_n). \tag{6.1}$$

We can use this to verify that the series $\sum_n \mu(A_n)$ is absolutely convergent. In fact, replacing the sequence $(A_n)_{n\in\mathbb{N}}$ by the sequence $(A'_n)_{n\in\mathbb{N}}$ defined by setting $A'_n = A_n$ if $\mu(A_n) > 0$ and $A'_n = \varnothing$ otherwise, we have

$$\sum_{n\in\mathbb{N}} \mu(A'_n) = \lim_{k\to\infty} \uparrow \sum_{n=1}^{k} \mu(A'_n) = \mu\Big(\bigcup_{n\in\mathbb{N}} A'_n\Big),$$

where the second equality holds by (6.1). It follows that $\sum_{n\in\mathbb{N}} \mu(A_n)^+ = \sum_{n\in\mathbb{N}} \mu(A'_n) < \infty$, and one similarly obtains that $\sum_{n\in\mathbb{N}} \mu(A_n)^- < \infty$. Once we know that the series $\sum_{n\in\mathbb{N}} \mu(A_n)$ is absolutely convergent, the equality $\mu(A) = \sum_{n\in\mathbb{N}} \mu(A_n)$ follows from (6.1), and we have proved that μ is a signed measure.

If $A \in \mathcal{A}$ and $\nu(A) = 0$, we have $\mathbf{1}_A = 0$ in $L^p(E, \mathcal{A}, \nu)$ and thus $\mu(A) = \Phi(\mathbf{1}_A) = 0$. Therefore $\mu \ll \nu$ and Theorem 6.6 shows that there exists a function $g \in L^1(E, \mathcal{A}, \nu)$ such that

$$\forall A \in \mathcal{A}\,, \qquad \Phi(\mathbf{1}_A) = \mu(A) = \int_A g \, \mathrm{d}\nu.$$

The equality

$$\Phi(f) = \int fg \, \mathrm{d}\nu$$

then holds when f is a simple function. We claim that it also holds if f is only measurable and bounded (which implies $f \in L^p(\nu)$ since ν is finite). In that case, Proposition 2.5 (1) applied to both f^+ and f^- allows us to find a sequence $(f_n)_{n\in\mathbb{N}}$ of simple functions such that $|f_n| \le |f|$ and $f_n \longrightarrow f$ pointwise as $n \to \infty$. Using dominated convergence, it follows that $f_n \longrightarrow f$ in $L^p(\nu)$, so that $\Phi(f_n) \longrightarrow \Phi(f)$. Since we have also $\int f_n g\, \mathrm{d}\nu \longrightarrow \int fg\, \mathrm{d}\nu$, again by dominated convergence, our claim follows by passing to the limit $n \to \infty$ in the equality $\Phi(f_n) = \int f_n g\, \mathrm{d}\nu$.

Let us now show that $g \in L^q(\nu)$.

- If $p = 1$, then, for every $A \in \mathcal{A}$,

$$\Big|\int_A g\, \mathrm{d}\nu\Big| = |\Phi(\mathbf{1}_A)| \le \|\Phi\|\, \|\mathbf{1}_A\|_1 = \|\Phi\|\, \nu(A)$$

which easily implies that $|g| \le \|\Phi\|$, ν a.e. (apply the last display to $A = \{g > \|\Phi\| + \varepsilon\}$ or $A = \{g < -\|\Phi\| - \varepsilon\}$), and thus $\|g\|_\infty \le \|\Phi\|$.

- If $p \in (1, \infty)$, we set $F_n = \{x \in E : |g(x)| \le n\}$ for every $n \in \mathbb{N}$, and then $h_n = \mathbf{1}_{F_n} |g|^{q-1}\, (\mathbf{1}_{\{g>0\}} - \mathbf{1}_{\{g<0\}})$. Since h_n is bounded, we have

$$\int_{F_n} |g|^q \mathrm{d}\nu = \int h_n g\, \mathrm{d}\nu = \Phi(h_n) \le \|\Phi\|\, \|h_n\|_p = \|\Phi\| \Big(\int_{F_n} |g|^q \mathrm{d}\nu\Big)^{1/p},$$

hence

$$\Big(\int_{F_n} |g|^q \mathrm{d}\nu\Big)^{1/q} \le \|\Phi\|.$$

Letting n tend to ∞, we get by monotone convergence that $\|g\|_q \le \|\Phi\|$.

So in both cases we have obtained that $g \in L^q(\nu)$ and $\|g\|_q \le \|\Phi\|$. Using the notation $\Phi_g(f) = \int fg\, \mathrm{d}\mu$ introduced before the theorem, we get that Φ and Φ_g are continuous linear forms on $L^p(\nu)$, which coincide on the dense subset of bounded measurable functions (cf. Theorem 4.8). It follows that $\Phi = \Phi_g$. As explained before the statement of the theorem, we have $\|\Phi\| = \|\Phi_g\| \le \|g\|_q$, by the Hölder inequality. The reverse inequality $\|\Phi\| \ge \|g\|_q$ has been obtained above, so that we get $\|\Phi\| = \|g\|_q$. Finally, the mapping $g \mapsto \Phi_g$ is linear and is an isometry from $L^q(\nu)$ onto the topological dual of $L^p(\nu)$. This mapping is necessarily one-to-one, which gives the uniqueness of g in the theorem. This completes the proof in the case where ν is finite.

It remains to consider the case when $\nu(E) = \infty$. In that case, write E as the union of a sequence $(E_n)_{n\in\mathbb{N}}$ of disjoint measurable subsets such that $\nu(E_n) < \infty$ for every $n \in \mathbb{N}$. Let ν_n be the restriction of ν to E_n. Then the mapping $f \mapsto f\mathbf{1}_{E_n}$ yields an isometry from $L^p(\nu_n)$ onto a subspace of $L^p(\nu)$. Replacing ν by ν_n, we can apply the first part of the proof to the continuous linear form Φ_n defined on $L^p(\nu_n)$ by

$$\Phi_n(f) := \Phi(f\mathbf{1}_{E_n}).$$

It follows that there exists a function $g_n \in L^q(\nu_n)$ such that, for every $f \in L^p(\nu_n)$,

$$\Phi(f\mathbf{1}_{E_n}) = \Phi_n(f) = \int f g_n \,\mathrm{d}\nu_n.$$

Up to replacing g_n by $g_n\mathbf{1}_{E_n}$, we may assume that $g_n = 0$ on E_n^c (more formally, we just say that the equivalence class of g_n in $L^p(\nu_n)$ contains a function that vanishes outside E_n), and then rewrite the preceding display as

$$\Phi(f\mathbf{1}_{E_n}) = \int f g_n \,\mathrm{d}\nu,$$

for every $f \in L^p(\nu)$.

If $f \in L^p(\nu)$, the dominated convergence theorem gives

$$f = \lim_{k\to\infty} \sum_{n=1}^{k} f\,\mathbf{1}_{E_n}, \quad \text{in } L^p(\nu),$$

and this implies

$$\Phi(f) = \lim_{k\to\infty} \int f\Big(\sum_{n=1}^{k} g_n\Big)\,\mathrm{d}\nu. \tag{6.2}$$

On the other hand, we have

$$\int f\Big(\sum_{n=1}^{k} g_n\Big)\mathrm{d}\nu = \Phi\Big(f\sum_{n=1}^{k} \mathbf{1}_{E_n}\Big) \leq \|\Phi\|\,\|f\|_p. \tag{6.3}$$

We can then apply the same arguments that we used in the case $\nu(E) < \infty$ (when we were establishing the bound $\|g\|_q \leq \|\Phi\|$). In particular, when $p \in (1,\infty)$, we apply (6.3) to $f = |\sum_{n=1}^{k} g_n|^{q-1}\,\mathrm{sgn}(\sum_{n=1}^{k} g_n)$, where $\mathrm{sgn}(y) = \mathbf{1}_{\{y>0\}} - \mathbf{1}_{\{y<0\}}$. In this way, we derive from (6.3) that, for every $k \geq 1$,

$$\|\sum_{n=1}^{k} g_n\|_q \leq \|\Phi\|. \tag{6.4}$$

Let us now set, for every $x \in E$,

$$g(x) = \sum_{n\in\mathbb{N}} g_n(x)$$

(note that the sum is well defined since, for every $x \in E$, there is at most one index n for which $g_n(x) \neq 0$). If $q = \infty$, (6.4) gives $\|g\|_\infty \leq \|\Phi\|$. If $q < \infty$, (6.4) and monotone convergence imply

$$\|g\|_q = \lim_{k\to\infty} \|\sum_{n\le k} g_n\|_q \le \|\Phi\|.$$

In both cases, we have $g \in L^q(\nu)$. Finally, recalling (6.2), we have if $f \in L^p(\nu)$,

$$\Phi(f) = \lim_{k\to\infty} \int f\Big(\sum_{n\le k} g_n\Big)\,\mathrm{d}\nu = \int fg\,\mathrm{d}\nu,$$

where the use of the dominated convergence theorem in the second equality is justified by the bound $|\sum_{n=1}^k g_n| \le |g|$.

The fact that $\|\Phi\| = \|g\|_q$ and the uniqueness of g are then established by the same arguments as in the case $\nu(E) < \infty$. This completes the proof of Theorem 6.7. □

Remark When $p = \infty$, the result of the theorem fails in general: there may exist continuous linear forms on $L^\infty(E, \mathcal{A}, \nu)$ that cannot be represented as $\Phi(f) = \int fg\,\mathrm{d}\nu$ with a function $g \in L^1(E, \mathcal{A}, \nu)$. Let us consider the case where $E = \mathbb{N}$ and ν is the counting measure. Then $L^\infty(E, \mathcal{A}, \nu) = \ell^\infty$ is the space of all bounded real sequences $a = (a_k)_{k\in\mathbb{N}}$ equipped with the norm $\|a\|_\infty = \sup\{|a_k| : k \in \mathbb{N}\}$. Let H be the closed linear subspace of ℓ^∞ defined by

$$H := \{a \in \ell^\infty : \lim_{k\to\infty} a_k \text{ exists in } \mathbb{R}\},$$

and for every $a \in H$, set

$$\Phi(a) = \lim_{k\to\infty} a_k.$$

Plainly, $|\Phi(a)| \le \|a\|_\infty$. The Hahn-Banach theorem (Theorem A.1 in the Appendix) allows us to extend Φ to a continuous linear form on ℓ^∞, whose operator norm is bounded by 1. It is easy to see that Φ cannot be represented in the form

$$\Phi(a) = \sum_{k\in\mathbb{N}} a_k b_k$$

with an element $b = (b_k)_{k\in\mathbb{N}}$ of ℓ^1. If such a representation did exist, then by considering, for every $n \in \mathbb{N}$, the element $a^{(n)}$ of ℓ^∞ defined by $a_k^{(n)} = \mathbf{1}_{\{k=n\}}$, we would get

$$b_n = \Phi(a^{(n)}) = 0,$$

for every $n \in \mathbb{N}$ and thus $\Phi = 0$ which is absurd.

6.4 The Riesz-Markov-Kakutani Representation Theorem for Signed Measures

In this section, we provide a version of the Riesz-Markov-Kakutani representation theorem (Theorem 3.13) for signed measures. Throughout this section, E is a separable locally compact metric space. We let $C_0(E)$ stand for the space of all real continuous functions on E that tend to 0 at infinity: a real continuous function f on E belongs to $C_0(E)$ if and only if, for every $\varepsilon > 0$, we can find a compact subset K of E such that $|f(x)| < \varepsilon$ for every $x \in E\backslash K$. The space $C_0(E)$ is a Banach space for the norm

$$\|f\| = \sup_{x\in E} |f(x)|.$$

If μ is a signed measure on $(E, \mathcal{B}(E))$, the formula

$$\Phi(f) := \int_E f \mathrm{d}\mu \ , \ \ f \in C_0(E),$$

defines a continuous linear form on $C_0(E)$. Moreover, this linear form is continuous since

$$|\Phi(f)| \leq \int_E |f| \, \mathrm{d}|\mu| \leq |\mu|(E) \, \|f\|,$$

which also shows that $\|\Phi\| \leq |\mu|(E)$.

Theorem 6.8 *Let Φ be a continuous linear form on $C_0(E)$. Then there exists a unique signed measure μ on $(E, \mathcal{B}(E))$ such that*

$$\forall f \in C_0(E) \ , \quad \Phi(f) = \int_E f \, \mathrm{d}\mu.$$

Moreover, $\|\Phi\| = |\mu|(E)$.

We refer the reader to Chapter 6 of Rudin [22] for a proof in the more general complex setting.

Remark The space $\mathcal{M}(E)$ of all signed measures on E is a linear space, and $\mu \mapsto |\mu|(E)$ defines a norm on this linear space. Moreover, it is easy to verify that $\mathcal{M}(E)$ is complete for this norm (cf. Exercise 6.2 below). The preceding theorem can be reformulated by saying that $\mathcal{M}(E)$ is the topological dual of $C_0(E)$.

When E is compact, $C_0(E)$ coincides with the space $C_b(E)$ of all bounded continuous functions on E, and therefore $\mathcal{M}(E)$ is the dual of $C_b(E)$. This last assertion becomes false when E is not compact, in particular when $E = \mathbb{R}$. An example can be given along the lines of the end of Section 6.3.

6.5 Exercises

Exercise 6.1 Let $a < b$ and let f be a real function defined on $[a, b]$. Prove that the following three assertions are equivalent:

(i) f is the difference of two increasing right-continuous functions.
(ii) There exists a signed measure μ on $[a, b]$ such that $f(x) = \mu([a, x])$ for every $x \in [a, b]$.
(iii) f is right-continuous and of bounded variation in the sense that

$$\sup_{a=a_0<a_1<\cdots<a_{n-1}<a_n=b} \sum_{i=1}^{n} |f(a_i) - f(a_{i-1})| < \infty,$$

where the supremum is over all choices of the integer $n \geq 1$ and the reals $a_0, \ldots, a_n$ such that $a = a_0 < a_1 < \cdots < a_{n-1} < a_n = b$.

Exercise 6.2 Let $\mathcal{M}(\mathbb{R})$ be the linear space of all signed measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and for every $\mu \in \mathcal{M}(\mathbb{R})$, set $\|\mu\| = |\mu|(\mathbb{R})$. Verify without using Theorem 6.8 that the mapping $\mu \mapsto \|\mu\|$ is a norm on $\mathcal{M}(\mathbb{R})$, and that $\mathcal{M}(\mathbb{R})$ equipped with this norm is a Banach space.

Exercise 6.3 We keep the notation $(\mathcal{M}(\mathbb{R}), \|\cdot\|)$ of the previous exercise.

(1) Let $\mu, \nu \in \mathcal{M}(\mathbb{R})$. Verify the existence of a unique signed measure $\mu \otimes \nu$ on $\mathbb{R}^2$ such that, for every $A, B \in \mathcal{B}(\mathbb{R})$,

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

(2) If $\mu, \nu \in \mathcal{M}(\mathbb{R})$), we set, for every $A \in \mathcal{B}(\mathbb{R})$,

$$\mu * \nu(A) = \mu \otimes \nu(\{(x, y) \in \mathbb{R}^2 : x + y \in A\}).$$

Verify that $\mu * \nu = \nu * \mu$ is a signed measure on $\mathbb{R}$, and $\|\mu * \nu\| \leq \|\mu\| \, \|\nu\|$. (*Hint:* Use the Jordan decomposition of μ and ν.)

(3) Verify that, for every bounded measurable function $f : \mathbb{R} \longrightarrow \mathbb{R}$,

$$\int f(z)\, \mu*\nu(\mathrm{d}z) = \int \Big(\int f(x+y)\, \mu(\mathrm{d}x) \Big) \nu(\mathrm{d}y) = \int \Big(\int f(x+y)\, \nu(\mathrm{d}y) \Big) \mu(\mathrm{d}x).$$

(4) Suppose that there exist two functions f and g in $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ such that $\mu = f \cdot \lambda$ and $\nu = g \cdot \lambda$, with the notation of Proposition 6.5. Show that $\mu * \nu = (f * g) \cdot \lambda$.

(5) Verify that the identities $\mu * (\nu * \gamma) = (\mu * \nu) * \gamma$, $\mu * (\nu + \gamma) = \mu * \nu + \mu * \gamma$ and $a(\mu * \nu) = (a\mu) * \nu$ hold for every $\mu, \nu, \gamma \in \mathcal{M}(\mathbb{R})$ and $a \in \mathbb{R}$. As we already know that $(\mathcal{M}(\mathbb{R}), \|\cdot\|)$ is a Banach space (Exercise 6.2) and that the

inequality $\|\mu * \nu\| \le \|\mu\|\,\|\nu\|$ holds, this means that $(\mathcal{M}(\mathbb{R}), \|\cdot\|)$ equipped with the operations $+$ and $*$ is a *Banach algebra*. Verify that the Dirac measure δ_0 is the unit element of this Banach algebra (compare with Exercise 5.7).

Exercise 6.4 Let $(E, \mathcal{A}, \mu)$ be a measure space. Assume that there exists a sequence $(A_n)_{n\in\mathbb{N}}$ of disjoint measurable sets such that $0 < \mu(A_n) < \infty$ for every $n \in \mathbb{N}$, and $E = \cup_{n\in\mathbb{N}} A_n$. Let F be the linear subspace of $L^\infty(E, \mathcal{A}, \mu)$ spanned by the functions 1 and $\mathbf{1}_{A_n}$ for every $n \in \mathbb{N}$. Show that, for every $f \in F$, the limit

$$\lim_{n\to\infty} \frac{1}{\mu(A_n)} \int \mathbf{1}_{A_n} f \, \mathrm{d}\mu$$

exists in $\mathbb{R}$. Using the Hahn-Banach theorem (Theorem A.1), construct a continuous linear form Φ on $L^\infty(E, \mathcal{A}, \mu)$ which cannot be represented as $\Phi(f) = \int fg \, \mathrm{d}\mu$ for some $g \in L^1(E, \mathcal{A}, \mu)$.

Exercise 6.5 Let $\mathcal{A}$ be the σ-field on [0, 1] that consists of all subsets of [0, 1] that are (at most) countable or whose complement is (at most) countable. Let μ be the counting measure on $([0, 1], \mathcal{A})$.

(1) Verify that a function $f : [0, 1] \longrightarrow \mathbb{R}$ is in $L^1([0, 1], \mathcal{A}, \mu)$ if and only if the set $E_f = \{x \in [0, 1] : f(x) \neq 0\}$ is at most countable, and

$$\sum_{x\in E_f} |f(x)| < \infty.$$

Furthermore, we have $\|f\|_1 = \sum_{x\in E_f} |f(x)| < \infty$ in that case.

(2) For every $f \in L^1([0, 1], \mathcal{A}, \mu)$, set

$$\Phi(f) = \sum_{x\in E_f} x\, f(x).$$

Prove that Φ is a continuous linear form on $L^1([0, 1], \mathcal{A}, \mu)$.

(3) Show that there exists no function $g \in L^\infty([0, 1], \mathcal{A}, \mu)$ such that $\Phi(f) = \int fg \, \mathrm{d}\mu$ for every $f \in L^1([0, 1], \mathcal{A}, \mu)$. This is not a contradiction with Theorem 6.7 because μ is not σ-finite.

Chapter 7
Change of Variables

This short chapter is devoted to the change of variables formula, which identifies the pushforward of Lebesgue measure on an open set of $\mathbb{R}^d$ under a diffeomorphism. Along with the Fubini theorem, the change of variables formula is an essential tool for the calculation of integrals, some of which play a crucial role in probability theory. An important application is the formula of integration in polar coordinates in the plane, and its generalization in higher dimensions involving Lebesgue measure on the unit sphere of $\mathbb{R}^d$. The latter measure will be instrumental in Chapter 14 when we study harmonic functions and their relations with Brownian motion.

7.1 The Change of Variables Formula

In view of proving the change of variables formula, we start with the important special case of an affine transformation. We use the notation λ_d for Lebesgue measure on $\mathbb{R}^d$.

Proposition 7.1 *Let $b \in \mathbb{R}^d$ and let M be a $d \times d$ invertible matrix with real coefficients. Define a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ by $f(x) = Mx + b$. Then, for every Borel subset A of $\mathbb{R}^d$,*

$$\lambda_d(f(A)) = |\det(M)|\, \lambda_d(A),$$

where $\det(M)$ *is the determinant of* M.

Remark If M is not invertible, $f(A) \subset f(\mathbb{R}^d)$ is contained in a hyperplane, which has zero Lebesgue measure.

Proof First notice that $f(A) = (f^{-1})^{-1}(A) \in \mathcal{B}(\mathbb{R}^d)$ if $A \in \mathcal{B}(\mathbb{R}^d)$ since f^{-1} is continuous hence measurable. Then, thanks to the translation invariance of

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_7

Lebesgue measure, it is enough to consider the case $b = 0$. In that case, we have, for every $a > 0$ and every Borel subset A of $\mathbb{R}^d$,

$$\lambda_d(f(a + A)) = \lambda_d(f(a) + f(A)) = \lambda_d(f(A)),$$

which shows that the measure $A \mapsto \lambda_d(f(A))$ (which is the pushforward of λ_d under f^{-1}) is invariant under translations. By Proposition 3.8, there exists a constant $c \geq 0$ such that, for every $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\lambda_d(f(A)) = c\,\lambda_d(A).$$

It remains to prove that $c = |\det(M)|$.

If M is an orthogonal matrix, and $\mathbb{B}_d$ denotes the closed unit ball of $\mathbb{R}^d$, we have $f(\mathbb{B}_d) = \mathbb{B}_d$, and it immediately follows that $c = 1 = |\det(M)|$.

Recall the notation ${}^tN$ for the transpose of a matrix N. Then, if M is a positive definite symmetric matrix, we can find an orthogonal matrix P such that ${}^tPMP$ is diagonal with diagonal coefficients $\alpha_i > 0$, $i \in \{1, \dots, d\}$. Then, using the fact that ${}^tP = P^{-1}$, we have

$$f(P([0,1]^d)) = \{MPx : x \in [0,1]^d\} = \{Py : y \in \prod_{i=1}^{d} [0, \alpha_i]\},$$

and thus, using the orthogonal case,

$$c = c\,\lambda_d(P([0,1]^d)) = \lambda_d(f(P([0,1]^d))) = \lambda_d\Big(\prod_{i=1}^{d} [0, \alpha_i]\Big) = \prod_{i=1}^{d} \alpha_i .$$

So we have again $c = |\det(M)|$.

Finally, in the general case, we observe that we can write $M = PS$, where P is orthogonal and S is positive definite symmetric (take S to be equal to the square root of the symmetric matrix ${}^tMM$ and $P = MS^{-1}$). Using the two special cases treated above, we immediately get

$$c = |\det(P)|\,|\det(S)| = |\det(M)|.$$

□

Let U and D be two open subsets of $\mathbb{R}^d$. We say that a function $\varphi : U \longrightarrow D$ is a C^1-diffeomorphism if φ is bijective and both φ and φ^{-1} are continuously differentiable. Then, the differential $\varphi'(u)$ is invertible, for every $u \in U$.

Theorem 7.2 *Let $\varphi : U \longrightarrow D$ be a C^1-diffeomorphism. Then, for every Borel function $f : D \longrightarrow \mathbb{R}_+$, we have*

$$\int_D f(x)\,\mathrm{d}x = \int_U f(\varphi(u))\,|J_\varphi(u)|\,\mathrm{d}u\,,$$

where $J_\varphi(u) = \det(\varphi'(u))$ is the Jacobian of φ at u.

Remark We state the theorem only for nonnegative functions, but of course the same formula holds for functions of arbitrary sign under appropriate integrability conditions (just decompose $f = f^+ - f^-$).

Proof Writing a nonnegative measurable function as an increasing limit of nonnegative simple functions, we see that it is enough to treat the case $f = \mathbf{1}_A$, where A is a Borel subset of U. In that case, the formula reduces to

$$\lambda_d(A) = \int_{\varphi^{-1}(A)} |J_\varphi(u)|\,\mathrm{d}u.$$

Replacing A by $\varphi^{-1}(A)$, we have to prove that, for every Borel subset A of U,

$$\lambda_d(\varphi(A)) = \int_A |J_\varphi(u)|\,\mathrm{d}u. \tag{7.1}$$

(note that $\varphi(A) = (\varphi^{-1})^{-1}(A)$ is a Borel set). In the next lemma, $\mathrm{dist}(F, F') = \inf\{|y - y'| : y \in F, y' \in F\}$ denotes the Euclidean distance between two closed subsets F and F' of $\mathbb{R}^d$. A cube with faces parallel to the coordinate axes is a subset of $\mathbb{R}^d$ of the form $C = I_1 \times I_2 \times \cdots \times I_d$, where $I_1, \ldots, I_d$ are intervals of the same length r, which is called the sidelength of C. The center of C is defined in the obvious manner.

Lemma 7.3 *Let K be a compact subset of U and let $\varepsilon > 0$. Then, we can choose $\delta > 0$ sufficiently small so that $d\delta < \mathrm{dist}(K, U^c)$ and, for every cube C with faces parallel to the coordinate axes, with center $u_0 \in K$ and sidelength smaller than δ, we have*

$$(1-\varepsilon)|J_\varphi(u_0)|\,\lambda_d(C) \le \lambda_d(\varphi(C)) \le (1+\varepsilon)|J_\varphi(u_0)|\,\lambda_d(C).$$

Proof of Lemma 7.3 Let $a = \sup\{\|\varphi'(v)^{-1}\| : v \in K\} < \infty$, where $\|\varphi'(v)^{-1}\|$ is the operator norm of the matrix $\varphi'(v)^{-1}$, and let $\eta > 0$ so small that $(1 + da\eta)^d < 1 + \varepsilon$. Since φ' is continuous, the mean value inequality allows us to choose $\delta > 0$ small enough so that $\delta < \frac{1}{d}\,\mathrm{dist}(K, U^c)$ and, for every $u_0 \in K$, for every $u \in \mathbb{R}^d$ such that $0 < |u - u_0| < d\delta$,

$$|\varphi(u) - \varphi(u_0) - \varphi'(u_0)(u - u_0)| < \eta|u - u_0|.$$

Fix $u_0 \in K$ and set $f(v) = \varphi(u_0) + \varphi'(u_0)v$ for every $v \in \mathbb{R}^d$. We get that, if $0 < |u - u_0| < d\delta$,

$$\varphi(u) = f(u - u_0) + h(u, u_0),$$

with $|h(u, u_0)| < \eta|u - u_0|$. Setting $g(u, u_0) = \varphi'(u_0)^{-1}h(u, u_0)$, we find

$$\varphi(u) = f(u - u_0 + g(u, u_0)),$$

where $|g(u, u_0)| < a\eta|u - u_0|$, and a was defined at the beginning of the proof.

Now let C be a cube (with faces parallel to the coordinate axes) centered at u_0 and of sidelength $r \leq \delta$, and let $\widetilde{C} = -u_0 + C$ be the "same" cube translated so that its center is the origin. We can apply the previous considerations to $u \in C$: we have then $u - u_0 \in \widetilde{C}$ and $|u - u_0| \leq dr/2$, which implies $|g(u, u_0)| < da\eta r/2$ and thus $g(u, u_0) \in da\eta\,\widetilde{C}$. It follows that

$$\varphi(C) \subset f((1 + da\eta)\widetilde{C}),$$

Thanks to Proposition 7.1, we then get

$$\begin{aligned}\lambda_d(\varphi(C)) \leq \lambda_d(f((1 + da\eta)\widetilde{C})) &= |\det\varphi'(u_0)|\,\lambda_d((1 + da\eta)\widetilde{C})\\ &= (1 + da\eta)^d|J_\varphi(u_0)|\,\lambda_d(C),\end{aligned}$$

which gives the desired upper bound since $(1 + da\eta)^d < 1 + \varepsilon$.

The proof of the lower bound is analogous, and we only sketch the arguments. We again fix $u_0 \in K$ and keep the same notation $f(v) = \varphi(u_0) + \varphi'(u_0)v$ as above. Given a positive constant γ, which can be taken arbitrarily small, we use the equality $(\varphi^{-1})'(\varphi(u_0)) = \varphi'(u_0)^{-1}$ and the mean value inequality to get, for $v \in \mathbb{R}^d$ such that $|v|$ is sufficiently small (independently of the choice of u_0),

$$\begin{aligned}\varphi^{-1}(f(v)) &= \varphi^{-1}(\varphi(u_0) + \varphi'(u_0)v)\\ &= u_0 + (\varphi^{-1})'(\varphi(u_0))\,\varphi'(u_0)v + k(u_0, v)\\ &= u_0 + v + k(u_0, v),\end{aligned}$$

where $|k(u_0, v)| \leq \gamma|v|$. It follows that, for $\delta > 0$ small enough and for every cube C of sidelength smaller than δ centered at u_0, we have, with the same notation $\widetilde{C}$ as above,

$$\varphi^{-1}(f((1 - d\gamma)\widetilde{C})) \subset C\,.$$

This implies that

$$f((1 - d\gamma)\widetilde{C}) \subset \varphi(C)$$

and we complete the proof in the same manner as for the upper bound. □

We return to the proof of the theorem. Let $n \in \mathbb{N}$. We call elementary cube of order n any cube of the form

$$C = \prod_{j=1}^{d} (k_j 2^{-n}, (k_j + 1)2^{-n}] \,, \quad k_j \in \mathbb{Z}.$$

We write $\mathcal{C}_n$ for the class of all elementary cubes of order n, and $\mathcal{C}_n^{(U)}$ for the class of all elementary cubes of order n whose closure is contained in U.

Let $n_0 \in \mathbb{N}$ such that $\mathcal{C}_{n_0}^{(U)}$ is not empty, and let $C_0 \in \mathcal{C}_{n_0}^{(U)}$. Also let $\varepsilon > 0$. Fix $n \geq n_0$ large enough so that the conclusion of Lemma 7.3 holds with $\delta = 2^{-n}$, when K is the closure of C_0. Taking n even larger if necessary, we can assume that, for every $u, v \in K$ such that $|u - v| \leq d2^{-n}$,

$$(1 - \varepsilon)|J_\varphi(u)| \leq |J_\varphi(v)| \leq (1 + \varepsilon)|J_\varphi(u)|. \tag{7.2}$$

Observe that C_0 is the disjoint union of those cubes $C \in \mathcal{C}_n$ that are contained in C_0. Then, writing x_C for the center of a cube C, we have

$$\begin{aligned}
\lambda_d(\varphi(C_0)) = \sum_{\substack{C \in \mathcal{C}_n \\ C \subset C_0}} \lambda_d(\varphi(C)) &\leq (1 + \varepsilon) \sum_{\substack{C \in \mathcal{C}_n \\ C \subset C_0}} |J_\varphi(x_C)|\, \lambda_d(C) \\
&\leq (1 + \varepsilon)^2 \sum_{\substack{C \in \mathcal{C}_n \\ C \subset C_0}} \int_C |J_\varphi(u)|\, \mathrm{d}u \\
&= (1 + \varepsilon)^2 \int_{C_0} |J_\varphi(u)|\, \mathrm{d}u.
\end{aligned}$$

We have used the bounds of Lemma 7.3 in the first inequality, and (7.2) in the second one. Similarly, we get the lower bound

$$\lambda_d(\varphi(C_0)) \geq (1 - \varepsilon)^2 \int_{C_0} |J_\varphi(u)|\, \mathrm{d}u.$$

As ε was arbitrary, we conclude that

$$\lambda_d(\varphi(C_0)) = \int_{C_0} |J_\varphi(u)|\, \mathrm{d}u.$$

We have thus obtained (7.1) when A is a cube of $\mathcal{C}_n^{(U)}$, for every $n \in \mathbb{N}$. The general case now follows from monotone class arguments. Let μ denote the pushforward of Lebesgue measure on D under φ^{-1}:

$$\mu(A) = \lambda_d(\varphi(A))$$

for every Borel subset A of U, and also set

$$\widetilde{\mu}(A) = \int_A |J_\varphi(u)|\,\mathrm{d}u.$$

We have thus obtained that $\mu(C) = \widetilde{\mu}(C)$ for every $C \in \mathcal{C}_n^{(U)}$ and every $n \in \mathbb{N}$. On the other hand, if U_n denotes the (disjoint) union of the cubes of $\mathcal{C}_n^{(U)}$ that are also contained in $[-n, n]^d$, we have $U_n \uparrow U$ as $n \to \infty$ and $\mu(U_n) = \widetilde{\mu}(U_n) < \infty$ for every n. Since the class of all elementary cubes whose closure is contained in U (we include $\varnothing$ in this class) is closed under finite intersections and generates the Borel σ-field on U, we can apply Corollary 1.19 to conclude that $\mu = \widetilde{\mu}$, which was the desired result. □

Application to Polar Coordinates

We take $d = 2$, $U = (0, \infty) \times (-\pi, \pi)$ and $D = \mathbb{R}^2 \backslash \{(x, 0); x \leq 0\}$. Then the function $\varphi : U \longrightarrow D$ defined by

$$\varphi(r, \theta) := (r \cos\theta, r \sin\theta)\,, \quad (r, \theta) \in U$$

is a C^1-diffeomorphism from U onto D. One easily computes

$$\varphi'(r, \theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

and thus $J_\varphi(r, \theta) = r$.

It follows from Theorem 7.2 that, for every Borel measurable function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}_+$,

$$\begin{aligned}\int_D f(x, y)\,\mathrm{d}x\mathrm{d}y &= \int_U f(r\cos\theta, r\sin\theta)\, r\,\mathrm{d}r\mathrm{d}\theta \\ &= \int_0^\infty \int_{-\pi}^{\pi} f(r\cos\theta, r\sin\theta)\, r\,\mathrm{d}r\mathrm{d}\theta.\end{aligned}$$

Since the negative half-line has zero Lebesgue measure in $\mathbb{R}^2$, we have also

$$\int_{\mathbb{R}^2} f(x, y)\,\mathrm{d}x\mathrm{d}y = \int_0^\infty \int_{-\pi}^{\pi} f(r\cos\theta, r\sin\theta)\, r\,\mathrm{d}r\mathrm{d}\theta.$$

This formula will be generalized to higher dimensions in Theorem 7.4 below.

Example If $f(x, y) = \exp(-x^2 - y^2)$, Theorem 5.3 gives

$$\int_{\mathbb{R}^2} e^{-x^2-y^2}\, \mathrm{d}x\mathrm{d}y = \Big(\int_{-\infty}^{+\infty} e^{-x^2}\, \mathrm{d}x\Big)^2$$

whereas

$$\int_0^\infty \int_{-\pi}^{\pi} f(r\cos\theta, r\sin\theta)\, r\, \mathrm{d}r\mathrm{d}\theta = 2\pi \int_0^\infty e^{-r^2}\, r\, \mathrm{d}r = \pi.$$

It follows that

$$\int_{-\infty}^{+\infty} e^{-x^2}\, \mathrm{d}x = \sqrt{\pi}. \tag{7.3}$$

This calculation is important in probability theory as it leads to the definition of the Gaussian (or normal) distribution!

7.2 The Gamma Function

The Gamma function is defined for $a > 0$ by

$$\Gamma(a) = \int_0^\infty x^{a-1}\, e^{-x}\, \mathrm{d}x.$$

An integration by parts shows that, for every $a > 0$,

$$\Gamma(a+1) = \int_0^\infty x^a\, e^{-x}\, \mathrm{d}x = \int_0^\infty a\, x^{a-1}\, e^{-x}\, \mathrm{d}x = a\Gamma(a).$$

Since $\Gamma(1) = 1$, it follows by induction that $\Gamma(n) = (n-1)!$ for every integer $n \geq 1$. We can also compute $\Gamma(\frac{1}{2} + n)$ for every integer $n \geq 0$. A change of variables shows that

$$\Gamma\Big(\frac{1}{2}\Big) = \int_0^\infty x^{-1/2}\, e^{-x}\, \mathrm{d}x = 2\int_0^\infty e^{-y^2}\, \mathrm{d}y = \sqrt{\pi},$$

by formula (7.3). Consequently,

$$\Gamma\Big(\frac{1}{2}+n\Big) = \frac{1\times 3\times\cdots\times(2n-1)}{2^n}\,\sqrt{\pi},$$

for every integer $n \geq 0$.

Recall our notation $\gamma_d = \lambda_d(\mathbb{B}_d)$ for the volume of the unit ball in $\mathbb{R}^d$, and formulas (5.6) giving the value of γ_d when $d \geq 2$. From the expression of $\Gamma(n)$ and $\Gamma(\frac{1}{2}+n)$, we can rewrite these formulas in the form

$$\gamma_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}. \tag{7.4}$$

7.3 Lebesgue Measure on the Unit Sphere

Let $\mathbb{S}^{d-1}$ be the unit sphere of $\mathbb{R}^d$:

$$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : |x| = 1\}.$$

If $A \in \mathcal{B}(\mathbb{S}^{d-1})$, we let $\Theta(A)$ be the Borel subset of $\mathbb{R}^d$ defined by

$$\Theta(A) := \{rx : r \in [0,1] \text{ and } x \in A\}.$$

Theorem 7.4 *For every $A \in \mathcal{B}(\mathbb{S}^{d-1})$, set*

$$\omega_d(A) := d\,\lambda_d(\Theta(A)).$$

Then ω_d is a finite positive measure on $\mathbb{S}^{d-1}$, which is invariant under vector isometries of $\mathbb{R}^d$. Moreover, for every nonnegative Borel measurable function f on $\mathbb{R}^d$, we have

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_0^\infty \int_{\mathbb{S}^{d-1}} f(rz)\,r^{d-1}\,\mathrm{d}r\,\omega_d(\mathrm{d}z). \tag{7.5}$$

The total mass of ω_d (volume of the unit sphere) is

$$\omega_d(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}.$$

Remark One can also show that any finite measure on $\mathbb{S}^{d-1}$ that is invariant under vector isometries is of the form $c\,\omega_d$, for some $c \geq 0$ (see Lemma 14.20 below).

Proof To see that ω_d is a finite positive measure on $\mathbb{S}^{d-1}$, we just observe that ω_d is (d times) the pushforward of the restriction of λ_d to the punctured unit ball $\mathbb{B}_d\backslash\{0\}$ under the mapping $x \mapsto |x|^{-1}x$. The fact that λ_d is invariant under vector isometries of $\mathbb{R}^d$ (Proposition 7.1) readily implies that the same property holds for ω_d. Indeed, if φ is a vector isometry of $\mathbb{R}^d$,

$$\lambda_d(\Theta(\varphi^{-1}(A))) = \lambda_d(\varphi^{-1}(\Theta(A))) = \lambda_d(\Theta(A)).$$

Using formula (7.4), the total mass of ω_d is

$$\omega_d(\mathbb{S}^{d-1}) = d\,\lambda_d(\mathbb{B}_d) = d\gamma_d = d\,\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}.$$

We still have to establish (7.5). It is enough to treat the case $f = \mathbf{1}_B$, where B is a Borel subset of $\mathbb{R}^d\backslash\{0\}$. The formula

$$\mu(B) := \int_0^\infty \int_{\mathbb{S}^{d-1}} \mathbf{1}_B(rz)\,r^{d-1}\,\mathrm{d}r\,\omega_d(dz)$$

defines a measure μ on $\mathbb{R}^d\backslash\{0\}$ and we have to verify that $\mu = \lambda_d$. Consider first the case of a Borel set B of the form

$$B = \{x \in \mathbb{R}^d\backslash\{0\} : a < |x| \le b \text{ and } \frac{x}{|x|} \in A\},$$

where A is a Borel subset of $\mathbb{S}^{d-1}$, and $0 < a < b$. Then,

$$\mu(B) = \omega_d(A)\int_a^b r^{d-1}\,\mathrm{d}r = \frac{b^d - a^d}{d}\omega_d(A).$$

To compute $\lambda_d(B)$, set $\alpha = a/b \in (0,1)$, and, for every integer $n \ge 0$,

$$\Theta_n(A) = \{y = rx : \alpha^{n+1} < r \le \alpha^n \text{ and } x \in A\}.$$

Then, $\lambda_d(\Theta_n(A)) = \alpha^{nd}\lambda_d(\Theta_0(A))$ and, on the other hand,

$$\lambda_d(\Theta(A)) = \sum_{n=0}^{\infty} \lambda_d(\Theta_n(A)).$$

It immediately follows that

$$\lambda_d(\Theta_0(A)) = (1-\alpha^d)\,\lambda_d(\Theta(A)) = \frac{1-\alpha^d}{d}\,\omega_d(A),$$

and, since $B = b\,\Theta_0(A)$,

$$\lambda_d(B) = b^d \lambda_d(\Theta_0(A)) = \frac{b^d - a^d}{d}\omega_d(A) = \mu(B).$$

Finally, the class of all sets B of the preceding type is closed under finite intersections and generates the Borel σ-field on $\mathbb{R}^d\backslash\{0\}$ (we leave the easy verification of the latter fact to the reader). We can again use Corollary 1.19 to conclude that $\mu = \lambda_d$. □

If $f : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ is a radial function, meaning that $f(x) = g(|x|)$ for some Borel measurable function $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, Theorem 7.4 shows that

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = c_d \int_0^\infty g(r)\, r^{d-1}\,\mathrm{d}r,$$

where $c_d = \omega_d(\mathbb{S}^{d-1})$.

7.4 Exercises

As we will see later for some of them, most of the calculations in the exercises below have natural interpretations in probability theory.

Exercise 7.1 Let $\mu(\mathrm{d}u\mathrm{d}v)$ be the probability measure on $(0,\infty) \times [0,1]$ with density e^{-u} with respect to Lebesgue measure $\mathrm{d}u\,\mathrm{d}v$. Verify that the pushforward of μ under the mapping $\Phi : (0,\infty) \times [0,1] \longrightarrow \mathbb{R}^2$ defined by

$$\Phi(u,v) = (\sqrt{u}\cos(2\pi u), \sqrt{u}\sin(2\pi v))$$

can be written as $\nu \otimes \nu$, where ν is a probability measure to be determined. (*A probabilistic interpretation of this exercise can be found in the example following Corollary 9.6 below.*)

Exercise 7.2 Prove that, for every $a, b > 0$,

$$\Gamma(a)\Gamma(b) = 4\int_{(0,\infty)^2} e^{-(u^2+v^2)} u^{2a-1} v^{2b-1}\,\mathrm{d}u\,\mathrm{d}v,$$

and use this and an integration in polar coordinates to derive the formula

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t. \tag{7.6}$$

The function $(a,b) \mapsto \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is called the Beta function.

Exercise 7.3

(1) Let $\mu(\mathrm{d}u)$ be the probability measure on $(0, \infty)$ with density e^{-u} with respect to Lebesgue measure. Prove that the pushforward of $\mu \otimes \mu$ under the mapping $\Phi : (0, \infty)^2 \longrightarrow (0, 1) \times (0, \infty)$ defined by

$$\Phi(u, v) = \Big(\frac{u}{u+v}, u+v\Big)$$

can be written as $\nu \otimes \theta$, where ν is Lebesgue measure on $(0, 1)$ and θ is a probability measure on $(0, \infty)$ to be determined. (*See Exercise 9.2 below for a probabilistic interpretation*.)

(2) More generally, for every $a > 0$ and $b > 0$, consider the measure $\mu_{a,b}(\mathrm{d}u\,\mathrm{d}v)$ on $(0, \infty)^2$ with density $u^{a-1}v^{b-1}e^{-u-v}$ with respect to Lebesgue measure on $(0, \infty)^2$. Compute the pushforward of $\mu_{a,b}$ under Φ and use this to reprove formula (7.6).

Exercise 7.4 Let $\mu(\mathrm{d}x\mathrm{d}y)$ be the measure on $\mathbb{R}^2$ with density $e^{-(x^2+y^2)}$ with respect to Lebesgue measure on $\mathbb{R}^2$. Compute the pushforward of μ under the mapping $\Psi : \mathbb{R}^2 \longrightarrow \mathbb{R}_+ \times [0, 1]$ defined by

$$\Psi(x, y) = \Big(x^2 + y^2, \frac{x^2}{x^2 + y^2}\Big).$$

Similarly, compute the pushforward of μ under the mapping $\Phi(x, y) = x/y$. Note that this makes sense since $\mu(\{(x, y) : y = 0\}) = 0$. (*See Exercise 9.3 below for a probabilistic interpretation*.)

Exercise 7.5 Let $\gamma > 0$. Show that, for any nonnegative measurable function f on $\mathbb{R}$,

$$\int_{\mathbb{R}} f(x - \frac{\gamma}{x})\,\mathrm{d}x = \int_{\mathbb{R}} f(x)\,\mathrm{d}x.$$

Hint: The mapping $x \mapsto x - \frac{\gamma}{x}$ is a diffeomorphism from $(0, \infty)$ onto $\mathbb{R}$, resp. from $(-\infty, 0)$ onto $\mathbb{R}$.

Part II
Probability Theory

Chapter 8
Foundations of Probability Theory

This chapter introduces the fundamental notions of probability theory: random variables, law, expected value, variance, moments of random variables, characteristic functions, etc. Since a probability space is nothing else than a measurable space equipped with a measure of total mass 1, many of these notions correspond to those we have introduced in the general setting of measure theory. For instance, a random variable is just a measurable function, and the expected value coincides with the integral of a measurable function.

However the point of view of probability theory, which is explained at the beginning of this chapter, is rather different, and it is important to have a good understanding of this point of view. In particular, the notion of the law of a random variable, which is a special case of the pushforward of a measure, is fundamental as it expresses the probability that a random variable "falls" into a given set. We list several classical probability laws, some of which will play a major role in the forthcoming chapters. We also restate some basic notions and facts of measure theory in the language of probability theory, and we apply the Hilbert space structure of L^2 to the linear regression problem, which consists in finding the best approximation of a (real) random variable as an affine function of several other random variables. The end of the chapter gives several ways of characterizing the law of a random variable with values in $\mathbb{R}$ or in $\mathbb{R}^d$. Of special importance is the characteristic function, which is defined as the Fourier transform of the law of the random variable in consideration.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_8

8.1 General Definitions

8.1.1 Probability Spaces

Let $(\Omega, \mathcal{A})$ be a measurable space, and let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{A})$, that is, a positive measure with total mass 1. We say that $(\Omega, \mathcal{A}, \mathbb{P})$ is a *probability space*.

A probability space is thus a special case of a measure space, when the total mass of the measure is equal to 1. The point of view of probability theory is however rather different from measure theory. In probability theory, one is aiming at a mathematical model for a "random experiment":

- Ω represents the set of all possible outcomes for the experiment. If we fix $\omega \in \Omega$, this means that we have completely determined all random quantities involved in the experiment.
- $\mathcal{A}$ is the set of all "events". Here events are the subsets of Ω whose probability can be evaluated (the measurable sets, in the language of measure theory). One should view an event A as the subset of all $\omega \in \Omega$ for which a certain property is satisfied.
- For every $A \in \mathcal{A}$, $\mathbb{P}(A)$ represents the probability that the event A occurs.

The idea of developing probability theory in the framework of measure theory goes back to Kolmogorov in 1933. For this reason, the properties of probability measures, in particular the σ-additivity of these measures, are often referred to as *Kolmogorov axioms*.

In the early age of probability theory (long before the invention of measure theory), the probability of an event was defined as the ratio of the number of favorable cases to the number of all possible cases (it was implicitly assumed that all possible cases were equiprobable). This is sometimes called the "classical" definition of probability, and it corresponds to the uniform probability measure in Example (1) below. The classical definition can be criticized in several respects, in particular because it does not apply to situations where there are infinitely many possible cases. In the nineteenth century, the so-called frequentist definition of probability became well established and even dominant. According to this definition, if one repeats the random experiment N times, and if N_A denotes the number of trials at which the event A occurs, the ratio N_A/N should be close to the probability $\mathbb{P}(A)$ when N is large. This is an informal statement of the *law of large numbers*, which we will discuss later.

One may naively ask why one could not consider the probability of "any" subset of Ω (why do we have to restrict to events, which belong to the σ-field $\mathcal{A}$?). The reason is the fact that it is in general not possible to define interesting probability measures on the set $\mathcal{P}(\Omega)$ of all subsets of Ω (except in simple situations where Ω is countable). Taking $\Omega = [0, 1]$, equipped with its Borel σ-field and with Lebesgue measure, already makes it possible to define many deep and interesting models of probability theory. However, the example treated in Section 3.4 indicates that there is no hope to extend Lebesgue measure to all subsets of [0, 1].

Examples (1) The experiment consists in throwing a (well-balanced) die twice. Then,

$$\Omega = \{1, 2, \ldots, 6\}^2 \ , \ \mathcal{A} = \mathcal{P}(\Omega) \ , \ \mathbb{P}(A) = \frac{\text{card}(A)}{36}.$$

The choice of the probability $\mathbb{P}$ corresponds to the (intuitive) idea that all outcomes have the same probability here. More generally, if Ω is a finite set, and $\mathcal{A} = \mathcal{P}(\Omega)$, the probability measure defined by $\mathbb{P}(\{\omega\}) := 1/\text{card}(\Omega)$ is called the uniform probability measure on Ω.
(2) We throw a die until we get 6. Here the choice of Ω is already less obvious. As the number of trials needed to get 6 is not bounded (even if you throw the die 1000 times there is a small probability that you do not get a 6), the right choice for Ω is to imagine that you throw the die infinitely many times

$$\Omega = \{1, 2, \ldots, 6\}^{\mathbb{N}}$$

in such a way that an element of Ω is a sequence $\omega = (\omega_1, \omega_2, \ldots)$ giving the outcomes of the successive trials. The σ-field $\mathcal{A}$ on Ω is defined as the smallest σ-field containing all sets of the form

$$\{\omega \in \Omega : \omega_1 = i_1, \omega_2 = i_2, \ldots, \omega_n = i_n\}$$

where $n \geq 1$ and $i_1, \ldots, i_n \in \{1, 2, \ldots, 6\}$ — we want at least to be able to compute the probability of given outcomes in a finite number of trials. Finally, $\mathbb{P}$ is the unique probability measure on Ω such that, for every choice of n and $i_1, \ldots, i_n$,

$$\mathbb{P}(\{\omega \in \Omega : \omega_1 = i_1, \omega_2 = i_2, \ldots, \omega_n = i_n\}) = \left(\frac{1}{6}\right)^n .$$

The uniqueness of $\mathbb{P}$ is a simple consequence of Corollary 1.19. The existence of $\mathbb{P}$ requires more work, but may be deduced from the fact that we can construct a sequence of independent random variables uniformly distributed over $\{1, \ldots, 6\}$ (this will be discussed in Section 9.4 below). Alternatively, one can view the existence of $\mathbb{P}$ as an extension of the construction of product measures in Chapter 5 to the case of infinite products.
(3) Suppose we want to model the displacement in space of a particle subject to random perturbations. If we restrict our attention to the time interval [0, 1], we choose the probability space $\Omega = C([0, 1], \mathbb{R}^3)$ consisting of all continuous functions $\omega : [0, 1] \longrightarrow \mathbb{R}^3$ (each such function corresponds to a possible trajectory of the particle). The σ-field $\mathcal{A}$ on Ω may be defined as the Borel σ-field if $C([0, 1], \mathbb{R}^3)$ is equipped with the topology of uniform convergence (equivalently, $\mathcal{A}$ is the smallest σ-field for which the "coordinate mappings" $\omega \mapsto$

$\omega(t)$ are measurable for every $t \in [0, 1]$, see Exercise 1.3). Then there are many possible choices for the probability measure $\mathbb{P}$. Perhaps the most important one, corresponding to a "purely random" motion, is the Wiener measure, corresponding to the law of Brownian motion, which we will construct in Chapter 14.

Important Remark Very often in what follows, the precise choice of the probability measure $\mathbb{P}$ will not be made explicit. The important data will be the properties of the (measurable) functions defined on $(\Omega, \mathcal{A})$, which are called random variables.

Terminology Even more than in measure theory, sets of measure zero are present in many statements of probability theory. We say that a property depending on $\omega \in \Omega$ holds almost surely (a.s. in short) if it holds for every ω belonging to an event of probability one. So almost surely has the same meaning as almost everywhere in measure theory.

8.1.2 Random Variables

In the remaining part of this chapter, we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and all random variables will be defined on this probability space.

Definition 8.1 Let $(E, \mathcal{E})$ be a measurable space. A random variable with values in E is a measurable function $X : \Omega \longrightarrow E$.

We will very often consider the case $E = \mathbb{R}$ or $E = \mathbb{R}^d$. The σ-field on E will then always be the Borel σ-field. We speak of a real random variable when $E = \mathbb{R}$ and of a random vector (also called a multivariate random variable) when $E = \mathbb{R}^d$. A real random variable X (or a random vector X) is said to be bounded if there exists a constant $K \geq 0$ such that $|X| \leq K$ a.s.

Examples Recall the three cases that have been discussed in the previous section, and let us give examples of random variables in each case.

(1) $X((i, j)) = i + j$ defines a random variable with values in $\{2, 3, \ldots, 12\}$.
(2) $X(\omega) = \inf\{j : \omega_j = 6\}$, with the convention $\inf \varnothing = \infty$, defines a random variable with values in $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$. To verify the measurability of $\omega \mapsto X(\omega)$, we write, for every $k \in \mathbb{N}$,

$$X^{-1}(\{k\}) = \{\omega \in \Omega : \omega_1 \neq 6, \omega_2 \neq 6, \ldots, \omega_{k-1} \neq 6, \omega_k = 6\}.$$

(3) For every fixed $t \in [0, 1]$, $X(\omega) = \omega(t)$ is a random variable with values in $\mathbb{R}^3$. The measurability is clear since $\omega \mapsto \omega(t)$ is even continuous. We note that we have not specified $\mathbb{P}$ in this example, but this is irrelevant for measurability questions.

We turn to a very important definition.

Definition 8.2 Let X be a random variable with values in $(E, \mathcal{E})$. The law $\mathbb{P}_X$ of the random variable X is the pushforward of the probability measure $\mathbb{P}$ under X. In other words, $\mathbb{P}_X$ is the probability measure on $(E, \mathcal{E})$ that is defined by

$$\mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B)) , \quad \forall B \in \mathcal{E}.$$

We also call $\mathbb{P}_X$ the distribution of X, and we say that X is distributed according to $\mathbb{P}_X$. Two random variables Y and Y' with values in $(E, \mathcal{E})$ are said to be identically distributed if they have the same law $\mathbb{P}_Y = \mathbb{P}_{Y'}$.

In probability theory, one prefers to write $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$ instead of $\mathbb{P}(X^{-1}(B))$. Here $X \in B$ is shorthand for $\{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B)$ (as another general rule, the parameter ω is often "hidden" in formulas of probability theory).

The law $\mathbb{P}_X$ gives the probability of events that "depend" on the random variable X. Informally, we may say that we associate a "random point" $X(\omega)$ with each $\omega \in \Omega$, and that $\mathbb{P}_X(B)$ is the probability that this random point "falls" in B.

Remark If μ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (or on a more general measurable space), there is a canonical way to define a random variable X whose law is μ. It suffices to take $\Omega = \mathbb{R}^d$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$, $\mathbb{P} = \mu$, and $X(\omega) = \omega$. It is then trivial that the law of X is μ.

Special Cases

- **Discrete random variables**. This is the case where E is finite or countable (and $\mathcal{E} = \mathcal{P}(E)$). The law of X is then a point measure, which can be written as

$$\mathbb{P}_X = \sum_{x \in E} p_x \, \delta_x$$

  where, for every $x \in E$, δ_x is the Dirac measure at x, and $p_x = \mathbb{P}(X = x)$ (as explained above $\mathbb{P}(X = x)$ stands for $\mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$). Indeed, for every subset B of E,

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\Big(\bigcup_{x \in B} \{X = x\}\Big) = \sum_{x \in B} \mathbb{P}(X = x) = \sum_{x \in E} p_x \, \delta_x(B).$$

  Note that the fact that E (hence B) is at most countable is used in the third equality. In practice, finding the law of a discrete random variable with values in E means computing the probabilities $\mathbb{P}(X = x)$ for every $x \in E$.

Example Let us consider example (2) above, with $X(\omega) = \inf\{j : \omega_j = 6\}$. Then, for every $k \geq 1$,

$$\begin{aligned}\mathbb{P}(X = k) &= \mathbb{P}\Big(\bigcup_{i_1,\ldots,i_{k-1}\neq 6} \{\omega_1 = i_1, \ldots, \omega_{k-1} = i_{k-1}, \omega_k = 6\}\Big)\\ &= 5^{k-1}\,(\frac{1}{6})^k\\ &= \frac{1}{6}(\frac{5}{6})^{k-1}.\end{aligned}$$

We note that $\sum_{k=1}^{\infty} \mathbb{P}(X = k) = 1$ and thus $\mathbb{P}(X = \infty) = 1 - \mathbb{P}(X \in \mathbb{N}) = 0$. However, the set $\{X = \infty\}$ is not empty as it contains all sequences $(i_1, i_2, \ldots)$ such that $i_k \neq 6$ for every $k \in \mathbb{N}$.

- **Random variables with a density**. A probability density function on $\mathbb{R}^d$ is a nonnegative Borel function $p : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ such that

  $$\int_{\mathbb{R}^d} p(x)\,\mathrm{d}x = 1.$$

  Then a random variable X with values in $\mathbb{R}^d$ is said to have density p if

  $$\mathbb{P}_X(B) = \int_B p(x)\,\mathrm{d}x,$$

  for every Borel subset B of $\mathbb{R}^d$. In other words, p is the density of $\mathbb{P}_X$ with respect to Lebesgue measure on $\mathbb{R}^d$.

Note that p is determined from $\mathbb{P}_X$ (or from X) only up to a set of zero Lebesgue measure. In most examples that we will encounter, p will be continuous on $\mathbb{R}^d$ (or, in the case $d = 1$, p will be continuous on $[0, \infty)$ and vanish on $(-\infty, 0)$) and under this additional condition p is determined uniquely by $\mathbb{P}_X$.

If $d = 1$, we have in particular, for every $\alpha \leq \beta$,

$$\mathbb{P}(\alpha \leq X \leq \beta) = \int_\alpha^\beta p(x)\,\mathrm{d}x.$$

8.1.3 Mathematical Expectation

Definition 8.3 Let X be a real random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$. We set

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathbb{P}(\mathrm{d}\omega),$$

provided that the integral makes sense, which (according to Chapter 2) is the case if one of the following two conditions holds:

- $X \geq 0$ (then $\mathbb{E}[X] \in [0, \infty]$),
- X is of arbitrary sign and $\mathbb{E}[|X|] = \int |X| \, \mathrm{d}\mathbb{P} < \infty$.

The quantity $\mathbb{E}[X]$ is called the expected value or expectation of X.

As we saw in Chapter 2, the case $X \geq 0$ can be extended to random variables with values in $[0, \infty]$, and this is often useful. As in Chapter 2, we can also define $\mathbb{E}[X]$ when X is a complex-valued random variable such that $\mathbb{E}[|X|] < \infty$ (we just deal separately with the real and the imaginary parts).

The preceding definition is extended to the case of a random vector $X = (X_1, \ldots, X_d)$ with values in $\mathbb{R}^d$ by setting $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$, provided of course that all quantities $\mathbb{E}[X_i]$, $1 \leq i \leq d$, are well defined. Similarly, if M is a random matrix (a random variable with values in the space of $n \times d$ real matrices), we can define the matrix $\mathbb{E}[M]$ by taking the expectation of each component of M, provided these expectations are well defined.

Remark If $X = \mathbf{1}_B$, $\mathbb{E}[X] = \mathbb{P}(B)$. In general, $\mathbb{E}[X]$ is interpreted as the mean value of X, and is also called the *mean* of X. In the special case where Ω is finite and $\mathbb{P}$ is the uniform probability measure on Ω, $\mathbb{E}[X]$ is the average (in the usual sense) of the values taken by X.

The following proposition is very easy, but nonetheless often useful.

Proposition 8.4 *Let X be a random variable with values in* $[0, \infty]$. *Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x) \, \mathrm{d}x.$$

Let Y be a random variable with values in $\mathbb{Z}_+$. Then,

$$\mathbb{E}[Y] = \sum_{k=0}^{\infty} k \, \mathbb{P}(Y = k) = \sum_{k=1}^{\infty} \mathbb{P}(Y \geq k).$$

Proof For the first assertion, we use the Fubini theorem to write

$$\mathbb{E}[X] = \mathbb{E}\Big[\int_0^\infty \mathbf{1}_{\{x \leq X\}} \, \mathrm{d}x\Big] = \int_0^\infty \mathbb{E}[\mathbf{1}_{\{x \leq X\}}] \, \mathrm{d}x = \int_0^\infty \mathbb{P}(X \geq x) \, \mathrm{d}x.$$

If Y is a random variable with values in $\mathbb{Z}_+$, the equality $\mathbb{E}[Y] = \sum_{k=0}^{\infty} k \, \mathbb{P}(Y = k)$ is immediate from the definition since $Y = \sum_{k=0}^{\infty} k \mathbf{1}_{\{Y=k\}}$. The other formula is also easy by writing $\mathbb{E}[Y] = \mathbb{E}[\sum_{k=1}^{\infty} \mathbf{1}_{\{Y \geq k\}}]$ and interchanging sum and expectation. □

The next proposition is a special case of Proposition 2.9. This special case deserves to be stated again because of its importance in probability theory.

Proposition 8.5 *Let X be a random variable with values in $(E, \mathcal{E})$. For every measurable function $f : E \longrightarrow [0, \infty]$, we have*

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

If f is of arbitrary sign (or even complex-valued), the formula of the proposition remains valid provided that both sides makes sense, which holds if $\mathbb{E}[|f(X)|] < \infty$. In particular, if X is a real random variable such that $\mathbb{E}[|X|] < \infty$, we have

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

Characterization of the Law of a Random Variable It will be useful to characterize the law of a random variable in different ways. It follows from Corollary 1.19 that a probability measure μ on $\mathbb{R}^d$ is characterized by its values on open rectangles, that is, by the quantities $\mu((a_1, b_1) \times \cdots \times (a_d, b_d))$, for any choice of the reals $a_1 < b_1, a_2 < b_2, \ldots, a_d < b_d$ (one could restrict to rational values of a_i, b_i, and/or replace open intervals by closed intervals). Since the indicator function of an open rectangle is the increasing limit of a sequence of continuous functions with compact support, the probability measure μ is also determined by the values of $\int \varphi(x)\, \mu(\mathrm{d}x)$ when φ varies in the space $C_c(\mathbb{R}^d)$ of all continuous functions with compact support from $\mathbb{R}^d$ into $\mathbb{R}$. In particular, for a random vector $X = (X_1, \ldots, X_d)$ with values in $\mathbb{R}^d$, the law $\mathbb{P}_X$ is characterized by the quantities

$$\mathbb{P}_X((a_1, b_1) \times \cdots \times (a_d, b_d)) = \mathbb{P}(a_1 < X_1 < b_1, \ldots, a_d < X_d < b_d)$$

for any reals $a_1 < b_1, a_2 < b_2, \ldots, a_d < b_d$, or alternatively by the quantities

$$\int_{\mathbb{R}^d} \varphi(x)\, \mathbb{P}_X(\mathrm{d}x) = \mathbb{E}[\varphi(X)],$$

when φ varies in $C_c(\mathbb{R}^d)$. In Sections 8.2.3 and 8.2.4, we will see other ways of characterizing the law of a random variable.

Proposition 8.5 shows that expected values of random variables of the form $f(X)$ (where f is a real function) can be computed from the law $\mathbb{P}_X$. On the other hand, the same proposition is often used to compute the law of X: if we are able to find a probability measure ν on E such that

$$\mathbb{E}[f(X)] = \int f\, \mathrm{d}\nu$$

for all real functions f belonging to a “sufficiently large” class (e.g., when $E = \mathbb{R}^d$, the class of all continuous functions with compact support) then we can conclude that $\mathbb{P}_X = \nu$. Let us give an illustration of this general principle.

Proposition 8.6 *Let $X = (X_1, \ldots, X_d)$ be a random vector with values in $\mathbb{R}^d$. Suppose that X has a density $p(x_1, \ldots, x_d)$. Then, for every $j \in \{1, \ldots, d\}$, X_j has a density given by*

$$p_j(x) = \int_{\mathbb{R}^{d-1}} p(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_d)\, \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1}\mathrm{d}x_{j+1} \ldots \mathrm{d}x_d$$

(for instance, if $d = 2$, $p_1(x) = \int_{\mathbb{R}} p(x, y)\, \mathrm{d}y$, $p_2(y) = \int_{\mathbb{R}} p(x, y)\, \mathrm{d}x$).

Remark If p_j is given as in the proposition, there may exist values of $x \in \mathbb{R}$ for which $p_j(x) = \infty$. However, the fact that $\int_{\mathbb{R}^d} p(x_1, \ldots, x_d)\mathrm{d}x_1 \ldots \mathrm{d}x_d = 1$ and the Fubini theorem readily imply that the Lebesgue measure of $\{x \in \mathbb{R} : p_j(x) = \infty\}$ is zero, and we can always change the values of a probability density function on a set of zero Lebesgue measure.

Proof Let π_j be the projection $\pi_j(x_1, \ldots, x_d) = x_j$. Using the Fubini theorem, we have, for every Borel function $f : \mathbb{R} \longrightarrow \mathbb{R}_+$,

$$\begin{aligned}\mathbb{E}[f(X_j)] &= \mathbb{E}[f(\pi_j(X))] \\ &= \int_{\mathbb{R}^d} f(x_j) p(x_1, \ldots, x_d)\, \mathrm{d}x_1 \ldots \mathrm{d}x_d \\ &= \int_{\mathbb{R}} f(x_j) \Big(\int_{\mathbb{R}^{d-1}} p(x_1, \ldots, x_d)\, \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1}\mathrm{d}x_{j+1} \ldots \mathrm{d}x_d \Big) \mathrm{d}x_j \\ &= \int_{\mathbb{R}} f(x_j) p_j(x_j)\, \mathrm{d}x_j,\end{aligned}$$

and the desired result follows. □

If $X = (X_1, \ldots, X_d)$ is a random vector with values in $\mathbb{R}^d$, the probability measures $\mathbb{P}_{X_j}$, $1 \leq j \leq d$ are called the marginal distributions of X (or simply the marginals of X). These laws are determined by the law $\mathbb{P}_X$ of X: it is immediate that $\mathbb{P}_{X_j}$ is the pushforward of $\mathbb{P}_X$ under π_j, with the notation introduced above. It is important to observe that the converse is false:

$\mathbb{P}_X$ **is not determined by the marginal distributions** $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$

except in the very special case where all but at most one of the probability measures $\mathbb{P}_{X_j}$ are Dirac measures. To give an example, consider a probability density function q on $\mathbb{R}$, and observe that the function $p(x_1, x_2) = q(x_1)q(x_2)$ is then also a probability density function on $\mathbb{R}^2$ (by the Fubini theorem). By a preceding remark, we can construct, on a suitable probability space, a random vector $X = (X_1, X_2)$ with density p. But then the two random vectors $X = (X_1, X_2)$ and $X' = (X_1, X_1)$ have the same marginal distributions (Proposition 8.6 shows that both $\mathbb{P}_{X_1}$ and $\mathbb{P}_{X_2}$ have density q), whereas $\mathbb{P}_X$ and $\mathbb{P}_{X'}$ are very different, simply because $\mathbb{P}_{X'}$ is singular with respect to Lebesgue measure ($\mathbb{P}_{X'}$ is supported on the diagonal of $\mathbb{R}^2$, which has zero Lebesgue measure), and $\mathbb{P}_X$ is absolutely continuous with respect to Lebesgue measure.

8.1.4 An Example: Bertrand's Paradox

To illustrate the notions introduced in the previous sections, let us consider the following problem. Suppose we throw a chord at random on a circle. What is the probability that this chord is longer than the side of the equilateral triangle inscribed in the circle ? Without loss of generality, we may assume that the circle is the unit circle of the plane. Bertrand suggested two different methods to compute this probability (Fig. 8.1).

(a) Both endpoints of the chord are chosen at random on the circle. Given the first endpoint, the length of the chord will be longer than the side of the equilateral triangle inscribed in the circle if and only if the second endpoint lies in an angular sector of opening $2\pi/3$. The probability is thus $\frac{2\pi/3}{2\pi} = \frac{1}{3}$.
(b) The center of the chord is chosen at random in the unit disk. The desired probability is the probability that this center falls in the disk of radius $\frac{1}{2}$ centered at the origin. Since the area of this disk is a quarter of the area of the unit disk, we get probability $\frac{1}{4}$.

We thus get a different result in cases (a) and (b). The explanation of this (apparent) paradox lies in the fact that we have two different random experiments, modeled by different choices of probability spaces and random variables. In fact, the

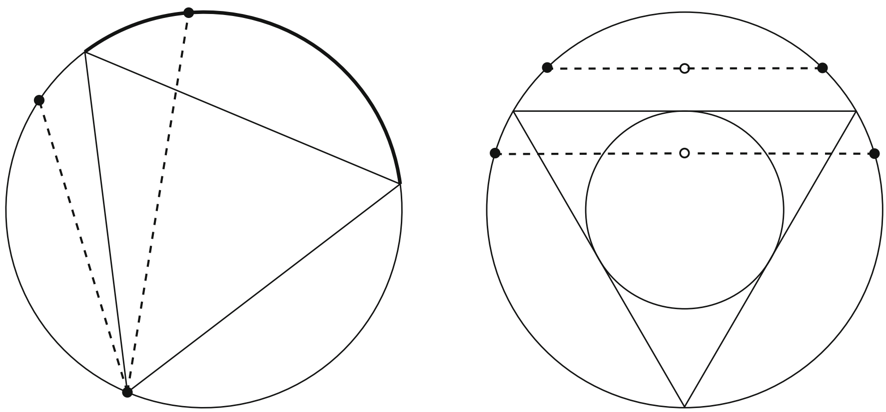

Fig. 8.1 Illustration of Bertrand's paradox. One compares the length of the chord (in dashed lines) to that of the side of the equilateral triangle inscribed in the circle. On the left, two cases depending on the choice of the second endpoint of the chord. On the right, two cases depending on the position of the center of the chord

assertion "we choose a chord at random" has no precise meaning if we do not specify the algorithm that is used to generate the random chord. The law of the random variable that represents the length of the chord will be different in the mathematical models corresponding to cases (a) and (b). Let us make this more explicit.
(a) In that case, the positions of the endpoints of the chord are represented by two angles θ and θ'. The probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is then defined by

$$\Omega = [0, 2\pi)^2\,, \quad \mathcal{A} = \mathcal{B}([0, 2\pi)^2)\,, \quad \mathbb{P}(\mathrm{d}\omega) = \frac{1}{4\pi^2}\mathrm{d}\theta\, \mathrm{d}\theta',$$

where we write $\omega = (\theta, \theta')$ for $\omega \in \Omega$. The length of the chord is

$$X(\omega) = 2|\sin(\frac{\theta - \theta'}{2})|.$$

It is easy to compute the law of X. Using the general principle given in the previous section (before Proposition 8.6), we write

$$\begin{aligned}
\mathbb{E}[f(X)] &= \int_\Omega f(X(\omega))\,\mathbb{P}(\mathrm{d}\omega) \\
&= \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} f(2|\sin(\frac{\theta - \theta'}{2})|)\,\mathrm{d}\theta\mathrm{d}\theta' \\
&= \frac{1}{\pi}\int_0^{\pi} f(2\sin(\frac{u}{2}))\,\mathrm{d}u \\
&= \frac{1}{\pi}\int_0^{2} f(x)\,\frac{1}{\sqrt{1-\frac{x^2}{4}}}\,\mathrm{d}x.
\end{aligned}$$

So we see that X has a density, $\mathbb{P}_X(\mathrm{d}x) = p(x)\mathrm{d}x$, with

$$p(x) = \frac{1}{\pi}\,\frac{1}{\sqrt{1-\frac{x^2}{4}}}\,\mathbf{1}_{(0,2)}(x).$$

In particular, the desired probability is

$$\mathbb{P}(X \geq \sqrt{3}) = \int_{\sqrt{3}}^{2} p(x)\,\mathrm{d}x = \frac{1}{3}.$$

(b) In that case, write (y, z) for the coordinates of the center of the chord. We take

$$\Omega = \{\omega = (y, z) \in \mathbb{R}^2 : y^2 + z^2 < 1\},\ \mathcal{A} = \mathcal{B}(\Omega),\ \mathbb{P}(d\omega) = \frac{1}{\pi}\,\mathbf{1}_\Omega(y, z)\,\mathrm{d}y\,\mathrm{d}z.$$

The length of the chord is

$$X(\omega) = 2\sqrt{1 - y^2 - z^2}$$

and to compute its law, we write

$$\begin{aligned}\mathbb{E}[f(X)] &= \frac{1}{\pi}\int_{\mathbb{R}^2} f(2\sqrt{1-y^2-z^2})\,\mathbf{1}_{\{y^2+z^2<1\}}\,\mathrm{d}y\mathrm{d}z\\ &= 2\int_0^1 f(2\sqrt{1-r^2})\,r\,\mathrm{d}r\\ &= \frac{1}{2}\int_0^2 f(x)\,x\,\mathrm{d}x.\end{aligned}$$

Hence, X has again a density, $\mathbb{P}_X(\mathrm{d}x) = q(x)\mathrm{d}x$, with

$$q(x) = \frac{1}{2}\,\mathbf{1}_{[0,2]}(x)\,x\,\mathrm{d}x.$$

We note that this density is very different from the density $p(x)$ obtained in case (a). In particular,

$$\mathbb{P}(X \geq \sqrt{3}) = \int_{\sqrt{3}}^2 q(x)\,\mathrm{d}x = \frac{1}{4}.$$

8.1.5 Classical Laws

In this section, we list a few important probability laws.

Discrete Distributions

(a) *The uniform distribution.* If E is a finite set, with $\mathrm{card}(E) = n$, a random variable X with values in E is said to be uniformly distributed over E if

$$\mathbb{P}(X = x) = \frac{1}{n}, \quad \forall x \in E.$$

(b) *The Bernoulli distribution with parameter $p \in [0, 1]$.* This is the law of a random variable with values in $\{0, 1\}$, such that

$$\mathbb{P}(X = 1) = p\,,\ \mathbb{P}(X = 0) = 1 - p.$$

We interpret X as the result of a coin-tossing experiment, where we get heads with probability p and tails with probability $1 - p$.

(c) *The binomial* $\mathcal{B}(n, p)$ distribution ($n \in \mathbb{N}$, $p \in [0, 1]$). This is the law of a random variable X taking values in $\{0, 1, \ldots, n\}$ and such that

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

We interpret X as the number of heads obtained in n trials of coin-tossing (with probability p of getting heads as above).

(d) *Geometric distribution of parameter* $p \in (0, 1)$. This is the law of a random variable X with values in $\mathbb{Z}_+$, such that

$$\mathbb{P}(X = k) = (1-p)\, p^k, \quad \forall k \in \mathbb{Z}_+.$$

In terms of coin-tossing with probability p of having heads, X corresponds to the number of heads before getting the first tails.

(e) *Poisson distribution with parameter* $\lambda > 0$. This is the law of a random variable X with values in $\mathbb{Z}_+$, such that

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \forall k \in \mathbb{Z}_+.$$

One easily computes $\mathbb{E}[X] = \lambda$. The Poisson distribution is very important both for practical applications and from the theoretical point of view. In practice, the Poisson distribution is used to model the number of "rare events" that occur during a long period. A precise mathematical statement is the binomial approximation of the Poisson distribution. If, for every $n \geq 1$, X_n has a binomial $\mathcal{B}(n, p_n)$ distribution, and if $np_n \longrightarrow \lambda$ when $n \to \infty$, then, for every $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

by a straightforward calculation. The interpretation is that, if every given day there is a small probability $p_n \approx \lambda/n$ that an earthquake occur, then the number of earthquakes that will occur during a long period of n days will be approximately Poisson distributed.

Continuous Distributions In the next five examples, X is a real random variable having a density p.

(a) *Uniform distribution on* $[a, b]$ ($a < b$):

$$p(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x).$$

It of course makes no difference to replace $[a, b]$ by (a, b).

(b) *Exponential distribution of parameter* $\lambda > 0$:

$$p(x) = \lambda\, e^{-\lambda x}\, \mathbf{1}_{\mathbb{R}_+}(x).$$

Note that $\mathbb{P}(X \geq a) = e^{-\lambda a}$ for every $a \geq 0$. Exponential distributions have the following important characteristic property. For every $a, b \geq 0$,

$$\mathbb{P}(X \geq a + b) = \mathbb{P}(X \geq a)\, \mathbb{P}(X \geq b), \tag{8.1}$$

which one can interpret by saying that the probability of having $X - a \geq b$ knowing that $X \geq a$ coincides with the probability of having $X \geq b$ (this interpretation can be stated more precisely in terms of conditional probabilities that will be discussed in Chapter 11 below).

Formula (8.1) thus corresponds to the property of "lack of memory" of the exponential distribution, which explains that this distribution is used to model the lifetime of a machine with no aging: knowing that the machine has been in service for 10 years, the no aging property means that the remaining time before it breaks down has the same probability distribution as if the machine was brand-new.

It is a simple exercise to check that a random variable X with values in $\mathbb{R}_+$ and such that (8.1) holds must either be equal to 0 a.s. or have an exponential distribution.

(c) *Gamma* $\Gamma(a, \lambda)$ *distribution* $(a > 0, \lambda > 0)$:

$$p(x) = \frac{\lambda^a}{\Gamma(a)}\, x^{a-1}\, e^{-\lambda x}\, \mathbf{1}_{\mathbb{R}_+}(x),$$

where $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \mathrm{d}t$ is the Gamma function (Section 7.2). These laws generalize the exponential distribution, which corresponds to $a = 1$.

(d) *Cauchy distribution with parameter* $a > 0$:

$$p(x) = \frac{1}{\pi}\, \frac{a}{a^2 + x^2}.$$

Note that $\mathbb{E}[|X|] = +\infty$ in this case.

(e) *Gaussian, or normal,* $\mathcal{N}(m, \sigma^2)$ *distribution* $(m \in \mathbb{R}, \sigma > 0)$:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \exp\Big(-\frac{(x-m)^2}{2\sigma^2}\Big).$$

The fact that p is a probability density function follows from the formula $\int_{\mathbb{R}} e^{-x^2} \mathrm{d}x = \sqrt{\pi}$ obtained in (7.3). Together with the Poisson distribution, the Gaussian distribution is the most important probability law. Its density has the famous bell shape. One easily verifies that the parameters m and σ correspond to $m = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{E}[(X - m)^2]$. If X has the $\mathcal{N}(m, \sigma^2)$ distribution,

$X - m$ is distributed according to $\mathcal{N}(0, \sigma^2)$. The Gaussian distribution will play a major role in Chapter 10.

It will be convenient to make the convention that a real random variable Y which is a.s. equal to the constant m follows the $\mathcal{N}(m, 0)$ distribution. Then, for every $a, b \in \mathbb{R}$, if X is distributed according to $\mathcal{N}(m, \sigma^2)$, $aX + b$ is distributed according to $\mathcal{N}(am + b, a^2\sigma^2)$.

8.1.6 Distribution Function of a Real Random Variable

Let X be a real random variable. The *distribution function* of X is the function $F_X : \mathbb{R} \longrightarrow [0, 1]$ defined by

$$F_X(t) := \mathbb{P}(X \le t) = \mathbb{P}_X((-\infty, t]) \ , \quad \forall t \in \mathbb{R}.$$

It follows from Corollary 1.19 that F_X characterizes the law $\mathbb{P}_X$.

The function F_X is increasing, right-continuous and has limit 0 at $-\infty$ and 1 at $+\infty$. Conversely, if $F : \mathbb{R} \longrightarrow [0, 1]$ is a function having these properties, Theorem 3.12 shows that there exists a (unique) probability measure μ on $\mathbb{R}$ such that $\mu((-\infty, t]) = F(t)$ for every $t \in \mathbb{R}$. This shows that F can be interpreted as the distribution function of a real random variable.

Let $F_X(a-)$ denote the left limit of F_X at $a \in \mathbb{R}$. It is easy to verify that

$$\begin{aligned} \mathbb{P}(a \le X \le b) &= F_X(b) - F_X(a-) \qquad \text{if } a \le b, \\ \mathbb{P}(a < X < b) &= F_X(b-) - F_X(a) \qquad \text{if } a < b. \end{aligned}$$

In particular, $\mathbb{P}(X = a) = F_X(a) - F_X(a-)$ and thus jump times of F_X are exactly the atoms of $\mathbb{P}_X$.

The following lemma uses distribution functions to construct a real random variable with a prescribed law from a random variable uniformly distributed over $(0, 1)$.

Lemma 8.7 *Let μ be a probability measure on $\mathbb{R}$. For every $x \in \mathbb{R}$ set $G_\mu(x) = \mu((-\infty, x])$, and for every $y \in (0, 1)$, set*

$$G_\mu^{-1}(y) = \inf\{x \in \mathbb{R} : G_\mu(x) \ge y\}.$$

Let Y be a random variable uniformly distributed on $(0, 1)$, and $X = G_\mu^{-1}(Y)$. Then $\mathbb{P}_X = \mu$.

Proof Note that the function G_μ is right-continuous. Then it follows from the definition of G_μ^{-1} that, for every $x \in \mathbb{R}$ and $y \in (0, 1)$, the property $G_\mu^{-1}(y) \le x$ holds if and only if $G_\mu(x) \ge y$. Hence, for every $a \in \mathbb{R}$,

$$\mathbb{P}(X \le a) = \mathbb{P}(G_\mu^{-1}(Y) \le a) = \mathbb{P}(Y \le G_\mu(a)) = G_\mu(a).$$

Therefore we have $\mathbb{P}_X((-\infty, a]) = \mu((-\infty, a])$ for every $a \in \mathbb{R}$ and this implies $\mathbb{P}_X = \mu$. □

8.1.7 The σ-Field Generated by a Random Variable

We start with an important definition.

Definition 8.8 Let X be a random variable with values in a measurable space $(E, \mathcal{E})$. The σ-field generated by X, which will be denoted by $\sigma(X)$, is the smallest σ-field on Ω for which the mapping $\omega \mapsto X(\omega)$ is measurable. This σ-field can be written as

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{E}\} \subset \mathcal{A}.$$

The formula for $\sigma(X)$ is easy, since, on one hand the σ-field generated by X must contain all events of the form $X^{-1}(B)$, $B \in \mathcal{E}$, and, on the other hand, the collection of all such events forms a σ-field.

The definition of $\sigma(X)$ can be extended to an arbitrary collection $(X_i)_{i\in I}$ of random variables, where X_i takes values in a measurable space $(E_i, \mathcal{E}_i)$. In that case,

$$\sigma\big((X_i)_{i\in I}\big) = \sigma(\{X_i^{-1}(B_i) : B_i \in \mathcal{E}_i, i \in I\}),$$

with the notation of Definition 1.2.

The next proposition shows that a real random variable is $\sigma(X)$-measurable if and only if it is a (measurable) function of X. This will be very useful when we study conditioning in Chapter 11.

Proposition 8.9 *Let X be a random variable with values in $(E, \mathcal{E})$, and let Y be a real random variable. The following are equivalent:*

(i) *Y is $\sigma(X)$-measurable.*
(ii) *There exists a measurable function f from $(E, \mathcal{E})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $Y = f(X)$.*

Proof The fact that (ii)⇒(i) is immediate since X is $\sigma(X)$-measurable (by definition) and a composition of measurable functions is measurable.

In the reverse direction, suppose that Y is $\sigma(X)$-measurable. Consider first the case where Y is a simple function, so that we can write

$$Y = \sum_{i=1}^{n} \lambda_i \, \mathbf{1}_{A_i}$$

where the λ_i's are distinct reals, and $A_i = Y^{-1}(\{\lambda_i\}) \in \sigma(X)$ for every $i \in \{1, \ldots, n\}$. Then, for every $i \in \{1, \ldots, n\}$, we can find $B_i \in \mathcal{E}$ such that $A_i = X^{-1}(B_i)$, and we have

$$Y = \sum_{i=1}^{n} \lambda_i \, \mathbf{1}_{A_i} = \sum_{i=1}^{n} \lambda_i \, \mathbf{1}_{B_i} \circ X = f \circ X,$$

where $f = \sum_{i=1}^{n} \lambda_i \, \mathbf{1}_{B_i}$ is $\mathcal{E}$-measurable.

In the general case, we know that Y is the pointwise limit of a sequence $(Y_n)_{n\in\mathbb{N}}$ of simple random variables which are $\sigma(X)$-measurable (Proposition 2.5). For every $n \in \mathbb{N}$, we can write $Y_n = f_n(X)$, where the function $f_n : E \longrightarrow \mathbb{R}$ is measurable. We then set, for every $x \in E$,

$$f(x) = \begin{cases} \lim_{n\to\infty} f_n(x) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases}$$

The function f is measurable by Lemma 1.15. Furthermore, for every $\omega \in \Omega$, $X(\omega)$ belongs to the set of all $x \in E$ such that $\lim f_n(x)$ exists (because we know that $\lim f_n(X(\omega)) = \lim Y_n(\omega) = Y(\omega)$), and moreover

$$f(X(\omega)) = \lim_{n\to\infty} f_n(X(\omega)) = Y(\omega)$$

which gives the desired formula $Y = f(X)$. □

8.2 Moments of Random Variables

8.2.1 Moments and Variance

Let X be a real random variable, and let $p \in \mathbb{N}$. The *p-th moment* of X is the quantity $\mathbb{E}[X^p]$, which is defined only if $\mathbb{E}[|X|^p] < \infty$ (or if $X \geq 0$). The quantity $\mathbb{E}[|X|^p]$ is sometimes called the p-th absolute moment. Note that the first moment of X is just the expected value of X, also called the mean of X. We say that X is *centered* if ($\mathbb{E}[|X|] < \infty$ and) $\mathbb{E}[X] = 0$ (similarly, a random vector is centered if its components are centered).

Since the expected value is a special case of the integral with respect to a positive measure, the results we saw in this more general setting apply and are of constant use. In particular, if X is a random variable with values in $[0, \infty]$, we have by Proposition 2.7,

- $\mathbb{E}[X] < \infty \;\Rightarrow\; X < \infty$ a.s.
- $\mathbb{E}[X] = 0 \;\Rightarrow\; X = 0$ a.s.

It is also worth restating the limit theorems of Chapter 2 in the language of probability theory.

- *Monotone convergence theorem.* If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables with values in $[0, \infty]$,

$$X_n \uparrow X \text{ a.s. } \Rightarrow\; \mathbb{E}[X_n] \uparrow \mathbb{E}[X].$$

- *Fatou's lemma.* If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables with values in $[0, \infty]$,

$$\mathbb{E}[\liminf X_n] \leq \liminf \mathbb{E}[X_n].$$

- *Dominated convergence theorem.* If $(X_n)_{n\in\mathbb{N}}$ is a sequence of real random variables,

$$|X_n| \leq Z,\; \mathbb{E}[Z] < \infty,\; X_n \longrightarrow X \text{ a.s. } \Rightarrow\; \mathbb{E}[|X_n - X|] \longrightarrow 0,\; \mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

Note that we stated a (straightforward) extension of the monotone convergence theorem by allowing almost sure convergence instead of the pointwise convergence in Theorem 2.4.

The spaces $L^p(\Omega, \mathcal{A}, \mathbb{P})$ are defined for every $p \in [1, \infty]$ as in Chapter 4. The Hölder inequality (Theorem 4.1) states that, for every real random variables X and Y, if $p, q \in (1, \infty)$ are conjugate exponents ($\frac{1}{p} + \frac{1}{q} = 1$),

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{1/p}\, \mathbb{E}[|Y|^q]^{1/q}.$$

Taking $Y = 1$, we have $\|X\|_1 \leq \|X\|_p$, which is immediately generalized to $\|X\|_r \leq \|X\|_p$ if $r \leq p$. Clearly, we have also $\|X\|_p \leq \|X\|_\infty$ for every $p \in [1, \infty)$. Consequently, $L^p(\Omega, \mathcal{A}, \mathbb{P}) \subset L^r(\Omega, \mathcal{A}, \mathbb{P})$ for any $1 \leq r \leq p \leq \infty$.

The scalar product on the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$ is given by $\langle X, Y\rangle_{L^2} = \mathbb{E}[XY]$. The Cauchy-Schwarz inequality gives

$$\mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2}\, \mathbb{E}[Y^2]^{1/2}$$

and the special case

$$\mathbb{E}[|X|]^2 \le \mathbb{E}[X^2] \, ,$$

is often useful.

Finally, Jensen's inequality (Theorem 4.3) states that, if $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, and if $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ is convex, we have

$$\mathbb{E}[f(X)] \ge f(\mathbb{E}[X]).$$

We now come to more "probabilistic" notions.

Definition 8.10 Let $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. The variance of X is

$$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \ge 0$$

and the standard deviation of X is

$$\sigma_X = \sqrt{\mathrm{var}(X)}.$$

Informally, $\mathrm{var}(X)$ measures how much X is dispersed around its mean $\mathbb{E}[X]$. We note that $\mathrm{var}(X) = 0$ if and only if X is a.s. constant (equal to $\mathbb{E}[X]$).

Proposition 8.11 *Let* $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. *Then* $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, *and, for every* $a \in \mathbb{R}$,

$$\mathbb{E}[(X - a)^2] = \mathrm{var}(X) + (\mathbb{E}[X] - a)^2.$$

Consequently,

$$\mathrm{var}(X) = \inf_{a \in \mathbb{R}} \mathbb{E}[(X - a)^2].$$

Proof By expanding $(X - a)^2$ we get

$$\mathbb{E}[(X - a)^2] = \mathbb{E}[X^2] - 2a\,\mathbb{E}[X] + a^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 + (\mathbb{E}[X] - a)^2.$$

We get the equality $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ by taking $a = \mathbb{E}[X]$, and the other two assertions immediately follow. □

Let us now state two famous (and useful) inequalities.

Markov's Inequality If X is a nonnegative random variable and $a > 0$,

$$\mathbb{P}(X \ge a) \le \frac{1}{a}\,\mathbb{E}[X].$$

This is a special case of Proposition 2.7 (1).

Bienaymé-Chebyshev Inequality If $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ and $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{1}{a^2} \operatorname{var}(X).$$

This follows from Markov's inequality applied to $(X - \mathbb{E}[X])^2$.

Definition 8.12 Let $X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. The covariance of X and Y is

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[X(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

If $Z = (Z_1, \ldots, Z_d)$ is a random vector with values in $\mathbb{R}^d$ whose components all belong to $L^2(\Omega, \mathcal{A}, \mathbb{P})$ (equivalently, $\mathbb{E}[|Z|^2] < \infty$), the covariance matrix of Z is

$$K_Z = \Big(\operatorname{cov}(Z_i, Z_j)\Big)_{1 \leq i \leq d, 1 \leq j \leq d}.$$

In a loose sense, the covariance of X and Y measures the correlation existing between X and Y. Note that $\operatorname{cov}(X, X) = \operatorname{var}(X)$ and that the Cauchy-Schwarz inequality gives

$$|\operatorname{cov}(X, Y)| \leq \sqrt{\operatorname{var}(X)}\,\sqrt{\operatorname{var}(Y)}.$$

The mapping $(X, Y) \longrightarrow \operatorname{cov}(X, Y)$ is symmetric and bilinear on $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

For a random vector $Z = (Z_1, \ldots, Z_d)$, the matrix K_Z is symmetric and nonnegative definite: for every $\lambda_1, \ldots, \lambda_d \in \mathbb{R}^d$,

$$\sum_{i,j=1}^{d} \lambda_i \lambda_j \, K_Z(i, j) = \operatorname{var}\Big(\sum_{i=1}^{d} \lambda_i Z_i\Big) \geq 0.$$

Set $\widetilde{Z} = Z - \mathbb{E}[Z]$. If we view $\widetilde{Z}$ as a column vector, we can write $K_Z = \mathbb{E}[\widetilde{Z}\,{}^t\widetilde{Z}]$, where, for any matrix M, ${}^tM$ denotes the transpose of M. Consequently, if A is an $n \times d$ real matrix, and $Z' = AZ$, we have

$$K_{Z'} = \mathbb{E}[\widetilde{Z}'\,{}^t\widetilde{Z}'] = \mathbb{E}[A\widetilde{Z}\,{}^t\widetilde{Z}\,{}^tA] = A\,\mathbb{E}[\widetilde{Z}\,{}^t\widetilde{Z}]\,{}^tA = AK_Z{}^tA. \tag{8.2}$$

As a special case, if $\xi \in \mathbb{R}^d$, the scalar product $\xi \cdot Z$ can be written as ${}^t\xi Z$ in matrix form, and we get

$$\operatorname{var}(\xi \cdot Z) = {}^t\xi K_Z \xi. \tag{8.3}$$

8.2.2 Linear Regression

Let $X, Y_1, \ldots, Y_n$ be real random variables in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. We write $Y = (Y_1, \ldots, Y_n)$. We are looking for the "best" approximation of X as an affine function of $Y_1, \ldots, Y_n$. More precisely, we are trying to minimize

$$\mathbb{E}[(X - (\beta_0 + \beta_1 Y_1 + \cdots + \beta_n Y_n))^2]$$

over all possible choices of the $(n+1)$-tuple of reals $(\beta_0, \ldots, \beta_n)$.

Proposition 8.13 *We have*

$$\inf_{\beta_0,\ldots,\beta_n \in \mathbb{R}} \mathbb{E}[(X - (\beta_0 + \beta_1 Y_1 + \cdots + \beta_n Y_n))^2] = \mathbb{E}[(X - Z)^2],$$

where

$$Z = \mathbb{E}[X] + \sum_{j=1}^{n} \alpha_j (Y_j - \mathbb{E}[Y_j]), \tag{8.4}$$

and the coefficients α_j, $1 \le j \le n$, are any solution of the linear system

$$\sum_{j=1}^{n} \alpha_j \operatorname{cov}(Y_j, Y_k) = \operatorname{cov}(X, Y_k) \ , \quad 1 \le k \le n.$$

In particular, if K_Y is non-degenerate, we have $\alpha = \operatorname{cov}(X, Y)\, K_Y^{-1}$ in matrix notation (here $\operatorname{cov}(X, Y)$ stands for the vector $(\operatorname{cov}(X, Y_j))_{1 \le j \le n}$).

Proof Let H be the (closed) linear subspace of $L^2(\Omega, \mathcal{A}, \mathbb{P})$ spanned by $1, Y_1, \ldots, Y_n$, or equivalently by $1, Y_1 - \mathbb{E}[Y_1], \ldots, Y_n - \mathbb{E}[Y_n]$. Then, by the classical theory of Hilbert spaces (Theorem A.3 in the Appendix), we know that the random variable that minimizes $\|X - U\|_2$ over all $U \in H$ is the orthogonal projection of X on H. A random variable $Z \in H$ can be written in the form

$$Z = \alpha_0 + \sum_{j=1}^{n} \alpha_j (Y_j - \mathbb{E}[Y_j]),$$

where $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$. Then Z is equal to the orthogonal projection of X on H if and only if $X - Z$ is orthogonal to H, which holds if and only if we have both

$$\mathbb{E}[(X - Z) \times 1] = 0, \tag{8.5}$$

and

$$\mathbb{E}[(X - Z) \times (Y_k - \mathbb{E}[Y_k])] = 0, \quad \forall k \in \{1, \dots, n\}. \tag{8.6}$$

Property (8.5) holds if and only if $\alpha_0 = \mathbb{E}[X]$, and the equalities in (8.6) are satisfied if and only if the coefficients α_j solve the linear system stated in the proposition. This completes the proof. □

Example If $n = 1$ and we assume that Y is not a.s. constant (so that $\mathrm{var}(Y) > 0$), we find that the best approximation (in the L^2 sense) of X by an affine function of Y is

$$Z = \mathbb{E}[X] + \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(Y)}(Y - \mathbb{E}[Y]).$$

This affine function of Y is sometimes called the regression line of X in the variable Y.

8.2.3 Characteristic Functions

We start with a basic definition. We use the notation $x \cdot y$ for the standard Euclidean scalar product in $\mathbb{R}^d$.

Definition 8.14 Let X be a random variable with values in $\mathbb{R}^d$. The characteristic function of X is the function $\Phi_X : \mathbb{R}^d \longrightarrow \mathbb{C}$ defined by

$$\Phi_X(\xi) := \mathbb{E}[\exp(\mathrm{i}\xi \cdot X)]\,, \qquad \xi \in \mathbb{R}^d.$$

Proposition 8.5 allows us to write

$$\Phi_X(\xi) = \int_{\mathbb{R}^d} e^{\mathrm{i}\xi \cdot x}\, \mathbb{P}_X(\mathrm{d}x)$$

and we can also view Φ_X as the Fourier transform of the law $\mathbb{P}_X$ of X. For this reason, we sometimes write $\Phi_X(\xi) = \widehat{\mathbb{P}_X}(\xi)$. As we already noticed in Chapter 2 (in the case $d = 1$), the dominated convergence theorem ensures that the function Φ_X is (bounded and) continuous on $\mathbb{R}^d$.

Our first goal is to show that the characteristic function determines the law $\mathbb{P}_X$ of X (as the name suggests !). This is equivalent to showing that the Fourier transform acting on probability measures on $\mathbb{R}^d$ is injective. We start with a useful calculation in a special case.

Lemma 8.15 *Let X be a real random variable following the Gaussian $\mathcal{N}(m, \sigma^2)$ distribution. Then,*

$$\Phi_X(\xi) = \exp(\mathrm{i}m\xi - \frac{\sigma^2\xi^2}{2}) , \qquad \xi \in \mathbb{R}.$$

Proof We may assume that $\sigma > 0$ and, replacing X by $X - m$, that $m = 0$. We have then

$$\Phi_X(\xi) = \int_{\mathbb{R}} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/(2\sigma^2)} e^{\mathrm{i}\xi x} \,\mathrm{d}x.$$

Clearly, it is enough to consider the case $\sigma = 1$. A parity argument shows that the imaginary part of $\Phi_X(\xi)$ is zero. It remains to compute

$$f(\xi) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cos(\xi x) \,\mathrm{d}x.$$

By differentiating under the integral sign (cf. Theorem 2.13), we get

$$f'(\xi) = -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} x\, e^{-x^2/2} \sin(\xi x) \,\mathrm{d}x$$

(the justification is easy since $|x \sin(\xi x)\, e^{-x^2/2}| \leq |x|\, e^{-x^2/2}$, which is integrable on $\mathbb{R}$). An integration by parts shows that

$$f'(\xi) = -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \xi \cos(\xi x) \,\mathrm{d}x = -\xi f(\xi).$$

The function f thus solves the differential equation $f'(\xi) = -\xi f(\xi)$, with initial condition $f(0) = 1$. It follows that $f(\xi) = \exp(-\xi^2/2)$. □

Theorem 8.16 *The characteristic function of a random variable X with values in $\mathbb{R}^d$ determines the law of this random variable. In other words, the Fourier transform acting on probability measures on $\mathbb{R}^d$ is injective.*

Proof Let us consider the case $d = 1$. For every $\sigma > 0$, let g_σ be the density of the Gaussian $\mathcal{N}(0, \sigma^2)$ distribution:

$$g_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}) , \qquad x \in \mathbb{R}.$$

Let μ be a probability measure on $\mathbb{R}$, and set

$$f_\sigma(x) = \int_{\mathbb{R}} g_\sigma(x-y)\,\mu(\mathrm{d}y) \stackrel{\text{(def)}}{=} g_\sigma * \mu(x),$$
$$\mu_\sigma(\mathrm{d}x) = f_\sigma(x)\,\mathrm{d}x.$$

Let $C_b(\mathbb{R})$ denote the space of all bounded continuous functions $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$, and recall that a probability measure ν on $\mathbb{R}$ is characterized by the values of $\int \varphi(x)\,\nu(\mathrm{d}x)$ for $\varphi \in C_b(\mathbb{R})$ (see the discussion after Proposition 8.5). The result of the theorem (when $d = 1$) will follow if we are able to prove that:

(1) μ_σ is determined by $\widehat{\mu}$.
(2) For every $\varphi \in C_b(\mathbb{R})$, $\int \varphi(x)\mu_\sigma(\mathrm{d}x) \longrightarrow \int \varphi(x)\mu(\mathrm{d}x)$ as $\sigma \to 0$.

In order to establish (1), we use Lemma 8.15 to write, for every $x \in \mathbb{R}$,

$$\sigma\sqrt{2\pi}\,g_\sigma(x) = \exp(-\frac{x^2}{2\sigma^2}) = \int_{\mathbb{R}} e^{\mathrm{i}\xi x}\,g_{1/\sigma}(\xi)\,\mathrm{d}\xi.$$

It follows that

$$\begin{aligned} f_\sigma(x) = \int_{\mathbb{R}} g_\sigma(x-y)\,\mu(\mathrm{d}y) &= (\sigma\sqrt{2\pi})^{-1}\int_{\mathbb{R}}\Big(\int_{\mathbb{R}} e^{\mathrm{i}\xi(x-y)}\,g_{1/\sigma}(\xi)\,\mathrm{d}\xi\Big)\mu(\mathrm{d}y) \\ &= (\sigma\sqrt{2\pi})^{-1}\int_{\mathbb{R}} e^{\mathrm{i}\xi x}\,g_{1/\sigma}(\xi)\Big(\int_{\mathbb{R}} e^{-\mathrm{i}\xi y}\mu(\mathrm{d}y)\Big)\mathrm{d}\xi \\ &= (\sigma\sqrt{2\pi})^{-1}\int_{\mathbb{R}} e^{\mathrm{i}\xi x}\,g_{1/\sigma}(\xi)\,\widehat{\mu}(-\xi)\mathrm{d}\xi. \end{aligned}$$

In the penultimate equality, we used the Fubini-Lebesgue theorem. The application of this theorem is easily justified since μ is a probability measure and the function $g_{1/\sigma}$ is integrable with respect to Lebesgue measure. The formula of the last display shows that f_σ (and hence μ_σ) is determined by $\widehat{\mu}$.

As for (2), we start by writing, for every $\varphi \in C_b(\mathbb{R})$,

$$\begin{aligned} \int \varphi(x)\mu_\sigma(\mathrm{d}x) = \int \varphi(x)\Big(\int g_\sigma(y-x)\mu(\mathrm{d}y)\Big)\mathrm{d}x &= \int\Big(\int g_\sigma(y-x)\varphi(x)\mathrm{d}x\Big)\mu(\mathrm{d}y) \\ &= \int g_\sigma * \varphi(y)\mu(\mathrm{d}y), \end{aligned}$$

where we have again applied the Fubini-Lebesgue theorem, with the same justification as above. Then, we note that the properties

$$\int g_\sigma(x)\,\mathrm{d}x = 1\,,$$

$$\lim_{\sigma\to 0}\int_{\{|x|>\varepsilon\}} g_\sigma(x)\,\mathrm{d}x = 0\,, \qquad \forall \varepsilon > 0,$$

imply that, for every $y \in \mathbb{R}$,

$$\lim_{\sigma\to 0} g_\sigma * \varphi(y) = \varphi(y)$$

thanks to Proposition 5.8 (i). By dominated convergence, using the fact that $|g_\sigma * \varphi| \le \sup\{|\varphi(x)| : x \in \mathbb{R}\}$, we get

$$\lim_{\sigma\to 0}\int \varphi(x)\mu_\sigma(\mathrm{d}x) = \lim_{\sigma\to 0}\int g_\sigma * \varphi(y)\mu(\mathrm{d}y) = \int \varphi(x)\mu(\mathrm{d}x),$$

which completes the proof of (2) and of the case $d = 1$ of the theorem.

The proof in the general case is similar. We now use the auxiliary functions

$$g_\sigma^{(d)}(x_1,\ldots,x_d) = \prod_{j=1}^{d} g_\sigma(x_j)$$

noting that, for $\xi = (\xi_1,\ldots,\xi_d) \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} g_\sigma^{(d)}(x)\,e^{\mathrm{i}\xi\cdot x}\,\mathrm{d}x = \prod_{j=1}^{d}\int_{\mathbb{R}} g_\sigma(x_j)\,e^{\mathrm{i}\xi_j x_j}\,\mathrm{d}x_j = (2\pi/\sigma^2)^{d/2} g_{1/\sigma}^{(d)}(\xi).$$

The arguments of the case $d = 1$ then apply with obvious changes. □

Proposition 8.17 *Let $X = (X_1,\ldots,X_d)$ be a random variable with values in $\mathbb{R}^d$. Assume that $\mathbb{E}[(X_j)^2] < \infty$ for every $j \in \{1,\ldots,d\}$. Then, Φ_X is twice continuously differentiable and*

$$\Phi_X(\xi) = 1 + \mathrm{i}\sum_{j=1}^{d}\xi_j\mathbb{E}[X_j] - \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\xi_j\xi_k\mathbb{E}[X_jX_k] + o(|\xi|^2)$$

when $\xi = (\xi_1,\ldots,\xi_d)$ tends to 0.

Proof By differentiating under the integral sign (Theorem 2.13), we get, for every $1 \leq j \leq d$,

$$\frac{\partial \Phi_X}{\partial \xi_j}(\xi) = \mathrm{i}\,\mathbb{E}[X_j e^{\mathrm{i}\xi \cdot X}].$$

The justification is easy since $|\mathrm{i}X_j e^{\mathrm{i}\xi \cdot X}| = |X_j|$ and $X_j \in L^2(\Omega, \mathcal{A}, \mathbb{P}) \subset L^1(\Omega, \mathcal{A}, \mathbb{P})$. Similarly, for every $j, k \in \{1, \ldots, d\}$, the fact that $X_j X_k \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ allows us to differentiate once more and to get

$$\frac{\partial^2 \Phi_X}{\partial \xi_j \partial \xi_k}(\xi) = -\,\mathbb{E}[X_j X_k e^{\mathrm{i}\xi \cdot X}].$$

Moreover Theorem 2.12 (or more directly the dominated convergence theorem) ensures that the quantity $\mathbb{E}[X_j X_k e^{\mathrm{i}\xi \cdot X}]$ is a continuous function of ξ. We have thus proved that Φ_X is twice continuously differentiable.

Finally, the last assertion of the proposition is just Taylor's expansion of Φ_X at order two at the origin. □

Remark If we assume that $\mathbb{E}[|X|^p] < \infty$, for some integer $p \geq 1$, the same argument shows that Φ_X is p times continuously differentiable. The case $p = 2$ will however be most useful in what follows.

8.2.4 Laplace Transform and Generating Functions

For a random variable with values in $\mathbb{R}_+$, one often uses the Laplace transform instead of the characteristic function.

Definition 8.18 Let X be a nonnegative random variable. The Laplace transform of (the law of) X is the function L_X defined on $\mathbb{R}_+$ by

$$L_X(\lambda) := \mathbb{E}[e^{-\lambda X}] = \int_{\mathbb{R}_+} e^{-\lambda x}\, \mathbb{P}_X(\mathrm{d}x).$$

The dominated convergence theorem (or Theorem 2.12) shows that L_X is continuous on $[0, \infty)$. An application of the theorem of differentiation under the integral sign shows that L_X is infinitely differentiable on $(0, \infty)$. The function L_X is also convex (as a consequence of the fact that the functions $\lambda \mapsto e^{-\lambda x}$ are convex, for every $x \geq 0$) and thus has a (possibly infinite) right derivative at 0, which is given by

$$L'_X(0) = -\lim_{\lambda \downarrow 0} \mathbb{E}\Big[\frac{1 - e^{-\lambda X}}{\lambda}\Big] = -\mathbb{E}[X]$$

where the second equality follows from the monotone convergence theorem.

Theorem 8.19 *If X is a nonnegative random variable, the Laplace transform of X determines the law $\mathbb{P}_X$ of X.*

Remark As in the case of characteristic functions, this statement is equivalent to the injectivity of the Laplace transform acting on probability measures on $\mathbb{R}_+$.

Proof Suppose X and X' are two nonnegative random variables, and that $L_X = L_{X'}$. For every $\lambda \geq 0$ and every $x \geq 0$, set $\psi_\lambda(x) = e^{-\lambda x}$. We extend the functions ψ_λ to $[0, \infty]$ by continuity, setting $\psi_\lambda(\infty) = 0$ if $\lambda > 0$, and $\psi_0(\infty) = 1$.

Let H be the vector space spanned by the functions ψ_λ, $\lambda \geq 0$. Then H is a subspace of the space $C([0, \infty], \mathbb{R})$ of all real continuous functions on $[0, \infty]$, which is equipped with the topology of uniform convergence. Moreover, the Stone-Weierstrass theorem ensures that H is dense in $C([0, \infty], \mathbb{R})$.

The equality $L_X = L_{X'}$ implies that $\mathbb{E}[\phi(X)] = \mathbb{E}[\phi(X')]$ for every $\phi \in H$. On the other hand, the mapping $\phi \mapsto \mathbb{E}[\phi(X)] = \int \phi \, \mathrm{d}\mathbb{P}_X$ is clearly continuous on $C([0, \infty], \mathbb{R})$. Since H is dense, it follows that we have $\mathbb{E}[\psi(X)] = \mathbb{E}[\psi(X')]$ for every $\psi \in C([0, \infty], \mathbb{R})$, and in particular $\mathbb{E}[\psi(X)] = \mathbb{E}[\psi(X')]$ for every continuous function $\psi : \mathbb{R}_+ \longrightarrow \mathbb{R}$ with compact support, which suffices to say that $\mathbb{P}_X = \mathbb{P}_{X'}$. □

Finally, in the case of integer-valued random variables, one often uses generating functions.

Definition 8.20 Let X be a random variable with values in $\mathbb{Z}_+$. The generating function of X is the function $g_X : [0, 1] \longrightarrow [0, 1]$ defined by

$$g_X(r) := \mathbb{E}[r^X] = \sum_{n=0}^{\infty} \mathbb{P}(X = n)\, r^n.$$

Note that $g_X(0) = \mathbb{P}(X = 0)$ and $g_X(1) = 1$. The function g_X is continuous on $[0, 1]$. Since the radius of convergence of the series in the definition is at least 1, the function g_X is infinitely differentiable on $[0, 1)$. Moreover, it is obvious that g_X determines $\mathbb{P}_X$, since the quantities $\mathbb{P}(X = n)$ appear in the Taylor expansion of g_X at the origin.

The derivative of g_X is

$$g'_X(r) = \sum_{n=1}^{\infty} n\, \mathbb{P}(X = n)\, r^{n-1}$$

for $r \in [0, 1)$. It follows that g_X has a (possibly infinite) left-derivative at $r = 1$, and

$$g'_X(1) = \sum_{n=1}^{\infty} n\,\mathbb{P}(X = n) = \mathbb{E}[X].$$

More generally, for every integer $p \geq 1$, if $g_X^{(p)}$ denotes the p-th derivative of g_X,

$$\lim_{r\uparrow 1} g_X^{(p)}(r) = \mathbb{E}[X(X-1)\cdots(X-p+1)]$$

which shows how to recover all moments of X from the knowledge of its generating function.

8.3 Exercises

Exercise 8.1 Consider a population of n individuals and $r \in \{1, \dots, n-1\}$.

(1) Give a probability space for the random experiment consisting in choosing at random a sample of r individuals in the population.
(2) Suppose that the population is composed of individuals of two types, with n_1 individuals of type 1 and n_2 individuals of type 2 (where $n_1 + n_2 = n$). Let X be the number of individuals of type 1 in the sample. Prove that the law of X is given by

$$\mathbb{P}(X = k) = \frac{\binom{n_1}{k}\binom{n_2}{r-k}}{\binom{n}{r}}$$

for every $k \in \{0, 1, \dots, r\}$ (we make the convention that $\binom{k}{j} = 0$ if $j > k$). This is the so-called *hypergeometric distribution*.
(3) Show that, when $n, n_1, n_2 \to \infty$ in such a way that n_1/n tends to $p \in (0, 1)$, and r remains fixed, the law of X becomes close to the binomial $\mathcal{B}(r, p)$ distribution. Interpret this result.

Exercise 8.2 Let $n \geq 1$ and $r \geq 1$ be integers. Suppose that we have n balls and r compartments numbered $1, 2, \dots, r$.

(1) Give a probability space for the random experiment consisting in placing the n balls at random in the r compartments (each ball is placed in one of the r compartments chosen at random). Compute the law $\mu_{r,n}$ of the number of balls placed in the first compartment.
(2) Show that, when $r, n \to \infty$ in such a way that $r/n \longrightarrow \lambda \in (0, \infty)$, the law $\mu_{r,n}$ becomes close to the Poisson distribution with parameter λ.

Exercise 8.3

(1) Let $A_1, \ldots, A_n$ be n events in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Prove that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le j_1<\ldots<j_k\le n} \mathbb{P}(A_{j_1} \cap \ldots \cap A_{j_k}).$$

This is called the *inclusion-exclusion formula*.

(2) Consider a group of n persons attending a lecture. Each person wears a hat and leaves it in a dark cloakroom before the lecture. After the lecture, the members of the group come successively to the cloakroom and each of them picks a hat at random among the remaining ones. What is the probability that at least one person of the group picks the hat he or she was wearing before the lecture ? What is the limit of this probability when $n \to \infty$? Interpret and reprove the result of the calculation in terms of the group of permutations of n elements.

Exercise 8.4 (Ballot Theorem) In an election, candidate A has obtained a votes and candidate B has obtained b votes, where $a > b$. The scrutineer proceeds to the counting of votes by reading the ballot papers one after the other in a random order. Prove that the probability that candidate A has strictly more votes than candidate B at each step of the counting process is $(a-b)/(a+b)$. (*Hint:* Represent the difference between votes for A and votes for B during the counting process by a discrete function from $\{0, 1, \ldots, a+b\}$ into $\mathbb{Z}$ that starts from 0, has jumps of size $+1$ or -1 and terminates at $a-b$. Then note that the probability of occurrence of any such function is the same, so that the problem reduces to enumerating those among these functions that stay positive on $\{1, \ldots, a+b\}$.)

Exercise 8.5 Following the description of Bertrand's paradox in Section 8.1.4, treat the third method that had been proposed by Bertrand: one first chooses the ray carrying the center of the chord, and then a point uniformly distributed on this ray to be the center of the chord. Give the probability space corresponding to this method and compute the law of the length of the chord.

Exercise 8.6 Let $X = (X_1, X_2, \ldots, X_d)$ be a random vector with values in $\mathbb{R}^d$. Assume that the law of X has a density $p_X(x_1, \ldots, x_d)$. Compute the density of the random variable $X_1 + X_2$ in terms of the function p_X.

Exercise 8.7 Let N be a Gaussian $\mathcal{N}(0, 1)$ random variable. Compute the law of $1/N^2$. (This is the so-called stable (1/2) distribution, which we shall encounter in Chapter 14).

Exercise 8.8 Let (X, Y) be a random variable with values in $\mathbb{R}^2$ whose law has a density given by $p(x, y) = \mathbf{1}_{\mathbb{R}_+^2}(x, y)\,\lambda\mu e^{-\lambda x-\mu y}$, where $\lambda, \mu > 0$. Compute the law of $U = \max(X, Y)$, of $V = \min(X, Y)$ and of the pair (U, V).

Exercise 8.9 Suppose that a light source is located at the point $(-1, 0)$ of the plane. Let θ be uniformly distributed over the interval $(-\pi/2, \pi/2)$. The light source emits a ray in the direction of the vertical coordinate axis making an angle θ with the horizontal axis. Determine the law of the point of the vertical axis that is hit by the ray.

Exercise 8.10 Let X be a real random variable, and let $F = F_X$ be its distribution function. Assume that the law of X has no atoms. What is the distribution of the random variable $Y = F(X)$? (Compare with Lemma 8.7).

Exercise 8.11 Determine the σ-field generated by X in the following two cases:

(i) $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $X(\omega) = \omega^2$.
(ii) $(\Omega, \mathcal{A}) = (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$, and

$$X(\omega_1, \omega_2) = \frac{\omega_1\omega_2}{\omega_1^2 + \omega_2^2} \text{ if } (\omega_1, \omega_2) \neq (0, 0), \text{ and } X(0, 0) = 0.$$

Exercise 8.12 Let X be a real random variable. Assume that X is integrable ($\mathbb{E}[|X|] < \infty$). Prove that

$$\lim_{x\to+\infty} \mathbb{E}[|X|\, \mathbf{1}_{\{|X|\geq x\}}] = 0.$$

Exercise 8.13

(1) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative random variables in L^2. Assume that the sequence $(X_n)_{n\in\mathbb{N}}$ is increasing and that

$$\mathbb{E}[X_n] \underset{n\to\infty}{\longrightarrow} +\infty\,, \quad \liminf_{n\to\infty} \frac{\mathrm{var}(X_n)}{\mathbb{E}[X_n]^2} = 0.$$

Prove that $X_n \longrightarrow +\infty$ a.s. as $n \to \infty$.
(2) Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events. Prove that the conditions

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty\,, \quad \liminf_{n\to\infty} \frac{\sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{P}(A_j \cap A_k)}{(\sum_{k=1}^{n} \mathbb{P}(A_k))^2} = 1$$

imply that $\displaystyle\sum_{n=1}^{\infty} \mathbf{1}_{A_n} = \infty$, a.s.

Exercise 8.14 Let $(X_1, X_2, \ldots, X_d)$ be a random vector with values in $\mathbb{R}^d$.

(1) Prove that one can uniquely define real random variables $Y_1, Y_2, \ldots, Y_d$ such that, for every $\omega \in \Omega$, $Y_1(\omega) \leq Y_2(\omega) \leq \cdots \leq Y_d(\omega)$, and, for every $\omega \in \Omega$

and $x \in \mathbb{R}$, the sets $\{i \in \{1, \ldots, d\} : X_i(\omega) = x\}$ and $\{i \in \{1, \ldots, d\} : Y_i(\omega) = x\}$ have the same cardinality. The random vector $(Y_1, \ldots, Y_d)$ is called the increasing reordering of $(X_1, \ldots, X_d)$.

(2) Suppose that $(X_1, \ldots, X_d)$ has density $p(x_1, \ldots, x_d) = \mathbf{1}_{[0,1]^d}(x_1, \ldots, x_d)$. Show that the random vector $(Y_1, \ldots, Y_d)$ has density

$$q(x_1, \ldots, x_d) = d!\,\mathbf{1}_{\{0 \le x_1 < x_2 < \cdots < x_d \le 1\}}.$$

(3) Suppose that $d = 3$ and that (X_1, X_2, X_3) has density $p(x) = \mathbf{1}_{[0,1]^3}(x)$ for $x \in \mathbb{R}^3$. Compute the law of the pair $(Y_1/Y_2, Y_2/Y_3)$.

Exercise 8.15 Let $(X_n)_{n \in \mathbb{Z}}$ be random variables in L^2 such that, for every $n, m \in \mathbb{Z}$,

$$\mathbb{E}[X_n] = a\,, \quad \operatorname{cov}(X_n, X_m) = b\,\rho^{|m-n|}\,,$$

where a, b, ρ are reals such that $b > 0$ and $|\rho| < 1$. Let F be the closed linear subspace of L^2 spanned by the variables X_n for $n \le 0$ and the constant variable 1. Show that, for every integer $m \ge 1$,

$$\inf_{Y \in F} \mathbb{E}[(X_m - Y)^2] = \mathbb{E}[(X_m - Y_m)^2],$$

where $Y_m = a + \rho^m (X_0 - a)$.

Exercise 8.16 Compute the generating function of the integer-valued random variable X in the following three cases:

(i) X is binomial $\mathcal{B}(n, p)$ where $n \in \mathbb{N}$ and $p \in [0, 1]$.
(ii) X is geometric with parameter $p \in (0, 1)$.
(iii) X is Poisson with parameter $\lambda > 0$.

Chapter 9
Independence

The concept of independence is the first important notion of probability theory that is not a simple adaptation of similar notions in measure theory. If it is easier to have an intuitive understanding of what the independence of two events or of two random variables means, the most fundamental concept is the independence of two (or several) σ-fields.

A key result relates the independence of two random variables to the fact that the law of the pair made of these two variables is equal to the product measure of the individual laws. Together with the Fubini theorem, this leads to useful reformulations of the notion of independence. We establish the famous Borel-Cantelli lemma, and as an application we derive surprising properties of the dyadic expansion of a real number chosen at random. We use Lebesgue measure and this dyadic expansion to give an elementary construction of sequences of independent (real) random variables, which suffices for our needs in the following chapters. We pay special attention to sums of independent and identically distributed random variables. In particular we give a first version of the law of large numbers, which provides a link between our axiomatic presentation of probability theory and the "historical" approach, where the probability of an event is the asymptotic frequency of occurence of this event when the random experiment is repeated a large number of times. The last section, which may be omitted at first reading, is a brief presentation of the Poisson process, which illustrates many of the notions introduced in this chapter.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_9

9.1 Independent Events

Throughout this chapter, we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. If $A, B \in \mathcal{A}$ are two events, we say that A and B are independent if

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

If $\mathbb{P}(B) > 0$, we can interpret this definition by saying that the conditional probability

$$\mathbb{P}(A \mid B) \overset{\text{(def)}}{=} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

is equal to $\mathbb{P}(A)$: knowing that B holds gives no information on whether A holds or not (and the roles of A and B can be interchanged).

Examples

(i) *Throwing two dice* : $\Omega = \{1, 2, \dots, 6\}^2$, $\mathbb{P}(\{\omega\}) = 1/36$ for every $\omega \in \Omega$. The events $A = \{6\} \times \{1, 2, \dots, 6\}$ (6 at the first trial) and $B = \{1, 2, \dots, 6\} \times \{6\}$ (6 at the second trial) are independent. In fact the probability measure $\mathbb{P}$ has been constructed so that any event concerning the first trial is independent of any event concerning the second trial.

(ii) *Throwing one die* : $\Omega = \{1, 2, \dots, 6\}$, $\mathbb{P}(\{\omega\}) = 1/6$ for every $\omega \in \Omega$. The events $A = \{1, 2\}$ and $B = \{1, 3, 5\}$ are independent.

Definition 9.1 We say that n events $A_1, \dots, A_n$ are independent if, for every nonempty subset $\{j_1, \dots, j_p\}$ of $\{1, \dots, n\}$, we have

$$\mathbb{P}(A_{j_1} \cap A_{j_2} \cap \dots \cap A_{j_p}) = \mathbb{P}(A_{j_1})\,\mathbb{P}(A_{j_2}) \dots \mathbb{P}(A_{j_p}).$$

Remark It is not enough to assume that, for every pair $\{i, j\} \subset \{1, \dots, n\}$, the events A_i and A_j are independent. To give an example, consider two trials of coin-tossing with a fair coin (probability $1/2$ of having heads or tails). Let

$$A_1 = \{\text{heads at the first trial}\}$$
$$A_2 = \{\text{heads at the second trial}\}$$
$$A_2 = \{\text{same outcome at both trials}\}.$$

Then A_i and A_j are independent if $i \neq j$, but $\mathbb{P}(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/8 = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.

Proposition 9.2 *The n events $A_1, \dots, A_n$ are independent if and only if we have*

$$\mathbb{P}(B_1 \cap \dots \cap B_n) = \mathbb{P}(B_1) \dots \mathbb{P}(B_n)$$

whenever $B_i \in \sigma(A_i) = \{\varnothing, A_i, A_i^c, \Omega\}$, for every $i \in \{1, \dots, n\}$.

Proof The condition given in the proposition is stronger than our definition of independence: just take $B_i = A_i$ if $i \in \{j_1, \dots, j_p\}$ and $B_i = \Omega$ otherwise. So we need to verify that the condition in the proposition holds if $A_1, \dots, A_n$ are independent. Clearly, we may assume that $B_i \neq \varnothing$ for every $i \in \{1, \dots, n\}$. Then, if $\{j_1, \dots, j_p\} = \{i : B_i \neq \Omega\}$, we have to check that

$$\mathbb{P}(B_{j_1} \cap B_{j_2} \cap \dots \cap B_{j_p}) = \mathbb{P}(B_{j_1})\, \mathbb{P}(B_{j_2}) \dots \mathbb{P}(B_{j_p}),$$

as soon as $B_{j_k} = A_{j_k}$ or $A_{j_k}^c$, for every $k \in \{1, \dots, p\}$. Finally, it suffices to verify that, if $C_1, C_2, \dots, C_p$ are independent, then so are $C_1^c, C_2, \dots, C_p$. This is easy since, for every subset $\{i_1, \dots, i_q\}$ de $\{2, \dots, p\}$,

$$\begin{aligned}\mathbb{P}(C_1^c \cap C_{i_1} \cap \dots \cap C_{i_q}) &= \mathbb{P}(C_{i_1} \cap \dots \cap C_{i_q}) - \mathbb{P}(C_1 \cap C_{i_1} \cap \dots \cap C_{i_q}) \\ &= \mathbb{P}(C_{i_1}) \dots \mathbb{P}(C_{i_q}) - \mathbb{P}(C_1)\mathbb{P}(C_{i_1}) \dots \mathbb{P}(C_{i_q}) \\ &= \mathbb{P}(C_1^c)\mathbb{P}(C_{i_1}) \dots \mathbb{P}(C_{i_q}).\end{aligned}$$

□

9.2 Independence for σ-Fields and Random Variables

We say that $\mathcal{B}$ is a sub-σ-field of $\mathcal{A}$ if $\mathcal{B}$ is a σ-field and $\mathcal{B} \subset \mathcal{A}$. Roughly speaking, a sub-σ-field $\mathcal{B}$ corresponds to some partial information in the probability space, namely the information given by knowing which events in $\mathcal{B}$ have occurred (for instance, if $\mathcal{B} = \sigma(X)$ where X is a random variable, the corresponding information is just the value of X). This suggests that the most general notion of independence is that of independent sub-σ-fields: informally, two sub-σ-fields $\mathcal{B}$ and $\mathcal{B}'$ will be independent if knowing which events of $\mathcal{B}$ have occurred does not give any information on the occurrence of events of $\mathcal{B}'$, and conversely. We make this more precise in the following definition.

Definition 9.3 Let $\mathcal{B}_1, \dots, \mathcal{B}_n$ be n sub-σ-fields of $\mathcal{A}$. We say that $\mathcal{B}_1, \dots, \mathcal{B}_n$ are independent if, for every $A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2, \dots, A_n \in \mathcal{B}_n$,

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1)\, \mathbb{P}(A_2) \dots \mathbb{P}(A_n).$$

Let $X_1, \dots, X_n$ be n random variables taking values in $(E_1, \mathcal{E}_1), \dots, (E_n, \mathcal{E}_n)$ respectively. We say that the random variables $X_1, \dots, X_n$ are independent if the σ-fields $\sigma(X_1), \dots, \sigma(X_n)$ are independent. This is equivalent to saying that the property

$$\mathbb{P}(\{X_1 \in F_1\} \cap \dots \cap \{X_n \in F_n\}) = \mathbb{P}(X_1 \in F_1) \dots \mathbb{P}(X_n \in F_n) \tag{9.1}$$

holds for every $F_1 \in \mathcal{E}_1, \dots, F_n \in \mathcal{E}_n$.

The fact that (9.1) is equivalent to the independence of $\sigma(X_1), \dots, \sigma(X_n)$ is clear since $\sigma(X_j) = \{X_j^{-1}(F) : F \in \mathcal{E}_j\}$, for every $j \in \{1, \dots, p\}$, as we saw in Section 8.1.7. Informally, the random variables $X_1, \dots, X_n$ are independent if the knowledge of a subset of these variables does not give information on the remaining ones.

It will also be useful to consider the independence of several (finite or infinite) collections of random variables, which is defined in the obvious manner. For instance, two collections $(X_i)_{i \in I}$ and $(Y_j)_{j \in J}$ of random variables are independent if the σ-fields $\sigma((X_i)_{i \in I})$ and $\sigma((Y_j)_{j \in J})$ are independent. Note that this is stronger than just saying that X_i and Y_j are independent for every $i \in I$ and $j \in J$.

Remarks

(i) If $\mathcal{B}_1, \dots, \mathcal{B}_n$ are n independent sub-σ-fields of $\mathcal{A}$, and if, for every $i \in \{1, \dots, n\}$, X_i is a $\mathcal{B}_i$-measurable random variable, then $X_1, \dots, X_n$ are independent (this is obvious since saying that X_i is $\mathcal{B}_i$-measurable is equivalent to $\sigma(X_i) \subset \mathcal{B}_i$).

(ii) The n events $A_1, \dots, A_n$ are independent if and only if the σ-fields $\sigma(A_1), \dots, \sigma(A_n)$ are independent (Proposition 9.2).

The next theorem gives a very important characterization of the independence of n random variables. Before stating this theorem, we observe that, if $X_1, \dots, X_n$ are n random variables taking values in $(E_1, \mathcal{E}_1), \dots, (E_n, \mathcal{E}_n)$ respectively, the n-tuple $(X_1, \dots, X_n)$ is a random variable with values in $E_1 \times \dots \times E_n$ equipped with the product σ-field $\mathcal{E}_1 \otimes \dots \otimes \mathcal{E}_n$ (see Lemma 1.12). Recall the notion of product measure introduced in Chapter 5.

Theorem 9.4 *Let $X_1, \dots, X_n$ be random variables taking values in $(E_1, \mathcal{E}_1), \dots, (E_n, \mathcal{E}_n)$ respectively. Then $X_1, \dots, X_n$ are independent if and only if the law of the n-tuple $(X_1, \dots, X_n)$ is the product measure of the respective laws of $X_1, \dots, X_n$,*

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

Moreover, we have then

$$\mathbb{E}\Big[\prod_{i=1}^{n} f_i(X_i)\Big] = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)] \tag{9.2}$$

whenever f_i is a nonnegative measurable function on $(E_i, \mathcal{E}_i)$, for every $i \in \{1, \dots, n\}$ (note that the convention $0 \times \infty = 0$ is used here as elsewhere).

Proof Let $F_i \in \mathcal{E}_i$, for every $i \in \{1, \dots, n\}$. On one hand,

$$\mathbb{P}_{(X_1,\dots,X_n)}(F_1 \times \cdots \times F_n) = \mathbb{P}(\{X_1 \in F_1\} \cap \ldots \cap \{X_n \in F_n\})$$

and on the other hand,

$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}(F_1 \times \cdots \times F_n) = \prod_{i=1}^{n} \mathbb{P}_{X_i}(F_i) = \prod_{i=1}^{n} \mathbb{P}(X_i \in F_i).$$

Comparing with (9.1), we set that $X_1, \dots, X_n$ are independent if and only if the two probability measures $\mathbb{P}_{(X_1,\dots,X_n)}$ and $\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ take the same values on boxes $F_1 \times \cdots \times F_n$. This is equivalent to saying that $\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$, since we know that a probability measure on a product space is characterized by its values on boxes (as a straightforward consequence of Corollary 1.19).

The second assertion is then a consequence of the Fubini theorem:

$$\begin{aligned}
\mathbb{E}\Big[\prod_{i=1}^{n} f_i(X_i)\Big] &= \int_{E_1 \times \cdots \times E_n} \prod_{i=1}^{n} f_i(x_i)\, \mathbb{P}_{X_1}(\mathrm{d}x_1) \ldots \mathbb{P}_{X_n}(\mathrm{d}x_n) \\
&= \prod_{i=1}^{n} \int_{E_i} f_i(x_i)\, \mathbb{P}_{X_i}(\mathrm{d}x_i) \\
&= \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)].
\end{aligned}$$

□

Let us briefly discuss some useful extensions of formula (9.2).

(i) As in Theorem 5.3, we can allow the functions f_i to take values in $[0, \infty]$ in (9.2).

(ii) For functions f_i of arbitrary sign, formula (9.2) is still valid provided $\mathbb{E}[|f_i(X_i)|] < \infty$ for every $i \in \{1, \dots, n\}$, noting that this implies

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

which justifies the existence of the expected value in the left-hand side of (9.2). Similarly, (9.2) holds for complex-valued functions f_i under the same integrability condition.

(iii) As a special case of (ii), if $X_1, \ldots, X_n$ are independent real random variables in L^1, we have also $X_1 \cdots X_n \in L^1$, and

$$\mathbb{E}[X_1 \cdots X_n] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

Note that, in general, a product of random variables in L^1 may not belong to L^1.

Theorem 9.4 also shows how to construct (finitely many) independent random variables. Consider the case of real random variables. Let $\mu_1, \ldots, \mu_n$ be n probability measures on $\mathbb{R}$. As we observed in the previous chapter after Definition 8.2, we can construct a random variable $Y = (Y_1, \ldots, Y_n)$ with values in $\mathbb{R}^n$ whose law is $\mu_1 \otimes \cdots \otimes \mu_n$. Then Theorem 9.4 implies that the components $Y_1, \ldots Y_n$ of Y are independent real random variables with respective laws $\mu_1, \ldots, \mu_n$.

Let us turn to the relation between independence and covariance.

Corollary 9.5 *If X_1, X_2 are two independent real random variables in L^2, then* $\mathrm{cov}(X_1, X_2) = 0$.

This follows from the preceding remarks since $\mathrm{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$.

The converse to the corollary is false. The property $\mathrm{cov}(X_1, X_2) = 0$ is much weaker than the independence of X_1 and X_2. To give a simple example, let X_1 be a real random variable having a symmetric density $p(x) = p(-x)$ and such that $\int x^2 p(x)\mathrm{d}x < \infty$ (so that $X_1 \in L^2$). Let Y be another random variable, with values in $\{-1, 1\}$, such that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}$, and assume that Y is independent of X_1. Then, setting $X_2 = \varepsilon X_1$, we easily get

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] = \mathbb{E}[Y(X_1)^2] = \mathbb{E}[Y]\mathbb{E}[(X_1)^2] = 0,$$

but X_1 and X_2 are not independent. Indeed, if X_1 and X_2 were independent, $|X_1|$ would be independent of $|X_2| = |X_1|$, and it is easy to see that a real random variable cannot be independent of itself unless it is constant a.s.

Corollary 9.6 *Let $X_1, \ldots, X_n$ be n real random variables.*

(i) *Suppose that, for every $i \in \{1, \ldots, n\}$, X_i has a density denoted by p_i, and that the random variables $X_1, \ldots, X_n$ are independent. Then, the n-tuple $(X_1, \ldots, X_n)$ has a density given by*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i).$$

(ii) *Conversely, suppose that* $(X_1, \ldots, X_n)$ *has a density* p *which can be written in the form*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} q_i(x_i),$$

where $q_1, \ldots, q_n$ *are nonnegative measurable functions on* $\mathbb{R}$. *Then, the random variables* $X_1, \ldots, X_n$ *are independent and, for every* $i \in \{1, \ldots, n\}$, X_i *has a density* p_i *of the form* $p_i = C_i\, q_i$, *where* $C_i > 0$ *is a constant.*

Proof Part (i) is an immediate consequence of Theorem 9.4, since the Fubini theorem readily shows that, under the assumption $\mathbb{P}_{X_i}(\mathrm{d}x_i) = p_i(x_i)\mathrm{d}x_i$ for $i \in \{1, \ldots, n\}$, we have

$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}(\mathrm{d}x_1 \ldots \mathrm{d}x_n) = \Big(\prod_{i=1}^{n} p_i(x_i)\Big)\mathrm{d}x_1 \ldots \mathrm{d}x_n.$$

As for (ii), we first observe that, thanks again to the Fubini theorem,

$$\prod_{i=1}^{n}\Big(\int q_i(x)\mathrm{d}x\Big) = \int_{\mathbb{R}^n} p(x_1, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_n = 1,$$

and in particular the quantity $K_i = \int q_i(x)\mathrm{d}x$ is positive and finite, for every $i \in \{1, \ldots, n\}$. Then, by Proposition 8.6, the density of X_i is

$$\begin{aligned} p_i(x_i) = \int_{\mathbb{R}^{n-1}} p(x_1, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \ldots, \mathrm{d}x_n &= \Big(\prod_{j\neq i} K_j\Big) q_i(x_i) \\ &= \frac{1}{K_i} q_i(x_i), \end{aligned}$$

giving the desired form $p_i = C_i\, q_i$, with $C_i = 1/K_i$. This also allows us to rewrite the density of $(X_1, \ldots, X_n)$ in the form

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} q_i(x_i) = \prod_{i=1}^{n} p_i(x_i)$$

and we see that $\mathbb{P}_{(X_1,\ldots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ so that $X_1, \ldots, X_n$ are independent. □

Example Let U be a random variable following the exponential distribution with parameter 1 and let V be uniformly distributed over $[0, 1]$. Assume that U and V are independent. Then, if we set

$$X = \sqrt{U}\,\cos(2\pi V)\ , \quad Y = \sqrt{U}\,\sin(2\pi V),$$

X and Y are independent real random variables. To verify this, let us compute the law of (X, Y). For every nonnegative measurable function φ on $\mathbb{R}^2$, we have

$$\begin{aligned}\mathbb{E}[\varphi(X,Y)] &= \int_0^\infty \int_0^1 \varphi(\sqrt{u}\cos(2\pi v), \sqrt{u}\sin(2\pi v))\, e^{-u}\,\mathrm{d}u\mathrm{d}v \\ &= \frac{1}{\pi}\int_0^\infty \int_0^{2\pi} \varphi(r\cos\theta, r\sin\theta)\, r e^{-r^2}\,\mathrm{d}r\mathrm{d}\theta \\ &= \frac{1}{\pi}\int_{\mathbb{R}^2} \varphi(x,y)\, e^{-x^2-y^2}\,\mathrm{d}x\mathrm{d}y.\end{aligned}$$

We thus obtain that the pair (X, Y) has density $\pi^{-1}\exp(-x^2 - y^2)$ which has a product form as in part (ii) of Proposition 9.6. It follows that X and Y are independent (and we also see that X and Y have the same density $\frac{1}{\sqrt{\pi}}e^{-x^2}$ and thus follow the $\mathcal{N}(0, 1/2)$ distribution).

We now state a very useful technical result.

Proposition 9.7 *Let $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be sub-σ-fields of $\mathcal{A}$. For every $i \in \{1, \ldots, n\}$, let $\mathcal{C}_i \subset \mathcal{B}_i$ be a class closed under finite intersections, containing Ω and such that $\sigma(\mathcal{C}_i) = \mathcal{B}_i$. Assume that*

$$\forall C_1 \in \mathcal{C}_1, \ldots, \forall C_n \in \mathcal{C}_n,\ \mathbb{P}(C_1 \cap C_2 \cap \ldots \cap C_n) = \mathbb{P}(C_1)\,\mathbb{P}(C_2) \ldots \mathbb{P}(C_n).$$

Then the σ-fields $\mathcal{B}_1, \ldots, \mathcal{B}_n$ are independent.

Proof Let us first fix $C_2 \in \mathcal{C}_2, \ldots, C_n \in \mathcal{C}_n$, and set

$$\mathcal{M}_1 = \{B_1 \in \mathcal{B}_1 : \mathbb{P}(B_1 \cap C_2 \cap \ldots \cap C_n) = \mathbb{P}(B_1)\,\mathbb{P}(C_2) \ldots \mathbb{P}(C_n)\}.$$

Then $\mathcal{C}_1 \subset \mathcal{M}_1$ by assumption, and on the other hand, it is easy to see that $\mathcal{M}_1$ is a monotone class as defined in Section 1.4. The monotone class theorem (Theorem 1.18) then implies that $\mathcal{M}_1$ contains $\sigma(\mathcal{C}_1) = \mathcal{B}_1$, and we have proved

$$\forall B_1 \in \mathcal{B}_1, C_2 \in \mathcal{C}_2, \ldots, C_n \in \mathcal{C}_n,\ \mathbb{P}(B_1 \cap C_2 \cap \ldots \cap C_n) = \mathbb{P}(B_1)\mathbb{P}(C_2) \ldots \mathbb{P}(C_n).$$

To proceed, we now fix $B_1 \in \mathcal{B}_1, C_3 \in \mathcal{C}_3, \ldots, C_n \in \mathcal{C}_n$ and we set

$$\mathcal{M}_2 = \{B_2 \in \mathcal{B}_2 : \mathbb{P}(B_1 \cap B_2 \cap C_3 \cap \ldots \cap C_n) = \mathbb{P}(B_1)\,\mathbb{P}(B_2)\,\mathbb{P}(C_3) \ldots \mathbb{P}(C_n)\}.$$

Again $\mathcal{M}_2$ is a monotone class, which contains $\mathcal{C}_2$, and it follows that $\mathcal{M}_2$ contains $\sigma(\mathcal{C}_2) = \mathcal{B}_2$. By induction, we arrive at the property

$$\forall B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2, \ldots, B_n \in \mathcal{B}_n,\ \mathbb{P}(B_1 \cap B_2 \cap \ldots \cap B_n) = \mathbb{P}(B_1)\mathbb{P}(B_2)\ldots\mathbb{P}(B_n),$$

which means that $\mathcal{B}_1, \ldots, \mathcal{B}_n$ are independent. □

Consequence: Grouping by Blocks Recall that, if $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are σ-fields, $\mathcal{A}_1 \vee \mathcal{A}_2 \vee \cdots \vee \mathcal{A}_k$ denotes the smallest σ-field that contains $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_k$. Let $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be independent σ-fields, and let $n_0 = 0 < n_1 < \cdots < n_p = n$. Then the σ-fields

$$\begin{aligned} \mathcal{D}_1 &= \mathcal{B}_1 \vee \cdots \vee \mathcal{B}_{n_1} \\ \mathcal{D}_2 &= \mathcal{B}_{n_1+1} \vee \cdots \vee \mathcal{B}_{n_2} \\ &\cdots \\ \mathcal{D}_p &= \mathcal{B}_{n_{p-1}+1} \vee \cdots \vee \mathcal{B}_{n_p} \end{aligned}$$

are independent. This can be deduced from Proposition 9.7. For every $j \in \{1, \ldots, p\}$, we let $\mathcal{C}_j$ be the class of all sets of the form

$$B_{n_{j-1}+1} \cap \cdots \cap B_{n_j}$$

where $B_i \in \mathcal{B}_i$ for every $i \in \{n_{j-1} + 1, \ldots, n_j\}$. Then the assumptions of Proposition 9.7 hold for $\mathcal{C}_1 \ldots, \mathcal{C}_p$, and since $\mathcal{D}_j = \sigma(\mathcal{C}_j)$ for every $j \in \{1, \ldots, p\}$, we get the desired result.

In particular, if $X_1, \ldots, X_n$ are independent, the random variables

$$Y_1 = (X_1, \ldots, X_{n_1}), Y_2 = (X_{n_1+1}, \ldots, X_{n_2}), \ldots, Y_p = (X_{n_{p-1}+1}, \ldots, X_{n_p})$$

are independent.

Example If $X_1, \ldots, X_4$ are independent real random variables, the random variables $Z_1 = X_1X_3$ and $Z_2 = X_2^3 + X_4$ are independent.

The next proposition uses Proposition 9.7 and Theorem 8.16 to give different characterizations of the independence of real random variables.

Proposition 9.8 *Let $X_1, \ldots, X_n$ be n real random variables. The following are equivalent:*

(i) *$X_1, \ldots, X_n$ are independent.*

(ii) *For every $a_1, \ldots, a_n \in \mathbb{R}$, $\mathbb{P}(X_1 \le a_1, \ldots, X_n \le a_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \le a_i)$.*

(iii) *Let $f_1, \dots, f_n$ be continuous functions with compact support from $\mathbb{R}$ into $\mathbb{R}_+$. Then,*

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

(iv) *The characteristic function of the random vector $X = (X_1, \dots, X_n)$ is*

$$\Phi_X(\xi_1, \dots, \xi_n) = \prod_{i=1}^{n} \Phi_{X_i}(\xi_i)$$

Proof The fact that (i) implies both (ii) and (iii) follows from Theorem 9.4. Conversely, to show that (iii) implies (i), we note that the indicator function of an open interval is the increasing limit of a sequence of continuous functions with compact support. By monotone convergence, it follows that the property $\mathbb{P}(\{X_1 \in F_1\} \cap \dots \cap \{X_n \in F_n\}) = \mathbb{P}(X_1 \in F_1) \cdots \mathbb{P}(X_n \in F_n)$ holds when $F_1, \dots, F_n$ are open intervals. Then we just have to apply Proposition 9.7, taking for $\mathcal{C}_j$ the class of all $\{X_j \in F\}$, F open interval of $\mathbb{R}$ (observe that this class generates $\sigma(X_j)$). The proof that (ii) implies (i) is similar.

To show the equivalence between (i) and (iv), we note that the Fourier transform of the product measure $\mathbb{P}_{X_1} \otimes \dots \otimes \mathbb{P}_{X_n}$ is

$$\begin{aligned}(\xi_1, \dots, \xi_n) \mapsto \int \exp\Big(\mathrm{i}\sum_{j=1}^{n} \xi_j x_j\Big)\mathbb{P}_{X_1}(\mathrm{d}x_1) \dots \mathbb{P}_{X_n}(\mathrm{d}x_n)\\ = \prod_{j=1}^{n} \int e^{\mathrm{i}\xi_j x_j}\mathbb{P}_{X_j}(\mathrm{d}x_j) = \prod_{j=1}^{n} \Phi_{X_j}(\xi_j)\end{aligned}$$

using the Fubini theorem in the first equality. Hence the property (iv) is equivalent to the fact that the Fourier transform of $\mathbb{P}_{(X_1,\dots,X_n)}$ coincides with the Fourier transform of $\mathbb{P}_{X_1} \otimes \dots \otimes \mathbb{P}_{X_n}$. By Theorem 8.16, this holds if and only if $\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \dots \otimes \mathbb{P}_{X_n}$, which is equivalent to the independence of $X_1, \dots, X_n$ by Theorem 9.4. □

In forthcoming applications, it will be important to consider the independence of infinitely many random variables.

Definition 9.9 (Independence of an Infinite Family) Let $(\mathcal{B}_i)_{i \in I}$ be an arbitrary collection of sub-σ-fields of $\mathcal{A}$. We say that the σ-fields $\mathcal{B}_i$, $i \in I$, are independent if, for every finite subset $\{i_1, \dots, i_p\}$ of I, the σ-fields $\mathcal{B}_{i_1}, \dots, \mathcal{B}_{i_p}$ are independent. If $(X_i)_{i \in I}$ is an arbitrary collection of random variables, we say that the random variables X_i, $i \in I$, are independent if the σ-fields $\sigma(X_i)$, $i \in I$, are independent.

Similarly, for an infinite collection $(A_i)_{i \in I}$ of events, we say that the events A_i, $i \in I$, are independent if the σ-fields $\sigma(A_i)$, $i \in I$, are independent.

The grouping by blocks principle can be extended to infinite collections of random variables. Rather than stating a general result, we give a simple example of such extensions, which will be useful later in this chapter.

Proposition 9.10 *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent random variables. Then, for every $p \in \mathbb{N}$, the σ-fields*

$$\mathcal{B}_1 = \sigma(X_1, \ldots, X_p)\,, \quad \mathcal{B}_2 = \sigma(X_{p+1}, X_{p+2}, \ldots)$$

are independent.

Proof We apply Proposition 9.7 with

$$\mathcal{C}_1 = \sigma(X_1, \ldots, X_p) = \mathcal{B}_1$$

$$\mathcal{C}_2 = \bigcup_{k=p+1}^{\infty} \sigma(X_{p+1}, X_{p+2}, \ldots, X_k) \subset \mathcal{B}_2$$

and we note that the assumption in this proposition holds thanks to the grouping by blocks principle. □

The reader will be able to formulate extensions of the last proposition involving disjoint (finite or infinite) blocks in an infinite collection of independent random variables. The proof always relies on applications of Proposition 9.7.

Recall that two (finite or infinite) collections $(X_i)_{i \in I}$ and $(Y_j)_{j \in J}$ of random variables are independent if the σ-fields $\sigma((X_i)_{i \in I})$ and $\sigma((Y_j)_{j \in J})$ are independent. This property holds if and only if, for every finite subset $\{i_1, \ldots, i_p\}$ of I and every finite subset $\{j_1, \ldots, j_q\}$ of J, the p-tuple $(X_{i_1}, \ldots, X_{i_p})$ and the q-tuple $(Y_{j_1}, \ldots, Y_{j_q})$ are independent. Once again, the latter fact is an easy application of Proposition 9.7 and we omit the details.

9.3 The Borel-Cantelli Lemma

Recall that, if $(A_n)_{n \in \mathbb{N}}$ is a sequence of events, we write

$$\limsup A_n = \bigcap_{n=1}^{\infty} \Big(\bigcup_{k=n}^{\infty} A_k \Big).$$

The event $\limsup A_n$ consists of those ω that belong to infinitely many of the sets A_n. Note that Lemma 1.7 gives $\mathbb{P}(\limsup A_n) \geq \limsup \mathbb{P}(A_n)$.

Lemma 9.11 *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events.*

(i) *If $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}(\limsup A_n) = 0$$

or equivalently,

$$\{n \in \mathbb{N} : \omega \in A_n\} \text{ is a.s. finite.}$$

(ii) *If $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$ and if the events A_n, $n \in \mathbb{N}$, are independent, then*

$$\mathbb{P}(\limsup A_n) = 1$$

or equivalently

$$\{n \in \mathbb{N} : \omega \in A_n\} \text{ est a.s. infinite.}$$

Remark The independence assumption, or another suitable assumption, is needed in (ii), as can be seen from the trivial example where $A_n = A$ for every $n \in \mathbb{N}$, with $0 < \mathbb{P}(A) < 1$. See Exercise 8.13 for an extension of part (ii) of the lemma where the independence assumption is replaced by a weaker one.

Proof

(i) If $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$, then

$$\mathbb{E}\Big[\sum_{n\in\mathbb{N}} \mathbf{1}_{A_n}\Big] = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$$

and thus $\sum_{n\in\mathbb{N}} \mathbf{1}_{A_n} < \infty$ a.s., which exactly means that $\{n \in \mathbb{N} : \omega \in A_n\}$ is a.s. finite.

(ii) Fix $n_0 \in \mathbb{N}$, and observe that, if $n \geq n_0$,

$$\mathbb{P}\Big(\bigcap_{k=n_0}^{n} A_k^c\Big) = \prod_{k=n_0}^{n} \mathbb{P}(A_k^c) = \prod_{k=n_0}^{n} (1 - \mathbb{P}(A_k)).$$

The fact that the series $\sum \mathbb{P}(A_k)$ diverges implies that the right-hand side tends to 0 as $n \to \infty$. Consequently,

$$\mathbb{P}\Big(\bigcap_{k=n_0}^{\infty} A_k^c\Big) = 0.$$

Since this holds for every $n_0 \in \mathbb{N}$, we have also

$$\mathbb{P}\Big(\bigcup_{n_0=1}^{\infty}\Big(\bigcap_{k=n_0}^{\infty} A_k^c\Big)\Big) = 0$$

and, by considering the complementary set,

$$\mathbb{P}\Big(\bigcap_{n_0=1}^{\infty}\Big(\bigcup_{k=n_0}^{\infty} A_k\Big)\Big) = 1,$$

which was the desired result.

□

Two Applications (1) One cannot find a probability measure $\mathbb{P}$ on $\mathbb{N}$ such that, for every integer $n \geq 1$, the probability of the set of all multiples of n is equal to $1/n$. Indeed suppose that there exists such a probability measure $\mathbb{P}$. Let $\mathcal{P}$ denote the set of all prime numbers, and, for every $p \in \mathcal{P}$, let $A_p = p\mathbb{N}$ be the set of all multiples of p. Then it is easy to see that the sets A_p, $p \in \mathcal{P}$, are independent under the probability measure $\mathbb{P}$. Indeed, if $p_1, \ldots, p_k$ are distinct prime numbers, we have

$$\begin{aligned}\mathbb{P}(A_{p_1} \cap \ldots \cap A_{p_k}) = \mathbb{P}(p_1\mathbb{N} \cap \ldots \cap p_k\mathbb{N}) &= \mathbb{P}((p_1 \ldots p_k)\mathbb{N}) \\ &= \frac{1}{p_1 \ldots p_k} = \prod_{j=1}^{k} \mathbb{P}(A_{p_j}).\end{aligned}$$

On the other hand, we have

$$\sum_{p\in\mathcal{P}} \mathbb{P}(A_p) = \sum_{p\in\mathcal{P}} \frac{1}{p} = \infty.$$

We can then apply part (ii) of the Borel-Cantelli lemma to obtain that $\mathbb{P}$-almost every integer $n \in \mathbb{N}$ belongs to infinitely many sets A_p, and is therefore a multiple of infinitely many prime numbers, which is absurd.
(2) Suppose now that

$$(\Omega, \mathcal{A}, \mathbb{P}) = ([0, 1), \mathcal{B}([0, 1)), \lambda),$$

where λ denotes Lebesgue measure. For every $n \in \mathbb{N}$, set

$$\forall \omega \in [0, 1), \quad X_n(\omega) = \lfloor 2^n\omega \rfloor - 2\lfloor 2^{n-1}\omega \rfloor,$$

where $\lfloor x \rfloor$ denotes the integer part of the real number x. Then $X_n(\omega) \in \{0, 1\}$ and one easily verifies by induction on n that, for every $\omega \in [0, 1)$,

$$0 \leq \omega - \sum_{k=1}^{n} X_k(\omega)2^{-k} < 2^{-n}.$$

which shows that

$$\omega = \sum_{k=1}^{\infty} X_k(\omega)\, 2^{-k}.$$

The numbers $X_k(\omega)$ are the coefficients of the (proper) dyadic expansion of ω. By writing explicitly the set $\{X_n = 1\}$ one checks that, for every $n \in \mathbb{N}$,

$$\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = \frac{1}{2}.$$

Finally, we observe that the random variables X_n, $n \in \mathbb{N}$ are independent. In fact, it is enough to verify that, for every $i_1, \ldots, i_p \in \{0, 1\}$, one has

$$\mathbb{P}(X_1 = i_1, \ldots, X_p = i_p) = \frac{1}{2^p} = \prod_{j=1}^{p} \mathbb{P}(X_j = i_j).$$

However, one immediately sees that

$$\{X_1 = i_1, \ldots, X_p = i_p\} = \Big[\sum_{j=1}^{p} i_j 2^{-j}, \sum_{j=1}^{p} i_j 2^{-j} + 2^{-p}\Big),$$

and the desired result follows.

Let $p \in \mathbb{N}$ and let $i_1, \ldots, i_p \in \{0, 1\}$. Then we can use the Borel-Cantelli lemma to prove that

$$\mathrm{card}\{k \geq 0 : X_{k+1} = i_1, \ldots, X_{k+p} = i_p\} = \infty \quad \text{a.s.} \tag{9.3}$$

This shows that a given finite sequence of 0 and 1 appears infinitely many times in the dyadic expansion of Lebesgue almost every real of $[0, 1)$. In order to establish (9.3), set, for every $n \in \mathbb{Z}_+$,

$$Y_n = (X_{np+1}, X_{np+2}, \ldots, X_{np+p}).$$

The grouping by blocks principle shows that the random variables Y_n, $n \in \mathbb{Z}_+$, are independent and then the desired result follows from an application of the Borel-Cantelli lemma to the events

$$A_n = \{Y_n = (i_1, \ldots, i_p)\},$$

which are independent and all have probability 2^{-p}.

Since a countable union of events of probability zero still has probability zero, we can reinforce (9.3) as follows: it is a.s. true that

$$\forall p \geq 1,\ \forall i_1, \ldots, i_p \in \{0, 1\},\ \operatorname{card}\{k \geq 0 : X_{k+1} = i_1, \ldots, X_{k+p} = i_p\} = \infty.$$

In other words, for a real x chosen at random in $[0, 1)$, any finite sequence of 0 and 1 appears infinitely many times in the dyadic expansion of x.

9.4 Construction of Independent Sequences

Many of the developments that follow are devoted to the study of sequences of independent and identically distributed random variables. An obvious question is the existence of such sequences on an appropriate probability space. The natural way to answer this question would be to extend the construction of product measures to infinite products, so that we can then proceed as outlined after the proof of Proposition 9.4, when we explained how to construct finitely many independent random variables. As the construction of measures on infinite products involves some measure-theoretic difficulties, we present here a more elementary approach, which is limited to real variables but will be sufficient for our needs in the present book (even for the construction of Brownian motion in Chapter 14).

We consider the probability space

$$(\Omega, \mathcal{A}, \mathbb{P}) = ([0, 1), \mathcal{B}([0, 1)), \lambda).$$

As we have just seen in the previous section, the (proper) dyadic expansion of a real $\omega \in [0, 1)$,

$$\omega = \sum_{n=1}^{\infty} X_n(\omega)\, 2^{-n}, \qquad X_n(\omega) \in \{0, 1\}$$

yields a sequence $(X_n)_{n\in\mathbb{N}}$ of independent random variables of Bernoulli distribution of parameter $1/2$. Let φ be any fixed one-to-one mapping from $\mathbb{N} \times \mathbb{N}$ into $\mathbb{N}$ and set $Y_{i,j} = X_{\varphi(i,j)}$ for every $i, j \in \mathbb{N}$. The random variables $Y_{i,j}$, $i, j \in \mathbb{N}$, are (obviously) independent. By setting, for every $i \in \mathbb{N}$,

$$U_i = \sum_{j=1}^{\infty} Y_{i,j}\, 2^{-j}$$

we get a sequence $U_1, U_2, \ldots$ of independent random variables uniformly distributed over $[0, 1]$.

The reason why $U_1, U_2, \ldots$ are independent is the fact that the σ-fields $\mathcal{G}_i := \sigma(Y_{i,j} : j \in \mathbb{N})$ for $i \in \mathbb{N}$, are independent, by an extension of the grouping by blocks principle, which is easily proved using Proposition 9.7. To see that U_i is uniformly distributed on [0, 1], note that $U_i^{(p)} := \sum_{j=1}^{p} Y_{i,j}\, 2^{-j}$ has the same law as $X^{(p)} := \sum_{n=1}^{p} X_n\, 2^{-n}$, for every integer $p \geq 1$. Then, if φ is an arbitrary bounded continuous function on $\mathbb{R}$, passing to the limit $p \to \infty$ in the equality $\mathbb{E}[\varphi(U_i^{(p)})] = \mathbb{E}[\varphi(X^{(p)})]$ shows that $\mathbb{E}[\varphi(U_i)] = \mathbb{E}[\varphi(X)]$ where $X(\omega) = \omega$ is uniformly distributed on [0, 1].

If we now consider a probability measure μ on $\mathbb{R}$, and $G_\mu(x) = \mu((-\infty, x])$, for every $x \in \mathbb{R}$, then Proposition 8.7 shows that the random variables $Z_i = G_\mu^{-1}(U_i)$ form a sequence of independent identically distributed random variables with law μ.

9.5 Sums of Independent Random Variables

Sums of independent random variables play an important role in probability theory and will be studied in the next chapter. We gather several useful properties of these sums in the next proposition.

We first need to introduce the convolution of probability measures on $\mathbb{R}^d$. If μ and ν are two probability measures on $\mathbb{R}^d$, the convolution $\mu * \nu$ is the pushforward of the product measure $\mu \otimes \nu$ on $\mathbb{R}^d \times \mathbb{R}^d$ under the mapping $(x, y) \mapsto x + y$: for any nonnegative measurable function φ on $\mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \varphi(z)\, \mu * \nu(\mathrm{d}z) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi(x + y)\, \mu(\mathrm{d}x)\nu(\mathrm{d}y).$$

In the special case, where μ has density f and ν has density g (with respect to Lebesgue measure), then $\mu * \nu$ has density $f * g$ (which makes sense by Proposition 5.5). This is easily checked by writing

$$\int \int \varphi(x + y)\, f(x)g(y)\mathrm{d}x\mathrm{d}y = \int \varphi(z)\Big(\int f(x)g(z - x)\mathrm{d}x\Big)\mathrm{d}z,$$

where we used the Fubini theorem and a simple change of variable.

If μ and ν are probability measures on $\mathbb{R}^d$, the Fourier transform of $\mu * \nu$ is the product $\widehat{\mu}\,\widehat{\nu}$ of the Fourier transforms of μ and ν. This is an immediate application of the Fubini theorem, since for every $\xi \in \mathbb{R}^d$,

$$\widehat{\mu * \nu}(\xi) = \int_{\mathbb{R}^d} e^{\mathrm{i}\xi \cdot z}\, \mu * \nu(\mathrm{d}z) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{\mathrm{i}\xi \cdot (x+y)}\, \mu(\mathrm{d}x)\nu(\mathrm{d}y) = \widehat{\mu}(\xi)\widehat{\nu}(\xi).$$

Proposition 9.12 *Let X and Y be two independent random variables with values in $\mathbb{R}^d$.*

(i) *The law of $X + Y$ is $\mathbb{P}_X * \mathbb{P}_Y$. In particular, if X has a density p_X and Y has a density p_Y, then $X + Y$ has density $p_X * p_Y$.*
(ii) *The characteristic function of $X + Y$ is $\Phi_{X+Y}(\xi) = \Phi_X(\xi)\Phi_Y(\xi)$.*
(iii) *If $\mathbb{E}[|X|^2] < \infty$ and $\mathbb{E}[|Y|^2] < \infty$, and K_X denotes the covariance matrix of X, we have $K_{X+Y} = K_X + K_Y$. In particular, if $d = 1$ and X and Y are in L^2,* $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

Proof

(i) We know that $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y$, and thus, for every nonnegative measurable function φ on $\mathbb{R}^d$,

$$\mathbb{E}[\varphi(X+Y)] = \int\int \varphi(x+y)\,\mathbb{P}_X(\mathrm{d}x)\mathbb{P}_Y(\mathrm{d}y) = \int \varphi(z)\,\mathbb{P}_X * \mathbb{P}_Y(\mathrm{d}z)$$

by the definition of $\mathbb{P}_X * \mathbb{P}_Y$. The case with density follows from the remarks before the proposition.
(ii) This follows from the equality $\mathbb{P}_{X+Y} = \mathbb{P}_X * \mathbb{P}_Y$ and the last observation before the proposition.
(iii) If $X = (X_1, \dots, X_d)$ and $Y = (Y_1, \dots, Y_d)$, Corollary 9.5 implies that $\mathrm{cov}(X_i, Y_j) = 0$ for every $i, j \in \{1, \dots, d\}$. Consequently, using bilinearity,

$$\mathrm{cov}(X_i + Y_i, X_j + Y_j) = \mathrm{cov}(X_i, X_j) + \mathrm{cov}(Y_i, Y_j)$$

which gives $K_{X+Y} = K_X + K_Y$.

□

Theorem 9.13 (Weak Law of Large Numbers) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables in L^2. Then,*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow[n\to\infty]{L^2} \mathbb{E}[X_1].$$

Proof By linearity,

$$\mathbb{E}\Big[\frac{1}{n}(X_1 + \cdots + X_n)\Big] = \mathbb{E}[X_1].$$

Furthermore, Proposition 9.12(iii) shows that $\mathrm{var}(X_1 + \cdots + X_n) = n\,\mathrm{var}(X_1)$. Consequently,

$$\mathbb{E}\Big[\Big(\frac{1}{n}(X_1 + \cdots + X_n) - \mathbb{E}[X_1]\Big)^2\Big] = \frac{1}{n^2}\mathrm{var}(X_1 + \cdots + X_n) = \frac{1}{n}\mathrm{var}(X_1)$$

which tends to 0 as $n \to \infty$. □

Remark The proof shows that the result holds under much weaker hypotheses. Instead of requiring that the X_n's have the same law, it is enough to assume that $\mathbb{E}[X_n] = \mathbb{E}[X_1]$ for every n and that the sequence $(\mathbb{E}[X_n^2])_{n\in\mathbb{N}}$ is bounded. We can replace the independence assumption by the property $\mathrm{cov}(X_n, X_m) = 0$ for every $n \neq m$, which is also much weaker.

The word "weak" in the weak law of large numbers refers to the fact that the convergence holds in L^2, whereas from the point of view of probability theory, it is more meaningful to have an almost sure convergence, that is, a pointwise convergence outside a set of probability zero (we then speak of a "strong" law). We give a first version of the strong law, which will be significantly improved in the next chapter.

Proposition 9.14 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables, and assume that $\mathbb{E}[(X_1)^4] < \infty$. Then, we have almost surely*

$$\frac{1}{n}(X_1 + \cdots + X_n) \underset{n\to\infty}{\longrightarrow} \mathbb{E}[X_1].$$

Proof Up to replacing X_n by $X_n - \mathbb{E}[X_n]$, we may assume that $\mathbb{E}[X_n] = 0$. Then, by expanding the fourth power of $X_1 + \cdots + X_n$, we have

$$\mathbb{E}[(\frac{1}{n}(X_1 + \cdots + X_n))^4] = \frac{1}{n^4} \sum_{i_1,\ldots,i_4\in\{1,\ldots,n\}} \mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}].$$

Using independence and the property $\mathbb{E}[X_k] = 0$, we see that the only nonzero terms in the sum in the right-hand side are those for which any value taken by a component of the 4-tuple (i_1, i_2, i_3, i_4) appears at least twice in this 4-tuple. Since the X_k's have the same distribution, we get

$$\mathbb{E}[(\frac{1}{n}(X_1 + \cdots + X_n))^4] = \frac{1}{n^4}\Big(n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2 X_2^2]\Big) \leq \frac{C}{n^2}$$

with some constant $C < \infty$ ($C = 3\,\mathbb{E}[X_1^4]$ works). It follows that

$$\sum_{n=1}^{\infty} \mathbb{E}[(\frac{1}{n}(X_1 + \cdots + X_n))^4] < \infty.$$

By interchanging sum and expected value, we obtain

$$\mathbb{E}\Big[\sum_{n=1}^{\infty} (\frac{1}{n}(X_1 + \cdots + X_n))^4\Big] < \infty,$$

and therefore

$$\sum_{n=1}^{\infty}(\frac{1}{n}(X_1 + \cdots + X_n))^4 < \infty \,, \quad \text{a.s.}$$

which implies that $\frac{1}{n}(X_1 + \cdots + X_n)$ converges a.s. to 0. $\square$

Corollary 9.15 *If $(A_n)_{n\in\mathbb{N}}$ is a sequence of independent events with the same probability, we have almost surely*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{A_i} \underset{n\to\infty}{\longrightarrow} \mathbb{P}(A_1).$$

Proof Just apply Proposition 9.14 to $X_n = \mathbf{1}_{A_n}$. $\square$

This corollary provides a link between our modern axiomatic presentation of probability theory and the "frequentist" definition of the probability of an event as the asymptotic frequency of its occurence when the same random experiment is repeated a large number of times. See the discussion at the beginning of Chapter 8.

Let us come back to the second application of the Borel-Cantelli lemma given in Section 9.3, which was concerned with the dyadic expansion

$$\omega = \sum_{k=1}^{\infty} X_k(\omega)\, 2^{-k}$$

of a real $\omega \in [0, 1)$. If $p \geq 1$ we saw that the random variables $Y_1 = (X_1, \ldots, X_p)$, $Y_2 = (X_{p+1}, \ldots, X_{2p}), \ldots$ are independent and identically distributed (recall that the underlying probability is Lebesgue measure on $[0, 1)$). It then follows from Corollary 9.15 that, for every choice of $i_1, \ldots, i_p \in \{0, 1\}$, we have a.s.

$$\frac{1}{n}\,\mathrm{card}\{j \leq n : Y_j(\omega) = (i_1, \ldots, i_p)\} \underset{n\to\infty}{\longrightarrow} \frac{1}{2^p}.$$

For every $\ell \in \{1, \ldots, p\}$, the same argument applied to the random variables $(X_\ell, X_{\ell+1}, \ldots, X_{\ell+p-1}), (X_{\ell+p}, X_{\ell+p+1}, \ldots, X_{\ell+2p-1}), \ldots$ gives, a.s.,

$$\frac{1}{n}\,\mathrm{card}\{j \leq n : X_{\ell+jp}(\omega) = i_1, \ldots, X_{\ell+(j+1)p-1}(\omega) = i_p\} \underset{n\to\infty}{\longrightarrow} \frac{1}{2^p}.$$

By combining these results, we have a.s.

$$\frac{1}{n}\,\mathrm{card}\{k \leq n : X_{k+1}(\omega) = i_1, \ldots, X_{k+p}(\omega) = i_p\} \underset{n\to\infty}{\longrightarrow} \frac{1}{2^p}. \tag{9.4}$$

Since a countable union of sets of zero probability has probability zero, we conclude that the following holds: for every $\omega \in [0, 1)$, except on a set of Lebesgue measure 0, the property (9.4) holds simultaneously for every $p \geq 1$ and every choice of $i_1, \ldots, i_p \in \{0, 1\}$.

In other words, for almost every real $\omega \in [0, 1)$, the asymptotic frequency of occurence of any finite block of 0 and 1 in the dyadic expansion of ω exists and is equal to 2^{-p}, where p is the length of the block. Note that it is not easy to exhibit just one real $\omega \in [0, 1)$ for which the latter property holds. In fact the fastest way to prove that such reals do exist is certainly the probabilistic argument we have given. This is typical of applications of probability theory to existence problems: to get the existence of an object having certain properties, one shows that an object chosen at random (according to an appropriate probability distribution) will almost surely satisfy the desired properties.

9.6 Convolution Semigroups

In this section, we briefly discuss convolution semigroups. Roughly speaking, this notion gives rise to certain collections of probability distributions, such that the convolution of any two elements of the collection belongs to the same collection. We consider an indexing set I which can be $I = \mathbb{Z}_+$ or $I = \mathbb{R}_+$.

Definition 9.16 Let $(\mu_t)_{t \in I}$ be a collection of probability measures on $\mathbb{R}$ (or on $\mathbb{R}^d$). We say that $(\mu_t)_{t \in I}$ is a convolution semigroup if $\mu_0 = \delta_0$ and

$$\mu_t * \mu_{t'} = \mu_{t+t'} \,, \qquad \forall t, t' \in I.$$

The probabilistic interpretation is as follows. If X and Y are two independent random variables, and if the law of X is μ_t, and the law of Y is $\mu_{t'}$, then the law of $X + Y$ is $\mu_{t+t'}$ (cf. Proposition 9.12 (i)).

Lemma 9.17 *In order for $(\mu_t)_{t \in I}$ to be a convolution semigroup, it is enough that there exists a function $\varphi : \mathbb{R} \longrightarrow \mathbb{C}$ such that, for every $\xi \in \mathbb{R}$,*

- *if $I = \mathbb{Z}_+$, $\widehat{\mu}_t(\xi) = \varphi(\xi)^t$, $\forall t \in I$;*
- *if $I = \mathbb{R}$, $\widehat{\mu}_t(\xi) = \exp(-t\varphi(\xi))$, $\forall t \in I$.*

The proof is immediate, since, if $\widehat{\mu}_t$ has the form given in the lemma,

$$\widehat{\mu_t * \mu_{t'}} = \widehat{\mu}_t \, \widehat{\mu}_{t'} = \widehat{\mu}_{t+t'}$$

and the injectivity of the Fourier transform on probability measures (Theorem 8.16) gives $\mu_{t+t'} = \mu_t * \mu_{t'}$.

In the case of probability measures on $\mathbb{R}_+$, respectively on $\mathbb{N}$, one can give an analogous criterion in terms of Laplace transforms, resp. of generating functions. If $(\mu_t)_{t \in \mathbb{R}_+}$ is a collection of probability measures on $\mathbb{R}_+$, the existence of a function

$\varphi : \mathbb{R}_+ \to (0, \infty)$ such that, for every $t \geq 0$ and $\lambda \geq 0$,

$$\int_{\mathbb{R}_+} e^{-\lambda x} \, \mu_t(\mathrm{d}x) = \varphi(\lambda)^t$$

implies that $(\mu_t)_{t \in \mathbb{R}_+}$ is a convolution semigroup (use Theorem 8.19).

Examples

(1) $I = \mathbb{Z}_+$ and, for every $n \in \mathbb{N}$, μ_n is the binomial $\mathcal{B}(n, p)$ distribution, where $p \in [0, 1]$ is fixed. The property $\mu_{n+m} = \mu_n * \mu_m$ is immediate from the interpretation of the binomial distribution given in Section 8.1.5. Alternatively, we can use Lemma 9.17 after checking that $\widehat{\mu}_n(\xi) = (pe^{\mathrm{i}\xi} + 1 - p)^n$.

(2) $I = \mathbb{R}_+$ and, for every $t \in \mathbb{R}_+$, μ_t is the Poisson distribution of parameter t. We can use Lemma 9.17 after verifying that

$$\widehat{\mu}_t(\xi) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \, e^{\mathrm{i}k\xi} \, e^{-t} = \exp(-t(1 - e^{\mathrm{i}\xi})).$$

(3) $I = \mathbb{R}_+$ and, for every $t > 0$, μ_t is the Gamma $\Gamma(t, \theta)$ distribution where $\theta > 0$ is a fixed parameter. Here it is easier to use Laplace transforms. From the form of the density of the $\Gamma(t, \theta)$ distribution, one computes

$$\int e^{-\lambda x} \, \mu_t(\mathrm{d}x) = \Big(\frac{\theta}{\theta + \lambda}\Big)^t.$$

(4) $I = \mathbb{R}_+$ and, for every $t > 0$, μ_t is the Gaussian $\mathcal{N}(0, t)$ distribution. Indeed, by Lemma 8.15, we have

$$\widehat{\mu}_t(\xi) = \exp(-\frac{t\xi^2}{2}).$$

Important Consequences Let X and X' be two independent real random variables. We can rephrase examples (2), (3), (4) as follows.

- If X is Poisson with parameter λ and X' is Poisson with parameter λ', then $X + X'$ is Poisson with parameter $\lambda + \lambda'$.
- If X follows the $\Gamma(a, \theta)$ law and X' follows the $\Gamma(a', \theta)$ law, then $X + X'$ follows the $\Gamma(a + a', \theta)$ law.
- If X is Gaussian $\mathcal{N}(m, \sigma^2)$ and X' is Gaussian $\mathcal{N}(m', \sigma'^2)$, then $X + X'$ is Gaussian $\mathcal{N}(m + m', \sigma^2 + \sigma'^2)$. (Considering $X - m$ and $X' - m'$ shows that it is enough to deal with the case $m = m' = 0$.) Since mutiplying a Gaussian variable by a constant again gives a Gaussian variable, we also see that any linear combination of independent Gaussian variables is Gaussian.

9.7 The Poisson Process

In this section, we introduce and study the Poisson process, which is one of the most important random processes in probability theory (together with Brownian motion, which will be studied in Chapter 14). The study of the Poisson process will illustrate many of the preceding results about independence.

Throughout this section, we fix a parameter $\lambda > 0$. Let $U_1, U_2, \ldots$ be a sequence of independent and identically distributed random variables following the exponential distribution with parameter λ, which has density $\lambda\, e^{-\lambda x}$ on $\mathbb{R}_+$ (Section 9.4 tells us that we can construct such a sequence !). We then set, for every $n \in \mathbb{N}$,

$$T_n = U_1 + U_2 + \cdots + U_n,$$

and, for every real $t \geq 0$,

$$N_t = \sum_{n=1}^{\infty} \mathbf{1}_{\{T_n \leq t\}} = \sup\{n \in \mathbb{N} : T_n \leq t\}$$

with the usual convention $\sup \varnothing = 0$. Note that $T_n \longrightarrow \infty$ a.s. as $n \to \infty$ (as a consequence of Proposition 9.14). Thus, by disgarding the set of zero probability where the sequence (T_n) is bounded, we can assume that $N_t < \infty$ for every $t \geq 0$ and $\omega \in \Omega$. Similarly, since the random variables U_i are positive a.s., we can also assume that $0 < T_1(\omega) < T_2(\omega) < \cdots$ for every $\omega \in \Omega$.

For every fixed ω, the function $t \longrightarrow N_t(\omega)$ vanishes at 0, is increasing and right-continuous, and increases only by jumps of size 1 (see Fig. 9.1). Such a function is called a counting function. We also note that $N_t \longrightarrow \infty$ as $t \to \infty$.

Definition 9.18 The collection $(N_t)_{t \geq 0}$ is called the Poisson process of parameter λ.

The Poisson process is very often used in applied probability models, for instance in queueing theory, where N_t represents the number of customers arriving at a server before time t. The fact that the exponential distribution is chosen to model the period between two successive arrivals of customers is related to the property of lack of memory of this distribution (see Section 8.1.5). Roughly speaking, this property says that, for any given time $t \geq 0$, the time between t and the next arrival of a customer has always the same distribution independently of what has happened before time t (we will formulate this in more precise mathematical terms in Theorem 9.21 below).

Proposition 9.19 *For every integer $n \geq 1$, T_n follows the Gamma $\Gamma(n, \lambda)$ distribution, with density*

$$p(x) = \frac{\lambda^n}{(n-1)!}\, x^{n-1}\, e^{-\lambda x}\, \mathbf{1}_{\mathbb{R}_+}(x).$$

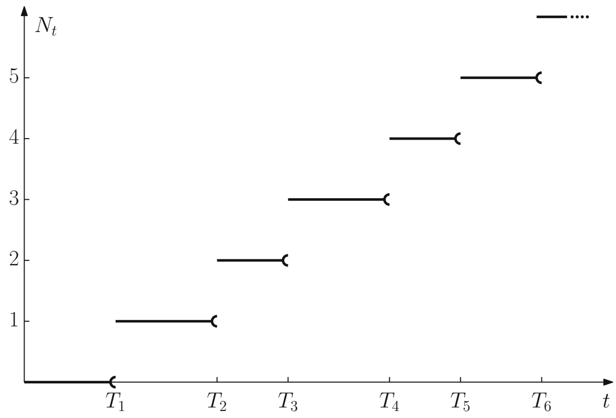

Fig. 9.1 Representation of the Poisson process. The quantities $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are independent exponential random variables with parameter λ

For every $t > 0$, N_t follows the Poisson distribution with parameter λt :

$$\mathbb{P}(N_t = k) = \frac{(\lambda t)^k}{k!}\, e^{-\lambda t} \ , \quad \forall k \in \mathbb{N}.$$

Proof For the first assertion, we note that the exponential distribution with parameter λ is just the the $\Gamma(1, \lambda)$ distribution, and we use the results about sums of independent random variables following Gamma distributions that are stated at the end of the preceding section.

To get the second assertion, we first write $\mathbb{P}(N_t = 0) = \mathbb{P}(T_1 > t) = e^{-\lambda t}$, and then, for every $k \geq 1$,

$$\begin{aligned}
\mathbb{P}(N_t = k) &= \mathbb{P}(T_k \leq t < T_{k+1}) \\
&= \mathbb{P}(T_k \leq t) - \mathbb{P}(T_{k+1} \leq t) \\
&= \int_0^t \frac{\lambda^k}{(k-1)!}\, x^{k-1}\, e^{-\lambda x}\, \mathrm{d}x - \int_0^t \frac{\lambda^{k+1}}{k!}\, x^k\, e^{-\lambda x}\, \mathrm{d}x \\
&= \frac{(\lambda t)^k}{k!}\, e^{-\lambda t},
\end{aligned}$$

where the last equality follows from an integration by parts. □

We will now state a first important result about the Poisson process. We need to introduce the notion of conditional probability given an event (much more about conditioning will be found in Chapter 11). If $B \in \mathcal{A}$ is such that $\mathbb{P}(B) > 0$, we define a new probability measure on $(\Omega, \mathcal{A})$, which is called the conditional probability knowing B and denoted by $\mathbb{P}(\cdot \mid B)$, by setting, for every $A \in \mathcal{A}$,

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

For every nonnegative random variable X, the expected value of X under $\mathbb{P}(\cdot \mid B)$ is denoted by $\mathbb{E}[X \mid B]$, and it is straightforward to verify that

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X\, \mathbf{1}_B]}{\mathbb{P}(B)}.$$

Proposition 9.20 *Let $t > 0$ and $n \in \mathbb{N}$. Under the conditional probability $\mathbb{P}(\cdot \mid N_t = n)$, the random vector $(T_1, T_2, \ldots, T_n)$ has density*

$$\frac{n!}{t^n}\, \mathbf{1}_{\{0<s_1<s_2<\cdots<s_n<t\}}.$$

Moreover, under the conditional probability $\mathbb{P}(\cdot \mid N_t = n)$, the random variable $T_{n+1} - t$ is exponentially distributed with parameter λ and is independent of $(T_1, \ldots, T_n)$.

Interpretation One easily verifies that the law of $(T_1, T_2, \ldots, T_n)$ is the distribution of the increasing reordering of n independent random variables uniformly distributed over $[0, t]$ (if $V_1, V_2, \ldots, V_n$ are n real random variables, their increasing reordering is the random vector $(\widetilde{V}_1, \widetilde{V}_2, \ldots, \widetilde{V}_n)$ such that $\widetilde{V}_1 \le \widetilde{V}_2 \le \cdots \le \widetilde{V}_n$ and, for every ω, values taken by $\widetilde{V}_1(\omega), \widetilde{V}_2(\omega), \ldots, \widetilde{V}_n(\omega)$, counted with their multiplicities, are the same as the values taken by $V_1(\omega), V_2(\omega), \ldots, V_n(\omega)$—see Exercise 8.14). We can thus reformulate the first part of the proposition by saying that, if we fix the number n of jumps of the Poisson process on the time interval $[0, t]$, the set of the corresponding jump times is distributed as the values of n independent random variables uniformly distributed over $[0, t]$.

Proof The density of the vector $(U_1, U_2, \ldots, U_{n+1})$ is the function

$$(x_1, \ldots, x_{n+1}) \mapsto \lambda^{n+1} \exp\Big(-\sum_{i=1}^{n+1} x_i\Big)\, \mathbf{1}_{\mathbb{R}_+^{n+1}}(x_1, \ldots, x_{n+1}).$$

If φ is a nonnegative measurable function on $\mathbb{R}_+^{n+1}$, we can then compute

$$\begin{aligned}
&\mathbb{E}[\varphi(T_1, T_2, \ldots, T_{n+1})\, \mathbf{1}_{\{N_t=n\}}]\\
&= \mathbb{E}[\varphi(T_1, T_2, \ldots, T_{n+1})\, \mathbf{1}_{\{T_n \le t < T_{n+1}\}}]\\
&= \mathbb{E}[\varphi(U_1, U_1 + U_2, \ldots, U_1 + \cdots + U_{n+1})\, \mathbf{1}_{\{U_1+\cdots+U_n \le t < U_1+\cdots+U_{n+1}\}}]
\end{aligned}$$

$$= \int_{\mathbb{R}_+^{n+1}} \varphi(x_1, x_1 + x_2, \dots, x_1 + \cdots + x_{n+1})$$

$$\times \mathbf{1}_{\{x_1+\cdots+x_n \le t < x_1+\cdots+x_{n+1}\}} \lambda^{n+1} \exp\Big(-\sum_{i=1}^{n+1} x_i\Big) \mathrm{d}x_1 \dots \mathrm{d}x_{n+1}$$

$$= \lambda^{n+1} \int_{\mathbb{R}_+^{n+1}} \mathbf{1}_{\{0<y_1<\cdots<y_n \le t < y_{n+1}\}} \exp(-\lambda y_{n+1}) \varphi(y_1, \dots, y_{n+1}) \mathrm{d}y_1 \dots \mathrm{d}y_{n+1}$$

using the change of variables $(x_1, \dots, x_n) \mapsto (x_1, x_1 + x_2, \dots, x_1 + \cdots + x_n)$ in the last equality. Since $\mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$, we get

$$\mathbb{E}[\varphi(T_1, T_2, \dots, T_{n+1}) \mid N_t = n] \tag{9.5}$$

$$= \mathbb{P}(N_t = n)^{-1} \, \mathbb{E}[\varphi(T_1, T_2, \dots, T_{n+1}) \, \mathbf{1}_{\{N_t = n\}}]$$

$$= \frac{n!}{t^n} \int_{\mathbb{R}_+^{n+1}} \mathbf{1}_{\{0<y_1<\cdots<y_n \le t < y_{n+1}\}} \, \lambda \, e^{-\lambda(y_{n+1}-t)} \, \varphi(y_1, \dots, y_{n+1}) \mathrm{d}y_1 \dots \mathrm{d}y_{n+1}.$$

In particular, if $\varphi(T_1, T_2, \dots, T_{n+1}) = \psi(T_1, T_2, \dots, T_n)$ only depends on $T_1, \dots, T_n$, we have

$$\mathbb{E}[\psi(T_1, T_2, \dots, T_n) \mid N_t = n]$$

$$= \frac{n!}{t^n} \int_{\{0<y_1<y_2<\cdots<y_n \le t\}} \psi(y_1, \dots, y_n) \Big(\int_t^\infty \lambda \, e^{-\lambda(y_{n+1}-t)} \, \mathrm{d}y_{n+1} \Big) \mathrm{d}y_1 \dots \mathrm{d}y_n$$

$$= \frac{n!}{t^n} \int_{\{0<y_1<\cdots<y_n \le t\}} \psi(y_1, \dots, y_n) \, \mathrm{d}y_1 \dots \mathrm{d}y_n,$$

giving the first assertion of the proposition.

To get the second assertion, we use (9.5) again, with the change of variable $z = y_{n+1} - t$ and the Fubini theorem, to get for any nonnegative measurable functions ψ and θ defined on $(\mathbb{R}_+)^n$ and on $\mathbb{R}_+$ respectively,

$$\mathbb{E}[\psi(T_1, T_2, \dots, T_n) \, \theta(T_{n+1} - t) \mid N_t = n]$$

$$= \frac{n!}{t^n} \int_{\mathbb{R}_+^n} \mathbf{1}_{\{0<y_1<\cdots<y_n \le t\}} \, \psi(y_1, y_2, \dots, y_n) \Big(\int_0^\infty \lambda \, e^{-\lambda z} \, \theta(z) \, \mathrm{d}z \Big) \mathrm{d}y_1 \dots \mathrm{d}y_n$$

$$= \mathbb{E}[\psi(T_1, T_2, \dots, T_n) \mid N_t = n] \times \int_0^\infty \lambda \, e^{-\lambda z} \, \theta(z) \, \mathrm{d}z,$$

In particular, under $\mathbb{P}(\cdot \mid N_t = n)$, $T_{n+1} - t$ is exponentially distributed with parameter λ (take $\psi = 1$), and the desired independence property also follows from the last display. $\square$

Remark The second part of the proposition also holds for $n = 0$: the fact that the law of $T_1 - t$ under $\mathbb{P}(\cdot \mid N_t = 0)$ is the exponential distribution with parameter λ is exactly the lack of memory of the exponential distribution.

We now state another very important theorem about the Poisson process.

Theorem 9.21 *Let $t > 0$. For every $r \geq 0$, set*

$$N_r^{(t)} := N_{t+r} - N_t.$$

Then the collection $(N_r^{(t)})_{r\geq 0}$ is again a Poisson process with parameter λ, and is independent of $(N_r)_{0\leq r\leq t}$.

Interpretation If we interpret the jump times of the Poisson process as the arrival times of customers at a server, the preceding theorem means that an observer arriving at time $t > 0$ and recording the arrivals of customers after that time will see (in distribution) the same thing as if he had arrived at time 0, and the knowledge of arrival times of customers between times 0 and t will give him no information on what happens after time t. This can be viewed as an aspect of the so-called "Markov property", which we will discuss in different settings in Chapters 13 and 14.

Proof Let us introduce the random variables

$$U'_1 = T_{N_t+1} - t$$

and, for every $i \geq 2$,

$$U'_i = T_{N_t+i} - T_{N_t+i-1} = U_{N_t+i}$$

(we leave it as an exercise to check that the U'_i's are indeed random variables). Also set $T'_n = U'_1 + U'_2 + \cdots + U'_n = T_{N_t+n} - t$ for every $n \geq 1$. By construction, for every $r \geq 0$,

$$N_r^{(t)} = N_{t+r} - N_t = \mathrm{card}\{i \geq 1 : t < T_i \leq t + r\} = \mathrm{card}\{n \geq 1 : T'_n \leq r\}.$$

To prove that $(N_r^{(t)})_{r\geq 0}$ is a Poisson process, it then suffices to verify that $U'_1, U'_2, \ldots$ are independent and exponentially distributed with parameter λ.

To this end, we fix an integer $n \geq 0$ and we will first condition on the event

$$\{N_t = n\} = \{U_1 + \cdots + U_n \leq n < U_1 + \cdots + U_{n+1}\},$$

which is measurable with respect to $\sigma(U_1, \dots, U_{n+1})$. Let $p \in \mathbb{N}$, and let $A_1, \dots, A_p$ be Borel subsets of $\mathbb{R}_+$. Then,

$$\begin{aligned}
&\mathbb{P}(U'_1 \in A_1, U'_2 \in A_2, \dots, U'_p \in A_p \mid N_t = n) \\
&= \frac{\mathbb{P}(\{N_t = n\} \cap \{T_{n+1} - t \in A_1\} \cap \{U_{n+2} \in A_2\} \cap \dots \cap \{U_{n+p} \in A_p\})}{\mathbb{P}(N_t = n)} \\
&= \frac{\mathbb{P}(\{N_t = n\} \cap \{T_{n+1} - t \in A_1\})}{\mathbb{P}(N_t = n)}\, \mathbb{P}(U_{n+2} \in A_2) \cdots \mathbb{P}(U_{n+p} \in A_p)
\end{aligned}$$

where we used the fact that the event $\{N_t = n\} \cap \{T_{n+1} - t \in A_1\}$ is $\sigma(U_1, \dots, U_{n+1})$-measurable, and thus independent of $(U_{n+1+k})_{k \in \mathbb{N}}$. From Proposition 9.20 (and the subsequent remark in the case $n = 0$), we know that the law of $T_{n+1} - t$ under $\mathbb{P}(\cdot \mid N_t = n)$ is the exponential distribution with parameter λ. So, writing μ_λ for this exponential distribution, we have obtained that

$$\mathbb{P}(U'_1 \in A_1, U'_2 \in A_2, \dots, U'_p \in A_p \mid N_t = n) = \mu_\lambda(A_1)\, \mu_\lambda(A_2) \cdots \mu_\lambda(A_p).$$

Finally, we have

$$\begin{aligned}
&\mathbb{P}(U'_1 \in A_1, U'_2 \in A_2, \dots, U'_p \in A_p) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N_t = n)\, \mathbb{P}(U'_1 \in A_1, U'_2 \in A_2, \dots, U'_p \in A_p \mid N_t = n) \\
&= \mu_\lambda(A_1)\, \mu_\lambda(A_2) \cdots \mu_\lambda(A_p),
\end{aligned}$$

which shows that $U'_1, U'_2, \dots$ are independent and exponentially distributed with parameter λ under $\mathbb{P}$.

We still have to prove that $(N_r^{(t)})_{r \geq 0}$ is independent of $(N_r)_{0 \leq r \leq t}$. It is clear that

$$\sigma(N_r^{(t)}, r \geq 0) \subset \sigma(T'_1, T'_2, \dots) = \sigma(U'_1, U'_2, \dots)$$

and therefore it is enough to prove that the sequence $(U'_1, U'_2, \dots)$ is independent of $(N_r)_{0 \leq r \leq t}$. Let us fix two integers $p \geq 2$ and $k \geq 1$. Let $r_1, r_2, \dots, r_k \in [0, t]$, let $\ell_1, \dots, \ell_k \in \mathbb{Z}_+$ and let $A_1, \dots, A_p$ be Borel subsets of $\mathbb{R}_+$. For every integer $n \geq 0$, we have

$$\begin{aligned}
&\mathbb{P}(\{N_{r_1} = \ell_1, \dots, N_{r_k} = \ell_k\} \cap \{U'_1 \in A_1, \dots, U'_p \in A_p\} \mid N_t = n) \\
&= \frac{1}{\mathbb{P}(N_t = n)} \times \mathbb{P}\Big(\{N_{r_1} = \ell_1, \dots, N_{r_k} = \ell_k, N_t = n\} \cap \{T_{n+1} - t \in A_1\} \\
&\qquad\qquad \cap \{U_{n+2} \in A_2\} \cap \dots \cap \{U_{n+p} \in A_p\}\Big)
\end{aligned}$$

$$= \frac{1}{\mathbb{P}(N_t = n)} \times \mathbb{P}\Big(\{N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k, N_t = n\} \cap \{T_{n+1} - t \in A_1\}\Big)$$
$$\times \mathbb{P}(U_{n+2} \in A_2) \times \cdots \times \mathbb{P}(U_{n+p} \in A_p)$$

where the last equality holds because the event

$$\{N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k, N_t = n\} \cap \{T_{n+1} - t \in A_1\}$$

is $\sigma(U_1, \ldots, U_{n+1})$-measurable (this event is empty if one of the numbers $\ell_1, \ldots, \ell_k$ is greater than n) and therefore independent of $(U_{n+2}, \ldots, U_{n+p})$. We have $\mathbb{P}(U_{n+2} \in A_2) \times \cdots \times \mathbb{P}(U_{n+p} \in A_p) = \mu_\lambda(A_2) \cdots \mu_\lambda(A_p)$, and then

$$\frac{1}{\mathbb{P}(N_t = n)} \times \mathbb{P}\Big(\{N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k, N_t = n\} \cap \{T_{n+1} - t \in A_1\}\Big)$$
$$= \mathbb{P}(\{N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k\} \cap \{T_{n+1} - t \in A_1\} \mid N_t = n)$$
$$= \mathbb{P}(N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k \mid N_t = n)\ \mu_\lambda(A_1),$$

using the second part of Proposition 9.20 together with the fact that, on the event $\{N_t = n\}$, $N_{r_1}, \ldots, N_{r_k}$ can be written as functions of $T_1, \ldots, T_n$. Finally, we have proved that

$$\mathbb{P}(\{N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k\} \cap \{U'_1 \in A_1, \ldots, U'_p \in A_p\} \mid N_t = n)$$
$$= \mathbb{P}(N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k \mid N_t = n)\ \mu_\lambda(A_1)\mu_\lambda(A_2)\ldots\mu_\lambda(A_p).$$

If we multiply both sides by $\mathbb{P}(N_t = n)$ and then sum over n, we get

$$\mathbb{P}(\{N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k\} \cap \{U'_1 \in A_1, \ldots, U'_p \in A_p\})$$
$$= \mathbb{P}(N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k)\ \mu_\lambda(A_1)\mu_\lambda(A_2)\ldots\mu_\lambda(A_p)$$
$$= \mathbb{P}(N_{r_1} = \ell_1, \ldots, N_{r_k} = \ell_k)\ \mathbb{P}(U'_1 \in A_1, \ldots, U'_p \in A_p).$$

Hence, $(U'_1, \ldots, U'_p)$ is independent of $(N_{r_1}, \ldots, N_{r_k})$. Recalling the end of Section 9.2, this suffices to get that the sequence $(U'_1, U'_2, \ldots)$ is independent of $(N_r)_{0 \le r \le t}$, which was the desired result. □

Corollary 9.22 *Let $t_0 = 0 \le t_1 \le \cdots \le t_k$. The random variables $N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}}$ are independent, and, for every $j \in \{1, \ldots, k\}$, $N_{t_j} - N_{t_{j-1}}$ follows the Poisson distribution with parameter $\lambda(t_j - t_{j-1})$.*

Proof By applying Proposition 9.19 and Theorem 9.21, we get that, for every $j \in \{1, \ldots, k\}$, $N_{t_j} - N_{t_{j-1}} = N^{(t_{j-1})}_{t_j - t_{j-1}}$ is a Poisson random variable with parameter $\lambda(t_j - t_{j-1})$ and is independent of $(N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_{j-1}} - N_{t_{j-2}})$. This last property easily implies the independence of the variables N_{t_1}, $N_{t_2} - N_{t_1}, \dots, N_{t_k} - N_{t_{k-1}}$. □

For any choice of $0 \le t_1 \le \dots \le t_k$, Corollary 9.22 describes the law of the vector $(N_{t_1}, N_{t_2} - N_{t_1}, \dots, N_{t_k} - N_{t_{k-1}})$ and therefore also (via a linear transformation) the law of $(N_{t_1}, N_{t_2}, \dots, N_{t_k})$. We say that we have described the *finite-dimensional marginal distributions* of the Poisson process.

9.8 Exercises

Exercise 9.1

(1) Let X and Y be two independent real random variables with the same law. Compute $\mathbb{P}(X = Y)$ in terms of the common law μ of X and Y. Show that $\mathbb{P}(X > Y) = \mathbb{P}(Y > X) > 0$, except in a particular case to be discussed.
(2) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{R}_+$. Show that $\sum_{n\in\mathbb{N}} X_n = \infty$ a.s., except in the special case where $X_n = 0$ a.s. for every n.
(3) Under the assumptions of the previous question, show that there is a constant $\ell \in [0, \infty]$ such that $\max\{X_1, \dots, X_n\} \longrightarrow \ell$ as $n \to \infty$, almost surely, and determine ℓ in terms of the distribution function of X_1.

Exercise 9.2 Let U and V be two independent real random variables distributed according to the exponential distribution with parameter $\lambda > 0$. Show that the variables $\frac{U}{U+V}$ and $U + V$ are independent and determine their law.

Exercise 9.3 Let N and N' be two independent Gaussian $\mathcal{N}(0, 1)$ random variables. Show that the random variable $N^2/(N^2 + N'^2)$ has density

$$\frac{1}{\pi} \frac{1}{\sqrt{t(1-t)}} \mathbf{1}_{(0,1)}(t).$$

This is the so-called arcsine distribution.

Exercise 9.4 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables uniformly distributed over $\{1, 2, \dots, p\}$. For every $n \in \mathbb{N}$, determine the law of $M_n = \max\{X_1, \dots, X_n\}$, and show that $\mathbb{E}[M_n]/p \longrightarrow n/(n + 1)$ when $p \to \infty$.

Exercise 9.5 Let $A_0, A_1, A_2, \dots$ be a sequence of independent events. For every $\omega \in \Omega$, set

$$T(\omega) = \inf\{n \ge 0 : \omega \in A_n\},$$

with the convention $\inf \varnothing = \infty$. Verify that T is a random variable and give its distribution in terms of the numbers $p_n = \mathbb{P}(A_n)$. What condition on the p_n's ensures that $T < \infty$ a.s. ? In the case where $p_n = p \in (0, 1)$ for every n, identify the distribution of T and compute $\mathbb{E}[T]$ and $\mathrm{var}(T)$.

Exercise 9.6 A real random variable X is called symmetric if X and $-X$ have the same law.

(1) Let X be a symmetric random variable, whose law has a density f. Show that f can be chosen such that $f(x) = f(-x)$ for every $x \in \mathbb{R}$.
(2) Show that a real random variable X is symmetric if and only if its characteristic function takes values in $\mathbb{R}$.
(3) Let Y and Y' be two independent real random variables with the same distribution. Show that $Y - Y'$ is symmetric. Does this still hold without the independence assumption ?
(4) Let ε be a random variable with values in $\{-1, 1\}$ such that $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$. Show that, if X is a symmetric random variable and X is independent of ε, then $\varepsilon|X|$ has the same distribution as X.

Exercise 9.7 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{R}_+$. Show that, if $\mathbb{E}[X_1] < \infty$,

$$\limsup_{n\to\infty} \frac{X_n}{n} = 0\,, \quad \text{a.s.,}$$

whereas, if $\mathbb{E}[X_1] = \infty$,

$$\limsup_{n\to\infty} \frac{X_n}{n} = \infty\,, \quad \text{a.s.}$$

Exercise 9.8 Let $\alpha > 0$ and let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with values in $\{0, 1\}$, such that, for every $n \in \mathbb{N}$,

$$\mathbb{P}(Z_n = 1) = \frac{1}{n^\alpha} \quad \text{and} \quad \mathbb{P}(Z_n = 0) = 1 - \frac{1}{n^\alpha}.$$

Verify that $Z_n \longrightarrow 0$ as $n \to \infty$ in L^1, but nonetheless we have a.s.

$$\limsup_{n\to\infty} Z_n = \begin{cases} 1 \text{ if } \alpha \leq 1, \\ 0 \text{ if } \alpha > 1 \end{cases}$$

Exercise 9.9 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables. Assume that there exists a constant C such that $\mathbb{E}[(X_n)^2] \leq C$ for every $n \in \mathbb{N}$, and that $\mathrm{cov}(X_n, X_m) = 0$ if $n \neq m$.

(1) Verify that

$$\frac{S_{n^2} - \mathbb{E}[S_{n^2}]}{n^2} \underset{n\to\infty}{\longrightarrow} 0\,, \quad \text{a.s.}$$

(2) Deduce from question (1) that we have also

$$\frac{S_n - \mathbb{E}[S_n]}{n} \underset{n\to\infty}{\longrightarrow} 0\,, \quad \text{a.s.}$$

Exercise 9.10 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables distributed according to the exponential distribution with parameter 1.

(1) Prove that

$$\limsup_{n\to\infty} (\log n)^{-1} X_n = 1, \quad \text{a.s.}$$

(2) Let $Z_n = \max\{X_1, \ldots, X_n\}$. Verify that

$$\liminf_{n\to\infty} (\log n)^{-1} Z_n \geq 1, \quad \text{a.s.}$$

(3) Verify that, for an appropriate sequence $n_k \uparrow \infty$, one has

$$\limsup_{k\to\infty} (\log n_k)^{-1} Z_{n_k} \leq 1, \quad \text{a.s.}$$

Then show that $\lim_{n\to\infty} (\log n)^{-1} Z_n = 1$, a.s.

Exercise 9.11

(1) Let N and N' be two independent Gaussian $\mathcal{N}(0, 1)$ random variables. Show that $X = N/N'$ follows a Cauchy distribution with density $(\pi(1 + x^2))^{-1}$.
(2) Compute the characteristic function of X. (*Hint:* Verify that $\mathbb{E}[e^{i\xi X}] = (2\pi)^{-1/2}\int_{\mathbb{R}} \exp\big(-\frac{1}{2}(y - \frac{|\xi|}{y})^2 - |\xi|\big)\mathrm{d}y$ and then use the result of Exercise 7.5.)
(3) Let $X_1, \ldots, X_n$ be n independent random variables with the same distribution as X. Show that $\frac{1}{n}(X_1 + \cdots + X_n)$ also has the same distribution as X. Why does this not contradict the weak law of large numbers ?
(4) Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables with a symmetric distribution (Y_n has the same law as $-Y_n$). Assume that $\frac{1}{n}(Y_1 + \cdots + Y_n)$ has the same distribution as Y_1, for every $n \in \mathbb{N}$. Show that Y_n follows a Cauchy distribution.

Exercise 9.12

(1) Let $a, b \in \mathbb{R}$ with $a < 0 < b$. If Y is a random variable with values in $[a, b]$, verify that $\mathrm{var}(Y) \leq (b - a)^2/4$.

(2) Let Z be a centered random variable with values in $[a, b]$, and, for every $\lambda \geq 0$, set $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$. Prove that, for every $\lambda \geq 0$,

$$\psi_Z(\lambda) \leq \frac{(b-a)^2}{8}\,\lambda^2.$$

(*Hint:* Verify that the second derivative $\psi_Z''(\lambda)$ makes sense and is equal to the variance of Z under a probability measure absolutely continuous with respect to $\mathbb{P}$).

(3) Let $X_1, \ldots, X_n$ be independent real random variables such that, for every $i \in \{1, \ldots, n\}$, X_i takes values in $[a_i, b_i]$, where $a_i < 0 < b_i$. Prove that, for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i - \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] \geq \varepsilon\Big) \leq \exp\Bigg(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Bigg).$$

This is known as *Hoeffding's inequality*.

Exercise 9.13 Let $U_1, \ldots, U_n$ be independent random variables with values in $\{-1, 1\}$ such that $\mathbb{P}(U_j = 1) = \mathbb{P}(U_j = -1) = 1/2$ for every $j \in \{1, \ldots, n\}$. Let $a_1, \ldots, a_n \in \mathbb{R}$. Prove that

$$\mathbb{E}\Big[\Big|\sum_{j=1}^{n} a_j U_j\Big|\Big] \geq \sqrt{\frac{1}{3}\sum_{j=1}^{n} a_j^2}.$$

This is a particular case of the *Khintchine inequality* (the constant $1/3$ can be replaced by $1/2$, but this requires more work). *Hint:* Verify that, if X is a real random variable in L^4,

$$\mathbb{E}[|X|] \geq \frac{\mathbb{E}[X^2]^{3/2}}{\mathbb{E}[X^4]^{1/2}}.$$

Chapter 10
Convergence of Random Variables

The first section of this chapter presents the different notions of convergence of random variables, and the relations between them. In particular, we introduce and discuss the convergence in probability of a sequence of random variables. For simplicity, we restrict our attention to random variables with values in $\mathbb{R}^d$, although many of the concepts and results that follow can be extended to random variables with values in more general metric spaces. We then prove the strong law of large numbers, which is one of the fundamental limit theorems of probability theory. A useful ingredient is Kolmogorov's zero-one law, which roughly speaking says that an event depending only on the asymptotic behavior of an independent sequence must have probability zero or one. The third section discusses the convergence in distribution of random variables. This convergence is more delicate to grasp, partly because it is a convergence of the laws of the random variables in consideration, and not of the random variables themselves. We provide a detailed proof of Lévy's theorem characterizing the convergence in distribution in terms of characteristic functions. This result yields an easy proof of the central limit theorem, which is another fundamental limit theorem of probability theory. The last section is devoted to the multidimensional central limit theorem, whose statement involves the important notion of a Gaussian vector.

10.1 The Different Notions of Convergence

Throughout this chapter, we argue on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $(X_n)_{n\in\mathbb{N}}$ and X be random variables with values in $\mathbb{R}^d$. We have already encountered several notions of convergence of the sequence $(X_n)_{n\in\mathbb{N}}$ towards X. In particular, the almost

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_10

sure convergence is defined by saying that

$$X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \quad \text{if } \mathbb{P}(\{\omega \in \Omega : X(\omega) = \lim_{n\to\infty} X_n(\omega)\}) = 1.$$

Let $p \in [1, \infty)$, and assume that $\mathbb{E}[|X|^p] < \infty$ and $\mathbb{E}[|X_n|^p] < \infty$ for every $n \in \mathbb{N}$. We can define the convergence of X_n to X in L^p by

$$X_n \xrightarrow[n\to\infty]{L^p} X \quad \text{if } \lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

This is equivalent to saying that each component sequence of $(X_n)_{n\in\mathbb{N}}$ converges to the corresponding component of X in the Banach space $L^p(\Omega, \mathcal{A}, \mathbb{P})$. One could also consider the convergence in L^∞, but this convergence is very rarely used in probability theory and we will not discuss it.

Definition 10.1 We say that the sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X, and we write

$$X_n \xrightarrow[n\to\infty]{(\mathbb{P})} X$$

if, for every $\varepsilon > 0$,

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

Proposition 10.2 *Let $\mathcal{L}^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ be the space of all random variables with values in $\mathbb{R}^d$, and let $L^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ be the quotient space of $\mathcal{L}^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ for the equivalence relation defined by saying that $X \sim Y$ if and only if $\mathbb{P}(X = Y) = 1$. Then the formula*

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

defines a distance on $L^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$, and this distance is compatible with the convergence in probability, meaning that a sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X if and only if $\mathrm{d}(X_n, X)$ tends to 0 as $n \to \infty$. Moreover, the space $L^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ is complete for the distance d.

Proof It is very easy to verify that d is a distance on $L^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ (in particular, $\mathrm{d}(X, Y) = 0$ implies $|X - Y| \wedge 1 = 0$ a.s. and thus $\mathbb{P}(X = Y) = 1$). Then, if the sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X, we have for every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}[|X_n - X| \wedge 1] &\le \mathbb{E}[|X_n - X| \mathbf{1}_{\{|X_n - X| \le \varepsilon\}}] + \mathbb{E}[(|X_n - X| \wedge 1)\mathbf{1}_{\{|X_n - X| > \varepsilon\}}] \\ &\le \varepsilon + \mathbb{P}(|X_n - X| > \varepsilon).\end{aligned}$$

From the definition of the convergence in probability, this implies that $\limsup \mathrm{d}(X_n, X) \le \varepsilon$, and since ε was arbitrary, we have proved that $\mathrm{d}(X_n, X) \longrightarrow 0$. Conversely, if $\mathrm{d}(X_n, X) \longrightarrow 0$, the Markov inequality gives, for every $\varepsilon \in (0, 1)$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \le \varepsilon^{-1}\mathbb{E}[|X_n - X| \wedge 1] = \varepsilon^{-1}\mathrm{d}(X_n, X) \underset{n\to\infty}{\longrightarrow} 0.$$

It remains to prove that $L^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ is complete for the distance d. So let $(X_n)_{n\in\mathbb{N}}$ be a Cauchy sequence for the distance d. We can then find a subsequence $Y_k = X_{n_k}$, $k \in \mathbb{N}$, such that for every $k \ge 1$,

$$\mathrm{d}(Y_k, Y_{k+1}) \le 2^{-k}.$$

Then,

$$\mathbb{E}\Big[\sum_{k=1}^{\infty}(|Y_{k+1} - Y_k| \wedge 1)\Big] = \sum_{k=1}^{\infty}\mathrm{d}(Y_k, Y_{k+1}) < \infty,$$

which implies that $\sum_{k=1}^{\infty}(|Y_{k+1} - Y_k| \wedge 1) < \infty$ a.s., and therefore also $\sum_{k=1}^{\infty}|Y_{k+1} - Y_k| < \infty$ a.s. (there are a.s. only finitely many values of k such that $|Y_{k+1} - Y_k| \ge 1$). We then define a random variable in $L^0_{\mathbb{R}^d}(\Omega, \mathcal{A}, \mathbb{P})$ by setting

$$X = Y_1 + \sum_{k=1}^{\infty}(Y_{k+1} - Y_k),$$

on the event where the series converges absolutely, and $X = 0$ on the complementary event, which has zero probability. By construction, the sequence $(Y_k)_{k\in\mathbb{N}}$ converges a.s. to X, and this implies

$$\mathrm{d}(Y_k, X) = \mathbb{E}[|Y_k - X| \wedge 1] \underset{k\to\infty}{\longrightarrow} 0,$$

by an application of the dominated convergence theorem. Hence the sequence $(Y_k)_{k\in\mathbb{N}}$ converges to X for the distance d, and since the sequence $(X_n)_{n\in\mathbb{N}}$ is Cauchy and has a convergent subsequence, it must also converge. □

A by-product of the preceding proof is the fact that any sequence that converges in probability has a subsequence that converges a.s. (compare with Proposition 4.6). We state this in the following proposition.

Proposition 10.3 *If the sequence $(X_n)_{n\in\mathbb{N}}$ converges to X a.s., or in L^p for some $p \in [1, \infty]$, then it also converges to X in probability. Conversely, if the sequence $(X_n)_{n\in\mathbb{N}}$ converges to X in probability, there is a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ that converges a.s. to X.*

Proof If X_n converge a.s. to X, then,

$$\mathrm{d}(X_n, X) = \mathbb{E}[|X_n - X| \wedge 1] \underset{n\to\infty}{\longrightarrow} 0,$$

by dominated convergence. If X_n converges to X in L^p, then

$$\mathrm{d}(X_n, X) \le \mathbb{E}[|X_n - X|] = \||X_n - X|\|_1 \le \||X_n - X|\|_p \underset{n\to\infty}{\longrightarrow} 0.$$

The last assertion has been derived in the proof of Proposition 10.2. □

Remark The second part of the proposition has the following interesting consequence. In the dominated convergence theorem applied to a sequence of random variables $(X_n)_{n\in\mathbb{N}}$, the assumption of almost sure convergence of the sequence can be replaced by convergence in probability. To see this, observe that the sequence (X_n) converges to X in L^1 if and only if, from any subsequence of (X_n), one can extract another subsequence that converges to X in L^1, and then note that the second part of Proposition 10.3 ensures that the latter property holds.

To summarize, the convergence in probability is weaker than both the almost sure convergence and the convergence in L^p for any $p \in [1, \infty]$. In the reverse direction, the convergence in probability implies the almost sure convergence along a subsequence, and the next proposition gives conditions that make it possible to derive the convergence in L^p from the convergence in probability.

Proposition 10.4 *Suppose that the sequence $(X_n)_{n\in\mathbb{N}}$ converges to X in probability, and that there exists $r \in (1, \infty)$ such that the sequence $(\mathbb{E}[|X_n|^r])_{n\in\mathbb{N}}$ is bounded. Then $\mathbb{E}[|X|^r] < \infty$, and, for every $p \in [1, r)$, the sequence $(X_n)_{n\in\mathbb{N}}$ converges to X in L^p.*

Proof By assumption, there is a constant C such that $\mathbb{E}[|X_n|^r] \le C$ for every n. By Proposition 10.3 we can find a subsequence (X_{n_k}) that converges a.s. to X. By Fatou's lemma, we have

$$\mathbb{E}[|X|^r] = \mathbb{E}\Big[\liminf_{k\to\infty} |X_{n_k}|^r\Big] \le \liminf_{k\to\infty} \mathbb{E}[|X_{n_k}|^r] \le C.$$

Then, using the Hölder inequality, we have, for every $p \in [1, r)$ and every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}[|X_n - X|^p] &= \mathbb{E}[|X_n - X|^p \mathbf{1}_{\{|X_n - X| \le \varepsilon\}}] + \mathbb{E}[|X_n - X|^p \mathbf{1}_{\{|X_n - X| > \varepsilon\}}] \\ &\le \varepsilon^p + \mathbb{E}[|X_n - X|^r]^{p/r}\, \mathbb{P}(|X_n - X| > \varepsilon)^{1-p/r} \\ &\le \varepsilon^p + (2C^{1/r})^p\, \mathbb{P}(|X_n - X| > \varepsilon)^{1-p/r}.\end{aligned}$$

Using the fact that X_n converges in probability to X, we get

$$\limsup_{n\to\infty} \mathbb{E}[|X_n - X|^p] \leq \varepsilon^p$$

which gives the desired result since ε was arbitrary. □

The next proposition known as Scheffé's lemma gives a useful criterion allowing one to deduce L^1-convergence from almost sure convergence in the case of nonnegative real random variables.

Proposition 10.5 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative random variables in L^1. Assume that X_n converges in probability to X as $n \to \infty$, and $X \in L^1$. Then the condition $\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X]$ as $n \to \infty$ implies that X_n converges to X in L^1.*

Proof We write

$$\mathbb{E}[|X_n - X|] = \mathbb{E}[X_n - X] + 2\mathbb{E}[(X_n - X)^-].$$

We have $\mathbb{E}[X_n - X] \longrightarrow 0$, and on the other hand, the bound $(X_n - X)^- \leq X$ allows us to apply the dominated convergence theorem in order to get that $\mathbb{E}[(X_n - X)^-] \longrightarrow 0$ (we use the fact that the assumption of almost sure convergence in the dominated convergence theorem can be replaced by convergence in probability, see the remark following the proof of Proposition 10.3). □

Figure 10.1 illustrates the relations between the different types of convergence that we have introduced. The dashed lines give partial converses (under the stated conditions) that correspond to results stated in Propositions 10.3 and 10.4, and to the dominated convergence theorem.

As a final remark, the condition required to apply the dominated convergence theorem ($|X_n| \leq Z$ with $\mathbb{E}[Z] < \infty$) is not the best one in order to deduce L^1-convergence from convergence in probability. In fact, this condition can be replaced by the (weaker) property of *uniform integrability* of the sequence $(X_n)_{n\in\mathbb{N}}$, which will be discussed in Section 12.5 below (see in particular Theorem 12.27).

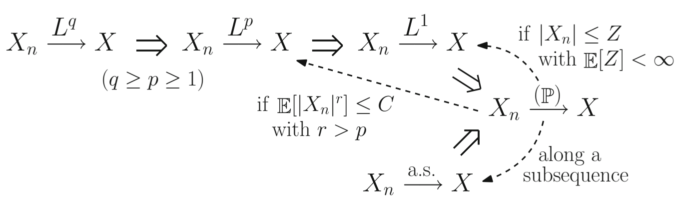

Fig. 10.1 Relations between the different convergences

10.2 The Strong Law of Large Numbers

We start with an important preliminary result.

Theorem 10.6 (Kolmogorov's Zero-One Law) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables taking values in arbitrary measurable spaces. For every $n \in \mathbb{N}$, let $\mathcal{B}_n$ be the σ-field*

$$\mathcal{B}_n = \sigma(X_k\,;\; k \geq n).$$

Define the asymptotic σ-field $\mathcal{B}_\infty$ by

$$\mathcal{B}_\infty := \bigcap_{n=1}^{\infty} \mathcal{B}_n.$$

Then $\mathbb{P}(B) = 0$ or 1, for every $B \in \mathcal{B}_\infty$.

Proof For every $n \in \mathbb{N}$, set

$$\mathcal{D}_n = \sigma(X_k\,;\; k \leq n).$$

Proposition 9.10 shows that $\mathcal{D}_n$ is independent of $\mathcal{B}_{n+1}$, hence a fortiori $\mathcal{D}_n$ is independent of $\mathcal{B}_\infty$. Consequently,

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

Since the class $\bigcup_{n=1}^{\infty} \mathcal{D}_n$ is closed under finite intersections, Proposition 9.7 implies that $\mathcal{B}_\infty$ is independent of

$$\sigma\Big(\bigcup_{n=1}^{\infty} \mathcal{D}_n\Big) = \sigma(X_n\,;\; n \geq 1).$$

In particular, $\mathcal{B}_\infty$ is independent of itself, but this means that, for every $B \in \mathcal{B}_\infty$, $\mathbb{P}(B) = \mathbb{P}(B \cap B) = \mathbb{P}(B)^2$, which is only possible if $\mathbb{P}(B) = 0$ or 1. □

Let us give a simple illustration of Theorem 10.6. In the setting of this theorem, a random variable Y taking values in $[-\infty, \infty]$ which is $\mathcal{B}_\infty$-measurable is necessarily equal to a constant a.s. (this can be verified by observing that the distribution function $x \mapsto \mathbb{P}(Y \leq x)$ takes values in $\{0, 1\}$). Let us then consider a sequence $(X_n)_{n\in\mathbb{N}}$ of independent real random variables. The random variable

$$Z := \limsup_{n\to\infty} \frac{1}{n}(X_1 + \cdots + X_n),$$

which takes values in $[-\infty, \infty]$, is $\mathcal{B}_\infty$-measurable, since we have also, for every $k \in \mathbb{N}$,

$$Z = \limsup_{n\to\infty} \frac{1}{n}(X_k + X_{k+1} + \cdots + X_n)$$

which shows that Z is $\mathcal{B}_k$-measurable for every $k \in \mathbb{N}$. Theorem 10.6 thus implies that Z is constant a.s. In particular, if we know that $\frac{1}{n}(X_1 + \cdots + X_n)$ converges a.s., the limit must be constant.

Before using the zero-one law to establish the strong law of large numbers, we give an easier application to the so-called coin-tossing process (also called simple random walk on $\mathbb{Z}$).

Proposition 10.7 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\{-1, 1\}$, such that $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = \frac{1}{2}$. For every $n \in \mathbb{N}$, set*

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then we have a.s.

$$\sup_{n\in\mathbb{N}} S_n = +\infty \quad \textit{and} \quad \inf_{n\in\mathbb{N}} S_n = -\infty.$$

Consequently, $\{n \in \mathbb{N} : S_n = 0\}$ is unbounded a.s.

In other words, we may imagine a coin-tossing game where at each step the player wins or loses one Euro with probability $1/2$, and S_n then represents the (positive, negative or zero) gain accumulated after n steps. The proposition shows that, as $n \to \infty$, the gain will alternate between positive and negative values, and therefore there will be arbitrarily large times at which the gain is 0.

Proof We start by proving that, for every $p \in \mathbb{N}$,

$$\mathbb{P}(-p \le \inf_{n\in\mathbb{N}} S_n \le \sup_{n\in\mathbb{N}} S_n \le p) = 0.$$

To this end, fix an integer $k > 2p$, and note that

$$\bigcup_{j=0}^{\infty} \{X_{jk+1} = X_{jk+2} = \cdots = X_{jk+k} = 1\} \subset (\{-p \le \inf_n S_n \le \sup_n S_n \le p\})^c.$$

However, an application of the Borel-Cantelli lemma (see Section 9.3 for a similar argument) shows that the set in the left-hand side has probability 1, and the desired result follows.

If we let p tend to ∞, we get

$$\mathbb{P}(\{\inf_{n\in\mathbb{N}} S_n > -\infty\} \cap \{\sup_{n\in\mathbb{N}} S_n < \infty\}) = 0,$$

and therefore,

$$\mathbb{P}(\{\inf_{n\in\mathbb{N}} S_n = -\infty\} \cup \{\sup_{n\in\mathbb{N}} S_n = \infty\}) = 1.$$

In particular,

$$\mathbb{P}(\{\inf_{n\in\mathbb{N}} S_n = -\infty\}) + \mathbb{P}(\{\sup_{n\in\mathbb{N}} S_n = \infty\}) \geq 1.$$

A symmetry argument shows that

$$\mathbb{P}(\{\inf_{n\in\mathbb{N}} S_n = -\infty\}) = \mathbb{P}(\{\sup_{n\in\mathbb{N}} S_n = \infty\})$$

and this probability is positive by the preceding display. To complete the proof, we observe that

$$\{\sup_{n\in\mathbb{N}} S_n = \infty\} \in \mathcal{B}_\infty.$$

Indeed, for every $k \in \mathbb{N}$,

$$\{\sup_{n\in\mathbb{N}} S_n = \infty\} = \{\sup_{n\geq k}(X_k + X_{k+1} + \cdots + X_n) = \infty\} \in \mathcal{B}_k$$

so that the event $\{\sup_n S_n = \infty\}$ is measurable with respect to the intersection of the σ-fields $\mathcal{B}_k$, which is the σ-field $\mathcal{B}_\infty$. The zero-one law now shows that $\mathbb{P}(\{\sup_n S_n = \infty\}) = 1$. □

We now come to the main result of this section. Notice that a first version of this result, under stronger integrability assumptions, already appeared as Proposition 9.14 (see also Exercise 9.9).

Theorem 10.8 (Strong law of large numbers) *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed real random variables in L^1. Then,*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1].$$

Remarks

(i) The L^1 integrability assumption is optimal in the sense that it is needed for the limit $\mathbb{E}[X_1]$ to be defined (and finite). If we remove the L^1 assumption but assume instead that the random variables X_n are nonnegative and $\mathbb{E}[X_1] = \infty$, it is easy to obtain that

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow[n\to\infty]{\text{a.s.}} +\infty$$

by applying the theorem to the variables $X_n \wedge K$, for every $K \in \mathbb{N}$.

(ii) One can show that the convergence of the theorem also holds in L^1. The proof will be given at the end of Chapter 12 as an application of martingale theory. As we already mentioned, the almost sure convergence is the most interesting one from a probabilistic point of view.

Proof To simplify notation, we set $S_n = X_1 + \cdots + X_n$ for every $n \in \mathbb{N}$, and $S_0 = 0$. Let $a > \mathbb{E}[X_1]$, and

$$M := \sup_{n\in\mathbb{Z}_+} (S_n - na),$$

which is a random variable with values in $[0, \infty]$. We will show that

$$M < \infty\ , \quad \text{a.s.} \tag{10.1}$$

Assume that (10.1) holds (for any choice of $a > \mathbb{E}[X_1]$). Since $S_n \le na + M$ for every $n \in \mathbb{N}$, it immediately follows from (10.1) that

$$\limsup_{n\to\infty} \frac{1}{n} S_n \le a\ , \quad \text{a.s.}$$

By considering a sequence of values of a that decreases to $\mathbb{E}[X_1]$, we find

$$\limsup_{n\to\infty} \frac{1}{n} S_n \le \mathbb{E}[X_1]\ , \quad \text{a.s.}$$

Replacing X_n by $-X_n$, we get the reverse inequality

$$\liminf_{n\to\infty} \frac{1}{n} S_n \ge \mathbb{E}[X_1]\ , \quad \text{a.s.}$$

and the convergence in the theorem follows.

It remains to prove (10.1). With the notation of Theorem 10.6, we first observe that $\{M < \infty\}$ belongs to $\mathcal{B}_\infty$. Indeed, we have, for every integer $k \in \mathbb{N}$,

$$\{M < \infty\} = \{\sup_{n\in\mathbb{Z}_+} (S_n - na) < \infty\} = \{\sup_{n\ge k}(S_n - S_k - na) < \infty\}$$

and the event in the right-hand side is $\mathcal{B}_{k+1}$-measurable. It will therefore be enough to prove that $\mathbb{P}(M < \infty) > 0$, or equivalently that $\mathbb{P}(M = \infty) < 1$. In the remaining part of the proof we verify that $\mathbb{P}(M = \infty) < 1$.

For every $k \in \mathbb{N}$, set

$$M_k := \sup_{0 \le n \le k} (S_n - na),$$

$$M'_k := \sup_{0 \le n \le k} (S_{n+1} - S_1 - na).$$

Then, M_k and M'_k are nonnegative random variables with the same distribution, as a consequence of the fact that the random vectors $(X_1, \dots, X_k)$ and $(X_2, \dots, X_{k+1})$ have the same law. It follows that

$$M = \lim_{k \to \infty} \uparrow M_k$$

and

$$M' := \lim_{k \to \infty} \uparrow M'_k$$

also have the same law (just observe that $\mathbb{P}(M' \le x) = \lim \downarrow \mathbb{P}(M'_k \le x) = \lim \downarrow \mathbb{P}(M_k \le x) = \mathbb{P}(M \le x)$ for every $x \in \mathbb{R}$).

On the other hand, it follows from the definitions that, for every integer $k \ge 0$,

$$M_{k+1} = \sup\Big(0, \sup_{1 \le n \le k+1} (S_n - na)\Big) = \sup(0, M'_k + X_1 - a),$$

which can also be written as

$$M_{k+1} = M'_k - \inf(a - X_1, M'_k).$$

Since M'_k has the same law as M_k (and both these variables are clearly in L^1), we get

$$\mathbb{E}[\inf(a - X_1, M'_k)] = \mathbb{E}[M'_k] - \mathbb{E}[M_{k+1}] = \mathbb{E}[M_k] - \mathbb{E}[M_{k+1}] \le 0$$

thanks to the trivial inequality $M_k \le M_{k+1}$. We may now apply the dominated convergence theorem to the sequence $\inf(a - X_1, M'_k)$, $k \in \mathbb{N}$, noting that these variables are bounded in absolute value by $|a - X_1|$ (recall that $M'_k \ge 0$). It follows that

$$\mathbb{E}[\inf(a - X_1, M')] = \lim_{k \to \infty} \mathbb{E}[\inf(a - X_1, M'_k)] \le 0.$$

We now argue by contradiction to prove that $\mathbb{P}(M = \infty) < 1$. Suppose that $\mathbb{P}(M = \infty) = 1$, then we have also $\mathbb{P}(M' = \infty) = 1$, since M and M' have the same law, and it follows that $\inf(a - X_1, M') = a - X_1$ a.s. But then the inequality in the last display gives $\mathbb{E}[a - X_1] \leq 0$, which is a contradiction since we chose $a > \mathbb{E}[X_1]$. This contradiction completes the proof. □

Remark The preceding proof can be found in Neveu [16]. It is different from the "classical" proofs that appear in most textbooks on probability theory, which involve making an appropriate truncation of the random variables X_n (typically, the first step of the argument is to replace X_n by $X_n \mathbf{1}_{\{X_n \leq n\}}$, in a way similar to Exercise 12.14 below). In Chapter 12, we will present three other proofs of the strong law of large numbers, which depend more or less on martingale theory. See Section 12.7 and Exercises 12.9 and 12.14. The proof presented in Exercise 12.9, which uses very little of martingale theory, is arguably the simplest one.

10.3 Convergence in Distribution

Recall that $C_b(\mathbb{R}^d)$ denote the set of all bounded continuous functions from $\mathbb{R}^d$ into $\mathbb{R}$. We equip $C_b(\mathbb{R}^d)$ with the supremum norm

$$\|\varphi\| = \sup_{x \in \mathbb{R}^d} |\varphi(x)|.$$

We let $M_1(\mathbb{R}^d)$ denote the set of all probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Definition 10.9 A sequence $(\mu_n)_{n\in\mathbb{N}}$ in $M_1(\mathbb{R}^d)$ converges weakly to $\mu \in M_1(\mathbb{R}^d)$ if

$$\forall \varphi \in C_b(\mathbb{R}^d) , \quad \int \varphi \, \mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int \varphi \, \mathrm{d}\mu.$$

A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables with values in $\mathbb{R}^d$ converges in distribution to a random variable X with values in $\mathbb{R}^d$ if the sequence $(\mathbb{P}_{X_n})_{n\in\mathbb{N}}$ converges weakly to $\mathbb{P}_X$. By Proposition 8.5, this is equivalent to saying that

$$\forall \varphi \in C_b(\mathbb{R}^d) , \quad \mathbb{E}[\varphi(X_n)] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[\varphi(X)].$$

We will write $\mu_n \overset{\text{(w)}}{\longrightarrow} \mu$, resp. $X_n \overset{\text{(d)}}{\longrightarrow} X$, if the sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ, resp. if $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X.

Remarks

(i) There is some abuse of terminology in saying that the sequence $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X, because the limiting random variable X is not determined uniquely (only its distribution $\mathbb{P}_X$ is determined). For this reason, we will sometimes write that a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables converges in distribution to a probability measure μ — one should of course understand that the laws $\mathbb{P}_{X_n}$ converge weakly to μ. We may also note that the convergence in distribution makes sense even if the random variables X_n, $n \in \mathbb{N}$ are defined on different probability spaces (in the present book, we will however always assume that they are defined on the same probability space). This makes the convergence in distribution very different from the other types of convergence studied in this chapter.

(ii) The space $M_1(\mathbb{R}^d)$ may be seen as a subspace of the topological dual of $C_b(\mathbb{R}^d)$. The weak convergence on $M_1(\mathbb{R}^d)$ then corresponds to the so-called weak*-topology on the dual.

Examples

(a) Suppose that the random variables $(X_n)_{n\in\mathbb{N}}$ and X take values in $\mathbb{Z}^d$. Then X_n converges in distribution to X if and only if

$$\forall x \in \mathbb{Z}^d \;, \quad \mathbb{P}(X_n = x) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X = x)$$

(the "if" part requires a little work, but will be immediate when we have established that $C_b(\mathbb{R}^d)$ can be replaced by $C_c(\mathbb{R}^d)$ in the definition of the weak convergence, see Proposition 10.12 below).

(b) Suppose that, for every $n \in \mathbb{N}$, X_n has a density $p_n(x)$, and that there exists a probability density function $p(x)$ on $\mathbb{R}^d$ such that

$$p_n(x) \underset{n\to\infty}{\longrightarrow} p(x) \;, \quad \text{a.e.}$$

where a.e. refers to Lebesgue measure on $\mathbb{R}^d$. Then X_n converge in distribution to a random variable X with law $\mathbb{P}_X(\mathrm{d}x) = p(x)\mathrm{d}x$. Indeed, consider a function $\varphi \in C_b(\mathbb{R}^d)$ such that $0 \le \varphi \le 1$ (clearly, we may restrict our attention to that case). Then, Fatou's lemma shows that

$$\liminf_{n\to\infty} \int \varphi(x)p_n(x)\mathrm{d}x \ge \int \varphi(x)p(x)\mathrm{d}x \;,$$

$$\liminf_{n\to\infty} \int (1-\varphi(x))p_n(x)\mathrm{d}x \ge \int (1-\varphi(x))p(x)\mathrm{d}x \;.$$

By combining these two bounds, and using the fact that $p(x)$ is a probability density function, we get

$$\lim_{n\to\infty} \int \varphi(x) p_n(x) \mathrm{d}x = \int \varphi(x) p(x) \mathrm{d}x.$$

(c) If, for every $n \in \mathbb{N}$, X_n is uniformly distributed on $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$, then X_n converges in distribution to the uniform distribution on $[0, 1]$. This is just a special case of the approximation of the integral of a continuous function by Riemann sums.

(d) If X_n is Gaussian $\mathcal{N}(0, \sigma_n^2)$ and $\sigma_n \longrightarrow 0$, then X_n converges in distribution to the random variable 0.

Proposition 10.10 *If the sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X, then it also converges in distribution to X.*

Proof Suppose first that X_n converges a.s. to X. Then, for every $\varphi \in C_b(\mathbb{R}^d)$, $\varphi(X_n)$ converges a.s. to $\varphi(X)$ and the dominated convergence theorem gives $\mathbb{E}[\varphi(X_n)] \longrightarrow \mathbb{E}[\varphi(X)]$.

If X_n only converges in probability to X, then we know from Proposition 10.3 that there is a subsequence of $(X_n)_{n\in\mathbb{N}}$ that converges a.s. to X, and thus, for every $\varphi \in C_b(\mathbb{R}^d)$, $\mathbb{E}[\varphi(X_n)]$ converges to $\mathbb{E}[\varphi(X)]$ along this subsequence. But then, the same argument shows that, from any subsequence of $(X_n)_{n\in\mathbb{N}}$, we can extract a sub-subsequence along which $\mathbb{E}[\varphi(X_n)]$ converges to $\mathbb{E}[\varphi(X)]$. This is only possible if $\mathbb{E}[\varphi(X_n)] \longrightarrow \mathbb{E}[\varphi(X)]$ as $n \to \infty$. □

Remark The converse to the proposition is (of course) false, because, as we already mentioned, the convergence in distribution of X_n to X does not determine the limiting variable X. There is however a very special case where the converse holds, namely the case where the limiting random variable X is constant. Indeed, if X_n converges in distribution to $a \in \mathbb{R}^d$, property (ii) of the next proposition shows that, for every $\varepsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_{X_n}(B(a, \varepsilon)) \geq 1$$

where $B(a, \varepsilon)$ denotes the open ball of radius ε centered at a. This is exactly saying that X_n converges in probability to a.

If $(X_n)_{n\in\mathbb{N}}$ is a sequence that converges in distribution to X, it is not always true that

$$\mathbb{P}(X_n \in B) \longrightarrow \mathbb{P}(X \in B)$$

when B is a Borel subset of $\mathbb{R}^d$ (take for B the set of all rational numbers in example (c) above). Still we have the following proposition, where ∂B denotes the topological boundary of a subset B of $\mathbb{R}^d$.

Proposition 10.11 (Portmanteau Theorem) *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence in $M_1(\mathbb{R}^d)$ and let $\mu \in M_1(\mathbb{R}^d)$. The following four assertions are equivalent:*

(i) *The sequence (μ_n) converges weakly to μ.*
(ii) *For every open subset G of $\mathbb{R}^d$,*

$$\liminf \mu_n(G) \geq \mu(G).$$

(iii) *For every closed subset F of $\mathbb{R}^d$,*

$$\limsup \mu_n(F) \leq \mu(F).$$

(iv) *For every Borel subset B of $\mathbb{R}^d$ such that $\mu(\partial B) = 0$,*

$$\lim \mu_n(B) = \mu(B).$$

Proof Let us start by proving that (i)⇒(ii). If G is an open subset of $\mathbb{R}^d$, we can find a sequence $(\varphi_p)_{p\in\mathbb{N}}$ of bounded continuous functions such that $0 \leq \varphi_p \leq \mathbf{1}_G$ and $\varphi_p \uparrow \mathbf{1}_G$ (for instance, $\varphi_p(x) = p\,\mathrm{dist}(x, G^c) \wedge 1$). Then

$$\liminf_{n\to\infty} \mu_n(G) \geq \sup_p \Big(\liminf_{n\to\infty} \int \varphi_p \mathrm{d}\mu_n \Big) = \sup_p \Big(\int \varphi_p \mathrm{d}\mu \Big) = \mu(G).$$

The equivalence (ii)⇔(iii) is immediate since complements of open sets are closed and conversely.

Let us show that (ii) and (iii) imply (iv). If $B \in \mathcal{B}(\mathbb{R}^d)$, write $\overline{B}$ for the closure of B and B° for the interior of B so that $\partial B = \overline{B} \backslash B^\circ$. By (ii) and (iii), we have

$$\limsup \mu_n(B) \leq \limsup \mu_n(\overline{B}) \leq \mu(\overline{B})$$
$$\liminf \mu_n(B) \geq \liminf \mu_n(B^\circ) \geq \mu(B^\circ).$$

If $\mu(\partial B) = 0$ then $\mu(\overline{B}) = \mu(B^\circ) = \mu(B)$ and we get (iv).

We still have to prove that (iv)⇒(i). Let $\varphi \in C_b(\mathbb{R}^d)$. Without loss of generality, we can assume that $\varphi \geq 0$. Then, let $K \geq 0$ such that $0 \leq \varphi \leq K$. By the Fubini theorem,

$$\int \varphi(x)\mu(\mathrm{d}x) = \int \Big(\int_0^K \mathbf{1}_{\{t \leq \varphi(x)\}} \mathrm{d}t \Big) \mu(\mathrm{d}x) = \int_0^K \mu(E_t^\varphi)\mathrm{d}t,$$

where E_t^{φ} is the closed set $E_t^{\varphi} := \{x \in \mathbb{R}^d : \varphi(x) \geq t\}$. Similarly, for every n,

$$\int \varphi(x)\mu_n(\mathrm{d}x) = \int_0^K \mu_n(E_t^{\varphi})\mathrm{d}t.$$

Now note that $\partial E_t^{\varphi} \subset \{x \in \mathbb{R}^d : \varphi(x) = t\}$, and that there are at most countably many values of t such that

$$\mu(\{x \in \mathbb{R}^d : \varphi(x) = t\}) > 0$$

(indeed, for every $k \in \mathbb{N}$, there are at most k distinct values of t such that $\mu(\{x \in \mathbb{R}^d : \varphi(x) = t\}) \geq \frac{1}{k}$, since the sets $\{x \in \mathbb{R}^d : \varphi(x) = t\}$ are disjoint when t varies, and μ is a probability measure). Hence (iv) implies

$$\mu_n(E_t^{\varphi}) \underset{n\to\infty}{\longrightarrow} \mu(E_t^{\varphi}),$$

for every $t \in [0, K]$, except possibly for a countable set of values of t, which has zero Lebesgue measure. By dominated convergence, we conclude that

$$\int \varphi(x)\mu_n(\mathrm{d}x) = \int_0^K \mu_n(E_t^{\varphi})\mathrm{d}t \underset{n\to\infty}{\longrightarrow} \int_0^K \mu(E_t^{\varphi})\mathrm{d}t = \int \varphi(x)\mu(\mathrm{d}x).$$

□

Consequence A sequence $(X_n)_{n\in\mathbb{N}}$ of real random variables converges to X in distribution if and only if the distribution functions $F_{X_n}(x)$ converge to $F_X(x)$ at every point x where F_X is continuous. The "only if" part immediately follows from property (iv) in Proposition 10.11. In the reverse direction, suppose that $F_{X_n}(x) \longrightarrow F_X(x)$ whenever F_X is continuous at x. Since the set of discontinuity points of F_X is at most countable, it follows that, for every $x \in \mathbb{R}$,

$$\liminf F_{X_n}(x-) \geq F_X(x-),$$
$$\limsup F_{X_n}(x) \leq F_X(x),$$

where $F_X(x-)$ denotes the left limit of F_X at x (to get the statement about the limsup, take a sequence $(x_p)_{p\in\mathbb{R}}$ that decreases to x, and such that F_X is continuous at x_p for every $p \in \mathbb{N}$, then write $\limsup F_{X_n}(x) \leq \lim F_{X_n}(x_p) = F_X(x_p)$ and note that $F_X(x_p) \longrightarrow F_X(x)$ as $p \to \infty$ since F_X is right-continuous — the statement about the liminf is derived in a similar manner). Recalling that $\mathbb{P}_X((a, b)) = F_X(b-) - F_X(a)$ for any $a < b$, it follows from the preceding display that the property (ii) of Proposition 10.11 holds for $\mu_n = \mathbb{P}_{X_n}$ and $\mu = \mathbb{P}_X$ when G is an open interval. Since any open subset of $\mathbb{R}$ is the disjoint union of at most countably many open intervals, we easily get that property (ii) holds for any open set, proving that X_n converges in distribution to X.

Recall the notation $C_c(\mathbb{R}^d)$ for the space of all real continuous functions with compact support on $\mathbb{R}^d$.

Proposition 10.12 *Let $(\mu_n)_{n\in\mathbb{N}}$ and μ be probability measures on $\mathbb{R}^d$. Let H be a subset of $C_b(\mathbb{R}^d)$ whose closure (with respect to the supremum norm) contains $C_c(\mathbb{R}^d)$. The following properties are equivalent.*

(i) *The sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ.*
(ii) *We have*

$$\forall \varphi \in C_c(\mathbb{R}^d)\,, \quad \int \varphi\, \mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int \varphi\, \mathrm{d}\mu.$$

(iii) *We have*

$$\forall \varphi \in H\,, \quad \int \varphi\, \mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int \varphi\, \mathrm{d}\mu.$$

Proof It is obvious that (i)⇒(ii) and (i)⇒(iii). Then suppose that (ii) holds. Let $\varphi \in C_b(\mathbb{R}^d)$ and let $(f_k)_{k\in\mathbb{N}}$ be a sequence in $C_c(\mathbb{R}^d)$ such that $0 \le f_k \le 1$ for every $k \in \mathbb{N}$, and $f_k \uparrow 1$ as $k \to \infty$. Then, for every $k \in \mathbb{N}$, $\varphi f_k \in C_c(\mathbb{R}^d)$ and thus

$$\int \varphi f_k\, \mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int \varphi f_k\, \mathrm{d}\mu. \tag{10.2}$$

On the other hand,

$$\Big|\int \varphi\, \mathrm{d}\mu_n - \int \varphi f_k\, \mathrm{d}\mu_n\Big| \le \|\varphi\| \int (1-f_k)\mathrm{d}\mu_n = \|\varphi\|\Big(1 - \int f_k \mathrm{d}\mu_n\Big),$$

and similarly

$$\Big|\int \varphi\, \mathrm{d}\mu - \int \varphi f_k\, \mathrm{d}\mu\Big| \le \|\varphi\|\Big(1 - \int f_k \mathrm{d}\mu\Big).$$

Hence, using (10.2), we get for every $k \in \mathbb{N}$,

$$\begin{aligned}
&\limsup_{n\to\infty} \Big|\int \varphi\, \mathrm{d}\mu_n - \int \varphi\, \mathrm{d}\mu\Big| \\
&\le \limsup_{n\to\infty} \Big|\int \varphi\, \mathrm{d}\mu_n - \int \varphi f_k\, \mathrm{d}\mu_n\Big| + \Big|\int \varphi\, \mathrm{d}\mu - \int \varphi f_k\, \mathrm{d}\mu\Big| \\
&\le \|\varphi\| \Big(\limsup_{n\to\infty} \Big(1 - \int f_k \mathrm{d}\mu_n\Big) + \Big(1 - \int f_k \mathrm{d}\mu\Big)\Big) \\
&= 2\|\varphi\|\Big(1 - \int f_k \mathrm{d}\mu\Big).
\end{aligned}$$

Now we just have to let k tend to ∞ (noting that $\int f_k \,\mathrm{d}\mu \uparrow 1$, by monotone convergence) and we find that $\int \varphi \,\mathrm{d}\mu_n$ converges to $\int \varphi \,\mathrm{d}\mu$, so that (i) holds.

We complete the proof by verifying that (iii)⇒(ii). So we assume that (iii) holds. If $\varphi \in C_c(\mathbb{R}^d)$, then, for every $k \in \mathbb{N}$, we can find a function $\varphi_k \in H$ such that $\|\varphi - \varphi_k\| \le 1/k$, and it follows that

$$\begin{aligned}
&\limsup_{n\to\infty} \Big| \int \varphi \,\mathrm{d}\mu_n - \int \varphi \,\mathrm{d}\mu \Big| \\
&\le \limsup_{n\to\infty} \Big(\Big| \int (\varphi - \varphi_k) \,\mathrm{d}\mu_n \Big| + \Big| \int \varphi_k \,\mathrm{d}\mu_n - \int \varphi_k \,\mathrm{d}\mu \Big| + \Big| \int (\varphi_k - \varphi) \,\mathrm{d}\mu \Big| \Big) \\
&\le \frac{2}{k}.
\end{aligned}$$

Since this holds for every k, we get that $\int \varphi \,\mathrm{d}\mu_n \longrightarrow \int \varphi \,\mathrm{d}\mu$. □

Recall our notation $\widehat{\mu}(\xi) = \int e^{\mathrm{i}\xi\cdot x}\,\mu(\mathrm{d}x)$ for the Fourier transform of $\mu \in M_1(\mathbb{R}^d)$, and $\Phi_X = \widehat{\mathbb{P}}_X$ for the characteristic function of a random variable X with values in $\mathbb{R}^d$.

Theorem 10.13 (Lévy's Theorem) *A sequence $(\mu_n)_{n\in\mathbb{N}}$ in $M_1(\mathbb{R}^d)$ converges weakly to $\mu \in M_1(\mathbb{R}^d)$ if and only if*

$$\forall \xi \in \mathbb{R}^d\,, \quad \widehat{\mu}_n(\xi) \underset{n\to\infty}{\longrightarrow} \widehat{\mu}(\xi).$$

Equivalently, a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables with values in $\mathbb{R}^d$ converges in distribution to X if and only if

$$\forall \xi \in \mathbb{R}^d\,, \quad \Phi_{X_n}(\xi) \underset{n\to\infty}{\longrightarrow} \Phi_X(\xi).$$

Proof It is enough to prove the first assertion. First, if we assume that the sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ, the definition of this convergence shows that

$$\forall \xi \in \mathbb{R}^d\,, \quad \widehat{\mu}_n(\xi) = \int e^{\mathrm{i}\xi\cdot x}\mu_n(\mathrm{d}x) \underset{n\to\infty}{\longrightarrow} \int e^{\mathrm{i}\xi\cdot x}\mu(\mathrm{d}x) = \widehat{\mu}(\xi).$$

Conversely, assume that $\widehat{\mu}_n(\xi) \to \widehat{\mu}(\xi)$ for every $\xi \in \mathbb{R}^d$ and let us show that (μ_n) converges weakly to μ. To simplify notation, we only treat the case $d = 1$, but the proof in the general case is exactly similar.

Let $f \in C_c(\mathbb{R})$, and, for every $\sigma > 0$, let

$$g_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\exp(-\frac{x^2}{2\sigma^2}).$$

be the density of the $\mathcal{N}(0, \sigma^2)$ distribution. We already observed in the proof of Theorem 8.16 that $g_\sigma * f$ converges pointwise to f as $\sigma \to 0$. Since f has compact support here, it is not hard to verify that this convergence is uniform on $\mathbb{R}$ (use Proposition 5.8 (i) and note that, for every $\varepsilon > 0$, we can find a compact set K such that $|g_\sigma * f| \leq \varepsilon$ for every $x \in \mathbb{R}\backslash K$ and $\sigma \in (0, 1]$). Thus, if we let H be the subset of $C_b(\mathbb{R}^d)$ defined by

$$H := \{\varphi = g_\sigma * f : \sigma > 0 \text{ and } f \in C_c(\mathbb{R}^d)\}$$

the closure of H in $C_b(\mathbb{R}^d)$ contains $C_c(\mathbb{R}^d)$

On the other hand, we also saw in the proof of Theorem 8.16 that, for any $\nu \in M_1(\mathbb{R})$,

$$\begin{aligned}\int g_\sigma * f \,\mathrm{d}\nu &= \int f(x)\, g_\sigma * \nu(x)\mathrm{d}x \\ &= \int f(x)\Big((\sigma\sqrt{2\pi})^{-1}\int e^{\mathrm{i}\xi x} g_{1/\sigma}(\xi)\widehat{\nu}(-\xi)\mathrm{d}\xi\Big)\mathrm{d}x. \end{aligned} \tag{10.3}$$

We apply this formula to $\nu = \mu_n$ and $\nu = \mu$. From our assumption $\widehat{\mu}_n(\xi) \longrightarrow \widehat{\mu}(\xi)$ for every $\xi \in \mathbb{R}$ and the trivial bound $|\widehat{\mu}_n(\xi)| \leq 1$, we can apply the dominated convergence theorem to get that, for every $x \in \mathbb{R}$,

$$\int e^{\mathrm{i}\xi x} g_{1/\sigma}(\xi)\widehat{\mu}_n(-\xi)\mathrm{d}\xi \underset{n\to\infty}{\longrightarrow} \int e^{\mathrm{i}\xi x} g_{1/\sigma}(\xi)\widehat{\mu}(-\xi)\mathrm{d}\xi.$$

Since the quantities in the last display are bounded by 1 and $f \in C_c(\mathbb{R})$, we can use (10.3) and again the dominated convergence theorem to get

$$\int g_\sigma * f \,\mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int g_\sigma * f \,\mathrm{d}\mu.$$

Finally, we have proved that $\int \varphi \,\mathrm{d}\mu_n \longrightarrow \int \varphi \,\mathrm{d}\mu$ for every $\varphi \in H$, and we know that the closure of H in $C_b(\mathbb{R}^d)$ contains $C_c(\mathbb{R}^d)$. By Proposition 10.12, this is enough to show that the sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ. □

10.4 Two Applications

10.4.1 The Convergence of Empirical Measures

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{R}^d$. One may think of these variables as giving the successive results of a random experiment which is repeated independently. A fundamental statistical problem is to estimate the law of X_1 from the data $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ for a *single* value of ω.

Example (Opinion Polls) Imagine that we have a large population of N individuals numbered $1, 2, \ldots, N$. The integer N is supposed to be very large, and one may think of the population of a country. With each individual $i \in \{1, \ldots, N\}$ we associate a trait $a(i) \in \mathbb{R}^d$ (this may correspond, for instance, to the voting intention of individual i at the next election). If $A \in \mathcal{B}(\mathbb{R}^d)$, we are then interested in the quantity

$$\mu(A) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_A(a(i))$$

which is the proportion of individuals whose trait lies in A (for instance, the proportion of individuals who intend to vote for a given candidate).

Since N is very large, it is impossible to compute the exact value of $\mu(A)$. The principle of an opinion poll is to choose a sample of the population, meaning that one selects at random n individuals (n will be large but small in comparison with N), hoping that the proportion of individuals whose trait belongs to A in the sample will be close to the same proportion in the whole population, that is, to $\mu(A)$. In more precise mathematical terms, the sample will consist of n independent random variables $Y_1, \ldots, Y_n$ uniformly distributed over $\{1, \ldots, N\}$. The trait associated with the j-th individual in the sample is $X_j = a(Y_j)$. The random variables $X_1, \ldots, X_n$ are also independent and have the same law given by

$$\mathbb{P}_{X_1}(A) = \mathbb{P}(a(Y_1) \in A) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_A(a(i)) = \mu(A).$$

On the other hand, the proportion of individuals of the sample whose trait belongs to A is

$$\frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_A(X_j(\omega)) = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}(A)$$

Finally, the question of whether proportions computed in the sample are close to the "real" proportions in the population boils down to verifying that the so-called *empirical measure*

$$\frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}$$

is close to $\mathbb{P}_{X_1}$ when n is large. The next theorem provides a theoretical answer to this question.

Theorem 10.14 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{R}^d$. For every $\omega \in \Omega$ and every $n \in \mathbb{N}$, let $\mu_{n,\omega}$ be the probability measure on $\mathbb{R}^d$ defined by*

$$\mu_{n,\omega} := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}.$$

Then we have

$$\mu_{n,\omega} \xrightarrow[n\to\infty]{\text{(w)}} \mathbb{P}_{X_1}, \quad \textit{a.s.}$$

Remark From the practical perspective, Theorem 10.14 is useless unless we have good bounds on the speed of convergence. In the case of opinion polls for instance, one expects that the empirical measure $\mu_{n,\omega}$ is "sufficiently close" to $\mathbb{P}_{X_1}$ when n is of order 10^3 (while N is typically of order 10^7).

Proof Let H a countable dense subset of $C_c(\mathbb{R}^d)$ (the existence of such a subset can be deduced, for instance, from the Stone-Weierstrass theorem), and let $\varphi \in H$. The strong law of large numbers ensures that

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

This can be reformulated as

$$\int \varphi \, \mathrm{d}\mu_{n,\omega} \xrightarrow[n\to\infty]{\text{a.s.}} \int \varphi \, \mathrm{d}\mathbb{P}_{X_1}.$$

Since H is countable, up to discarding a countable union of sets of probability zero, we get

$$\int \varphi \, \mathrm{d}\mu_{n,\omega} \xrightarrow[n\to\infty]{} \int \varphi \, \mathrm{d}\mathbb{P}_{X_1}, \ \forall \varphi \in H, \ \text{a.s.}$$

By Proposition 10.12, this is enough to conclude that $\mu_{n,\omega}$ converges weakly to $\mathbb{P}_{X_1}$, a.s. □

10.4.2 The Central Limit Theorem

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables in L^1. The strong law of large numbers asserts that

$$\frac{1}{n}(X_1+\cdots+X_n) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1].$$

We are now interested in the speed of this convergence, meaning that we want information about the typical size of

$$\frac{1}{n}(X_1+\cdots+X_n) - \mathbb{E}[X_1]$$

when n is large.

Under the additional assumption that the variables X_n are in L^2, it is easy to guess the answer, because a simple calculation made in the proof of the weak law of large numbers (Theorem 9.13) gives

$$\mathbb{E}[(X_1+\cdots+X_n-n\,\mathbb{E}[X_1])^2] = \text{var}(X_1+\cdots+X_n) = n\,\text{var}(X_1).$$

This shows that the mean value of $(X_1+\cdots+X_n-n\,\mathbb{E}[X_1])^2$ grows linearly with n, and thus the typical size of $X_1+\cdots+X_n-n\,\mathbb{E}[X_1]$ is expected to be $\sqrt{n}$, or equivalently the typical size of $\frac{1}{n}(X_1+\cdots+X_n)-\mathbb{E}[X_1]$ should be $1/\sqrt{n}$. The central limit theorem gives a much more precise statement.

Theorem 10.15 (Central Limit Theorem) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables in L^2. Set $\sigma^2 = \text{var}(X_1)$ and assume that $\sigma > 0$. Then*

$$\frac{1}{\sqrt{n}}\Big(X_1+\cdots+X_n-n\,\mathbb{E}[X_1]\Big) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0,\sigma^2)$$

where $\mathcal{N}(0,\sigma^2)$ is the Gaussian distribution with mean 0 and variance σ^2. Hence, for every $a, b \in \overline{\mathbb{R}}$ with $a < b$,

$$\lim_{n\to\infty} \mathbb{P}\big(X_1+\cdots+X_n \in [n\mathbb{E}[X_1]+a\sqrt{n}, n\mathbb{E}[X_1]+b\sqrt{n}]\big)$$
$$= \frac{1}{\sigma\sqrt{2\pi}}\int_a^b \exp(-\frac{x^2}{2\sigma^2})\,\mathrm{d}x.$$

Remark When $\sigma = 0$ there is not much to say since the variables X_n are all equal a.s. to the constant $\mathbb{E}[X_1]$.

Proof The second part of the statement follows from the first one, thanks to Proposition 10.11. To prove the first assertion, note that we can assume $\mathbb{E}[X_1] = 0$

(just replace X_n by $X_n - \mathbb{E}[X_n]$). Then set

$$Z_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n).$$

The characteristic function of Z_n is

$$\Phi_{Z_n}(\xi) = \mathbb{E}\Big[\exp\Big(\mathrm{i}\xi(\frac{X_1 + \cdots + X_n}{\sqrt{n}})\Big)\Big] = \mathbb{E}\Big[\exp\Big(\mathrm{i}\frac{\xi}{\sqrt{n}}X_1\Big)\Big]^n = \Phi_{X_1}(\frac{\xi}{\sqrt{n}})^n, \tag{10.4}$$

where, in the second equality, we used the fact that the random variables X_i are independent and identically distributed. On the other hand, by Proposition 8.17, we have

$$\Phi_{X_1}(\xi) = 1 + \mathrm{i}\xi\mathbb{E}[X_1] - \frac{1}{2}\xi^2\mathbb{E}[X_1^2] + o(\xi^2) = 1 - \frac{\sigma^2\xi^2}{2} + o(\xi^2)$$

as $\xi \to 0$. For every fixed $\xi \in \mathbb{R}$, we have thus

$$\Phi_{X_1}(\frac{\xi}{\sqrt{n}}) = 1 - \frac{\sigma^2\xi^2}{2n} + o(\frac{1}{n})$$

as $n \to \infty$. Thanks to (10.4) and the last display, we have, for every $\xi \in \mathbb{R}$,

$$\lim_{n\to\infty} \Phi_{Z_n}(\xi) = \lim_{n\to\infty} \Big(1 - \frac{\sigma^2\xi^2}{2n} + o(\frac{1}{n})\Big)^n = \exp(-\frac{\sigma^2\xi^2}{2}),$$

and, by Lemma 8.15, the limit is the characteristic function of the Gaussian $\mathcal{N}(0, \sigma^2)$ distribution. An application of Lévy's theorem (Theorem 10.13) now completes the proof.

□

Special Case: de Moivre Theorem We assume that the random variables X_n follow the Bernoulli distribution with parameter $1/2$, meaning that $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = 0) = \frac{1}{2}$ (this was the case considered by de Moivre in 1733, and later Laplace dealt with the more general case of the Bernoulli distribution with parameter $p \in (0, 1)$), Then, $S_n = X_1 + \cdots + X_n$ has the binomial $\mathcal{B}(n, 1/2)$ distribution:

$$\mathbb{P}(S_n = k) = 2^{-n}\binom{n}{k}, \quad 0 \le k \le n.$$

Since $\sigma^2 = 1/4$ in this special case, the central limit theorem implies that, for every $a < b$,

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

This last convergence can be verified directly (with some technical work!) via Stirling's formula. In fact the following more precise asymptotics hold when $n \to \infty$,

$$\sqrt{n}\, 2^{-n} \binom{n}{k} = \sqrt{\frac{2}{\pi}} \exp\Big(-\frac{2}{n}(k - \frac{n}{2})^2 \Big) + o(1)$$

where the remainder $o(1)$ is uniform when k varies in $\{0, 1, \dots, n\}$.

10.4.3 The Multidimensional Central Limit Theorem

We now assume that $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent and identically distributed random variables with values in $\mathbb{R}^d$, such that $\mathbb{E}[|X_1|] < \infty$. We can apply the strong law of large numbers to each component of X_n and we have

$$\frac{1}{n}\,(X_1 + \cdots + X_n) \underset{n\to\infty}{\overset{\text{a.s.}}{\longrightarrow}} \mathbb{E}[X_1],$$

as in the real case. If we assume furthermore that $\mathbb{E}[|X_1|^2] < \infty$, then we expect to have also an analog of the central limit theorem. However, in contrast with the law of large numbers, this is not immediate because the convergence in distribution of a sequence of random vectors does not follow from the convergence in distribution of each component sequence (in fact, the law of the limiting vector is not determined by its marginal distributions, as we already noticed in Chapter 8).

To extend the central limit theorem to the case of random variables with values in $\mathbb{R}^d$, we must first generalize the Gaussian distribution in higher dimensions.

Definition 10.16 A random variable X with values in $\mathbb{R}^d$ is called a Gaussian vector if any linear combination of the components of X is a (real) Gaussian variable. This is equivalent to the existence of a vector $m \in \mathbb{R}^d$ and a $d \times d$ symmetric nonnegative definite matrix C such that, for every $\xi \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(\mathrm{i}\xi \cdot X)] = \exp\Big(\mathrm{i}\xi \cdot m - \frac{1}{2}{}^t\xi C \xi\Big). \tag{10.5}$$

Moreover, we have then $\mathbb{E}[X] = m$ and $K_X = C$, and we say that X follows the Gaussian $\mathcal{N}_d(m, C)$ distribution.

To establish the equivalence of the two forms of the definition, first assume that any linear combination of the components of X is Gaussian. In particular, these components are in L^2 and thus the expectation $\mathbb{E}[X]$ and the covariance matrix K_X are well defined. Then, for every $\xi \in \mathbb{R}^d$, we have $\mathbb{E}[\xi \cdot X] = \xi \cdot \mathbb{E}[X]$ and $\mathrm{var}(\xi \cdot X) = {}^t\xi K_X \xi$ (by (8.3)), so that $\xi \cdot X$ has the Gaussian $\mathcal{N}(\xi \cdot \mathbb{E}[X], {}^t\xi K_X \xi)$ distribution. Formula (10.5), with $m = \mathbb{E}[X]$ and $C = K_X$, now follows from Lemma 8.15 giving the characteristic function of a real Gaussian variable. Conversely, if (10.5) holds, then by replacing ξ with $\lambda\xi$ for $\lambda \in \mathbb{R}$, and using again Lemma 8.15, we get that the characteristic function of $\xi \cdot X$ coincides with the characteristic function of the Gaussian $\mathcal{N}(\xi \cdot m, {}^t\xi C \xi)$ distribution, so that, in particular, $\xi \cdot X$ is Gaussian.

Remark In order for X to be a Gaussian vector, it is not sufficient that each component of X is a Gaussian (real) random variable. Let us give a simple example with $d = 2$. Let X_1 be Gaussian $\mathcal{N}(0, 1)$ and let U be a random variable with law $\mathbb{P}(U = 1) = \mathbb{P}(U = -1) = 1/2$, which is independent of X_1. If we set $X_2 = UX_1$, then one immediately checks that X_2 is also Gaussian $\mathcal{N}(0, 1)$. On the other hand, (X_1, X_2) is certainly not a Gaussian vector, because $\mathbb{P}(X_1 + X_2 = 0) = \mathbb{P}(U = -1) = 1/2$, which makes it impossible for $X_1 + X_2$ to be Gaussian.

Let $Y_1, \ldots, Y_d$ be independent Gaussian (real) random variables. Then $(Y_1, \ldots, Y_d)$ is a Gaussian vector. Indeed we noticed at the end of Section 9.5 that any linear combination of independent Gaussian random variables is Gaussian. Alternatively, we can also notice that formula (10.5) holds, with a diagonal matrix C. Conversely, if the covariance matrix of a Gaussian vector is diagonal, then its components are independent. This is an immediate application of Proposition 9.8 (iv), and we will come back to this result in the next chapter.

Proposition 10.17 *Let $m \in \mathbb{R}^d$ and let C be a symmetric nonnegative definite $d \times d$ matrix. One can construct a Gaussian $\mathcal{N}_d(m, C)$ random vector.*

Proof It is enough to treat the case $m = 0$ (if X is Gaussian $\mathcal{N}_d(0, C)$, $X + m$ is Gaussian $\mathcal{N}_d(m, C)$). Set $A = \sqrt{C}$ in such a way that A is a symmetric nonnegative definite matrix and $A^2 = C$. Let $Y_1, \ldots, Y_d$ be independent Gaussian $\mathcal{N}(0, 1)$ variables, and $Y = (Y_1, \ldots, Y_d)$, so that the covariance matrix K_Y is the identity matrix. Then $X = AY$ follows the Gaussian $\mathcal{N}_d(0, C)$ distribution. Indeed any linear combination of the components of X is also a linear combination of $Y_1, \ldots, Y_d$ and is therefore Gaussian. Furthermore, by (8.2),

$$K_X = A K_Y {}^t A = A^2 = C,$$

which shows that X is Gaussian $\mathcal{N}_d(0, C)$. □

Theorem 10.18 (Multidimensional Central Limit Theorem) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{R}^d$, such that $\mathbb{E}[|X_1|^2] < \infty$. Then,*

$$\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n - n\,\mathbb{E}[X_1]) \underset{n\to\infty}{\overset{\text{(d)}}{\longrightarrow}} \mathcal{N}_d(0, K_{X_1})$$

Proof Replacing X_n by $X_n - \mathbb{E}[X_n]$ allows us to assume that $\mathbb{E}[X_1] = 0$. For every $\xi \in \mathbb{R}^d$, the central limit theorem in the real case (Theorem 10.15) shows that

$$\frac{1}{\sqrt{n}}(\xi \cdot X_1 + \cdots + \xi \cdot X_n) \underset{n\to\infty}{\overset{\text{(d)}}{\longrightarrow}} \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = \mathbb{E}[(\xi \cdot X_1)^2] = {}^t\xi K_{X_1}\xi$. This implies that

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}\xi \cdot (\frac{X_1 + \cdots + X_n}{\sqrt{n}})\Big)\Big] \underset{n\to\infty}{\longrightarrow} \exp(-\frac{1}{2}\,{}^t\xi K_{X_1}\xi)$$

and Lévy's theorem (Theorem 10.13) gives the desired result. □

10.5 Exercises

Exercise 10.1 In this exercise, to emphasize the dependence on the underlying probability measure, we speak of convergence in $\mathbb{P}$-probability instead of convergence in probability. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables such that X_n converges in $\mathbb{P}$-probability to X when $n \to \infty$. Suppose that $\mathbb{P}'$ is another probability measure on $(\Omega, \mathcal{A})$ and that $\mathbb{P}'$ is absolutely continuous with respect to $\mathbb{P}$. Show that X_n also converges in $\mathbb{P}'$-probability to X when $n \to \infty$.

Exercise 10.2 Let $f : [0, 1] \longrightarrow \mathbb{R}$ be a continuous function. Prove that

$$\lim_{n\to\infty} \int_{[0,1]^n} f\Big(\frac{x_1 + x_2 + \cdots + x_n}{n}\Big)\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n = f(\frac{1}{2}).$$

Exercise 10.3 (Bernstein Polynomials) Let $f : [0, 1] \longrightarrow \mathbb{R}$ be a continuous function. Prove that, for every $p \in [0, 1]$,

$$f(p) = \lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(k/n),$$

and that the convergence holds uniformly in $p \in [0, 1]$.

Exercise 10.4

(1) Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a bounded continuous function. Prove that, for every $\lambda > 0$,

$$\lim_{n\to\infty} e^{-\lambda n} \sum_{k=0}^{\infty} \frac{(\lambda n)^k}{k!} f(k/n) = f(\lambda), \tag{10.6}$$

and that the convergence is uniform when λ varies in a bounded subset of $(0, \infty)$.

(2) Suppose that f is only bounded and measurable, and is continuous at $x = \lambda$. Prove that the convergence in (10.6) still holds.

(3) Let μ be a probability measure on $\mathbb{R}_+$, and let $L(\lambda) = \int e^{-\lambda y} \mu(\mathrm{d}y)$, for $\lambda \geq 0$, be its Laplace transform. For every $\lambda > 0$, write $L^{(k)}(\lambda)$ for the k-th derivative of L at λ (justify its existence!). Prove that, for every $x \in [0, \infty)$ such that $\mu(\{x\}) = 0$,

$$\mu([0, x]) = \lim_{n\to\infty} \sum_{k=0}^{[nx]} \frac{(-1)^k}{k!} n^k L^{(k)}(n).$$

Exercise 10.5 (Coupon Collector Problem) Suppose that, for every $n \in \mathbb{N}$, $(X_k^{(n)})_{k\geq 1}$ is a sequence of independent random variables uniformly distributed on $\{1, 2, \dots, n\}$. For every integer $m \geq 1$, let $N_m^{(n)}$ be the number of distinct values in the finite sequence $X_1^{(n)}, X_2^{(n)}, \dots, X_m^{(n)}$, and set

$$T_n = \inf\{m \geq 1 : N_m^{(n)} = n\}.$$

If we observe the values $X_1^{(n)}, X_2^{(n)}, \dots$ one after the other, T_n is the first time when all possible values have been observed.

(1) For every $k \in \{1, \dots, n\}$, set $\tau_k^{(n)} = \inf\{m \geq 1 : N_m^{(n)} = k\}$, so that in particular $T_n = \tau_n^{(n)}$. Show that the random variables $\tau_k^{(n)} - \tau_{k-1}^{(n)}$, for $k \in \{2, \dots, n\}$, are independent, and the distribution of $\tau_k^{(n)} - \tau_{k-1}^{(n)} - 1$ is geometric of parameter $(k-1)/n$.

(2) Prove that

$$\frac{T_n}{n \log n} \underset{n\to\infty}{\longrightarrow} 1$$

in probability. (*Hint:* Estimate the expected value and the variance of T_n).

Exercise 10.6 Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be two sequences of real random variables, and let X and Y be two real random variables. Is it always true that the properties

$$X_n \xrightarrow[n\to\infty]{(d)} X \quad \text{and} \quad Y_n \xrightarrow[n\to\infty]{(d)} Y$$

imply that $(X_n, Y_n) \xrightarrow[n\to\infty]{(d)} (X, Y)$? Show that this fact holds in each of the following two cases:

(i) The random variable Y is constant a.s.
(ii) For every $n \in \mathbb{N}$, X_n and Y_n are independent, and X and Y are independent.

Exercise 10.7 Let μ be a probability measure on $\mathbb{R}$, and for every $n \in \mathbb{N}$,

$$\mu_n(\mathrm{d}x) = \sum_{k\in\mathbb{Z}} \mu([k2^{-n}, (k+1)2^{-n}))\, \delta_{k2^{-n}}(\mathrm{d}x).$$

Show that the sequence μ_n converges weakly to μ.

Exercise 10.8 Suppose that, for every $n \in \mathbb{N}$, Y_n is a Gaussian $\mathcal{N}(m_n, \sigma_n^2)$ random variable, where $m_n \in \mathbb{R}$ and $\sigma_n > 0$. Prove that the sequence $(Y_n)_{n\in\mathbb{N}}$ converges in distribution if and only if the two sequences (m_n) and (σ_n) converge, and identify the limiting distribution in that case.

Exercise 10.9 Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of random variables with values in $\mathbb{R}^d$, and let $(a_n)_{n\in\mathbb{N}}$ be a sequence of reals. Assume that Z_n converges in distribution to Z and a_n converges to a as $n \to \infty$. Prove that $a_n Z_n$ converges in distribution to aZ. (*Hint:* Use Proposition 10.12, and note that the space H in this proposition can be chosen to contain only Lipschitz functions).

Exercise 10.10 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables. For every $n \in \mathbb{N}$, set $M_n = \max\{X_1, \ldots, X_n\}$.

(1) Suppose that X_n is uniformly distributed over $[0, 1]$. Prove that $n(1 - M_n)$ converges in distribution and identify the limit.
(2) Suppose that X_n is distributed according to the Cauchy distribution of parameter 1. Show that n/M_n converges in distribution and identify the limit.

Exercise 10.11

(1) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables in L^2, such that $\mathbb{E}[X_n] = 1$ and $\mathrm{var}(X_n) > 0$. Set $S_n = X_1 + \cdots + X_n$. Show that the limit

$$\lim_{n\to\infty} \mathbb{P}(S_n \le n)$$

exists and compute it.

(2) Compute $\lim_{n\to\infty} e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!}$.

Exercise 10.12 (Glivenko-Cantelli Theorem) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables with distribution function F. For every integer $n \in \mathbb{N}$ and every $x \in \mathbb{R}$, define the random variable

$$F_n(x) = \frac{1}{n}\,\mathrm{card}\{j \in \{1,\dots,n\} : X_j \le x\},$$

which is the distribution function of the empirical measure associated with $X_1, \dots, X_n$. Prove that a.s.,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0.$$

Compare with Theorem 10.14. (*Hint:* It may be useful to assume that the random variables X_n are represented as in Lemma 8.7).

Exercise 10.13 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables in L^2, such that $\mathbb{E}[X_n] = 0$ and $\mathrm{var}(X_n) > 0$. Set $S_n = X_1 + \cdots + X_n$.

(1) Prove that

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} = \infty\,, \quad \text{a.s.}$$

(2) Prove that the sequence $S_n/\sqrt{n}$ does not converge in probability.
(3) Prove that the limit

$$\lim_{n\to\infty} \mathbb{P}(S_n > 0, S_{2n} < 0)$$

exists and compute this limit.

Chapter 11
Conditioning

This chapter is devoted to the construction and properties of the conditional expectation with respect to a sub-σ-field. Such a sub-σ-field $\mathcal{B}$ may be interpreted as containing some partial information, and the conditional expectation of a real random variable X with respect to $\mathcal{B}$ can be understood intuitively as the mean value of X knowing this partial information. In more precise terms, and assuming that X is in L^2, the conditional expectation of X with respect to $\mathcal{B}$ is the $\mathcal{B}$-measurable random variable that is closest to X in the L^2 sense. The motivation for considering conditional expectations comes from many problems of applied probability, including prediction, filtering, risk theory, etc.

In the first section, we discuss the discrete setting, which helps to understand the axiomatic definition of conditional expectation in the general case. This axiomatic definition involves a fundamental "characteristic property", which plays an essential role in the study of properties of the conditional expectation. The effective calculation of conditional expectations is in general a difficult problem, but we provide statements showing that it sometimes becomes tractable under appropriate independence assumptions. We also discuss several "concrete" cases where explicit formulas can be given. The last section introduces the notion of conditional law, which is useful in this book mainly from a conceptual point of view.

11.1 Discrete Conditioning

As in the previous chapters, we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We already mentioned in Chapter 9 that, if $B \in \mathcal{A}$ is an event of positive probability, we can define a new probability measure on $(\Omega, \mathcal{A})$ by setting, for every $A \in \mathcal{A}$,

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

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_11

This probability measure $A \mapsto \mathbb{P}(A \mid B)$ is called the conditional probability given B. Similarly, for every nonnegative random variable X, or for $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, the *conditional expectation* of X given B is defined by

$$\mathbb{E}[X \mid B] := \frac{\mathbb{E}[X\,\mathbf{1}_B]}{\mathbb{P}(B)}.$$

and one verifies that $\mathbb{E}[X \mid B]$ is also the expected value of X under $\mathbb{P}(\cdot \mid B)$. We can interpret $\mathbb{E}[X \mid B]$ as the mean value of X knowing that the event B occurs.

Let us now define the conditional expectation knowing a discrete random variable. We consider a random variable Y taking values in a countable space E equipped as usual with the σ-field of all its subsets. Let $E' = \{y \in E : \mathbb{P}(Y = y) > 0\}$. If $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, we may consider, for every $y \in E'$,

$$\mathbb{E}[X \mid Y = y] = \frac{\mathbb{E}[X\,\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)},$$

as a special case of the conditional expectation $\mathbb{E}[X \mid B]$ when $\mathbb{P}(B) > 0$.

Definition 11.1 Let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. The conditional expectation of X knowing Y is the real random variable

$$\mathbb{E}[X \mid Y] = \varphi(Y),$$

where the function $\varphi : E \longrightarrow \mathbb{R}$ is given by

$$\varphi(y) = \begin{cases} \mathbb{E}[X \mid Y = y] & \text{if } y \in E', \\ 0 & \text{if } y \in E \backslash E'. \end{cases}$$

Remark In this definition, the choice of the value of $\varphi(y)$ when $y \in E \backslash E'$ is irrelevant. Indeed this choice influences the definition of $\mathbb{E}[X \mid Y] = \varphi(Y)$ only on a set of zero probability, since

$$\mathbb{P}(Y \in E \backslash E') = \sum_{y \in E \backslash E'} \mathbb{P}(Y = y) = 0.$$

In what follows, we will consider conditional expectations in a much more general setting, but they will always be defined up to a set of zero probability.

Comparing with the conditioning with respect to an event, we note that $\mathbb{E}[X \mid Y]$ is now a *random variable*. This is the random variable that gives the mean value of X when the value of Y is known. Indeed, by definition, we have a.s.

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = y] \qquad \text{if } Y(\omega) = y.$$

We also note that $\mathbb{E}[X \mid Y]$ is a function of Y hence is $\sigma(Y)$-measurable. In a sense that will be made precise later, this is the best approximation of X by a function of Y.

Example Throwing a die. We take $\Omega = \{1, 2, \ldots, 6\}$ and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$ for every $\omega \in \Omega$. Set

$$Y(\omega) = \begin{cases} 1 \text{ if } \omega \text{ is odd,} \\ 0 \text{ if } \omega \text{ even,} \end{cases}$$

and $X(\omega) = \omega$. Then,

$$\mathbb{E}[X \mid Y](\omega) = \begin{cases} 3 \text{ if } \omega \in \{1, 3, 5\}, \\ 4 \text{ if } \omega \in \{2, 4, 6\}. \end{cases}$$

Proposition 11.2 *Let* $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. *We have* $\mathbb{E}[|\mathbb{E}[X \mid Y]|] \leq \mathbb{E}[|X|]$, *and thus* $\mathbb{E}[X \mid Y] \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. *Moreover, for every bounded* $\sigma(Y)$*-measurable real random variable* Z*,*

$$\mathbb{E}[Z X] = \mathbb{E}[Z \, \mathbb{E}[X \mid Y]].$$

Proof From the definition of $\mathbb{E}[X \mid Y]$, we have

$$\begin{aligned} \mathbb{E}[|\mathbb{E}[X \mid Y]|] = \sum_{y \in E'} \mathbb{P}(Y = y) \, \frac{|\mathbb{E}[X \, \mathbf{1}_{\{Y=y\}}]|}{\mathbb{P}(Y = y)} &\leq \sum_{y \in E} \mathbb{E}[|X| \, \mathbf{1}_{\{Y=y\}}] \\ &= \mathbb{E}\Big[|X| \sum_{y \in E} \mathbf{1}_{\{Y=y\}}\Big] \\ &= \mathbb{E}[|X|]. \end{aligned}$$

If Z is $\sigma(Y)$-measurable and bounded, Proposition 8.9 allows us to find a bounded function $\psi : E \longrightarrow \mathbb{R}$ such that $Z = \psi(Y)$, and then,

$$\begin{aligned} \mathbb{E}[\psi(Y) \, \mathbb{E}[X \mid Y]] = \sum_{y \in E} \psi(y) \, \mathbb{E}[X \, \mathbf{1}_{\{Y=y\}}] &= \sum_{y \in E} \mathbb{E}[\psi(Y) X \, \mathbf{1}_{\{Y=y\}}] \\ &= \mathbb{E}\Big[\psi(Y) X \sum_{y \in E} \mathbf{1}_{\{Y=y\}}\Big] \\ &= \mathbb{E}[\psi(Y) X]. \end{aligned}$$

In the third equality, the interchange of the expected value and the sum is justified by the Fubini theorem, noting that $\mathbb{E}[|\psi(Y) X|] < \infty$. $\square$

Consequence Let Y' be another discrete random variable such that $\sigma(Y) = \sigma(Y')$. We claim that

$$\mathbb{E}[X \,|\, Y] = \mathbb{E}[X \,|\, Y'] \qquad \text{a.s.}$$

To prove this, apply Proposition 11.2 with $Z = \mathbf{1}_{\{\mathbb{E}[X|Y] > \mathbb{E}[X|Y']\}}$, which is measurable with respect to $\sigma(Y) = \sigma(Y')$ since so are both $\mathbb{E}[X \,|\, Y]$ and $\mathbb{E}[X \,|\, Y']$. It follows that

$$\mathbb{E}[\mathbf{1}_{\{\mathbb{E}[X|Y] > \mathbb{E}[X|Y']\}}(\mathbb{E}[X \,|\, Y] - \mathbb{E}[X \,|\, Y'])] = 0$$

which is only possible if $\mathbf{1}_{\{\mathbb{E}[X|Y] > \mathbb{E}[X|Y']\}}(\mathbb{E}[X \mid Y] - \mathbb{E}[X \mid Y']) = 0$, a.s., hence $\mathbb{E}[X \mid Y] \leq \mathbb{E}[X \mid Y']$ a.s. By interchanging the roles of Y and Y', we obtain the reverse inequality, and we have proved that $\mathbb{E}[X \,|\, Y] = \mathbb{E}[X \,|\, Y']$ a.s. This argument also shows that the last assertion of Proposition 11.2 characterizes $\mathbb{E}[X \,|\, Y]$ among random variables that are both integrable and $\sigma(Y)$-measurable.

The preceding discussion suggests that the "good" notion of conditioning is conditioning with respect to a sub-σ-field of $\mathcal{A}$. We will define this notion in the next section, and the analog of the last assertion of Proposition 11.2 will serve as a major ingredient for the generalizations that follow.

11.2 The Definition of Conditional Expectation

11.2.1 Integrable Random Variables

The next theorem provides the definition of the conditional expectation of an integrable random variable with respect to a sub-σ-field.

Theorem and definition 11.3 *Let $\mathcal{B}$ be a sub-σ-field of $\mathcal{A}$, and let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. There exists a unique element of $L^1(\Omega, \mathcal{B}, \mathbb{P})$, which is denoted by $\mathbb{E}[X \,|\, \mathcal{B}]$, such that*

$$\forall B \in \mathcal{B}\,, \quad \mathbb{E}[X\,\mathbf{1}_B] = \mathbb{E}[\mathbb{E}[X \,|\, \mathcal{B}]\,\mathbf{1}_B]. \tag{11.1}$$

We have more generally, for every bounded $\mathcal{B}$-measurable real random variable Z,

$$\mathbb{E}[X\,Z] = \mathbb{E}[\mathbb{E}[X \,|\, \mathcal{B}]\,Z]. \tag{11.2}$$

If $X \geq 0$, we have $\mathbb{E}[X \,|\, \mathcal{B}] \geq 0$ a.s.

The crucial point is the fact that $\mathbb{E}[X \mid \mathcal{B}]$ is $\mathcal{B}$-measurable. Either of properties (11.1) and (11.2) characterizes the conditional expectation $\mathbb{E}[X \mid \mathcal{B}]$ among random variables of $L^1(\Omega, \mathcal{B}, \mathbb{P})$. In what follows, we will refer to (11.1) or (11.2) as the *characteristic property* of $E[X \,|\, \mathcal{B}]$.

In the special case where $\mathcal{B} = \sigma(Y)$ is generated by a random variable Y, we will write indifferently

$$\mathbb{E}[X \,|\, \mathcal{B}] = \mathbb{E}[X \,|\, \sigma(Y)] = \mathbb{E}[X \,|\, Y].$$

This notation is consistent with the discrete case considered in the previous section (compare (11.2) and Proposition 11.2).

Proof Let us start by proving uniqueness. Let X' and X'' be two random variables in $L^1(\Omega, \mathcal{B}, \mathbb{P})$ such that

$$\forall B \in \mathcal{B}, \quad \mathbb{E}[X' \, \mathbf{1}_B] = \mathbb{E}[X \, \mathbf{1}_B] = \mathbb{E}[X'' \, \mathbf{1}_B].$$

Taking $B = \{X' > X''\}$ (which is in $\mathcal{B}$ since both X' and X'' are $\mathcal{B}$-measurable), we get

$$\mathbb{E}[(X' - X'')\mathbf{1}_{\{X' > X''\}}] = 0$$

which implies that $X' \leq X''$ a.s., and we have similarly $X' \geq X''$ a.s. Thus $X' = X''$ a.s., which means that X' and X'' are equal as elements of $L^1(\Omega, \mathcal{B}, \mathbb{P})$.

Let us now turn to existence. We first assume that $X \geq 0$, and we let $\mathbf{Q}$ be the finite measure on $(\Omega, \mathcal{B})$ defined by

$$\forall B \in \mathcal{B}, \quad \mathbf{Q}(B) := \mathbb{E}[X \, \mathbf{1}_B].$$

Let us emphasize that we define $\mathbf{Q}(B)$ only for $B \in \mathcal{B}$. We may also view $\mathbb{P}$ as a probability measure on $(\Omega, \mathcal{B})$, by restricting the mapping $B \mapsto \mathbb{P}(B)$ to $B \in \mathcal{B}$, and it is immediate that $\mathbf{Q} \ll \mathbb{P}$. The Radon-Nikodym theorem (Theorem 4.11) applied to the probability measures $\mathbb{P}$ and $\mathbf{Q}$ on the measurable space $(\Omega, \mathcal{B})$ yields the existence of a nonnegative $\mathcal{B}$-measurable random variable $\widetilde{X}$ such that

$$\forall B \in \mathcal{B}, \quad \mathbb{E}[X \, \mathbf{1}_B] = \mathbf{Q}(B) = \mathbb{E}[\widetilde{X} \, \mathbf{1}_B].$$

Taking $B = \Omega$, we have $\mathbb{E}[\widetilde{X}] = \mathbb{E}[X] < \infty$, and thus $\widetilde{X} \in L^1(\Omega, \mathcal{B}, \mathbb{P})$. The random variable $\mathbb{E}[X \,|\, \mathcal{B}] = \widetilde{X}$ satisfies (11.1). When X is of arbitrary sign, we just have to take

$$\mathbb{E}[X \,|\, \mathcal{B}] = \mathbb{E}[X^+ \,|\, \mathcal{B}] - \mathbb{E}[X^- \,|\, \mathcal{B}]$$

and it is clear that (11.1) also holds in that case.

Finally, to see that (11.1) implies (11.2), we rely on the usual measure-theoretic arguments. (11.2) follows from (11.1) when Z is a simple random variable (taking only finitely many values), and in the general case Proposition 2.5 allows us to write

Z as the pointwise limit of a sequence $(Z_n)_{n\in\mathbb{N}}$ of simple $\mathcal{B}$-measurable random variables that are uniformly bounded by the same constant K (such that $|Z| \le K$) and the dominated convergence theorem yields the desired result. □

We emphasize that the preceding statement defines $\mathbb{E}[X \mid \mathcal{B}]$ as an element of $L^1(\Omega, \mathcal{B}, \mathbb{P})$, that is, as an equivalence class of $\mathcal{B}$-measurable random variables that are almost surely equal. In many of the forthcoming formulas involving conditional expectations, it would therefore be appropriate to include the mention "almost surely", but we will often omit this mention.

Example Let $\Omega = (0, 1]$, $\mathcal{A} = \mathcal{B}((0, 1])$ and $\mathbb{P}(\mathrm{d}\omega) = \mathrm{d}\omega$. Let $\mathcal{B}$ be the σ-field generated by the intervals $(\frac{i-1}{n}, \frac{i}{n}]$, $i \in \{1, \dots, n\}$, where $n \ge 1$ is fixed. If $f \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ we have $\int_0^1 |f(\omega)|\mathrm{d}\omega < \infty$, and one easily verifies that

$$\mathbb{E}[f \mid \mathcal{B}] = \sum_{i=1}^{n} f_i \, \mathbf{1}_{(\frac{i-1}{n}, \frac{i}{n}]},$$

where $f_i = n \int_{(i-1)/n}^{i/n} f(\omega) d\omega$ is the mean value of f on $(\frac{i-1}{n}, \frac{i}{n}]$.

Properties of Conditional Expectation

(a) If $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ and X is $\mathcal{B}$-measurable, then $\mathbb{E}[X \mid \mathcal{B}] = X$.
(b) The mapping $X \mapsto \mathbb{E}[X \mid \mathcal{B}]$ is linear on $L^1(\Omega, \mathcal{A}, \mathbb{P})$.
(c) If $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, $\mathbb{E}[\mathbb{E}[X \mid \mathcal{B}]] = \mathbb{E}[X]$.
(d) If $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, then $|\mathbb{E}[X \mid \mathcal{B}]| \le \mathbb{E}[|X| \mid \mathcal{B}]$ a.s. and, consequently, $\mathbb{E}[|\mathbb{E}[X \mid \mathcal{B}]|] \le \mathbb{E}[|X|]$. Therefore the mapping $X \mapsto \mathbb{E}[X \mid \mathcal{B}]$ is a contraction of $L^1(\Omega, \mathcal{A}, \mathbb{P})$.
(e) If $X, X' \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ and $X \ge X'$, then $\mathbb{E}[X \mid \mathcal{B}] \ge \mathbb{E}[X' \mid \mathcal{B}]$ a.s.

Proof (a) immediately folllows from uniqueness in Theorem 11.3. Similarly, for (b), we observe that, if $X, X' \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ and $\alpha, \alpha' \in \mathbb{R}$, the random variable

$$\alpha \, \mathbb{E}[X \mid \mathcal{B}] + \alpha' \mathbb{E}[X' \mid \mathcal{B}]$$

satisfies the characteristic property (11.1) for the conditional expectation of $\alpha X + \alpha' X'$. Property (c) is the special case $B = \Omega$ in (11.1). As for (d), using the fact that $X \ge 0$ implies $\mathbb{E}[X \mid \mathcal{B}] \ge 0$, we have

$$|\mathbb{E}[X \mid \mathcal{B}]| = |\mathbb{E}[X^+ \mid \mathcal{B}] - \mathbb{E}[X^- \mid \mathcal{B}]| \le \mathbb{E}[X^+ \mid \mathcal{B}] + \mathbb{E}[X^- \mid \mathcal{B}] = \mathbb{E}[|X| \mid \mathcal{B}].$$

Finally, (e) is immediate by linearity. □

11.2.2 Nonnegative Random Variables

We now turn to the definition of $\mathbb{E}[X \mid \mathcal{B}]$ for a nonnegative random variable X. In a way similar to the integral of nonnegative functions in Chapter 2, it will be convenient to allow X to take the value $+\infty$.

Theorem 11.4 *Let X be a random variable with values in $[0, \infty]$. There exists a $\mathcal{B}$-measurable random variable with values in $[0, \infty]$, which is denoted by $\mathbb{E}[X \mid \mathcal{B}]$ and is such that, for every nonnegative $\mathcal{B}$-measurable random variable Z,*

$$\mathbb{E}[X Z] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{B}] Z]. \tag{11.3}$$

Furthermore $\mathbb{E}[X \mid \mathcal{B}]$ is unique up to a $\mathcal{B}$-measurable set of probability zero.

Remark The uniqueness part of the theorem means that (as in the integrable case) we should view $\mathbb{E}[X \mid \mathcal{B}]$ as an equivalence class of $\mathcal{B}$-measurable random variables that are equal a.s. But of course it will be more convenient to speak about "the random variable" $\mathbb{E}[X \mid \mathcal{B}]$.

In the case where X is also integrable, the definition of the preceding theorem is consistent with that of Theorem 11.3. Indeed (11.3) with $Z = 1$ shows that $\mathbb{E}[\mathbb{E}[X \mid \mathcal{B}]] = \mathbb{E}[X]$, and in particular $\mathbb{E}[X \mid \mathcal{B}] < \infty$ a.s. (so that we can take a representative with values in $[0, \infty)$) and $\mathbb{E}[X \mid \mathcal{B}] \in L^1$. We then just have to note that (11.1) is the special case of (11.3) where $Z = \mathbf{1}_B$, $B \in \mathcal{B}$.

Similarly as in the case of integrable variables, we will refer to (11.3) (or to its variant when $Z = \mathbf{1}_B$, $B \in \mathcal{B}$) as the characteristic property of $\mathbb{E}[X \mid \mathcal{B}]$.

Proof We define $\mathbb{E}[X \mid \mathcal{B}]$ by setting

$$\mathbb{E}[X \mid \mathcal{B}] := \lim_{n\to\infty} \uparrow \mathbb{E}[X \wedge n \mid \mathcal{B}] \qquad \text{a.s.}$$

This definition makes sense because, for every $n \in \mathbb{N}$, $X \wedge n$ is bounded hence integrable and thus $\mathbb{E}[X \wedge n \mid \mathcal{B}]$ is well defined by the previous section. Furthermore, the fact that the sequence $(\mathbb{E}[X \wedge n \mid \mathcal{B}])_{n\in\mathbb{N}}$ is (a.s.) increasing follows from property (e) above. Then, if Z is nonnegative and $\mathcal{B}$-measurable, the monotone convergence theorem implies that

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{B}]Z] = \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[X \wedge n \mid \mathcal{B}](Z \wedge n)] = \lim_{n\to\infty} \mathbb{E}[(X \wedge n)(Z \wedge n)] = \mathbb{E}[XZ].$$

It remains to establish uniqueness. Let X' and X'' be two $\mathcal{B}$-measurable random variables with values in $[0, \infty]$, such that

$$\mathbb{E}[X'Z] = \mathbb{E}[X''Z]$$

for every nonnegative $\mathcal{B}$-measurable random variable Z. Let us fix two nonnegative rationals $a < b$, and take

$$Z = \mathbf{1}_{\{X' \le a < b \le X''\}}.$$

It follows that

$$\begin{aligned} a\,\mathbb{P}(X' \le a < b \le X'') &\ge \mathbb{E}[\mathbf{1}_{\{X' \le a < b \le X''\}} X'] \\ &= \mathbb{E}[\mathbf{1}_{\{X' \le a < b \le X''\}} X''] \\ &\ge b\,\mathbb{P}(X' \le a < b \le X'') \end{aligned}$$

which is only possible if $\mathbb{P}(X' \le a < b \le X'') = 0$. Hence,

$$\mathbb{P}\Big(\bigcup_{\substack{a,b \in \mathbb{Q}_+ \\ a<b}} \{X' \le a < b \le X''\} \Big) = 0$$

which implies $X' \ge X''$ a.s., and interchanging the roles of X' and X'' also gives $X'' \ge X'$ a.s. □

Remark We may have $X < \infty$ a.s. and at the same time $\mathbb{E}[X \,|\, \mathcal{B}] = \infty$ with positive probability. For instance, if $\mathcal{B} = \{\varnothing, \Omega\}$, it is immediate to verify that $\mathbb{E}[X \,|\, \mathcal{B}] = \mathbb{E}[X]$, which may be infinite even if $X < \infty$ a.s. To give a (slightly) less trivial example, consider again the case $\Omega = (0, 1]$, $\mathcal{B} = \sigma((\frac{i-1}{n}, \frac{i}{n}]; i \in \{1, \dots, n\})$ and $\mathbb{P}(d\omega) = d\omega$. Then, if $X(\omega) = \frac{1}{\omega}$, we have

$$\mathbb{E}[X \,|\, \mathcal{B}] = \infty\, \mathbf{1}_{(0,\frac{1}{n}]} + \sum_{i=2}^{n} n \, \log(\frac{i}{i-1})\, \mathbf{1}_{(\frac{i-1}{n}, \frac{i}{n}]}.$$

Properties . In the statement of the following properties, "nonnegative" means "with values in $[0, \infty]$".

(a) If X and X' are nonnegative random variables, and $a, b \ge 0$,

$$\mathbb{E}[aX + bX' \,|\, \mathcal{B}] = a\,\mathbb{E}[X \,|\, \mathcal{B}] + b\,\mathbb{E}[X' \,|\, \mathcal{B}].$$

(b) If X is nonnegative and $\mathcal{B}$-measurable, $\mathbb{E}[X \,|\, \mathcal{B}] = X$.
(c) For any nonnegative random variable X, $\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{B}]] = \mathbb{E}[X]$.
(d) If $(X_n)_{n \in \mathbb{N}}$ is an increasing sequence of nonnegative random variables, and $X = \lim \uparrow X_n$, then

$$\mathbb{E}[X \,|\, \mathcal{B}] = \lim_{n \to \infty} \uparrow \mathbb{E}[X_n \,|\, \mathcal{B}] \,, \qquad \text{a.s.}$$

As a useful consequence, if $(Y_n)_{n\in\mathbb{N}}$ is a sequence of nonnegative random variables, we have

$$\mathbb{E}\Big[\sum_{n\in\mathbb{N}} Y_n \,\Big|\, \mathcal{B}\Big] = \sum_{n\in\mathbb{N}} \mathbb{E}[Y_n \,|\, \mathcal{B}].$$

(e) If $(X_n)_{n\in\mathbb{N}}$ is any sequence of nonnegative random variables

$$\mathbb{E}[\liminf X_n \,|\, \mathcal{B}] \leq \liminf \mathbb{E}[X_n \,|\, \mathcal{B}] \,, \quad \text{a.s.}$$

(f) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of integrable random variables that converges a.s. to X. Assume that there exists a nonnegative random variable Z such that $|X_n| \leq Z$ a.s. for every $n \in \mathbb{N}$, and $\mathbb{E}[Z] < \infty$. Then,

$$\mathbb{E}[X \,|\, \mathcal{B}] = \lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{B}] \,, \quad \text{a.s. and in } L^1.$$

(g) (*Jensen's Inequality for Conditional Expectations*) If $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ is convex, and if $X \in L^1$,

$$\mathbb{E}[f(X) \,|\, \mathcal{B}] \geq f(\mathbb{E}[X \,|\, \mathcal{B}]).$$

Remarks (i) Of course (d)—(g) are the analogs for conditional expectations of properties of integrals with respect to a measure that were established in Chapter 2. The proofs below are indeed very similar to the proofs given in Chapter 2.

(iii) As we already mentioned, the words "almost surely" should appear in any statement involving a conditional expectation (in particular in (a), (b), (g) above).

Proof (a) and (b) are easy using the characteristic property (11.3), and (c) is the special case $Z = 1$ in (11.3).

(d) It follows from (a) that we have $\mathbb{E}[X \,|\, \mathcal{B}] \geq \mathbb{E}[Y \,|\, \mathcal{B}]$ if $X \geq Y \geq 0$. Under the assumptions of (d), we can therefore set $X' = \lim \uparrow \mathbb{E}[X_n \mid \mathcal{B}]$, which is a $\mathcal{B}$-measurable random variable with values in $[0, \infty]$. Then, for every nonnegative $\mathcal{B}$-measurable random variable Z, the monotone convergence theorem gives

$$\mathbb{E}[ZX'] = \lim \uparrow \mathbb{E}[Z\,\mathbb{E}[X_n \,|\, \mathcal{B}]] = \lim \uparrow \mathbb{E}[Z\,X_n] = \mathbb{E}[ZX]$$

which implies $X' = \mathbb{E}[X \mid \mathcal{B}]$ thanks to the characteristic property (11.3). The second assertion in (d) follows by applying the first one to $X_n = Y_1 + \cdots + Y_n$.

(e) Using (d), we have

$$\begin{aligned}\mathbb{E}[\liminf X_n \,|\, \mathcal{B}] &= \mathbb{E}\Big[\lim_{k\uparrow\infty} \uparrow \Big(\inf_{n\geq k} X_n\Big) \,\Big|\, \mathcal{B}\Big] \\ &= \lim_{k\uparrow\infty} \uparrow \mathbb{E}\Big[\inf_{n\geq k} X_n \,\Big|\, \mathcal{B}\Big]\end{aligned}$$

$$\leq \lim_{k\uparrow\infty}\Big(\inf_{n\geq k}\mathbb{E}[X_n \mid \mathcal{B}]\Big)$$
$$= \liminf \mathbb{E}[X_n \,|\, \mathcal{B}].$$

(f) It suffices to apply (e) twice:

$$\mathbb{E}[Z - X \,|\, \mathcal{B}] = \mathbb{E}[\liminf(Z - X_n) \,|\, \mathcal{B}] \leq \mathbb{E}[Z \,|\, \mathcal{B}] - \limsup \mathbb{E}[X_n \,|\, \mathcal{B}]$$
$$\mathbb{E}[Z + X \,|\, \mathcal{B}] = \mathbb{E}[\liminf(Z + X_n) \,|\, \mathcal{B}] \leq \mathbb{E}[Z \,|\, \mathcal{B}] + \liminf \mathbb{E}[X_n \,|\, \mathcal{B}]$$

which leads to

$$\mathbb{E}[X \,|\, \mathcal{B}] \leq \liminf \mathbb{E}[X_n \,|\, \mathcal{B}] \leq \limsup \mathbb{E}[X_n \,|\, \mathcal{B}] \leq \mathbb{E}[X \,|\, \mathcal{B}],$$

giving the desired almost sure convergence. The convergence in L^1 is now a consequence of the dominated convergence theorem, since we have $|\mathbb{E}[X_n \mid \mathcal{B}]| \leq \mathbb{E}[|X_n| \,|\, \mathcal{B}] \leq \mathbb{E}[Z \,|\, \mathcal{B}]$ and $\mathbb{E}[\mathbb{E}[Z \,|\, \mathcal{B}]] = \mathbb{E}[Z] < \infty$.

(g) Set

$$E_f = \{(a, b) \in \mathbb{R}^2 : \forall x \in \mathbb{R},\ f(x) \geq ax + b\}.$$

Then,

$$\forall x \in \mathbb{R}^2\,, \qquad f(x) = \sup_{(a,b)\in E_f} (ax + b) = \sup_{(a,b)\in E_f\cap\mathbb{Q}^2} (ax + b).$$

We can take advantage of the fact that $\mathbb{Q}^2$ is countable to disgard a countable collection of sets of probability zero and to get that, a.s.,

$$\mathbb{E}[f(X) \,|\, \mathcal{B}] = \mathbb{E}\Big[\sup_{(a,b)\in E_f\cap\mathbb{Q}^2} (aX + b) \,\Big|\, \mathcal{B}\Big]$$
$$\geq \sup_{(a,b)\in E_f\cap\mathbb{Q}^2} \mathbb{E}[aX + b \,|\, \mathcal{B}] = f(\mathbb{E}[X \,|\, \mathcal{B}]). \qquad \square$$

As a consequence of Jensen's inequality, we obtain that, for every $p \geq 1$, $X \mapsto \mathbb{E}[X \,|\, \mathcal{B}]$ maps $L^p(\Omega, \mathcal{A}, \mathbb{P})$ into itself, and is a contraction of $L^p(\Omega, \mathcal{A}, \mathbb{P})$. Indeed, for every $X \in L^p(\Omega, \mathcal{A}, \mathbb{P})$, we have using property (d) in Section 11.2.1,

$$\mathbb{E}[|\mathbb{E}[X \,|\, \mathcal{B}]|^p] \leq \mathbb{E}[\mathbb{E}[|X| \mid \mathcal{B}]^p] \leq \mathbb{E}[\mathbb{E}[|X|^p \,|\, \mathcal{B}]] = \mathbb{E}[|X|^p].$$

This contraction property also holds for $p = \infty$.

Remark By analogy with the formula $\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A]$, we often write, for every $A \in \mathcal{A}$,

$$\mathbb{P}(A \,|\, \mathcal{B}) = \mathbb{E}[\mathbf{1}_A \,|\, \mathcal{B}].$$

Beware that $\mathbb{P}(A \,|\, \mathcal{B})$ is a *random variable* and not a constant !

11.2.3 The Special Case of Square Integrable Variables

In the case where X is in L^2, there is another remarkable interpretation of $\mathbb{E}[X \mid \mathcal{B}]$, which involves the Hilbert space structure of L^2. Before giving the precise statement, let us observe that $L^2(\Omega, \mathcal{B}, \mathbb{P})$ is isometrically identified to a closed subspace of $L^2(\Omega, \mathcal{A}, \mathbb{P})$, namely the subspace consisting of all elements of $L^2(\Omega, \mathcal{A}, \mathbb{P})$ that have at least one representative which is $\mathcal{B}$-measurable. We can thus make sense of the orthogonal projection of an element of the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$ on the (closed) subspace $L^2(\Omega, \mathcal{B}, \mathbb{P})$, see Theorem A.3 in the Appendix.

Theorem 11.5 *If $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, then $\mathbb{E}[X \mid \mathcal{B}]$ is the orthogonal projection of X on $L^2(\Omega, \mathcal{B}, \mathbb{P})$.*

Proof Jensen's inequality shows that $\mathbb{E}[X \,|\, \mathcal{B}]^2 \leq \mathbb{E}[X^2 \,|\, \mathcal{B}]$, a.s. This implies that $\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{B}]^2] \leq \mathbb{E}[\mathbb{E}[X^2 \,|\, \mathcal{B}]] = \mathbb{E}[X^2] < \infty$, and thus the random variable $\mathbb{E}[X \,|\, \mathcal{B}]$ belongs to $L^2(\Omega, \mathcal{B}, \mathbb{P})$.

On the other hand, for every bounded $\mathcal{B}$-measurable random variable Z,

$$\mathbb{E}[Z(X - \mathbb{E}[X \,|\, \mathcal{B}])] = \mathbb{E}[ZX] - \mathbb{E}[Z\mathbb{E}[X \,|\, \mathcal{B}]] = 0,$$

by the characteristic property (11.2). Hence $X - \mathbb{E}[X \mid \mathcal{B}]$ is orthogonal to the space of all bounded $\mathcal{B}$-measurable random variables, and the latter space is dense in $L^2(\Omega, \mathcal{B}, \mathbb{P})$ (for instance by Theorem 4.8). It follows that $X - \mathbb{E}[X \mid \mathcal{B}]$ is orthogonal to $L^2(\Omega, \mathcal{B}, \mathbb{P})$, which gives the desired result. □

Remark We could have used the property stated in the last theorem to give an alternative construction of the conditional expectation (starting with the case of square integrable random variables) and this construction would have avoided the use of the Radon-Nikodym theorem—however some work would have been necessary to extend the definition to the integrable and the nonnegative case.

We observe that Theorem 11.5 and the property of orthogonal projections stated in Theorem A.3 lead to an interesting interpretation of the conditional expectation of a square integrable random variable: $\mathbb{E}[X \mid \mathcal{B}]$ is the best approximation of X by a $\mathcal{B}$-measurable random variable, in the sense that, for any other $\mathcal{B}$-measurable random variable Y, we have $\mathbb{E}[(Y - X)^2] \geq \mathbb{E}[(\mathbb{E}[X \,|\, \mathcal{B}] - X)^2]$.

To illustrate the last observation, let $Y_1, \ldots, Y_k$ be real random variables. By Proposition 8.9, any $\sigma(Y_1, \ldots, Y_k)$-measurable random variable can be written as a Borel measurable function of $(Y_1, \ldots, Y_k)$. Consequently, if $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, there exists a measurable function $\varphi : \mathbb{R}^k \longrightarrow \mathbb{R}$ such that $\mathbb{E}[\varphi(Y_1, \ldots, Y_k)^2] < \infty$ and

$$\mathbb{E}[X \,|\, Y_1, \ldots, Y_k] = \varphi(Y_1, \ldots, Y_k).$$

Moreover, we have

$$\mathbb{E}[(X - \varphi(Y_1, \ldots, Y_k))^2] = \inf_{g:\mathbb{R}^k \longrightarrow \mathbb{R}} \mathbb{E}[(X - g(Y_1, \ldots, Y_k))^2], \tag{11.4}$$

where the infimum is over all Borel measurable functions $g : \mathbb{R}^k \longrightarrow \mathbb{R}$. Note that we should a priori restrict the infimum in (11.4) to functions g such that $\mathbb{E}[g(Y_1, \ldots, Y_k)^2] < \infty$, but in fact, if g does not satisfy this condition, one has $\mathbb{E}[(X - g(Y_1, \ldots, Y_k))^2] = \infty$.

It is interesting to compare (11.4) with Proposition 8.13, which was concerned with the best approximation of X by an affine function of $(Y_1, \ldots, Y_k)$. Here it seems that we are doing much better, since we can consider any (measurable) function of $(Y_1, \ldots, Y_k)$. However, computing a conditional expectation is usually a hard problem, whereas finding the approximation in Proposition 8.13 only involves solving a finite-dimensional linear system.

11.3 Specific Properties of the Conditional Expectation

Most of the properties of conditional expectations that we have derived until now were very similar to the properties of the integral of measurable functions. In this section we derive more specific properties of conditional expectations. The next two propositions are extremely useful when manipulating conditional expectations.

Proposition 11.6 *Let X and Y be real random variables and assume that Y is $\mathcal{B}$-measurable. Then*

$$\mathbb{E}[YX \,|\, \mathcal{B}] = Y\, \mathbb{E}[X \,|\, \mathcal{B}]$$

provided X and Y are both nonnegative, or X and YX are both integrable.

Proof Suppose that $X \geq 0$ and $Y \geq 0$. Then, for every nonnegative $\mathcal{B}$-measurable random variable Z, we have

$$\mathbb{E}[Z(Y\mathbb{E}[X \,|\, \mathcal{B}])] = \mathbb{E}[(ZY)\mathbb{E}[X \,|\, \mathcal{B}]] = \mathbb{E}[ZYX],$$

using (11.3) and the fact that ZY is $\mathcal{B}$-measurable. Since $Y\mathbb{E}[X\,|\,\mathcal{B}]$ is a nonnegative $\mathcal{B}$-measurable random variable, the characteristic property (11.3) implies that $Y\mathbb{E}[X\,|\,\mathcal{B}] = \mathbb{E}[YX\,|\,\mathcal{B}]$.

In the case when X and YX are integrable, we get the desired result by decomposing $X = X^+ - X^-$ and $Y = Y^+ - Y^-$. □

Proposition 11.7 (Nested σ-Fields) *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two sub-σ-fields of $\mathcal{A}$ such that $\mathcal{B}_1 \subset \mathcal{B}_2$. Then, for every nonnegative (or integrable) random variable X, we have*

$$\mathbb{E}[\mathbb{E}[X\,|\,\mathcal{B}_2]\,|\,\mathcal{B}_1] = \mathbb{E}[X\,|\,\mathcal{B}_1].$$

Remark We have also $\mathbb{E}[\mathbb{E}[X \mid \mathcal{B}_1] \mid \mathcal{B}_2] = \mathbb{E}[X \mid \mathcal{B}_1]$ under the same assumptions, but this is trivial since $\mathbb{E}[X\,|\,\mathcal{B}_1]$ is $\mathcal{B}_2$-measurable.

Proof Consider the case where $X \geq 0$. Let Z be a nonnegative $\mathcal{B}_1$-measurable random variable. Then, since Z is also $\mathcal{B}_2$-measurable,

$$\mathbb{E}[Z\,\mathbb{E}[\mathbb{E}[X\,|\,\mathcal{B}_2]\,|\,\mathcal{B}_1]] = \mathbb{E}[Z\,\mathbb{E}[X\,|\,\mathcal{B}_2]] = \mathbb{E}[ZX].$$

Hence $\mathbb{E}[\mathbb{E}[X \mid \mathcal{B}_2] \mid \mathcal{B}_1]$ satisfies the characteristic property of $\mathbb{E}[X \mid \mathcal{B}_1]$, and the desired result follows. □

The next theorem provides a characterization of independence in terms of conditional expectations.

Theorem 11.8 *Two sub-σ-fields $\mathcal{B}_1$ and $\mathcal{B}_2$ of $\mathcal{A}$ are independent if and only if, for every $B \in \mathcal{B}_2$, we have $\mathbb{E}[\mathbf{1}_B \mid \mathcal{B}_1] = \mathbb{P}(B)$. Furthermore, if $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent, we have also $\mathbb{E}[X\,|\,\mathcal{B}_1] = \mathbb{E}[X]$ for every nonnegative $\mathcal{B}_2$-measurable random variable X and for every $X \in L^1(\Omega, \mathcal{B}_2, \mathbb{P})$.*

Proof Suppose that $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent, and let X be a nonnegative $\mathcal{B}_2$-measurable random variable Then, for every nonnegative $\mathcal{B}_1$-measurable random variable Z, Z and X are independent and thus

$$\mathbb{E}[ZX] = \mathbb{E}[Z]\mathbb{E}[X] = \mathbb{E}[Z\,\mathbb{E}[X]],$$

and thus the constant random variable $\mathbb{E}[X]$ satisfies the characteristic property (11.3) of $\mathbb{E}[X \mid \mathcal{B}_1]$, which implies that $\mathbb{E}[X \mid \mathcal{B}_1] = \mathbb{E}[X]$. We obtain the same result in the case where X integrable by decomposing $X = X^+ - X^-$.

Conversely, suppose that

$$\forall B \in \mathcal{B}_2\,, \quad \mathbb{E}[\mathbf{1}_B\,|\,\mathcal{B}_1] = \mathbb{P}(B).$$

Then, for every $A \in \mathcal{B}_1$, $B \in \mathcal{B}_2$,

$$\mathbb{P}(A \cap B) = \mathbb{E}[\mathbf{1}_A \mathbf{1}_B] = \mathbb{E}[\mathbf{1}_A\, \mathbb{E}[\mathbf{1}_B \,|\, \mathcal{B}_1]] = \mathbb{E}[\mathbf{1}_A\, \mathbb{P}(B)] = \mathbb{P}(A)\mathbb{P}(B)$$

which shows that the σ-fields $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent. □

Remark Let X and Y be two real random variables. Since the real random variables that are $\sigma(X)$-measurable are exactly the measurable functions of X (Proposition 8.9), the preceding theorem shows that X and Y are independent if and only if

$$\mathbb{E}[h(X) \,|\, Y] = \mathbb{E}[h(X)]$$

for every Borel function $h : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\mathbb{E}[|h(X)|] < \infty$. Assuming that X is integrable, we have in particular

$$\mathbb{E}[X \,|\, Y] = \mathbb{E}[X].$$

However, this last property alone is not sufficient to infer that X and Y are independent. Let us illustrate this on an example. Suppose that X is Gaussian $\mathcal{N}(0, 1)$, and that $Y = |X|$. Any bounded $\sigma(Y)$-measurable random variable Z is of the form $Z = g(Y)$, where g is measurable and bounded, and thus

$$\mathbb{E}[ZX] = \mathbb{E}[g(|X|)X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(|y|)\, y\, e^{-y^2/2}\, \mathrm{d}y = 0,$$

by a symmetry argument. It follows that $\mathbb{E}[X \mid Y] = 0 = \mathbb{E}[X]$, but X and Y are of course not independent.

We conclude this section with two other results relating the notions of conditional expectation and independence, which are often useful for explicit calculations of conditional expectations.

Theorem 11.9 *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces, and let X and Y be random variables taking values in E and F respectively. Assume that X is independent of $\mathcal{B}$ and that Y is $\mathcal{B}$-measurable.Then, for every $\mathcal{E} \otimes \mathcal{F}$-measurable function $g : E \times F \longrightarrow \mathbb{R}_+$,*

$$\mathbb{E}[g(X, Y) \mid \mathcal{B}] = \int g(x, Y)\, \mathbb{P}_X(\mathrm{d}x),$$

where we recall that $\mathbb{P}_X$ denotes the law of X. The right-hand side is the composition of the random variable Y with the mapping $\Psi : F \longrightarrow \mathbb{R}_+$ defined by $\Psi(y) = \int g(x, y)\, \mathbb{P}_X(\mathrm{d}x)$.

Remark The mapping Ψ is measurable thanks to the Fubini theorem. The theorem can be explained informally as follows. If we condition with respect to the σ-field $\mathcal{B}$, the random variable Y, which is $\mathcal{B}$-measurable, behaves like a constant, but on the other hand $\mathcal{B}$ gives no information on the random variable X. Thus the best approximation of $g(X, Y)$ knowing $\mathcal{B}$ is obtained by integrating $g(\cdot, Y)$ with respect to the law of X.

Proof We have to prove that, for any nonnegative $\mathcal{B}$-measurable random variable Z, we have

$$\mathbb{E}[g(X, Y)Z] = \mathbb{E}[\Psi(Y)Z].$$

Write $\mathbb{P}_{(X,Y,Z)}$ for the law of the triple (X, Y, Z), which is a probability measure on $E \times F \times \mathbb{R}_+$. Since X is independent of $\mathcal{B}$, X is independent of the pair (Y, Z), and therefore $\mathbb{P}_{(X,Y,Z)} = \mathbb{P}_X \otimes \mathbb{P}_{(Y,Z)}$. Then, using the Fubini theorem, we have

$$\begin{aligned}
\mathbb{E}[g(X, Y)Z] &= \int g(x, y)z\, \mathbb{P}_{(X,Y,Z)}(\mathrm{d}x\mathrm{d}y\mathrm{d}z) \\
&= \int g(x, y)z\, \mathbb{P}_X(\mathrm{d}x)\mathbb{P}_{(Y,Z)}(\mathrm{d}y\mathrm{d}z) \\
&= \int_{F\times\mathbb{R}_+} z \left(\int_E g(x, y)\mathbb{P}_X(\mathrm{d}x)\right)\mathbb{P}_{(Y,Z)}(\mathrm{d}y\mathrm{d}z) \\
&= \int_{F\times\mathbb{R}_+} z\Psi(y)\, \mathbb{P}_{(Y,Z)}(\mathrm{d}y\mathrm{d}z) \\
&= \mathbb{E}[\Psi(Y)Z]
\end{aligned}$$

which was the desired result. □

Recall that, if $\mathcal{B}$ and $\mathcal{B}'$ are two sub-σ-fields of $\mathcal{A}$, $\mathcal{B} \vee \mathcal{B}'$ denotes the smallest σ-field that contains both $\mathcal{B}$ and $\mathcal{B}'$.

Proposition 11.10 *Let Z be a random variable in L^1, and let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two sub-σ-fields of $\mathcal{A}$. Assume that $\mathcal{H}_2$ is independent of $\sigma(Z) \vee \mathcal{H}_1$. Then,*

$$\mathbb{E}[Z \,|\, \mathcal{H}_1 \vee \mathcal{H}_2] = \mathbb{E}[Z \,|\, \mathcal{H}_1]$$

Proof It suffices to prove that the equality

$$\mathbb{E}[\mathbf{1}_A Z] = \mathbb{E}[\mathbf{1}_A \mathbb{E}[Z \,|\, \mathcal{H}_1]] \tag{11.5}$$

holds for every $A \in \mathcal{H}_1 \vee \mathcal{H}_2$. Consider first the case where $A = B \cap C$, with $B \in \mathcal{H}_1$ and $C \in \mathcal{H}_2$. Then, we have

$$\begin{aligned}\mathbb{E}[\mathbf{1}_A Z] = \mathbb{E}[\mathbf{1}_C \times \mathbf{1}_B Z] = \mathbb{P}(C)\mathbb{E}[\mathbf{1}_B Z] &= \mathbb{P}(C)\mathbb{E}[\mathbf{1}_B \mathbb{E}[Z \mid \mathcal{H}_1]] \\ &= \mathbb{E}[\mathbf{1}_C \times \mathbf{1}_B \mathbb{E}[Z \mid \mathcal{H}_1]] \\ &= \mathbb{E}[\mathbf{1}_A \mathbb{E}[Z \mid \mathcal{H}_1]].\end{aligned}$$

Thus the class of all sets $A \in \mathcal{H}_1 \vee \mathcal{H}_2$ that satisfy (11.5) contains a class closed under finite intersections that generates the σ-field $\mathcal{H}_1 \vee \mathcal{H}_2$. An easy application of the monotone class theorem (Theorem 1.18) shows that (11.5) holds for *every* $A \in \mathcal{H}_1 \vee \mathcal{H}_2$. □

11.4 Evaluation of Conditional Expectation

11.4.1 Discrete Conditioning

Let Y be a random variable with values in a countable space E, and let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Then, we have already seen that

$$\mathbb{E}[X \mid Y] = \varphi(Y)$$

where

$$\varphi(y) = \frac{\mathbb{E}[X\, \mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)}$$

for every $y \in E$ such that $\mathbb{P}(Y = y) > 0$ (and $\varphi(y)$ can be chosen in an arbitrary manner when $\mathbb{P}(Y = y) = 0$).

11.4.2 Random Variables with a Density

Let X and Y be two random variables with values in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. We assume that the law of the pair (X, Y) has a density with respect to Lebesgue measure, which is denoted by $p(x, y)$, for $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. Then, for every Borel measurable function $f : \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}_+$,

$$\mathbb{E}[f(X, Y)] = \int_{\mathbb{R}^m \times \mathbb{R}^n} f(x, y)\, p(x, y)\, \mathrm{d}x\mathrm{d}y.$$

An easy generalization of Proposition 8.6 shows that Y has a density given by

$$q(y) = \int_{\mathbb{R}^m} p(x, y)\,\mathrm{d}x\,, \quad \text{for } y \in \mathbb{R}^n.$$

The right-hand side may be equal to ∞ on a set of values of y of zero Lebesgue measure: to ensure that q takes values in $\mathbb{R}_+$, we agree to take $q(y) = 0$ if $\int p(x, y)\,\mathrm{d}x = \infty$.

Next let $h : \mathbb{R}^m \longrightarrow \mathbb{R}_+$ be a measurable function. We may compute $\mathbb{E}[h(X) \,|\, Y]$ as follows. For every measurable function $g : \mathbb{R}^n \longrightarrow \mathbb{R}_+$, we have

$$\begin{aligned}
\mathbb{E}[h(X)g(Y)] &= \int_{\mathbb{R}^m \times \mathbb{R}^n} h(x)\, g(y)\, p(x, y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{\mathbb{R}^n} \Big(\int_{\mathbb{R}^m} h(x)\, p(x, y)\,\mathrm{d}x \Big) g(y)\,\mathrm{d}y \\
&= \int_{\mathbb{R}^n} \Big(\int_{\mathbb{R}^m} h(x)\, p(x, y)\,\mathrm{d}x \Big) g(y)\, \mathbf{1}_{\{q(y)>0\}}\mathrm{d}y,
\end{aligned}$$

where the last equality is justified by observing that for Lebesgue a.e. y such that $q(y) = 0$, we have $p(x, y) = 0$, $\mathrm{d}x$ a.e., and thus $\int h(x)\, p(x, y)\,\mathrm{d}x = 0$. Then,

$$\begin{aligned}
\mathbb{E}[h(X)g(Y)] &= \int_{\mathbb{R}^n} \Big(\frac{\int_{\mathbb{R}^m} h(x)\, p(x, y)\,\mathrm{d}x}{q(y)} \Big) g(y)\, q(y) \mathbf{1}_{\{q(y)>0\}}\,\mathrm{d}y \\
&= \int_{\mathbb{R}^n} \varphi(y)\, g(y)\, q(y) \mathbf{1}_{\{q(y)>0\}}\,\mathrm{d}y \\
&= \mathbb{E}[\varphi(Y)\, g(Y)],
\end{aligned}$$

where we have set

$$\varphi(y) = \begin{cases} \dfrac{1}{q(y)} \displaystyle\int_{\mathbb{R}^m} h(x)\, p(x, y)\,\mathrm{d}x & \text{if } q(y) > 0, \\ h(0) & \text{if } q(y) = 0 \end{cases}$$

(the value of $\varphi(y)$ when $q(y) = 0$ can be chosen in an arbitrary way: the choice of the value $h(0)$ will be convenient for the next statement).

It follows from the preceding calculation and the characteristic property of the conditional expectation that we have

$$\mathbb{E}[h(X) \,|\, Y] = \varphi(Y).$$

We reformulate this in a slightly different manner.

Proposition 11.11 *For every $y \in \mathbb{R}^n$, let $\nu(y, \mathrm{d}x)$ be the probability measure on $\mathbb{R}^m$ defined by*

$$\nu(y, \mathrm{d}x) := \begin{cases} \dfrac{1}{q(y)}\, p(x, y)\, \mathrm{d}x & \text{if } q(y) > 0, \\ \delta_0(\mathrm{d}x) & \text{if } q(y) = 0. \end{cases}$$

Then, for every measurable function $h : \mathbb{R}^m \longrightarrow \mathbb{R}_+$, we have

$$\mathbb{E}[h(X) \,|\, Y] = \int h(x)\, \nu(Y, \mathrm{d}x).$$

We often write, in a slightly abusive manner, for every $y \in \mathbb{R}$ such that $q(y) > 0$,

$$\mathbb{E}[h(X) \,|\, Y = y] = \int h(x)\, \nu(y, \mathrm{d}x) = \frac{1}{q(y)} \int h(x)\, p(x, y)\, \mathrm{d}x,$$

and we say that

$$x \mapsto \frac{p(x, y)}{q(y)}$$

is the conditional density of X knowing that $Y = y$.

Remark In the setting of Proposition 11.11, one has more generally, for any measurable function $h : \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}_+$,

$$\mathbb{E}[h(X, Y) \,|\, Y] = \int h(x, Y)\, \nu(Y, \mathrm{d}x).$$

See Exercise 11.13.

11.4.3 Gaussian Conditioning

Let $X, Y_1, \ldots, Y_p$ be $p + 1$ real random variables in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. As we saw in Section 11.2.3, the conditional expectation

$$\mathbb{E}[X \,|\, Y_1, \ldots, Y_p]$$

is the orthogonal projection of X on the space $L^2(\Omega, \sigma(Y_1, \ldots, Y_p), \mathbb{P})$, which is typically an infinite-dimensional linear space. This orthogonal projection is the best approximation of X (in the L^2 sense) by a random variable of the form $\varphi(Y_1, \ldots, Y_p)$ (see the remarks following Theorem 11.5).

On the other hand, we also studied (in Section 8.2.2) the best approximation of X by an affine function of $Y_1, \dots, Y_p$, which is the orthogonal projection of X on the vector space spanned by $1, Y_1, \dots, Y_p$. In general, the latter projection is very different from the conditional expectation $\mathbb{E}[X \mid Y_1, \dots, Y_p]$ which provides a much better approximation of X. We will however study a Gaussian setting where the conditional expectation coincides with the best approximation by an affine function. This has the enormous advantage of reducing calculations of conditional expectations to projections in finite dimension.

Recall from Section 10.4.3 that a random variable $Z = (Z_1, \dots, Z_k)$ with values in $\mathbb{R}^k$ is a Gaussian vector if any linear combination of the components $Z_1, \dots, Z_k$ is a Gaussian random variable, and this is equivalent to saying that $Z_1, \dots, Z_k$ are in L^2 and

$$\forall \xi \in \mathbb{R}^k\,, \quad \mathbb{E}[\exp(\mathrm{i}\,\xi \cdot Z)] = \exp\Big(\mathrm{i}\,\xi \cdot \mathbb{E}[Z] - \frac{1}{2}\,{}^t\xi K_Z \xi\Big), \tag{11.6}$$

where we recall that K_Z denotes the covariance function of Z. This property holds in particular if $Z_1, \dots, Z_k$ are independent Gaussian real random variables.

Proposition 11.12 *Let $(X_1, \dots, X_m, Y_1, \dots, Y_n)$ be a Gaussian vector. Then the vectors $(X_1, \dots, X_m)$ and $(Y_1, \dots, Y_n)$ are independent if and only if*

$$\mathrm{cov}(X_i, Y_j) = 0\,, \qquad \forall i \in \{1, \dots, m\},\ j \in \{1, \dots, n\}. \tag{11.7}$$

Consequently, if $(Z_1, \dots, Z_k)$ is a Gaussian vector whose covariance matrix is diagonal, its components $Z_1, \dots, Z_k$ are independent.

Proof For the first assertion, we only need to prove that the condition (11.7) implies the independence of $(X_1, \dots, X_m)$ and $(Y_1, \dots, Y_n)$ (the converse is true by Corollary 9.5). For every $\xi = (\eta_1, \dots, \eta_m, \zeta_1, \dots, \zeta_n) \in \mathbb{R}^{m+n}$, we have by (11.6),

$$\begin{aligned}&\mathbb{E}[\exp(\mathrm{i}\,\xi \cdot (X_1, \dots, X_m, Y_1, \dots, Y_n))]\\ &\quad = \exp(\mathrm{i}\,\xi \cdot \mathbb{E}[(X_1, \dots, X_m, Y_1, \dots, Y_n)] - \frac{1}{2}\,{}^t\xi K_{(X_1,\dots,X_m,Y_1,\dots,Y_n)}\xi).\end{aligned}$$

Writing $\eta = (\eta_1, \dots, \eta_m)$ and $\zeta = (\zeta_1, \dots, \zeta_n)$, we have

$$\xi \cdot \mathbb{E}[(X_1, \dots, X_m, Y_1, \dots, Y_n)] = \eta \cdot \mathbb{E}[(X_1, \dots, X_m)] + \zeta \cdot \mathbb{E}[(Y_1, \dots, Y_m)],$$

and, under the condition (11.7),

$${}^t\xi K_{(X_1,\dots,X_m,Y_1,\dots,Y_n)}\xi = {}^t\eta K_{(X_1,\dots,X_m)}\eta + {}^t\zeta K_{(Y_1,\dots,Y_n)}\zeta.$$

From the last two displays, it follows that

$$\begin{aligned}&\mathbb{E}[\exp(\mathrm{i}\,\xi\cdot(X_1,\ldots,X_m,Y_1,\ldots,Y_n))]\\&\quad=\mathbb{E}[\exp(\mathrm{i}\,\eta\cdot(X_1,\ldots,X_m))]\times\mathbb{E}[\exp(\mathrm{i}\,\zeta\cdot(Y_1,\ldots,Y_n))],\end{aligned}$$

or equivalently

$$\begin{aligned}&\widehat{\mathbb{P}}_{(X_1,\ldots,X_m,Y_1,\ldots,Y_n)}(\eta_1,\ldots,\eta_m,\zeta_1,\ldots,\zeta_n)\\&\quad=\widehat{\mathbb{P}}_{(X_1,\ldots,X_m)}(\eta_1,\ldots,\eta_m)\,\widehat{\mathbb{P}}_{(Y_1,\ldots,Y_n)}(\zeta_1,\ldots,\zeta_n),\end{aligned}$$

and the right-hand side is the Fourier transform of $\mathbb{P}_{(X_1,\ldots,X_m)}\otimes\mathbb{P}_{(Y_1,\ldots,Y_n)}$ evaluated at $(\eta_1,\ldots,\eta_m,\zeta_1,\ldots,\zeta_n)$. The injectivity of the Fourier transform (Theorem 8.16) now gives

$$\mathbb{P}_{(X_1,\ldots,X_m,Y_1,\ldots,Y_n)}=\mathbb{P}_{(X_1,\ldots,X_m)}\otimes\mathbb{P}_{(Y_1,\ldots,Y_n)}$$

which was the desired result.

The second assertion is a consequence of the first one, since we obtain first that Z_k is independent of $(Z_1,\ldots,Z_{k-1})$, then that Z_{k-1} is independent of $(Z_1,\ldots,Z_{k-2}),\ldots,Z_2$ is independent of Z_1. This implies the independence of $Z_1,\ldots,Z_k$. □

Remark If the covariance matrix of a Gaussian vector $(X_1,\ldots,X_n)$ is diagonal by blocks of respective sizes $i_1,\ldots,i_\ell$ (such that $i_1+\cdots+i_\ell=n$) the "sub-vectors" $(X_1,\ldots,X_{i_1}),(X_{i_1+1},\ldots,X_{i_1+i_2}),\ldots,(X_{i_1+\cdots+i_{\ell-1}+1},\ldots,X_n)$ are independent. This immediately follows from the first assertion of the proposition.

A random vector $(Z_1,\ldots,Z_n)$ is said to be centered if its components are in L^1 and $\mathbb{E}[Z_j]=0$ for every $j\in\{1,\ldots,n\}$. For the sake of simplicity, we consider a centered Gaussian vector in the following statement. The reader will be able to state the similar result in the non-centered case.

Theorem 11.13 *Let $(Y_1,\ldots,Y_n,X)$ be a centered Gaussian vector. Then $\mathbb{E}[X\mid Y_1,\ldots,Y_n]$ is equal to the orthogonal projection (in L^2) of X on the vector space spanned by $Y_1,\ldots,Y_n$. Hence there exist reals $\lambda_1,\ldots,\lambda_n$ such that*

$$\mathbb{E}[X\,|\,Y_1,\ldots,Y_n]=\sum_{j=1}^{n}\lambda_jY_j.$$

Let $\sigma^2=\mathbb{E}[(X-\mathbb{E}[X\mid Y_1,\ldots,Y_n])^2]$, and assume that $\sigma>0$. Then, for every measurable function $h:\mathbb{R}\longrightarrow\mathbb{R}_+$,

$$\mathbb{E}[h(X)\,|\,Y_1,\ldots,Y_n]=\int_{\mathbb{R}}h(x)\,q_{\sum_{j=1}^{n}\lambda_jY_j,\sigma^2}(x)\,\mathrm{d}x,\tag{11.8}$$

where, for every $m \in \mathbb{R}$*,*

$$q_{m,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-m)^2}{2\sigma^2})$$

is the density of the Gaussian $\mathcal{N}(m,\sigma^2)$ *distribution, and the right-hand side of* (11.8) *is the composition of the random variable* $\sum_{j=1}^n \lambda_j Y_j$ *with the function* $m \mapsto \int_{\mathbb{R}} h(x)\, q_{m,\sigma^2}(x)\,\mathrm{d}x$.

Remark We have $\sigma = 0$ if and only if $X = \sum_{j=1}^n \lambda_j Y_j$ a.s., and in that case $\mathbb{E}[X \mid Y_1,\ldots,Y_n] = X$ and $\mathbb{E}[h(X) \mid Y_1,\ldots,Y_n] = h(X)$.

Proof Let $\widehat{X} = \sum_{j=1}^n \lambda_j Y_j$ be the orthogonal projection of X on the vector space spanned by $Y_1,\ldots,Y_n$. Then, for every $j \in \{1,\ldots,n\}$,

$$\mathrm{cov}(X-\widehat{X}, Y_j) = \mathbb{E}[(X-\widehat{X})Y_j] = 0$$

by the definition of the orthogonal projection. Since $(Y_1,\ldots,Y_n,X-\widehat{X})$ is a Gaussian vector (any linear combination of its components if also a linear combination of $Y_1,\ldots,Y_n,X$), Proposition 11.12 shows that $X-\widehat{X}$ is independent of $(Y_1,\ldots,Y_n)$. Therefore,

$$\mathbb{E}[X \mid Y_1,\ldots,Y_n] = \mathbb{E}[X-\widehat{X} \mid Y_1,\ldots,Y_n] + \widehat{X} = \mathbb{E}[X-\widehat{X}] + \widehat{X} = \widehat{X}.$$

We used the fact that $\widehat{X}$ is measurable with respect to $\sigma(Y_1,\ldots,Y_n)$, and then the independence of $X-\widehat{X}$ and $(Y_1,\ldots,Y_n)$, which, thanks to Theorem 11.8, implies $\mathbb{E}[X-\widehat{X} \mid Y_1,\ldots,Y_n] = \mathbb{E}[X-\widehat{X}] = 0$.

For the last assertion, set $Z = X - \widehat{X}$, so that Z is independent of $(Y_1,\ldots,Y_n)$ and distributed according to $\mathcal{N}(0,\sigma^2)$ (by definition, $\sigma^2 = \mathbb{E}[Z^2]$). We then use Theorem 11.9, which shows that

$$\begin{aligned}\mathbb{E}[h(X) \mid Y_1,\ldots,Y_n] &= \mathbb{E}\Big[h\Big(\sum_{j=1}^n \lambda_j Y_j + Z\Big)\,\Big|\, Y_1,\ldots,Y_n\Big] \\ &= \int h\Big(\sum_{j=1}^n \lambda_j Y_j + z\Big)\,\mathbb{P}_Z(\mathrm{d}z).\end{aligned}$$

Writing $\mathbb{P}_Z(\mathrm{d}z) = q_{0,\sigma^2}(z)\mathrm{d}z$ and making the obvious change of variable, we arrive at the stated formula. □

11.5 Transition Probabilities and Conditional Distributions

The calculations made in Section 11.4 can be reformulated in a more convenient manner thanks to the notion of a transition probability.

Definition 11.14 Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces. A transition probability (or transition kernel) from E into F is a mapping

$$\nu : E \times \mathcal{F} \longrightarrow [0, 1]$$

which satisfies the following two properties:

(i) for every $x \in E$, $A \mapsto \nu(x, A)$ is a probability measure on $(F, \mathcal{F})$;
(ii) for every $A \in \mathcal{F}$, the function $x \mapsto \nu(x, A)$ is $\mathcal{E}$-measurable.

At an intuitive level, each time one fixes a "starting point" $x \in E$, the probability measure $\nu(x, \cdot)$ gives a way to choose an "arrival point" $y \in F$ in a random manner (but with a law depending on the starting point x). This notion plays a fundamental role in the theory of Markov chains that we will study in Chapter 13.

Example Let μ be a σ-finite measure on $(F, \mathcal{F})$, and let $f : E \times F \longrightarrow \mathbb{R}_+$ be an $\mathcal{E} \otimes \mathcal{F}$-measurable function such that

$$\int_F f(x, y)\, \mu(\mathrm{d}y) = 1 , \qquad \forall x \in E.$$

Then

$$\nu(x, A) = \int_A f(x, y)\, \mu(\mathrm{d}y)$$

defines a transition probability from E into F. In particular, property (ii) of the definition follows from the Fubini theorem.

In part (i) of the following statement, "nonnegative" means "with values in $[0, \infty]$".

Proposition 11.15 *Let ν be a transition probability from E into F.*

(i) *If h is a nonnegative (resp. bounded) measurable function on $(F, \mathcal{F})$, then*

$$\varphi(x) = \int \nu(x, \mathrm{d}y)\, h(y) , \;\; x \in E$$

is a nonnegative (resp. bounded) measurable function on E.

(ii) *If* γ *is a probability measure on* $(E, \mathcal{E})$*, then*

$$\mu(A) = \int \gamma(\mathrm{d}x)\, \nu(x, A)\;,\;\; A \in \mathcal{F}$$

is a probability measure on $(F, \mathcal{F})$.

We omit the easy proof of the proposition. When treating the nonnegative case of (i), we first consider simple functions, and then use an increasing passage to the limit.

We now come to the relation between the notion of a transition probability and conditional expectations.

Definition 11.16 Let X and Y be two random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively. Any transition probability ν from E into F such that, for every nonnegative measurable function h on F,

$$\mathbb{E}[h(Y) \,|\, X] = \int \nu(X, \mathrm{d}y)\, h(y)\;, \quad \text{a.s.},$$

is called a conditional distribution of Y knowing X.

Remark The random variable $\int \nu(X, \mathrm{d}y)\, h(y)$ is the composition of X with the function $x \mapsto \int \nu(x, \mathrm{d}y)\, h(y)$, which is measurable by Proposition 11.15. In particular, this random variable is a function of X, as $\mathbb{E}[h(Y) \,|\, X]$ should be.

By definition, if ν is a conditional distribution of Y knowing X, we have, for every $A \in \mathcal{F}$,

$$\mathbb{P}(Y \in A \,|\, X) = \nu(X, A)\;,\;\; \text{a.s.}$$

It is tempting to replace this equality of random variables by an equality of real numbers,

$$\mathbb{P}(Y \in A \,|\, X = x) = \nu(x, A),$$

for every $x \in E$. Although it gives the intuition behind the notion of conditional distribution, this last equality has no mathematical meaning in general, because one has typically $\mathbb{P}(X = x) = 0$ for every $x \in E$, which makes it impossible to define the conditional probability given the event $\{X = x\}$. The only rigorous formulation is thus the first equality $\mathbb{P}(Y \in A \,|\, X) = \nu(X, A)$.

Let us briefly discuss the uniqueness of the conditional distribution of Y knowing X. If ν and ν' are two such conditional distributions, we have, for every $A \in \mathcal{F}$,

$$\nu(X, A) = \mathbb{P}(Y \in A \,|\, X) = \nu'(X, A)\;,\;\; \text{a.s.}$$

which is equivalent to saying that, for every $A \in \mathcal{F}$,

$$\nu(x, A) = \nu'(x, A) \ , \ \mathbb{P}_X(\mathrm{d}x) \text{ a.s.}$$

Suppose that the measurable space $(F, \mathcal{F})$ is such that any probability measure on $(F, \mathcal{F})$ is characterized by its values on a (given) countable collection of measurable sets (this property holds for $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, by considering the collection of all rectangles with rational coordinates). Then, we conclude from the preceding display that

$$\nu(x, \cdot) = \nu'(x, \cdot) \ , \ \mathbb{P}_X(\mathrm{d}x) \text{ a.s.}$$

So uniqueness holds in this sense (and clearly we cannot expect more). By abuse of language, we will nonetheless speak about *the* conditional distribution of Y knowing X.

Let us now consider the problem of the existence of conditional distributions.

Theorem 11.17 *Let X and Y be two random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively. Suppose that $(F, \mathcal{F})$ is a complete separable metric space equipped with its Borel σ-field. Then, there exists a conditional distribution of Y knowing X.*

We will not prove this theorem, see Theorem 6.3 in [10] for the proof of a more general statement. In what follows, we will not need Theorem 11.17, because a direct construction allows one to avoid using the existence statement. As an illustration, let us consider the examples treated in the preceding section.

(1) If X is a discrete variable (E is countable), then we may define $\nu(x, A)$ by

$$\begin{aligned} &\nu(x, A) := \mathbb{P}(Y \in A \mid X = x) && \text{if } x \in E' := \{y \in E : \mathbb{P}(X = y) > 0\} \\ &\nu(x, A) = \delta_{y_0}(A) && \text{if } x \notin E' \end{aligned}$$

where y_0 is an arbitrary fixed point of F.

(2) Suppose that X and Y take values in $\mathbb{R}^m$ and in $\mathbb{R}^n$ respectively, and that the pair (X, Y) has density $p(x, y)$, $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$. The density of X is then

$$q(x) = \int_{\mathbb{R}^n} p(x, y)\, \mathrm{d}y,$$

where we take $q(x) = 0$ if the integral is infinite. Proposition 11.11 shows that we may define the conditional distribution of Y knowing X by

$$\begin{aligned} &\nu(x, A) = \frac{1}{q(x)} \int_A p(x, y)\, \mathrm{d}y && \text{if } q(x) > 0, \\ &\nu(x, A) = \delta_0(A) && \text{if } q(x) = 0. \end{aligned}$$

(3) Suppose that $(X_1, \dots, X_n, Y)$ is a centered Gaussian vector, and let

$$\sum_{j=1}^{n} \lambda_j X_j$$

be the orthogonal projection in L^2 of Y on the vector space spanned by $X_1, \dots, X_n$. Also set

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

and suppose that $\sigma > 0$. Theorem 11.13 shows that the conditional distribution of Y knowing $X = (X_1, \dots, X_n)$ is

$$\nu(x_1, \dots, x_n; A) = \int_A q_{\sum_{j=1}^{n} \lambda_j x_j, \sigma^2}(y)\, \mathrm{d}y$$

where q_{m,σ^2} is the density of $\mathcal{N}(m, \sigma^2)$. In a slightly abusive way, we say that, conditionally on $(X_1, \dots, X_n)$, Y follows the Gaussian $\mathcal{N}(\sum_{j=1}^{n} \lambda_j X_j, \sigma^2)$ distribution.

11.6 Exercises

Exercise 11.1 (Bayes Formula)

(1) Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $(A_1, A_2, \dots, A_n)$ be a measurable partition of Ω such that $\mathbb{P}(A_i) > 0$ for every $i \in \{1, \dots, n\}$. Prove that, for every $B \in \mathcal{A}$ such that $\mathbb{P}(B) > 0$, for every $i \in \{1, \dots, n\}$,

$$\mathbb{P}(A_i \,|\, B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B \,|\, A_i)}{\sum_{j=1}^{n} \mathbb{P}(A_j)\mathbb{P}(B \,|\, A_j)}.$$

(2) Suppose that we have n boxes numbered $1, 2, \dots, n$, and that the i-th box contains r_i red balls and n_i black boxes, where $r_i, n_i \geq 1$. Imagine that one chooses a box uniformly at random, and then picks a ball (again at random) in the chosen box. Compute the probability that the i-th box was chosen knowing that a red box was picked.

Exercise 11.2 Let $X_1, \dots, X_n$ be independent Bernoulli random variables with parameter $p \in (0, 1)$, and $S_n = X_1 + \cdots + X_n$. Prove that, for every $k \in \{0, 1, \dots, n\}$, the conditional distribution of $(X_1, \dots, X_n)$ knowing that $S_n = k$

(that is, the law of $(X_1, \ldots, X_n)$ under $\mathbb{P}(\cdot \mid S_n = k)$) is the uniform distribution on $\{(x_1, \ldots, x_n) \in \{0,1\}^n : x_1 + \cdots + x_n = k\}$.

Exercise 11.3 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent real random variables unifomly distributed over $[0, 1]$. Define the *record times* of the sequence by $T_1 = 1$ and, for every $p \geq 2$,

$$T_p := \inf\{n > T_{p-1} : X_n > X_{T_{p-1}}\}$$

with the convention $\inf \varnothing = \infty$. Show that $\mathbb{P}(T_p < \infty) = 1$ for every $p \in \mathbb{N}$. Then determine the law of T_2, and prove that, for every $p \geq 2$ and $k \in \mathbb{N}$,

$$\mathbb{E}[\mathbf{1}_{\{T_p=k\}} \mid (T_1, \ldots, T_{p-1})] = \mathbb{E}[\mathbf{1}_{\{T_p=k\}} \mid T_{p-1}] = \frac{T_{p-1}}{k(k-1)} \mathbf{1}_{\{k>T_{p-1}\}}.$$

Exercise 11.4 Let $\mathcal{B}$ be a sub-σ-field of $\mathcal{A}$, and let X be a nonnegative real random variable. Prove that the set

$$A = \{\mathbb{E}[X \mid \mathcal{B}] > 0\}$$

is the smallest $\mathcal{B}$-measurable set containing $\{X > 0\}$, in the sense that:

- $\mathbb{P}(\{X > 0\} \backslash A) = 0$;
- if $B \in \mathcal{B}$ is such that $\{X > 0\} \subset B$, then $\mathbb{P}(A \backslash B) = 0$.

Exercise 11.5 Let X and Y be two independent Gaussian $\mathcal{N}(0, 1)$ random variables. Compute

$$\mathbb{E}[X \mid X^2 + Y^2].$$

Exercise 11.6 Let X be a d-dimensional Gaussian vector. Prove that the law $\mathbb{P}_X$ of X is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$ if and only if the covariance matrix K_X is invertible, and in that case the density of $\mathbb{P}_X$ is

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(K_X)}} \exp\Big(-\frac{1}{2}\,{}^t(x-m)K_X^{-1}(x-m)\Big), \quad x \in \mathbb{R}^d,$$

where $m = \mathbb{E}[X]$ and $\det(K_X)$ is the determinant of K_X.

Exercise 11.7 Let $(\mathcal{A}_n)_{n\in\mathbb{N}}$ be a sequence of sub-σ-fields of $\mathcal{A}$, and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative random variables.

(1) Prove that the condition "$\mathbb{E}[X_n \mid \mathcal{A}_n]$ converges in probability to 0" implies that X_n converges in probability to 0.
(2) Show that the converse is false.

Exercise 11.8 Let $(\mathcal{A}_n)_{n\in\mathbb{N}}$ be a decreasing sequence of sub-σ-fields of $\mathcal{A}$, with $\mathcal{A}_1 = \mathcal{A}$, and let $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$.

(1) Prove that the random variables $\mathbb{E}[X \mid \mathcal{A}_n] - \mathbb{E}[X \mid \mathcal{A}_{n+1}]$, for $n \in \mathbb{N}$, are orthogonal in L^2, and that the series

$$\sum_{n\in\mathbb{N}} \big(\mathbb{E}[X\,|\,\mathcal{A}_n] - \mathbb{E}[X\,|\,\mathcal{A}_{n+1}]\big)$$

converges in L^2.

(2) Let $\mathcal{A}_\infty = \cap_{n\in\mathbb{N}}\mathcal{A}_n$. Prove that

$$\lim_{n\to\infty} \mathbb{E}[X\,|\,\mathcal{A}_n] = \mathbb{E}[X\,|\,\mathcal{A}_\infty], \quad \text{in } L^2.$$

Exercise 11.9 Let X and Y be to nonnegative random variables in L^1. We assume that we have both $\mathbb{E}[X\,|\,Y] = Y$ and $\mathbb{E}[Y\,|\,X] = X$.

(1) Under the additional assumption that $X \in L^2$, prove that $X = Y$.
(2) We come back to the general case. Prove that, for every $a > 0$,

$$\mathbb{E}[X\,|\,X \wedge a] \wedge a = X \wedge a.$$

(3) Verify that, for every $a > 0$, the pair $(X \wedge a, Y \wedge a)$ satisfies the same assumptions as the pair (X, Y), and conclude that $X = Y$. (*Hint:* Start by verifying that $\mathbb{E}[X \wedge a\,|\,Y \wedge a] \le Y \wedge a$).

Exercise 11.10 Let $\mathcal{B}$ be a sub-σ-field of $\mathcal{A}$, and let X and Y be two random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively. We say that X and Y are conditionally independent given $\mathcal{B}$ if, for any nonnegative measurable functions f and g defined respectively on E and on F, we have

$$\mathbb{E}[f(X)g(Y)\,|\,\mathcal{B}] = \mathbb{E}[f(X)\,|\,\mathcal{B}]\,\mathbb{E}[g(Y)\,|\,\mathcal{B}].$$

(1) Discuss the special cases $\mathcal{B} = \{\varnothing, \Omega\}$ and $\mathcal{B} = \mathcal{A}$.
(2) Prove that X and Y are conditionally independent given $\mathcal{B}$ if and only if, for any nonnegative $\mathcal{B}$-measurable random variable Z and any functions f and g as above,

$$\mathbb{E}[f(X)g(Y)Z] = \mathbb{E}\big[f(X)Z\,\mathbb{E}[g(Y)\,|\,\mathcal{B}]\big],$$

and that this property is also equivalent to saying that, for any nonnegative measurable function g on F,

$$\mathbb{E}[g(Y)\,|\,\mathcal{B} \vee \sigma(X)] = \mathbb{E}[g(Y)\,|\,\mathcal{B}].$$

(3) We now assume that $E = F = \mathbb{R}$, and that $\mathcal{B} = \sigma(Z)$, where Z is a real random variable. Furthermore, we assume that the random vector (X, Y, Z) has a density which is positive on $\mathbb{R}^3$. Prove that X and Y are conditionally independent given $\mathcal{B}$ if and only the density of (X, Y, Z) can be written in the form

$$p(x, y, z) = q(z)\, r(z, x)\, s(z, y)$$

where q is the density of Z and r, s are positive measurable functions on $\mathbb{R}^2$.

Exercise 11.11 Let $a, b \in (0, \infty)$, and let (X, Y) be a random variable with values in $\mathbb{Z}_+ \times \mathbb{R}_+$, whose distribution is characterized by the formula

$$\mathbb{P}(X = n, Y \leq t) = b \int_0^t \frac{(ay)^n}{n!} \exp(-(a+b)y)\, \mathrm{d}y,$$

for every $n \in \mathbb{Z}_+$ and $t \in \mathbb{R}_+$.

(1) Compute $\mathbb{P}(X = n)$ for every $n \in \mathbb{Z}_+$, and then determine the conditional distribution of Y knowing X. Compute $\mathbb{E}[\frac{Y}{X+1}]$.
(2) Compute the law of Y and then $\mathbb{E}[\mathbf{1}_{\{X=n\}} \,|\, Y]$. Give the conditional distribution of X knowing Y and compute $\mathbb{E}[X \,|\, Y]$.

Exercise 11.12 Let $\lambda > 0$, and let X be a Gamma $\Gamma(2, \lambda)$ random variable (with density $\lambda^2 x\, e^{-\lambda x}$ on $\mathbb{R}_+$). Let Y be another real random variable, and assume that the conditional distribution of Y knowing X is the uniform distribution over $[0, X]$. Prove that Y and $X - Y$ are two independent exponential variables with parameter λ.

Exercise 11.13 Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces, and let X and Y be two random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively. Assume that the conditional distribution of Y knowing X is the transition kernel $\nu(x, \mathrm{d}y)$. Prove that, for any nonnegative measurable function h on $(E \times F, \mathcal{E} \otimes \mathcal{F})$,

$$\mathbb{E}[h(X, Y) \,|\, X] = \int_F \nu(X, \mathrm{d}y)\, h(X, y).$$

(*Hint:* Consider first the case where $h = \mathbf{1}_{A \times B}$, with $A \in \mathcal{E}$ and $B \in \mathcal{F}$, and then use a monotone class argument).

Part III
Stochastic Processes

Chapter 12
Theory of Martingales

This chapter is devoted to the study of martingales, which form a very important class of random processes. A (discrete-time) martingale is a sequence $(X_n)_{n\in\mathbb{Z}_+}$ of integrable real random variables such that, for every integer n, the conditional expectation of X_{n+1} knowing $X_0, \ldots, X_n$ is equal to X_n. Intuitively, we may think of the evolution of the fortune of a player in a fair game: the mean value of the fortune at time $n+1$ knowing what has happened until time n is equal to the fortune at time n. Martingales play an important role in many developments of advanced probability theory. We present here the basic convergence theorems for martingales. As a special case of these theorems, if $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ is an increasing sequence (or a decreasing sequence) of sub-σ-fields and Z is an integrable real random variable, the conditional expectations $\mathbb{E}[Z \mid \mathcal{F}_n]$ converge a.s. and in L^1 as $n \to \infty$. We also discuss optional stopping theorems, which roughly speaking say that no strategy of the player can ensure a positive gain with probability one in a fair game. All these results have important applications to other random processes, including random walks and branching processes.

12.1 Definitions and Examples

We consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. By definition, a (discrete-time) *random process* is a sequence $(X_n)_{n\in\mathbb{Z}_+}$ of random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in the same measurable space. In this chapter, all random processes will take values in $\mathbb{R}$.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_12

Definition 12.1 A filtration on $(\Omega, \mathcal{A}, \mathbb{P})$ is an increasing sequence $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ of sub-σ-fields of $\mathcal{A}$. We have thus

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{A}$$

We also say that $(\Omega, \mathcal{A}, (\mathcal{F}_n)_{n\in\mathbb{Z}_+}, \mathbb{P})$ is a filtered probability space.

The parameter $n \in \mathbb{Z}_+$ is usually interpreted as time. The σ-field $\mathcal{F}_n$ gathers the information available at time n (events that are $\mathcal{F}_n$-measurable are interpreted as those that depend only on what has happened up to time n). We will write

$$\mathcal{F}_\infty = \bigvee_{n\in\mathbb{Z}_+} \mathcal{F}_n := \sigma\Big(\bigcup_{n\in\mathbb{Z}_+} \mathcal{F}_n \Big).$$

Examples

(a) If $(X_n)_{n\in\mathbb{Z}_+}$ is a random process, then, for every $n \in \mathbb{Z}_+$, we may define $\mathcal{F}_n^X$ as the smallest σ-field on Ω for which $X_0, X_1, \ldots, X_n$ are measurable:

$$\mathcal{F}_n^X = \sigma(X_0, X_1, \ldots, X_n).$$

Then $(\mathcal{F}_n^X)_{n\in\mathbb{Z}_+}$ is a filtration called the canonical filtration of $(X_n)_{n\in\mathbb{Z}_+}$.

(b) Suppose that $\Omega = [0, 1)$, $\mathcal{A}$ is the Borel σ-field on $[0, 1)$, and $\mathbb{P}$ is Lebesgue measure. For every $n \in \mathbb{Z}_+$, set

$$\mathcal{F}_n = \sigma([\frac{i-1}{2^n}, \frac{i}{2^n}); i \in \{1, 2, \ldots, 2^n\}).$$

Then $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ is a filtration called the dyadic filtration on $[0, 1)$.

Definition 12.2 A random process $(X_n)_{n\in\mathbb{Z}_+}$ is said to be adapted to the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ if, for every $n \in \mathbb{Z}_+$, X_n is $\mathcal{F}_n$-measurable.

By definition, the canonical filtration of $(X_n)_{n\in\mathbb{Z}_+}$ is the smallest filtration to which the random process $(X_n)_{n\in\mathbb{Z}_+}$ is adapted.

Throughout the remaining part of this chapter (Section 12.7 excepted), we fix the filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_n)_{n\in\mathbb{Z}_+}, \mathbb{P})$. The choice of this space, or only of the filtration, will sometimes be made explicit in examples. The notions that we will introduce, in particular in the next definition, are relative to this filtered probability space.

Definition 12.3 Let $(X_n)_{n\in\mathbb{Z}_+}$ be an adapted (real-valued) random process, such that $\mathbb{E}[|X_n|] < \infty$ for every $n \in \mathbb{Z}_+$. We say that the process $(X_n)_{n\in\mathbb{Z}_+}$ is:

- a martingale if, for every $n \in \mathbb{Z}_+$,

$$\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n\,;$$

- a supermartingale if, for every $n \in \mathbb{Z}_+$,

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leq X_n\,;$$

- a submartingale if, for every $n \in \mathbb{Z}_+$,

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \geq X_n\,.$$

An immediate consequence of the definition of a martingale is the (apparently stronger) property: for every $0 \leq n \leq m$,

$$\mathbb{E}[X_m \mid \mathcal{F}_n] = X_n \tag{12.1}$$

This is easy to verify by induction on $m - n$. If $m = n$, (12.1) is trivial (since X_n is $\mathcal{F}_n$-measurable), if $m = n + 1$, (12.1) is the definition of a martingale, and if $m - n \geq 2$, Proposition 11.7 gives

$$\mathbb{E}[X_m \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X_m \mid \mathcal{F}_{m-1}] \mid \mathcal{F}_n] = \mathbb{E}[X_{m-1} \mid \mathcal{F}_n].$$

Notice that (12.1) implies $\mathbb{E}[X_m] = \mathbb{E}[X_n]$, and thus we have $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ for every $n \in \mathbb{Z}_+$.

Similarly, if $(X_n)_{n\in\mathbb{Z}_+})$ is a supermartingale (resp. a submartingale), we get, for every $0 \leq n \leq m$,

$$\mathbb{E}[X_m \mid \mathcal{F}_n] \leq X_n \qquad (\text{resp. } \mathbb{E}[X_m \mid \mathcal{F}_n] \geq X_n),$$

and thus $\mathbb{E}[X_m] \leq \mathbb{E}[X_n]$ (resp. $\mathbb{E}[X_m] \geq \mathbb{E}[X_n]$). So, for a supermartingale, the sequence $(\mathbb{E}[X_n])_{n\in\mathbb{Z}_+}$ is decreasing, while it is increasing for a submartingale.

It is often useful to interpret a martingale as modelling the evolution of a fair game: X_n corresponds to the (possibly negative) fortune of the player at time n, and $\mathcal{F}_n$ is the information available at that time (including the outcomes of the preceding games). The martingale property $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ reflects the fact that the average value of the fortune at time $n + 1$ knowing the past up to time n is equal to the fortune at time n (on average, the player does not earn or lose money). In the same way, a supermartingale corresponds to a disadvantageous game.

If $(X_n)_{n\in\mathbb{Z}_+}$ is a supermartingale, then $(-X_n)_{n\in\mathbb{Z}_+}$ is a submartingale, and conversely. For this reason, many of the results that are stated below for supermartingales have an immediate analog for submartingales (and conversely).

Examples

(i) Let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. For every $n \in \mathbb{Z}_+$, set

$$X_n = \mathbb{E}[X \mid \mathcal{F}_n].$$

Then $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale. Indeed Proposition 11.7 shows that

$$\mathbb{E}[X_{n+1}\,|\,\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X\,|\,\mathcal{F}_{n+1}]\,|\,\mathcal{F}_n] = \mathbb{E}[X\,|\,\mathcal{F}_n] = X_n.$$

A martingale of this type is called *closed*.

(ii) Let $(X_n)_{n\in\mathbb{Z}_+}$ be a decreasing sequence of integrable random variables, and assume that $(X_n)_{n\in\mathbb{Z}_+}$ is adapted. Then $(X_n)_{n\in\mathbb{Z}_+}$ is a supermartingale:

$$\mathbb{E}[X_{n+1}\,|\,\mathcal{F}_n] \leq \mathbb{E}[X_n\,|\,\mathcal{F}_n] = X_n.$$

(iii) *Random walk on* $\mathbb{R}$. Let $x \in \mathbb{R}$ and let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed real random variables with distribution μ, such that $\mathbb{E}[|Y_1|] < \infty$. Set

$$X_0 = x \text{ and } X_n = x + Y_1 + Y_2 + \ldots + Y_n \text{ if } n \geq 1.$$

Also define the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ on Ω by

$$\mathcal{F}_0 = \{\varnothing, \Omega\} \text{ and } \mathcal{F}_n = \sigma(Y_1, \ldots, Y_n) \text{ if } n \geq 1$$

(this is the canonical filtration of $(X_n)_{n\in\mathbb{Z}_+}$). Then $(X_n)_{n\in\mathbb{Z}_+}$ is

- a martingale if $\mathbb{E}[Y_1] = 0$;
- a supermartingale if $\mathbb{E}[Y_1] \leq 0$;
- a submartingale if $\mathbb{E}[Y_1] \geq 0$.

Indeed, considering for instance the case $\mathbb{E}[Y_1] = 0$, we have

$$\mathbb{E}[X_{n+1}\,|\,\mathcal{F}_n] = \mathbb{E}[X_n + Y_{n+1}\,|\,\mathcal{F}_n] = X_n + \mathbb{E}[Y_{n+1}] = X_n,$$

since by construction Y_{n+1} is independent of $\mathcal{F}_n$, and we use Theorem 11.8.

We say that $(X_n)_{n\in\mathbb{Z}_+}$ is a *random walk* on $\mathbb{R}$ started at x with jump distribution μ. In the special case where $\mu(1) = \mu(-1) = 1/2$ and $x \in \mathbb{Z}$, we call $(X_n)_{n\in\mathbb{Z}_+}$ *simple random walk* on $\mathbb{Z}$ (this is the coin-tossing process already considered in Proposition 10.7).

(iv) Consider the example (b) of a filtered probability space given above. Let μ be a finite measure on $[0, 1)$, and recall that $\mathbb{P} = \lambda$ is Lebesgue measure on $[0, 1)$ in this example. For every integer $n \in \mathbb{Z}_+$, set

$$f_n = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}_{|\mathcal{F}_n},$$

which denotes the Radon-Nikodym density of μ with respect to λ, when both μ and λ are viewed as measures on the σ-field $\mathcal{F}_n$ (notice that, on the σ-field $\mathcal{F}_n$, all measures are absolutely continuous with respect to λ!). It is easy to verify that f_n is given by

$$f_n(\omega) = \sum_{i=1}^{2^n} \frac{\mu([(i-1)2^{-n}, i2^{-n}))}{2^{-n}} \mathbf{1}_{[(i-1)2^{-n}, i2^{-n})}(\omega), \quad \forall \omega \in [0,1).$$

Then, $(f_n)_{n\in\mathbb{Z}_+}$ is a martingale. Indeed, if $A \in \mathcal{F}_n$, we have

$$\mathbb{E}[\mathbf{1}_A f_{n+1}] = \int \mathbf{1}_A(\omega) f_{n+1}(\omega)\,\mathrm{d}\omega = \mu(A) = \int \mathbf{1}_A(\omega) f_n(\omega)\,\mathrm{d}\omega = \mathbb{E}[\mathbf{1}_A f_n],$$

and, recalling the characteristic property of conditional expectations, this implies that $f_n = \mathbb{E}[f_{n+1} \,|\, \mathcal{F}_n]$.

In the special case where μ is absolutely continuous with respect to λ (on the σ-field $\mathcal{A}$), the martingale $(f_n)_{n\in\mathbb{Z}_+}$ is closed, in the sense explained in example (i). Indeed, one checks by the same argument as above that

$$f_n = \mathbb{E}[f \,|\, \mathcal{F}_n], \tag{12.2}$$

where f is the Radon-Nikodym density of μ with respect to λ.

Two Martingale Transforms

Proposition 12.4 *Let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}_+$ be a convex function, and let $(X_n)_{n\in\mathbb{Z}_+}$ be an adapted process, such that $\mathbb{E}[\varphi(X_n)] < \infty$ for every $n \in \mathbb{Z}_+$.*

(i) *If $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale, $(\varphi(X_n))_{n\in\mathbb{Z}_+}$ is a submartingale.*
(ii) *If $(X_n)_{n\in\mathbb{Z}_+}$ is a submartingale, and if we assume in addition that φ is increasing, $(\varphi(X_n))_{n\in\mathbb{Z}_+}$ is a submartingale.*

In particular, if X_n is a martingale, $|X_n|$ is a submartingale (and so is X_n^2, provided that $\mathbb{E}[X_n^2] < \infty$ for every n) and, if X_n is a submartingale, $(X_n)^+$ is also a submartingale.

Proof

(i) By Jensen's inequality for conditional expectations,

$$\mathbb{E}[\varphi(X_{n+1}) \,|\, \mathcal{F}_n] \geq \varphi(\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n]) = \varphi(X_n).$$

(ii) Similary, since $X_n \leq \mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n]$ and φ is increasing,

$$\mathbb{E}[\varphi(X_{n+1}) \,|\, \mathcal{F}_n] \geq \varphi(\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n]) \geq \varphi(X_n).$$

□

Definition 12.5 A sequence $(H_n)_{n\in\mathbb{N}}$ of real random variables is called predictable if H_n is bounded and $\mathcal{F}_{n-1}$-measurable, for every $n \in \mathbb{N}$.

Proposition 12.6 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be an adapted process, and let $(H_n)_{n\in\mathbb{N}}$ be a predictable sequence. We set $(H \bullet X)_0 = 0$ and, for every integer $n \geq 1$,*

$$(H \bullet X)_n = H_1(X_1 - X_0) + H_2(X_2 - X_1) + \cdots + H_n(X_n - X_{n-1}).$$

Then,

(i) *If $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale, $((H \bullet X)_n)_{n\in\mathbb{Z}_+}$ is also a martingale.*
(ii) *If $(X_n)_{n\in\mathbb{Z}_+}$ is a supermartingale (resp. a submartingale), and $H_n \geq 0$ for every $n \in \mathbb{N}$, then $((H \bullet X)_n)_{n\in\mathbb{Z}_+}$ is a supermartingale (resp. a submartingale).*

Proof Let us prove (i). Since the random variables H_n are bounded, one immediately checks that the random variables $(H \bullet X)_n$ are integrable. It is also straightforward to verify that the process $((H \bullet X)_n)_{n\in\mathbb{Z}_+}$ is adapted (all random variables entering the definition of $(H \bullet X)_n$ are $\mathcal{F}_n$-measurable). Then it suffices to verify that, for every $n \in \mathbb{Z}_+$,

$$\mathbb{E}[(H \bullet X)_{n+1} - (H \bullet X)_n \,|\, \mathcal{F}_n] = 0.$$

However, $(H \bullet X)_{n+1} - (H \bullet X)_n = H_{n+1}(X_{n+1} - X_n)$ and since H_{n+1} is $\mathcal{F}_n$-measurable, Proposition 11.6 shows that

$$\begin{aligned}\mathbb{E}[H_{n+1}(X_{n+1} - X_n) \,|\, \mathcal{F}_n] &= H_{n+1}\,\mathbb{E}[X_{n+1} - X_n \,|\, \mathcal{F}_n] \\ &= H_{n+1}(\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] - X_n) = 0.\end{aligned}$$

The proof of (ii) is similar. □

Recall our interpretation of a martingale X_n as the fortune of the player at time n in a fair game. The difference $X_{n+1} - X_n$ is then the gain obtained by the player between time n and time $n + 1$. We can imagine that, at time n, the player decides to change his stake by multiplying it by the factor H_{n+1} (which has to be $\mathcal{F}_n$-measurable, since the player only knows what has happened up to time n). Then the new gain obtained between time n and time $n + 1$ will be $H_{n+1}(X_{n+1} - X_n)$ and this will still correspond to a fair game. The preceding lines give an intuitive explanation of the definition of $H \bullet X$.

12.2 Stopping Times

Definition 12.7 A random variable $T : \Omega \longrightarrow \overline{\mathbb{Z}}_+ := \mathbb{Z}_+ \cup \{+\infty\}$ is called a stopping time (of the filtration $(\mathcal{F}_n)_{n \in \mathbb{Z}_+}$) if, for every integer $n \in \mathbb{Z}_+$, one has

$$\{T = n\} \in \mathcal{F}_n.$$

This is equivalent to saying that, for every $n \in \mathbb{Z}_+$, one has $\{T \leq n\} \in \mathcal{F}_n$.

The equivalence between the two forms of the definition is very easy and left to the reader as an exercise.

It is important to note that the value $T = +\infty$ is allowed. By writing

$$\{T = +\infty\} = \Omega \backslash \bigcup_{n \in \mathbb{Z}_+} \{T = n\}$$

we see that $\{T = +\infty\} \in \mathcal{F}_\infty$.

Back to our game-theoretic interpretation, one may imagine that the player decides to leave the game at a random time (depending on the evolution of his fortune). This random time has to be a stopping time: for the player to decide to leave the game at time n, the only information available is the past up to this time (the player has no information about future events). Similarly, on the stock market, one can decide to sell shares from the knowledge of the evolution of prices up to the current time, but their future evolution is in principle unknown!

Examples

(i) Let $k \in \overline{\mathbb{Z}}_+$. The constant time $T = k$ is obviously a stopping time.

(ii) Let $(Y_n)_{n \in \mathbb{Z}_+}$ be an adapted process, and let $A \in \mathcal{B}(\mathbb{R})$. Set

$$T_A = \inf\{n \in \mathbb{Z}_+ : Y_n \in A\}$$

with the convention $\inf \varnothing = +\infty$ (this convention will be made systematically in what follows). Then T_A is a stopping time. Indeed, for every $n \in \mathbb{Z}_+$,

$$\{T_A = n\} = \{Y_0 \notin A, Y_1 \notin A, \ldots, Y_{n-1} \notin A, Y_n \in A\} \in \mathcal{F}_n.$$

We call T_A the first hitting time of A (by the process $(Y_n)_{n \in \mathbb{Z}_+}$).

(iii) Under the same assumptions as in (ii), if we fix $N \in \mathbb{N}$ and set

$$L_A := \sup\{n \leq N : Y_n \in A\} \quad (\sup \varnothing = 0 \quad \text{by convention}),$$

L_A is usually not a stopping time. Indeed, for every $n \in \{1, \ldots, N-1\}$, the event

$$\{L_A = n\} = \{Y_n \in A, Y_{n+1} \notin A, \ldots, Y_N \notin A\}$$

is a priori not in $\mathcal{F}_n$, since this event involves the random variables $Y_{n+1}, \ldots, Y_N$, which need not be $\mathcal{F}_n$-measurable.

Definition 12.8 Let T be a stopping time. The σ-field of the past up to time T is defined by

$$\mathcal{F}_T := \{A \in \mathcal{F}_\infty : \forall n \in \mathbb{Z}_+, A \cap \{T = n\} \in \mathcal{F}_n\},$$

or, equivalently,

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty : \forall n \in \mathbb{Z}_+, A \cap \{T \le n\} \in \mathcal{F}_n\},$$

We leave the fact that $\mathcal{F}_T$ is a σ-field and the equivalence of the two forms of the definition as an exercise for the reader. Our notation is consistent in the sense that $\mathcal{F}_T = \mathcal{F}_n$ if T is the constant stopping time equal to n. Note that the random variable T is $\mathcal{F}_T$-measurable: for every $k \in \mathbb{Z}_+$, $\{T = k\} \cap \{T = n\}$ is empty if $n \neq k$ and is equal to $\{T = n\} \in \mathcal{F}_n$ if $n = k$, proving that $\{T = k\} \in \mathcal{F}_T$.

Proposition 12.9 *Let S and T be two stopping times such that $S \le T$. Then, $\mathcal{F}_S \subset \mathcal{F}_T$.*

Proof Let $A \in \mathcal{F}_S$. Then, for every $n \in \mathbb{Z}_+$,

$$A \cap \{T = n\} = \bigcup_{k=0}^{n} (A \cap \{S = k\}) \cap \{T = n\} \in \mathcal{F}_n.$$

□

Proposition 12.10 *Let S and T be two stopping times. Then, $S \wedge T$ and $S \vee T$ are also stopping times. Moreover $\mathcal{F}_{S \wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$.*

Proof For the first assertion, we write $\{S \wedge T \le n\} = \{S \le n\} \cup \{T \le n\}$ and similarly $\{S \vee T \le n\} = \{S \le n\} \cap \{T \le n\}$. Then, the fact that $\mathcal{F}_{S \wedge T} \subset \mathcal{F}_S \cap \mathcal{F}_T$ follows from Proposition 12.9. Conversely, if $A \in \mathcal{F}_S \cap \mathcal{F}_T$, we have $A \cap \{S \wedge T \le n\} = (A \cap \{S \le n\}) \cup (A \cap \{T \le n\}) \in \mathcal{F}_n$ for every n, so that $A \in \mathcal{F}_{S \wedge T}$. □

The next proposition shows that the evaluation of a random process at a stopping time T gives an $\mathcal{F}_T$-measurable random variable (this will play an important role in the stopping theorems that are discussed below). A minor technical difficulty occurs because the evaluation at T makes sense only when $T < \infty$. For this reason, it is convenient to use a notion of measurability for a real function Z which is defined only on a measurable subset A of Ω: if $\mathcal{G}$ is a sub-σ-field of $\mathcal{A}$ such that $A \in \mathcal{G}$, we (obviously) say that Z is $\mathcal{G}$-measurable if and only if $Z^{-1}(B) \in \mathcal{G}$ for every $B \in \mathcal{B}(\mathbb{R})$, as in the usual definition.

Proposition 12.11 *Let $(Y_n)_{n\in\mathbb{Z}_+}$ be an adapted process, and let T be a stopping time. Then the random variable Y_T, which is defined on the event $\{T < \infty\}$ by $Y_T(\omega) = Y_{T(\omega)}(\omega)$, is $\mathcal{F}_T$-measurable.*

Proof Let B be a Borel subset of $\mathbb{R}$. Then, for every $n \in \mathbb{Z}_+$,

$$\{Y_T \in B\} \cap \{T = n\} = \{Y_n \in B\} \cap \{T = n\} \in \mathcal{F}_n,$$

which shows that $\{Y_T \in B\} \in \mathcal{F}_T$. □

If T is a stopping time, then, for every $n \in \mathbb{Z}_+$, $n \wedge T$ is also a stopping time (Lemma 12.10) and it follows from the preceding proposition that $Y_{n\wedge T}$ is $\mathcal{F}_{n\wedge T}$-measurable, hence also $\mathcal{F}_n$-measurable by Proposition 12.9.

Theorem 12.12 (Optional stopping theorem, first version) *Suppose that $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale (resp. a supermartingale) and let T be a stopping time. Then $(X_{n\wedge T})_{n\in\mathbb{Z}_+}$ is also a martingale (resp. a supermartingale). In particular, if the stopping time T is bounded, one has $X_T \in L^1$, and*

$$\mathbb{E}[X_T] = \mathbb{E}[X_0] \qquad (\textit{resp. } \mathbb{E}[X_T] \le \mathbb{E}[X_0]).$$

Proof For every $n \geq 1$, set

$$H_n = \mathbf{1}_{\{T\geq n\}} = 1 - \mathbf{1}_{\{T\leq n-1\}}.$$

Then the sequence $(H_n)_{n\geq 1}$ is predictable, since $\{T \leq n-1\} \in \mathcal{F}_{n-1}$ by the definition of a stopping time. Moreover, for every $n \in \mathbb{Z}_+$,

$$X_{n\wedge T} = X_0 + \sum_{i=1}^{n\wedge T}(X_i - X_{i-1}) = X_0 + \sum_{i=1}^{n} \mathbf{1}_{\{T\geq i\}}(X_i - X_{i-1}) = X_0 + (H \bullet X)_n.$$

The first assertion of the theorem then follows from Proposition 12.6 (it is obvious that adding an integrable $\mathcal{F}_0$-measurable random variable to a martingale, resp. to a supermartingale, yields again a martingale, resp. a supermartingale).

If the stopping time T is bounded above by the constant N, $X_T = X_{N\wedge T}$ is in L^1, and $\mathbb{E}[X_T] = \mathbb{E}[X_{N\wedge T}] = \mathbb{E}[X_0]$ (resp. $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$ in the case of a supermartingale). □

In the last assertion of the theorem, the assumption that T is bounded cannot be omitted, as shown by the following simple example. Consider a simple random walk $(X_n)_{n\in\mathbb{Z}_+}$ started from 0. Then $X_n = Y_1 + \cdots + Y_n$ where the random variables $Y_1, Y_2, \ldots$ are independent withe the same distribution $\mathbb{P}(Y_1 = 1) = \mathbb{P}(Y_1 = -1) = 1/2$. Since $\mathbb{E}[Y_1] = 0$, we know that $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale. Then,

$$T = \inf\{n \geq 0 : X_n = 1\}$$

is a stopping time, and we have $T < \infty$ a.s. as a consequence of Proposition 10.7 (we will later provide alternative proofs of this fact). However,

$$1 = \mathbb{E}[X_T] \neq \mathbb{E}[X_0] = 0.$$

Since the stopping time T is not bounded, this does not contradict Theorem 12.12.

12.3 Almost Sure Convergence of Martingales

Our goal is now to study the almost sure convergence of a martingale or a supermartingale when $n \to \infty$. Consider first a deterministic real sequence $\alpha = (\alpha_n)_{n\in\mathbb{Z}_+}$. Fix two reals $a < b$ and define two increasing sequences $(S_k(\alpha))_{k\in\mathbb{Z}_+}$ and $(T_k(\alpha))_{k\in\mathbb{Z}_+}$ in $\overline{\mathbb{Z}}_+$ inductively as follows:

$$S_1(\alpha) := \inf\{n \geq 0 : \alpha_n \leq a\}$$
$$T_1(\alpha) := \inf\{n \geq S_1(\alpha) : \alpha_n \geq b\}$$

and then, for every $k \geq 1$,

$$S_{k+1}(\alpha) := \inf\{n \geq T_k(\alpha) : \alpha_n \leq a\}$$
$$T_{k+1}(\alpha) := \inf\{n \geq S_{k+1}(\alpha) : \alpha_n \geq b\}.$$

The convention $\inf \varnothing = +\infty$ is used in these definitions. We then set, for every integer $n \geq 0$,

$$N_n([a,b],\alpha) = \sup\{k \geq 1 : T_k(\alpha) \leq n\} = \sum_{k=1}^{\infty} \mathbf{1}_{\{T_k(\alpha)\leq n\}},$$

$$N_\infty([a,b],\alpha) = \sup\{k \geq 1 : T_k(\alpha) < \infty\} = \sum_{k=1}^{\infty} \mathbf{1}_{\{T_k(\alpha)<\infty\}},$$

where $\sup \varnothing = 0$. The quantity $N_\infty([a,b],\alpha)$ is the total upcrossing number along $[a,b]$ for the sequence $(\alpha_n)_{n\in\mathbb{Z}_+}$. We will use the following simple analytic lemma, whose proof is left to the reader.

Lemma 12.13 *The sequence $(\alpha_n)_{n\in\mathbb{Z}_+}$ converges in $\overline{\mathbb{R}}$ if and only if, for every choice of the rationals a and b with $a < b$, we have $N_\infty([a,b],\alpha) < \infty$.*

Note that we do not exclude the possibility $\alpha_n \longrightarrow +\infty$ or $\alpha_n \longrightarrow -\infty$.

Let us replace the sequence $(\alpha_n)_{n\in\mathbb{Z}_+}$ by a random process $X = (X_n)_{n\in\mathbb{Z}_+}$, assuming that this process is adapted to the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$. The previous definitions then apply to define the quantities $S_k(X)$, $T_k(X)$, for every $k \geq 0$, and these quantities are now random. More precisely, we observe that $S_k(X)$ and $T_k(X)$ are stopping times, for every $k \geq 0$. Indeed, we can write

$$\{T_k(X) \leq n\}$$
$$= \bigcup_{0\leq m_1<n_1<\cdots<m_k<n_k\leq n} \{X_{m_1} \leq a, X_{n_1} \geq b, \ldots, X_{m_k} \leq a, X_{n_k} \geq b\},$$

which implies that $\{T_k(X) \leq n\} \in \mathcal{F}_n$, and a similar argument shows that $\{S_k(X) \leq n\} \in \mathcal{F}_n$.

As a consequence, we get that $N_n([a,b],X) = \sum_{k=1}^{\infty} \mathbf{1}_{\{T_k(\alpha)\leq n\}}$ is $\mathcal{F}_n$-measurable.

Lemma 12.14 (Doob's Upcrossing Inequality) *Let $X = (X_n)_{n\in\mathbb{Z}_+}$ be a submartingale. Then, for every reals $a < b$ and every $n \in \mathbb{Z}_+$,*

$$(b-a)\,\mathbb{E}[N_n([a,b],X)] \leq \mathbb{E}[(X_n-a)^+ - (X_0-a)^+].$$

Proof Fix $a < b$ and, to simplify notation, set $N_n = N_n([a,b],X)$, and also write S_k, T_k instead of $S_k(X), T_k(X)$. We define a nonnegative predictable sequence $(H_n)_{n\in\mathbb{N}}$ by setting, for every $n \in \mathbb{N}$,

$$H_n = \sum_{k=1}^{\infty} \mathbf{1}_{\{S_k<n\leq T_k\}} \leq 1$$

To verify that this sequence is predictable, we note that the event

$$\{S_k < n \leq T_k\} = \{S_k \leq n-1\}\backslash\{T_k \leq n-1\}$$

belongs to $\mathcal{F}_{n-1}$, because both S_k and T_k are stopping times.

Next set $Y_n = (X_n - a)^+$ for every $n \in \mathbb{Z}_+$. By Proposition 12.4, $(Y_n)_{n\in\mathbb{Z}_+}$ is a submartingale. Then, we observe that

$$\begin{aligned}(H \bullet Y)_n &= \sum_{k=1}^{N_n}(Y_{T_k} - Y_{S_k}) + \mathbf{1}_{\{S_{N_n+1}<n\}}(Y_n - Y_{S_{N_n+1}}) \\ &\geq \sum_{k=1}^{N_n}(Y_{T_k} - Y_{S_k}) \\ &\geq N_n(b-a).\end{aligned}$$

The first inequality holds because $Y_{S_{N_n+1}} = 0$ on the event $\{S_{N_n+1} < \infty\}$ (on this event, we have $X_{S_{N_n+1}} \leq a$), and $Y_n \geq 0$. We get in particular

$$\mathbb{E}[(H \bullet Y)_n] \geq (b-a)\,\mathbb{E}[N_n].$$

On the other hand, if $K_n = 1 - H_n$, $(K_n)_{n\in\mathbb{Z}_+}$ is also a nonnegative predictable sequence, and Proposition 12.6 implies that $(K \bullet Y)_n$ is a submartingale, hence $\mathbb{E}[(K \bullet Y)_n] \geq \mathbb{E}[(K \bullet Y)_0] = 0$.

We finally observe that

$$(K \bullet Y)_n + (H \bullet Y)_n = ((K + H) \bullet Y)_n = Y_n - Y_0 \,,$$

and thus

$$(b-a)\,\mathbb{E}[N_n] \leq \mathbb{E}[(H \bullet Y)_n] \leq \mathbb{E}[(K \bullet Y)_n + (H \bullet Y)_n] = \mathbb{E}[Y_n - Y_0]$$

which is the desired inequality. □

Theorem 12.15 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a submartingale or a supermartingale bounded in L^1, i.e. such that*

$$\sup_{n\in\mathbb{Z}_+} \mathbb{E}[|X_n|] < \infty. \tag{12.3}$$

Then the sequence $(X_n)_{n\in\mathbb{Z}_+}$ converges almost surely as $n \to \infty$. Furthermore, its limit X_∞ satisfies $\mathbb{E}[|X_\infty|] < \infty$.

Proof It is enough to treat the case of a submartingale. Let $a, b \in \mathbb{Q}$ such that $a < b$. By Lemma 12.14, we have, for every $n \geq 1$,

$$(b-a)\,\mathbb{E}[N_n([a,b],X)] \leq \mathbb{E}[(X_n - a)^+] \leq |a| + \mathbb{E}[(X_n)^+] \leq |a| + \sup_{k\in\mathbb{Z}_+} \mathbb{E}[|X_k|].$$

By letting $n \to +\infty$, and using (12.3), we get

$$(b-a)\,\mathbb{E}[N_\infty([a,b],X)] < \infty$$

and thus $N_\infty([a,b],X) < \infty$ a.s. Up to discarding a countable union of events of zero probability, we obtain that, almost surely, for every rationals $a < b$, $N_\infty([a,b],X) < \infty$. By Lemma 12.13, this is enough to get that X_n converges almost surely in $\overline{\mathbb{R}}$.

Then, thanks to Fatou's lemma, we have

$$\mathbb{E}[|X_\infty|] \leq \liminf_{n\to\infty} \mathbb{E}[|X_n|] \leq \sup_{n\in\mathbb{Z}_+} \mathbb{E}[|X_n|] < \infty$$

and in particular $|X_\infty| < \infty$ a.s. □

We note that a nonnegative supermartingale is always bounded in L^1 since $\mathbb{E}[X_n] \leq \mathbb{E}[X_0]$ for every $n \geq 0$.

Corollary 12.16 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a nonnegative supermartingale. Then, X_n converges a.s. as $n \to \infty$. Its limit X_∞ is in L^1 and we have $X_n \geq \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ for every $n \in \mathbb{Z}_+$.*

Proof This follows from Theorem 12.15, as explained before the statement of the corollary, except for the last assertion, which is a consequence of Fatou's lemma for conditional expectations (property (e) in Section 11.2.2):

$$X_n \geq \liminf_{m\to\infty} \mathbb{E}[X_m \mid \mathcal{F}_n] \geq \mathbb{E}[\liminf_{m\to\infty} X_m \mid \mathcal{F}_n] = \mathbb{E}[X_\infty \mid \mathcal{F}_n].$$

□

Examples

(1) Let $(Y_n)_{n\in\mathbb{Z}_+}$ be a simple random walk on $\mathbb{Z}$ started from $Y_0 = 1$. We have seen that $(Y_n)_{n\in\mathbb{Z}_+}$ is a martingale with respect to its canonical filtration. Set

$$T = \inf\{n \geq 0 : Y_n = 0\}.$$

Then T is a stopping time, and Theorem 12.12 implies that $X_n = Y_{n\wedge T}$ is a nonnegative martingale, to which we can apply Corollary 12.16. Hence X_n converges a.s. to a limiting random variable X_∞ such that $X_\infty < \infty$. However, on the event $\{T = \infty\}$, we have $|X_{n+1} - X_n| = [Y_{n+1} - Y_n| = 1$ for every n, which makes it impossible for the sequence X_n to converge. It follows that $T < \infty$ a.s. (this is the property we used in the example following the proof of Theorem 12.12). Also note that we have $X_\infty = 0$ a.s. and thus the inequality $X_n \geq \mathbb{E}[X_\infty \mid \mathcal{F}_n] = 0$ is not an equality despite the fact that $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale.

This example also shows that the convergences of Theorem 12.15 or Corollary 12.16 do not necessarily hold in L^1. Indeed, we have here $\mathbb{E}[X_n] = 1$ for every n since X is a martingale, but $\mathbb{E}[X_\infty] = 0$.

(2) *Branching processes*. Let μ be a probability measure on $\mathbb{Z}_+$, such that

$$m := \sum_{k=0}^{\infty} k\,\mu(k) < \infty.$$

To avoid trivialities, we exclude the cases $\mu = \delta_1$ and $\mu = \delta_0$.

Then let $(\xi_{n,j})_{n\in\mathbb{Z}_+, j\in\mathbb{N}}$ be a collection of independent identically distributed random variables with law μ. Also fix an integer $\ell \geq 1$. We define a sequence $(Z_n)_{n\in\mathbb{Z}_+}$ of random variables with values in $\mathbb{Z}_+$ by induction, by setting

$$Z_0 = \ell$$

$$Z_{n+1} = \sum_{j=1}^{Z_n} \xi_{n,j}\ , \qquad \forall n \in \mathbb{Z}_+.$$

In particular, $Z_{n+1} = 0$ if $Z_n = 0$. We interpret Z_n as the number of individuals at generation n in a population that evolves according to the following rules. The initial size of the population if ℓ, and then, at every generation, each individual has a random number of offspring distributed according to μ (and these numbers are independent for all individuals). The random process $(Z_n)_{n\in\mathbb{Z}_+}$ is called the Galton-Watson process with offspring distribution μ (and initial size ℓ).

We claim that $m^{-n}Z_n$ is a nonnegative martingale with respect to the filtration

$$\mathcal{F}_0 = \{\varnothing, \Omega\}$$
$$\mathcal{F}_n = \sigma(\xi_{k,j} : k \in \{0, 1, \ldots, n-1\}, j \in \mathbb{N}), \quad \text{if } n \geq 1.$$

The random process (Z_n) is adapted with respect to this filtration, because the definition of Z_n only involves the random variables $\xi_{k,j}$ for indices k such that $k < n$. Furthermore, for every $n \geq 0$,

$$\mathbb{E}[Z_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}\Big[\sum_{j=1}^{\infty} \mathbf{1}_{\{j \leq Z_n\}}\, \xi_{n,j} \,\Big|\, \mathcal{F}_n\Big] = \sum_{j=1}^{\infty} \mathbf{1}_{\{j \leq Z_n\}}\, \mathbb{E}[\xi_{n,j} \,|\, \mathcal{F}_n] = m\, Z_n$$

since $\xi_{n,j}$ is independent of $\mathcal{F}_n$, and thus $\mathbb{E}[\xi_{n,j} \mid \mathcal{F}_n] = \mathbb{E}[\xi_{n,j}] = m$. Consequently,

$$\mathbb{E}[m^{-(n+1)} Z_{n+1} \,|\, \mathcal{F}_n] = m^{-n} Z_n.$$

This implies $\mathbb{E}[m^{-(n+1)} Z_{n+1}] = \mathbb{E}[m^{-n} Z_n]$, and by induction we immediately get $\mathbb{E}[Z_n] = \ell\, m^n$, showing that the random variables Z_n are in L^1. Then the last display exactly gives the martingale property of the process $(m^{-n} Z_n)_{n\in\mathbb{Z}_+}$.

By Corollary 12.16, we get that $m^{-n} Z_n$ converges a.s. to a finite limit Z as $n \to \infty$. Let us distinguish three cases:

- $m < 1$. Since Z_n is integer-valued, the convergence of $m^{-n} Z_n$ is only possible if $Z_n = 0$ for every sufficiently large n (this means that the population becomes extinct a.s.).
- $m = 1$. In that case, Z_n is a nonnegative martingale and we can prove that the extinction of the population again occurs a.s. To verify this, since we know that Z_n converges a.s. as $n \to \infty$, we have to exclude the possibility that $Z_n = p$ for every sufficiently large n, for some integer $p \geq 1$. Put differently, we have to check that, for every $p \geq 1$,

$$\mathbb{P}(\exists N \geq 1 : \forall n \geq N, Z_n = p) = 0.$$

This is obtained by a straightforward application of the Borel-Cantelli lemma (recall that we excluded $\mu = \delta_1$, so that $\mu(1) < 1$).

- $m > 1$. We have

$$m^{-n} Z_n \xrightarrow[n\to\infty]{\text{a.s.}} Z \tag{12.4}$$

so that, on the event $\{Z > 0\}$, we see that the population size Z_n grows like m^n when n is large. Note that, if $\mu(0) > 0$, the event of extinction of the population still has positive probability. Two natural questions are then:

— Do we have $\mathbb{P}(Z > 0) > 0$?
— Do we have $Z > 0$ a.s. on the non-extinction event $\bigcap_{n\geq 0}\{Z_n > 0\}$?

A classical result known as the Kesten-Stigum theorem shows that the answer to both questions is yes if

$$\sum_{k=1}^{\infty} k \log(k)\, \mu(k) < \infty,$$

and moreover the convergence (12.4) holds in L^1 under this condition. We will prove a weaker result in Section 12.4 below.

Let $(X_n)_{n\in\mathbb{Z}_+}$ be a martingale. If the sequence $(X_n)_{n\in\mathbb{Z}_+}$ is bounded in L^1, Theorem 12.15 shows that X_n converges a.s. as $n \to \infty$. We have seen that this convergence does not always hold in L^1. The next theorem characterizes those martingales for which the L^1-convergence holds.

Theorem 12.17 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a martingale. The following two conditions are equivalent:*

(i) *X_n converges a.s. and in L^1 to a random variable denoted by X_∞.*
(ii) *The martingale $(X_n)_{n\in\mathbb{Z}_+}$ is closed, in the sense that there exists a random variable $Z \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ such that $X_n = \mathbb{E}[Z \,|\, \mathcal{F}_n]$ for every $n \in \mathbb{Z}_+$.*

Moreover, if these conditions hold, we can take $Z = X_\infty$ in (ii).

Proof Assume first that (i) holds. For integers $0 \leq n < m$, we have

$$X_n = \mathbb{E}[X_m \,|\, \mathcal{F}_n].$$

The mapping $Y \longrightarrow \mathbb{E}[Y \mid \mathcal{F}_n]$ is continuous from L^1 into L^1 (it is even a contraction of L^1 since $\mathbb{E}[|\mathbb{E}[Y \mid \mathcal{F}_n]|] \leq \mathbb{E}[|Y|]$). Therefore, by letting $m \to \infty$ in the last display, we get $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$.

Conversely, assume that (ii) holds. Then the sequence $(X_n)_{n\in\mathbb{Z}_+}$ is bounded in L^1 and thus converges a.s. by Theorem 12.15. Consider first the case where the random variable $|Z|$ is bounded by a constant $K < \infty$. Then we have also $|X_n| \leq K$ for every n, and the dominated convergence theorem gives $\mathbb{E}[|X_n - X_\infty|] \longrightarrow 0$. In the general case, fix $\varepsilon > 0$, and choose $M > 0$ large enough so that

$$\mathbb{E}[|Z - Z\,\mathbf{1}_{\{|Z|\leq M\}}|] < \varepsilon.$$

(note that $\mathbb{E}[|Z|\,\mathbf{1}_{\{|Z|>M\}}] \longrightarrow 0$ as $M \to \infty$ by dominated convergence). Then, for every n,

$$\begin{aligned}\mathbb{E}[|X_n - \mathbb{E}[Z\,\mathbf{1}_{\{|Z|\le M\}}\,|\,\mathcal{F}_n]|] &= \mathbb{E}[|\mathbb{E}[Z - Z\,\mathbf{1}_{\{|Z|\le M\}}\,|\,\mathcal{F}_n]|]\\ &\le \mathbb{E}[|Z - Z\,\mathbf{1}_{\{|Z|\le M\}}|]\\ &< \varepsilon.\end{aligned}$$

From the bounded case, the martingale $\mathbb{E}[Z\,\mathbf{1}_{\{|Z|\le M\}}\,|\,\mathcal{F}_n]$ converges in L^1. Hence, we can choose n_0 large enough so that, for every $m, n \ge n_0$,

$$\mathbb{E}[|\mathbb{E}[Z\,\mathbf{1}_{\{|Z|\le M\}}\,|\,\mathcal{F}_m] - \mathbb{E}[Z\,\mathbf{1}_{\{|Z|\le M\}}\,|\,\mathcal{F}_n]|] < \varepsilon.$$

By combining this with the preceding display, we get, for every $m, n \ge n_0$,

$$\mathbb{E}[|X_m - X_n|] < 3\varepsilon.$$

Since ε was arbitrary, we have obtained that the sequence $(X_n)_{n\in\mathbb{Z}_+}$ is Cauchy in L^1, and thus converges in L^1. □

Recall that $\mathcal{F}_\infty = \bigvee_{n=1}^{\infty} \mathcal{F}_n = \sigma\Big(\bigcup_{n=1}^{\infty} \mathcal{F}_n\Big)$.

Corollary 12.18 *Let $Z \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. The martingale $X_n = \mathbb{E}[Z\,|\,\mathcal{F}_n]$ converges a.s. and in L^1 to $X_\infty = \mathbb{E}[Z\,|\,\mathcal{F}_\infty]$.*

Proof From the preceding theorem, we already know that X_n converges a.s. and in L^1 to a random variable denoted by X_∞. It remains to prove that $X_\infty = \mathbb{E}[Z\,|\,\mathcal{F}_\infty]$. We first note that X_∞ is $\mathcal{F}_\infty$-measurable as a limit of $\mathcal{F}_\infty$-measurable random variables. Then, for every $n \in \mathbb{Z}_+$ and $A \in \mathcal{F}_n$, the characteristic property of conditional expectation gives

$$\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[X_n\,\mathbf{1}_A].$$

By letting $n \to \infty$, and using the L^1-convergence, we get $\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[X_\infty\,\mathbf{1}_A]$. A monotone class argument (Theorem 1.18) shows that the equality $\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[X_\infty\,\mathbf{1}_A]$, which holds for $A \in \bigcup_{n=1}^{\infty} \mathcal{F}_n$, remains valid for $A \in \sigma(\bigcup_{n=1}^{\infty} \mathcal{F}_n) = \mathcal{F}_\infty$. The characteristic property of conditional expectation then gives the desired result. □

Example Consider example (iv) in Section 12.1: $\Omega = [0, 1)$, $\mathcal{A} = \mathcal{B}([0, 1))$ is the Borel σ-field on $[0, 1)$, and $\mathbb{P} = \lambda$ is Lebesgue measure. The dyadic filtration is defined by

$$\mathcal{F}_n := \sigma\Big([\frac{i-1}{2^n}, \frac{i}{2^n}) : i \in \{1, 2, \ldots, 2^n\}\Big).$$

Let μ be a finite measure on $[0, 1)$, and, for every integer $n \in \mathbb{Z}_+$,

$$f_n(\omega) = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda_{|\mathcal{F}_n}}(\omega) = \sum_{i=1}^{2^n} \frac{\mu([(i-1)2^{-n}, i2^{-n}))}{2^{-n}} \mathbf{1}_{[(i-1)2^{-n}, i2^{-n})}(\omega).$$

We already noticed that $(f_n)_{n\in\mathbb{Z}_+}$ is a nonnegative martingale, and thus (Corollary 12.16),

$$f_n \xrightarrow[n\to\infty]{\text{a.s.}} f_\infty$$

where $\int f_\infty \mathrm{d}\lambda < \infty$. Moreover $f_n \geq \mathbb{E}[f_\infty \mid \mathcal{F}_n]$, which shows that, for every $A \in \mathcal{F}_n$,

$$\mu(A) = \int f_n \mathbf{1}_A \,\mathrm{d}\lambda \geq \int \mathbb{E}[f_\infty \,|\, \mathcal{F}_n] \mathbf{1}_A \,\mathrm{d}\lambda = \int f_\infty \mathbf{1}_A \,\mathrm{d}\lambda.$$

Recall our notation $f_\infty \cdot \lambda$ for the measure with density f_∞ with respect to λ. By the preceding display, we have $\mu(A) \geq f_\infty \cdot \lambda\,(A)$ for every $A \in \cup_{n=0}^\infty \mathcal{F}_n$. This implies that $\mu(O) \geq f_\infty \cdot \lambda\,(O)$ for every open subset O of $[0, 1)$ (every such set is a disjoint union of at most countably many sets in $\cup_{n=0}^\infty \mathcal{F}_n$). Furthermore, the regularity properties of finite measures on $[0, 1)$ (Proposition 3.10) imply that, for every Borel subset A of $[0, 1)$, $\mu(A) = \inf\{\mu(O) : O \text{ open}, A \subset O\}$ and it readily follows that the inequality $\mu(A) \geq f_\infty \cdot \lambda\,(A)$ holds for every $A \in \mathcal{B}([0, 1))$. We can then set $\nu = \mu - f_\infty \cdot \lambda$, and ν is a finite (positive) measure on $[0, 1)$.

Let us show that ν is singular with respect to λ. For every $n \geq 0$, set

$$h_n = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda_{|\mathcal{F}_n}} = f_n - \mathbb{E}[f_\infty \,|\, \mathcal{F}_n],$$

using (12.2) in the second equality. We now use Corollaire 12.18. In the special case we are considering, we have $\mathcal{F}_\infty = \mathcal{A}$ and thus we get $\mathbb{E}[f_\infty \,|\, \mathcal{F}_n] \longrightarrow f_\infty$ a.s. Consequently $h_n \longrightarrow 0$ a.s. and

$$\lambda\Big(\Big\{x \in [0, 1) : \limsup_{n\to\infty} h_n(x) > 0\Big\}\Big) = 0. \tag{12.5}$$

On the other hand, for every $\varepsilon > 0$, and $n \in \mathbb{Z}_+$,

$$\nu(\{x \in [0, 1) : h_n(x) \leq \varepsilon\}) = \int \mathbf{1}_{\{h_n \leq \varepsilon\}} h_n \,\mathrm{d}\lambda \leq \varepsilon,$$

which implies

$$\nu\Big(\Big\{x \in [0,1) : \limsup_{n\to\infty} h_n(x) < \varepsilon\Big\}\Big) \le \nu\Big(\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\{h_n \le \varepsilon\}\Big)$$
$$= \lim_{N\to\infty} \uparrow \nu\Big(\bigcap_{n=N}^{\infty}\{h_n \le \varepsilon\}\Big)$$
$$\le \varepsilon.$$

We have thus obtained

$$\nu\Big(\Big\{x \in [0,1) : \limsup_{n\to\infty} h_n(x) = 0\Big\}\Big) = 0$$

and comparing with (12.5) we see that λ and ν are supported on disjoint sets.

From the preceding considerations, we see that $\mu = f_\infty \cdot \lambda + \nu$ is the decomposition of the measure μ as the sum of a measure absolutely continuous with respect to λ and a singular measure. We have thus recovered a special case of the Radon-Nikodym theorem (Theorem 4.11). Note that μ is absolutely continuous with respect to λ if and only if $\nu = 0$, which holds if and only if the martingale $(f_n)_{n\in\mathbb{Z}_+}$ is closed.

12.4 Convergence in L^p When $p > 1$

Our goal is now to study which conditions ensure that a martingale converges in L^p when $p > 1$. The arguments will depend on some important estimates for the distribution of the maximal value of a martingale. We start with a simple lemma.

Lemma 12.19 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a submartingale, and let S and T be two bounded stopping times such that $S \le T$. Then*

$$\mathbb{E}[X_S] \le \mathbb{E}[X_T].$$

Remark The case $S = 0$ is contained in Theorem 12.12.

Proof We already know that X_S and X_T are in L^1. We define a predictable sequence by setting, for every $n \ge 1$,

$$H_n := \mathbf{1}_{\{S<n\le T\}} = \mathbf{1}_{\{S\le n-1\}} - \mathbf{1}_{\{T\le n-1\}}.$$

Let N be an integer such that $S \leq T \leq N$. Then, with the notation of Proposition 12.6,

$$(H \bullet X)_N = X_T - X_S$$

and $\mathbb{E}[(H \bullet X)_N] \geq 0$ since $H \bullet X$ is a submartingale (Proposition 12.6). □

Theorem 12.20 (Doob's maximal inequality) *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a submartingale. Then, for every $a > 0$ and every $n \in \mathbb{Z}_+$,*

$$a\,\mathbb{P}\Big(\sup_{0\leq k\leq n} X_k \geq a\Big) \leq \mathbb{E}\Big[X_n\,\mathbf{1}_{\{\sup_{0\leq k\leq n} X_k \geq a\}}\Big] \leq \mathbb{E}[X_n^+].$$

Similarly, if $(Y_n)_{n\in\mathbb{Z}_+}$ is a supermartingale, then, for every $a > 0$ and every $n \in \mathbb{Z}_+$,

$$a\,\mathbb{P}\Big(\sup_{0\leq k\leq n} Y_k \geq a\Big) \leq \mathbb{E}[Y_0] + \mathbb{E}[Y_n^-].$$

Proof Let us prove the first assertion. Consider the stopping time

$$T = \inf\{k \geq 0 : X_k \geq a\}.$$

Then, if

$$A = \Big\{\sup_{0\leq k\leq n} X_k \geq a\Big\},$$

we have $A = \{T \leq n\}$. On the other hand, by applying the preceding lemma to the stopping times $T \wedge n$ and n, we have

$$\mathbb{E}[X_{T\wedge n}] \leq \mathbb{E}[X_n].$$

Furthermore,

$$X_{T\wedge n} \geq a\,\mathbf{1}_A + X_n\,\mathbf{1}_{A^c}.$$

By combining these two inequalities, we get

$$\mathbb{E}[X_n] \geq a\mathbb{P}(A) + \mathbb{E}[X_n\,\mathbf{1}_{A^c}]$$

and it follows that $a\mathbb{P}(A) \leq \mathbb{E}[X_n\,\mathbf{1}_A]$, which gives the first part of the theorem.

The proof of the second part is very similar. We introduce the stopping time

$$R = \inf\{k \geq 0 : Y_k \geq a\},$$

and the event $B = \{R \le n\}$. We have then $\mathbb{E}[Y_{R\wedge n}] \le \mathbb{E}[Y_0]$, and $\mathbb{E}[Y_{R\wedge n}] \ge a\mathbb{P}(B) + \mathbb{E}[Y_n \mathbf{1}_{B^c}]$. It follows that $a\mathbb{P}(B) \le \mathbb{E}[Y_0] - \mathbb{E}[Y_n \mathbf{1}_{B^c}]$, which leads to the desired result. □

Proposition 12.21 *Let $p \in (1,\infty)$ and let $(X_n)_{n\in\mathbb{Z}_+}$ be a nonnegative submartingale. For every $n \in \mathbb{Z}_+$, set*

$$\widetilde{X}_n = \sup_{0\le k\le n} X_k.$$

Then,

$$\mathbb{E}[(\widetilde{X}_n)^p] \le \Big(\frac{p}{p-1}\Big)^p\, \mathbb{E}[(X_n)^p].$$

Consequently, if $(Y_n)_{n\in\mathbb{Z}_+}$ is a martingale, and

$$Y_n^* = \sup_{0\le k\le n} |Y_k|$$

we have, for every $n \in \mathbb{Z}_+$,

$$\mathbb{E}[(Y_n^*)^p] \le \Big(\frac{p}{p-1}\Big)^p\, \mathbb{E}[|Y_n|^p].$$

Remark In the first part of the proposition, we do not exclude the possibility that $\mathbb{E}[(X_n)^p] = \infty$, in which case the bound of the proposition reduces to $\infty \le \infty$. A similar remark applies to the second part.

Proof The second part of the proposition follows from the first part applied to the submartingale $X_n = |Y_n|$. In order to prove the first part, we may assume that $\mathbb{E}[(X_n)^p] < \infty$. Then, by Jensen's inequality for conditional expectations, we have, for every $0 \le k \le n$,

$$\mathbb{E}[(X_k)^p] \le \mathbb{E}[\mathbb{E}[X_n \,|\, \mathcal{F}_k]^p] \le \mathbb{E}[\mathbb{E}[(X_n)^p \,|\, \mathcal{F}_n]] = \mathbb{E}[(X_n)^p]. \tag{12.6}$$

Since $\widetilde{X}_n \le X_0 + \cdots + X_n$, it follows that we have also $\mathbb{E}[(\widetilde{X}_n)^p] < \infty$.

By Theorem 12.20, for every $a > 0$,

$$a\,\mathbb{P}(\widetilde{X}_n \ge a) \le \mathbb{E}[X_n \mathbf{1}_{\{\widetilde{X}_n \ge a\}}].$$

We multiply each side of this inequality by a^{p-2} and integrate with respect to Lebesgue measure on $(0,\infty)$. For the left side, we get

$$\int_0^\infty a^{p-1}\mathbb{P}(\widetilde{X}_n \ge a)\,\mathrm{d}a = \mathbb{E}\Big[\int_0^{\widetilde{X}_n} a^{p-1}\mathrm{d}a\Big] = \frac{1}{p}\,\mathbb{E}[(\widetilde{X}_n)^p]$$

using the Fubini theorem. Similarly, we get for the right side

$$\begin{aligned}\int_0^\infty a^{p-2}\mathbb{E}[X_n \mathbf{1}_{\{\widetilde{X}_n \geq a\}}]\mathrm{d}a &= \mathbb{E}\Big[X_n \int_0^{\widetilde{X}_n} a^{p-2}\mathrm{d}a\Big] \\ &= \frac{1}{p-1}\mathbb{E}[X_n(\widetilde{X}_n)^{p-1}] \\ &\leq \frac{1}{p-1}\mathbb{E}[(X_n)^p]^{\frac{1}{p}}\mathbb{E}[(\widetilde{X}_n)^p]^{\frac{p-1}{p}}\end{aligned}$$

by the Hölder inequality. We thus get

$$\frac{1}{p}\mathbb{E}[(\widetilde{X}_n)^p] \leq \frac{1}{p-1}\mathbb{E}[(X_n)^p]^{\frac{1}{p}}\mathbb{E}[(\widetilde{X}_n)^p]^{\frac{p-1}{p}}$$

giving the first part of the proposition (we use the fact that $\mathbb{E}[(\widetilde{X}_n)^p] < \infty$). □

If $(X_n)_{n\in\mathbb{Z}_+}$ is a random process, we set $X^*_\infty = \sup_{n\in\mathbb{Z}_+} |X_n|$.

Theorem 12.22 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a martingale and $p \in (1, \infty)$. Suppose that*

$$\sup_{n\in\mathbb{Z}_+} \mathbb{E}[|X_n|^p] < \infty,$$

Then, X_n converges a.s. and in L^p to a random variable X_∞ such that

$$\mathbb{E}[|X_\infty|^p] = \sup_{n\in\mathbb{Z}_+} \mathbb{E}[|X_n|^p]$$

and we have

$$\mathbb{E}[(X^*_\infty)^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_\infty|^p].$$

Proof Since the martingale (X_n) is bounded in L^p, it is also bounded in L^1, and we know from Theorem 12.15 that X_n converges a.s. to a limiting random variable X_∞. Moreover, Proposition 12.21 shows that, for every $n \in \mathbb{Z}_+$,

$$\mathbb{E}[(X^*_n)^p] \leq \left(\frac{p}{p-1}\right)^p \sup_{k\in\mathbb{Z}_+} \mathbb{E}[|X_k|^p].$$

Since $X^*_n \uparrow X^*_\infty$ when $n \uparrow \infty$, the monotone convergence theorem gives

$$\mathbb{E}[(X^*_\infty)^p] \leq \left(\frac{p}{p-1}\right)^p \sup_{k\in\mathbb{Z}_+} \mathbb{E}[|X_k|^p] < \infty$$

and thus $X^*_\infty \in L^p$. Since all random variables $|X_n|^p$ are dominated by $|X^*_\infty|^p$, the dominated convergence theorem ensures that X_n converges to X_∞ in L^p. Finally, we know from (12.6) that the sequence $\mathbb{E}[|X_n|^p]$ is increasing, and therefore

$$\mathbb{E}[|X_\infty|^p] = \lim_{n\to\infty} \mathbb{E}[|X_n|^p] = \sup_{n\in\mathbb{Z}_+} \mathbb{E}[|X_n|^p].$$

□

Example Let $(Z_n)_{n\in\mathbb{Z}_+}$ be the Galton-Watson branching process with offspring distribution μ and initial population $\ell \geq 1$ considered in the previous section. We now assume that

$$m = \sum_{k=0}^{\infty} k\,\mu(k) \in (1, \infty)$$

and

$$\sum_{k=0}^{\infty} k^2 \mu(k) < \infty.$$

We also set $\sigma^2 = \mathrm{var}(\mu) = (\sum k^2 \mu(k)) - m^2$. We have already observed that $m^{-n} Z_n$ is a martingale. We claim that, under the preceding assumptions, this martingale is bounded in L^2. To verify this property, we compute

$$\begin{aligned}\mathbb{E}[Z_{n+1}^2 \,|\, \mathcal{F}_n] &= \mathbb{E}\Big[\sum_{j,k=1}^{\infty} \mathbf{1}_{\{j\leq Z_n, k\leq Z_n\}} \xi_{n,j}\xi_{n,k} \,\Big|\, \mathcal{F}_n \Big] \\ &= \sum_{j,k=1}^{\infty} \mathbf{1}_{\{j\leq Z_n, k\leq Z_n\}}\, \mathbb{E}[\xi_{n,j}\xi_{n,k}] \\ &= \sum_{j,k=1}^{\infty} \mathbf{1}_{\{j\leq Z_n, k\leq Z_n\}}\, (m^2 + \sigma^2 \mathbf{1}_{\{j=k\}}) \\ &= m^2 Z_n^2 + \sigma^2 Z_n.\end{aligned}$$

Since we know that $\mathbb{E}[Z_n] = m^n \mathbb{E}[Z_0] = \ell\, m^n$, it follows that

$$\mathbb{E}[Z_{n+1}^2] = m^2 \mathbb{E}[Z_n^2] + \ell\sigma^2 m^n.$$

Setting $a_n = m^{-2n}\mathbb{E}[Z_n^2]$, we get

$$a_{n+1} = a_n + \ell\sigma^2 m^{-n-2}$$

and, since $m > 1$, the sequence $(a_n)_{n\in\mathbb{Z}_+}$ converges to a finite limit, and in particular this sequence is bounded. Consequently, we have obtained that the martingale $m^{-n}Z_n$ is bounded in L^2. By Theorem 12.22, this martingale converges a.s. and in L^2 to a random variable Z_∞. In particular, $\mathbb{E}[Z_\infty] = \mathbb{E}[Z_0] = \ell$ and therefore $\mathbb{P}(Z_\infty > 0) > 0$ (it is not difficult to see that we have $Z_\infty > 0$ a.s. on the event of non-extinction, but we omit the details).

We conclude this section with an important application to series of independent random variables.

Theorem 12.23 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a sequence of independent real random variables in L^2. Assume that $\mathbb{E}[X_n] = 0$ for every $n \in \mathbb{Z}_+$. Then the following two conditions are equivalent:*

(i) $\displaystyle\sum_{n=0}^{\infty} \mathbb{E}[(X_n)^2] < \infty$.

(ii) *The series $\displaystyle\sum_{n=0}^{\infty} X_n$ converges a.s. and in L^2.*

Proof For every integer $n \geq 0$, set $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ and

$$M_n = \sum_{k=0}^{n} X_k.$$

Since

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = \sum_{k=0}^{n} X_k + \mathbb{E}[X_{n+1}] = \sum_{k=0}^{n} X_k\,,$$

$(M_n)_{n\in\mathbb{Z}_+}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$. Furthermore, since $\mathbb{E}[X_j X_k] = 0$ if $j \neq k$, one has

$$\mathbb{E}[(M_n)^2] = \sum_{k=0}^{n} \mathbb{E}[(X_n)^2].$$

Hence, if M_n converges in L^2 as $n \to \infty$, condition (i) of the theorem must hold (we have just recovered a special case of a classical result of Hilbert space theory, see Proposition A.4 in the Appendix). Conversely, if (i) holds, the martingale M_n is bounded in L^2, and by Theorem 12.22 we get that M_n converges a.s. and in L^2. □

Example Suppose that $(\varepsilon_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables, such that $\mathbb{P}(\varepsilon_n = 1) = \mathbb{P}(\varepsilon_n = -1) = 1/2$. Then the series

$$\sum_{n=1}^{\infty} \frac{\varepsilon_n}{n^\alpha}$$

converges a.s. as soon as $\alpha > 1/2$. Note that absolute convergence holds only if $\alpha > 1$.

12.5 Uniform Integrability and Martingales

Definition 12.24 A collection $(X_i)_{i\in I}$ of random variables in $L^1(\Omega, \mathcal{A}, \mathbb{P})$ is said to be uniformly integrable (u.i. in short) if

$$\lim_{a\to+\infty} \Big(\sup_{i\in I} \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i|>a\}}] \Big) = 0.$$

A uniformly integrable collection $(X_i)_{i\in I}$ is bounded in L^1. To verify this, take a large enough so that

$$\sup_{i\in I} \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i|>a\}}] \leq 1$$

and write $\mathbb{E}[|X_i|] \leq \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i|\leq a\}}] + \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i|>a\}}] \leq a + 1$. The converse is false: a collection $(X_i)_{i\in I}$ which is bounded in L^1 may not be uniformly integrable.

Examples

(1) The collection consisting of a single random variable Z in L^1 is uniformly integrable (the dominated convergence theorem implies that $\mathbb{E}[|Z| \mathbf{1}_{\{|Z|>a\}}]$ tends to 0 as $a \to \infty$). More generally, any finite subset of $L^1(\Omega, \mathcal{A}, \mathbb{P})$ is u.i.

(2) If Z is a nonnegative random variable in $L^1(\Omega, \mathcal{A}, \mathbb{P})$, the set of all real random variables X such that $|X| \leq Z$ is u.i. (just bound $\mathbb{E}[|X| \mathbf{1}_{\{|X|>a\}}] \leq \mathbb{E}[Z \mathbf{1}_{\{Z>a\}}]$ and use (1)).

(3) Let $\Phi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ be a monotone increasing function such that $x^{-1}\Phi(x) \longrightarrow +\infty$ as $x \to +\infty$. Then, for every $C > 0$, the set $\{X \in L^1(\Omega, \mathcal{A}, \mathbb{P}) : \mathbb{E}[\Phi(|X|)] \leq C\}$ is u.i. Indeed, it suffices to write

$$\mathbb{E}[|X| \mathbf{1}_{\{|X|>a\}}] \leq (\sup_{x>a} \frac{x}{\Phi(x)})\, \mathbb{E}[\Phi(|X|)],$$

and to note that the supremum in the right-hand side tends to 0 as $a \to \infty$.

(4) If $p \in (1, \infty)$, any bounded subset of $L^p(\Omega, \mathcal{A}, \mathbb{P})$ is u.i. This is the special case of (3) where $\Phi(x) = x^p$.

The name "uniformly integrable" is justified by the following proposition.

Proposition 12.25 *Let $(X_i)_{i \in I}$ be a bounded subset of L^1. The following two conditions are equivalent:*

(i) *The collection $(X_i)_{i \in I}$ is u.i.*
(ii) *For every $\varepsilon > 0$, one can find $\delta > 0$ such that, for every event $A \in \mathcal{A}$ of probability $\mathbb{P}(A) < \delta$, one has*

$$\forall i \in I, \quad \mathbb{E}[|X_i| \mathbf{1}_A] < \varepsilon.$$

Proof (i)⇒(ii) Let $\varepsilon > 0$. We can first fix $a > 0$ such that

$$\sup_{i \in I} \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i| > a\}}] < \frac{\varepsilon}{2}.$$

If we set $\delta = \varepsilon/(2a)$, then the condition $\mathbb{P}(A) < \delta$ implies, that, for every $i \in I$,

$$\mathbb{E}[|X_i| \mathbf{1}_A] \leq \mathbb{E}[|X_i| \mathbf{1}_{A \cap \{|X_i| \leq a\}}] + \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i| > a\}}] \leq a\mathbb{P}(A) + \frac{\varepsilon}{2} < \varepsilon.$$

(ii)⇒(i) Set $C = \sup_{i \in I} \mathbb{E}[|X_i|]$. By the Markov inequality, we have for every $a > 0$,

$$\forall i \in I, \quad \mathbb{P}(|X_i| > a) \leq \frac{C}{a}.$$

Let $\varepsilon > 0$ and choose $\delta > 0$ so that the property stated in (ii) holds. Then, if a is large enough so that $C/a < \delta$, we have $\mathbb{P}(|X_i| > a) < \delta$ for every $i \in I$, and thus, by (ii),

$$\forall i \in I, \quad \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i| > a\}}] < \varepsilon.$$

We conclude that the collection $(X_i)_{i \in I}$ is u.i.

□

Corollary 12.26 *Let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Then the collection of all conditional expectations $\mathbb{E}[X \mid \mathcal{G}]$ where $\mathcal{G}$ varies among sub-σ-fields of $\mathcal{A}$ is u.i.*

Proof Let $\varepsilon > 0$. Since the singleton $\{X\}$ is u.i., Proposition 12.25 allows us to find $\delta > 0$ such that, for every $A \in \mathcal{A}$ with $\mathbb{P}(A) < \delta$, we have

$$\mathbb{E}[|X| \mathbf{1}_A] \leq \varepsilon.$$

Then, for every $a > 0$, and every sub-σ-field $\mathcal{G}$,

$$\mathbb{P}(|\mathbb{E}[X \,|\, \mathcal{G}]| > a) \leq \frac{1}{a}\,\mathbb{E}[|\mathbb{E}[X \,|\, \mathcal{G}]|] \leq \frac{\mathbb{E}[|X|]}{a}.$$

Hence, if a is sufficiently large so that $\mathbb{E}[|X|]/a < \delta$, we get

$$\mathbb{E}[|\mathbb{E}[X \,|\, \mathcal{G}]|\mathbf{1}_{\{|\mathbb{E}[X|\mathcal{G}]|>a\}}] \leq \mathbb{E}[\mathbb{E}[|X| \,|\, \mathcal{G}]\mathbf{1}_{\{|\mathbb{E}[X|\mathcal{G}]|>a\}}] = \mathbb{E}[|X|\mathbf{1}_{\{|\mathbb{E}[X|\mathcal{G}]|>a\}}] < \varepsilon,$$

where we used the characteristic property of the conditional expectation $\mathbb{E}[|X| \,|\, \mathcal{G}]$. The desired uniform integrability follows. □

Theorem 12.27 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables in L^1 that converges to X_∞ in probability. The following two conditions are equivalent:*

(i) *The sequence $(X_n)_{n\in\mathbb{N}}$ converges to X_∞ in L^1.*
(ii) *The sequence $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.*

Remark The dominated convergence theorem asserts that a sequence $(X_n)_{n\in\mathbb{N}}$ that converges a.s. (hence also in probability) converges in L^1 provided that $|X_n| \leq Z$ for every n, where the random variable $Z \geq 0$ is such that $\mathbb{E}[Z] < \infty$. This domination assumption is stronger than the uniform integrability (cf. example (2) above).

Proof

(i)⇒(ii) If the sequence $(X_n)_{n\in\mathbb{Z}_+}$ converges in L^1, then it is bounded in L^1. Then, let $\varepsilon > 0$. We can choose N large enough so that, for every $n \geq N$,

$$\mathbb{E}[|X_n - X_N|] < \frac{\varepsilon}{2}.$$

Since the finite set $\{X_1, \ldots, X_N\}$ is u.i., Proposition 12.25 allows us to choose $\delta > 0$ such that, for every event A of probability $\mathbb{P}(A) < \delta$, we have

$$\forall n \in \{1, \ldots, N\}, \quad \mathbb{E}[|X_n|\mathbf{1}_A] < \frac{\varepsilon}{2}.$$

Then, if $n > N$, we have also

$$\mathbb{E}[|X_n|\mathbf{1}_A] \leq \mathbb{E}[|X_N|\mathbf{1}_A] + \mathbb{E}[|X_n - X_N|] < \varepsilon.$$

By Proposition 12.25, the sequence $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.
(ii)⇒(i) Using the characterization of uniform integrability in Proposition 12.25(ii), it is straightforward to verify that the collection $(X_n - X_m)_{n,m\in\mathbb{N}}$ is also u.i. Hence, if $\varepsilon > 0$ is fixed, we can choose $a > 0$ large enough so that,

for every $m, n \in \mathbb{N}$,

$$\mathbb{E}[|X_n - X_m|\mathbf{1}_{\{|X_n - X_m| > a\}}] < \varepsilon.$$

Then, for every $m, n \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}[|X_n - X_m|] \\
&\quad \le \mathbb{E}[|X_n - X_m|\mathbf{1}_{\{|X_n - X_m| \le \varepsilon\}}] + \mathbb{E}[|X_n - X_m|\mathbf{1}_{\{\varepsilon < |X_n - X_m| \le a\}}] \\
&\qquad + \mathbb{E}[|X_n - X_m|\mathbf{1}_{\{|X_n - X_m| > a\}}] \\
&\quad \le 2\varepsilon + a\,\mathbb{P}(|X_n - X_m| > \varepsilon).
\end{aligned}$$

The convergence in probability of the sequence $(X_n)_{n\in\mathbb{Z}_+}$ implies that

$$\mathbb{P}(|X_n - X_m| > \varepsilon) \le \mathbb{P}(|X_n - X_\infty| > \frac{\varepsilon}{2}) + \mathbb{P}(|X_m - X_\infty| > \frac{\varepsilon}{2}) \underset{n,m\to\infty}{\longrightarrow} 0.$$

It follows that

$$\limsup_{m,n\to\infty} \mathbb{E}[|X_n - X_m|] \le 2\varepsilon$$

and, since ε was arbitrary, this shows that the sequence $(X_n)_{n\in\mathbb{Z}_+}$ is Cauchy in L^1, hence converges in L^1.

□

Remark As a consequence of the theorem, we recover the result of Proposition 10.4: if a sequence $(X_n)_{n\in\mathbb{Z}_+}$ converges in probability and is bounded in L^p for some $p \in (1, \infty)$, then it converges in L^1, and even in L^q for every $1 \le q < p$ (apply the theorem to $|X_n - X_\infty|^q$).

Application to martingales . By combining Theorems 12.15 and 12.17 with Theorem 12.27, we obtain that the following three conditions are equivalent for a martingale $(X_n)_{n\in\mathbb{Z}_+}$:

(i) X_n converges a.s. and in L^1 as $n \to \infty$.
(ii) The sequence $(X_n)_{n\in\mathbb{Z}_+}$ is uniformly integrable.
(iii) The martingale $(X_n)_{n\in\mathbb{Z}_+}$ is closed: there exists a random variable $Z \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ such that $X_n = \mathbb{E}[Z \,|\, \mathcal{F}_n]$ for every $n \in \mathbb{Z}_+$.

In particular, we get that a martingale is uniformly integrable if and only if it is closed, and in that case we have $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ for every n, where X_∞ is the almost sure limit of X_n as $n \to \infty$.

The fact that (iii) implies (ii) can also be derived from Corollary 12.26. We could have used this corollary and Theorem 12.27 to get a direct proof of the implication (ii)⇒(i) in Theorem 12.17.

12.6 Optional Stopping Theorems

Let $(X_n)_{n\in\mathbb{Z}_+}$ be an adapted process. Assume that X_n converges a.s. to a limiting random variable X_∞, which is $\mathcal{F}_\infty$-measurable. For every stopping time T (possibly taking the value $+\infty$, we extend the definition of X_T, which was given only on the event $\{T < \infty\}$ in Proposition 12.11, by setting

$$X_T = \sum_{n=0}^{\infty} \mathbf{1}_{\{T=n\}} X_n + \mathbf{1}_{\{T=\infty\}} X_\infty .$$

Then it is again true that X_T is $\mathcal{F}_T$-measurable (we leave the proof as an exercise).

Theorem 12.28 (Optional Stopping Theorem for a Uniformly Integrable Martingale) *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a uniformly integrable martingale, and let X_∞ be the almost sure limit of X_n as $n \to \infty$. Then, for every stopping time T, we have*

$$X_T = \mathbb{E}[X_\infty \,|\, \mathcal{F}_T],$$

and in particular, $X_T \in L^1$ and $\mathbb{E}[X_T] = \mathbb{E}[X_\infty] = \mathbb{E}[X_n]$ for every $n \in \mathbb{Z}_+$. If S and T are two stopping times such that $S \le T$, we have

$$X_S = \mathbb{E}[X_T \,|\, \mathcal{F}_S].$$

Remarks

(i) As a consequence of the theorem and Corollary 12.26, the collection $\{X_T : T \text{ stopping time}\}$ is u.i.
(ii) If $(X_n)_{n\in\mathbb{Z}_+}$ is any martingale, we can apply Theorem 12.28 to the "stopped martingale" $(X_{n\wedge N})_{n\in\mathbb{Z}_+}$, for every fixed integer $N \ge 0$, noting that this stopped martingale is u.i., and we recover the last part of Theorem 12.12.

Proof From the preceding section, we know that the martingale $(X_n)_{n\in\mathbb{Z}_+}$ is closed and $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$ for every n. Let us verify that $X_T \in L^1$:

$$\begin{aligned}
\mathbb{E}[|X_T|] &= \sum_{n=0}^{\infty} \mathbb{E}[\mathbf{1}_{\{T=n\}}|X_n|] + \mathbb{E}[\mathbf{1}_{\{T=\infty\}}|X_\infty|] \\
&= \sum_{n=0}^{\infty} \mathbb{E}[\mathbf{1}_{\{T=n\}}|\mathbb{E}[X_\infty \,|\, \mathcal{F}_n]|] + \mathbb{E}[\mathbf{1}_{\{T=\infty\}}|X_\infty|] \\
&\le \sum_{n=0}^{\infty} \mathbb{E}[\mathbf{1}_{\{T=n\}}\mathbb{E}[|X_\infty| \,|\, \mathcal{F}_n]] + \mathbb{E}[\mathbf{1}_{\{T=\infty\}}|X_\infty|] \\
&= \sum_{n=0}^{\infty} \mathbb{E}[\mathbf{1}_{\{T=n\}}|X_\infty|] + \mathbb{E}[\mathbf{1}_{\{T=\infty\}}|X_\infty|] \\
&= \mathbb{E}[|X_\infty|] < \infty.
\end{aligned}$$

Then, if $A \in \mathcal{F}_T$,

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_A X_T] &= \sum_{n\in\mathbb{Z}_+\cup\{\infty\}} \mathbb{E}[\mathbf{1}_{A\cap\{T=n\}} X_T] \\
&= \sum_{n\in\mathbb{Z}_+\cup\{\infty\}} \mathbb{E}[\mathbf{1}_{A\cap\{T=n\}} X_n] \\
&= \sum_{n\in\mathbb{Z}_+\cup\{\infty\}} \mathbb{E}[\mathbf{1}_{A\cap\{T=n\}} X_\infty] \\
&= \mathbb{E}[\mathbf{1}_A X_\infty].
\end{aligned}$$

In the first equality, we use the fact that $X_T \in L^1$ to justify the interchange of sum and expectation via the Fubini theorem. In the third equality, we use the properties $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$ and $A \cap \{T = n\} \in \mathcal{F}_n$ for $A \in \mathcal{F}_T$. Since X_T is $\mathcal{F}_T$-measurable and in L^1, the preceding display shows that X_T satisfies the characteristic property of the conditional expectation $\mathbb{E}[X_\infty \,|\, \mathcal{F}_T]$.

Finally, if S and T are two stopping times such that $S \le T$, the property $\mathcal{F}_S \subset \mathcal{F}_T$ (Proposition 12.9) and Proposition 11.7 imply that

$$X_S = \mathbb{E}[X_\infty \,|\, \mathcal{F}_S] = \mathbb{E}[\mathbb{E}[X_\infty \,|\, \mathcal{F}_T] \,|\, \mathcal{F}_S] = \mathbb{E}[X_T \,|\, \mathcal{F}_S].$$

□

Examples

(a) *Gambler's ruin*. Consider a simple random walk $(X_n)_{n\in\mathbb{Z}_+}$ on $\mathbb{Z}$ (coin-tossing process) with $X_0 = k \geq 0$. Let $m \geq 1$ be an integer such that $0 \leq k \leq m$. We set

$$T = \inf\{n \geq 0 : X_n = 0 \text{ or } X_n = m\}.$$

We know that $T < \infty$ a.s. (for instance by Proposition 10.7). From Theorem 12.12, the process $Y_n = X_{n\wedge T}$ is a martingale, which is uniformly integrable because it is bounded. By Theorem 12.28, we have thus $\mathbb{E}[Y_\infty] = \mathbb{E}[Y_0] = k$, or equivalently

$$m\,\mathbb{P}(X_T = m) = k.$$

It follows that

$$\mathbb{P}(X_T = m) = \frac{k}{m} \quad , \quad \mathbb{P}(X_T = 0) = 1 - \frac{k}{m}.$$

This can be generalized to the "biased" coin-tossing $X_n = k + Y_1 + \ldots + Y_n$, where the random variables Y_i, $i \in \mathbb{N}$, are independent with the same distribution

$$\mathbb{P}(Y_1 = 1) = p \quad , \quad \mathbb{P}(Y_1 = -1) = 1 - p.$$

Let us assume that $p \in (0, 1)\backslash\{\frac{1}{2}\}$. It is then straightforward to verify that

$$Z_n = \Big(\frac{1-p}{p}\Big)^{X_n}$$

is a martingale. Let the stopping time T be defined as above. Since $\mathbb{E}[Y_1] = 2p - 1 \neq 0$, the strong law of large numbers implies that $T < \infty$ a.s. Then, by applying Theorem 12.28 to the bounded martingale $Z_{n\wedge T}$, we get

$$\Big(\frac{1-p}{p}\Big)^k = \mathbb{E}\Big[\Big(\frac{1-p}{p}\Big)^{X_T}\Big] = \Big(\frac{1-p}{p}\Big)^m \mathbb{P}(X_T = m) + \mathbb{P}(X_T = 0)$$

and, consequently,

$$\mathbb{P}(X_T = m) = \frac{(\frac{1-p}{p})^k - 1}{(\frac{1-p}{p})^m - 1} \quad , \quad \mathbb{P}(X_T = 0) = \frac{(\frac{1-p}{p})^m - (\frac{1-p}{p})^k}{(\frac{1-p}{p})^m - 1}.$$

(b) *Law of hitting times*. Consider again a simple random walk $(X_n)_{n\in\mathbb{Z}_+}$ on $\mathbb{Z}$, now with $X_0 = 0$, and write $X_n = U_1 + \cdots + U_n$ for $n \in \mathbb{N}$. Our goal is to compute the law of

$$T = \inf\{n \geq 0 : X_n = -1\}.$$

To this end, we will compute the generating function $\mathbb{E}[r^T]$ for $r \in (0, 1)$. We fix $r \in (0, 1)$ and try to find $\rho > 0$ such that the process Z_n defined by $Z_n = r^n \rho^{X_n}$ is a martingale with respect to the canonical filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ of the process $(X_n)_{n\in\mathbb{Z}_+}$. We observe that

$$\mathbb{E}[Z_{n+1} \,|\, \mathcal{F}_n] = r^{n+1}\, \mathbb{E}[\rho^{X_n+U_{n+1}} \,|\, \mathcal{F}_n] = r^{n+1}\, \rho^{X_n}\, \mathbb{E}[\rho^{U_{n+1}}] = r(\frac{\rho}{2} + \frac{1}{2\rho})\, Z_n.$$

Hence, $(Z_n)_{n\geq 0}$ is a martingale if $r(\frac{\rho}{2} + \frac{1}{2\rho}) = 1$ or equivalently if $\rho^2 r - 2\rho + r = 0$. This condition holds if $\rho = \rho_1$ or $\rho = \rho_2$, where

$$\rho_1 = \frac{1 - \sqrt{1 - r^2}}{r}\ ,\ \ \rho_2 = \frac{1 + \sqrt{1 - r^2}}{r}.$$

We note that $0 < \rho_1 < 1 < \rho_2$. Consider then the nonnegative martingale $Z_n = r^n \rho_1^{X_n}$, and note that $Z_n \leq 1/\rho_1$ when $0 \leq n \leq T$ (because $X_n \geq -1$ when $0 \leq n \leq T$). The stopped martingale $(Z_{n\wedge T})_{n\geq 0}$ is thus bounded, hence uniformly integrable. Theorem 12.28 implies that $\mathbb{E}[Z_0] = \mathbb{E}[Z_T]$, and therefore

$$1 = \mathbb{E}[r^T (\rho_1)^{-1}],$$

which gives

$$\mathbb{E}[r^T] = \rho_1 = \frac{1 - \sqrt{1 - r^2}}{r}.$$

We have computed the generating function of T, and a Taylor expansion gives $\mathbb{P}(T = n) = 0$ if n is even (this is obvious for parity reasons) and, for every $n \in \mathbb{N}$,

$$\mathbb{P}(T = 2n - 1) = \frac{1 \times 3 \times \cdots \times (2n - 3)}{2^n n!}.$$

It is instructive to try reproducing the preceding argument with the martingale Z_n replaced by $Z'_n = r^n \rho_2^{X_n}$, which is also a martingale. If we apply Theorem 12.28 to this martingale without a proper justification and write $\mathbb{E}[Z'_0] = \mathbb{E}[Z'_T]$, we arrive at an absurd result: this is due to the fact that the martingale $Z'_{n\wedge T}$ is *not* uniformly integrable. When applying Theorem 12.28, it is crucial to **verify the uniform integrability assumption!**

We now state the optional stopping theorem for a supermartingale. Recall that, if a supermartingale $(X_n)_{n\in\mathbb{Z}_+}$ is bounded in L^1, the almost sure limit X_∞ exists (Theorem 12.15) and thus we can make sense of X_T even on the event $\{T=\infty\}$, for any stopping time T.

Theorem 12.29 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a supermartingale. Assume that at least one of the following two conditions holds:*

(i) $X_n \geq 0$ *for every* $n \in \mathbb{Z}_+$.
(ii) *The sequence* $(X_n)_{n\in\mathbb{Z}_+}$ *is uniformly integrable.*

Then we have $X_T \in L^1$ for every stopping time T. Moreover, if S and T are two stopping times such that $S \leq T$, we have

$$X_S \geq \mathbb{E}[X_T \,|\, \mathcal{F}_S].$$

Proof We first note that either (i) or (ii) implies that $(X_n)_{n\in\mathbb{Z}_+}$ is bounded in L^1. Theorem 12.15 then shows that X_n converges a.s. to X_∞, so that the definition of X_T makes sense for any stopping time, even on the event $\{T=\infty\}$.

Consider first case (i), where $X_n \geq 0$ for every n. If T is a bounded stopping time, we know that $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$ (Theorem 12.12). Fatou's lemma then shows that, for any stopping time T,

$$\mathbb{E}[X_T] \leq \liminf_{k\to\infty} \mathbb{E}[X_{T\wedge k}] \leq \mathbb{E}[X_0]$$

and thus $X_T \in L^1$.

Then let S and T be two stopping times such that $S \leq T$. Let us first assume that $T \leq N$ for some fixed integer N. According to Lemma 12.19, we have then $\mathbb{E}[X_S] \geq \mathbb{E}[X_T]$. More generally, let $A \in \mathcal{F}_S$, and set

$$S^A(\omega) = \begin{cases} S(\omega) & \text{if } \omega \in A, \\ N & \text{if } \omega \notin A. \end{cases}$$

It is straightforward to verify that S^A is a stopping time (just note that $\{S^A \leq n\} = \Omega$ if $n \geq N$, and $\{S^A \leq n\} = A \cap \{S \leq n\} \in \mathcal{F}_n$ if $n < N$, using the definition of $\mathcal{F}_S$). We define T^A in a similar manner, and note that T^A is also a stopping time since $A \in \mathcal{F}_S \subset \mathcal{F}_T$. Since $S^A \leq T^A$, we have $\mathbb{E}[X_{S^A}] \geq \mathbb{E}[X_{T^A}]$, and it follows that

$$\mathbb{E}[X_S \mathbf{1}_A] \geq \mathbb{E}[X_T \mathbf{1}_A].$$

Let us come back to the general case where S and T satisfy $S \leq T$ but are no longer assumed to be bounded. Let $A \in \mathcal{F}_S$, and $k \in \mathbb{N}$. Then $A \cap \{S \leq k\}$ belongs both to $\mathcal{F}_k$ (by the definition of $\mathcal{F}_S$) and to $\mathcal{F}_S$ (recall that S is $\mathcal{F}_S$-measurable). From Proposition 12.10, $A \cap \{S \le k\}$ belongs to $\mathcal{F}_{S\wedge k}$. Hence, by applying the bounded case to the stopping times $S \wedge k$ and $T \wedge k$, we get

$$\mathbb{E}[X_S \mathbf{1}_{A\cap\{S\le k\}}] = \mathbb{E}[X_{S\wedge k}\mathbf{1}_{A\cap\{S\le k\}}] \ge \mathbb{E}[X_{T\wedge k}\mathbf{1}_{A\cap\{S\le k\}}].$$

By monotone convergence,

$$\lim_{k\to\infty} \uparrow \mathbb{E}[X_S\mathbf{1}_{A\cap\{S\le k\}}]] = \mathbb{E}[X_S\mathbf{1}_{A\cap\{S<\infty\}}].$$

On the other hand, Fatou's lemma gives

$$\liminf_{k\to\infty} \mathbb{E}[X_{T\wedge k}\mathbf{1}_{A\cap\{S\le k\}}] \ge \mathbb{E}[X_T\mathbf{1}_{A\cap\{S<\infty\}}].$$

So we have obtained that

$$\mathbb{E}[X_S\mathbf{1}_{A\cap\{S<\infty\}}] \ge \mathbb{E}[X_T\mathbf{1}_{A\cap\{S<\infty\}}].$$

Since $X_S = X_T = X_\infty$ on the event $\{S = \infty\}$, it follows that

$$\mathbb{E}[X_S\mathbf{1}_A] \ge \mathbb{E}[X_T\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X_T \,|\, \mathcal{F}_S]\mathbf{1}_A].$$

Since this holds for every $A \in \mathcal{F}_S$, and both X_S and $\mathbb{E}[X_T \mid \mathcal{F}_S]$ are $\mathcal{F}_S$-measurable, it follows that $X_S \ge \mathbb{E}[X_T \,|\, \mathcal{F}_S]$ a.s. (apply the last display with the set $A = \{X_S < \mathbb{E}[X_T \,|\, \mathcal{F}_S]\}$). This completes the proof of case (i) of the theorem.

Consider then case (ii). In that case, Theorem 12.27 shows that X_n converges to X_∞ both a.s. and in L^1. Since the conditional expectation is a contraction of L^1, the L^1-convergence allows us to pass to the limit $m \to \infty$ in the inequality $X_n \ge \mathbb{E}[X_{n+m} \,|\, \mathcal{F}_n]$, and to obtain that, for every $n \in \mathbb{Z}_+$,

$$X_n \ge \mathbb{E}[X_\infty \,|\, \mathcal{F}_n].$$

Consider then the closed martingale $Z_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$, and set $Y_n = X_n - Z_n$. Then $(Y_n)_{n\in\mathbb{Z}_+}$ is a nonnegative supermartingale. By case (i) and Theorem 12.28, we get first that $X_T = Y_T + Z_T$ is in L^1, and then that

$$Y_S \ge \mathbb{E}[Y_T \,|\, \mathcal{F}_S],$$
$$Z_S = \mathbb{E}[Z_T \,|\, \mathcal{F}_S].$$

It readily follows that $X_S \ge \mathbb{E}[X_T \,|\, \mathcal{F}_S]$. □

12.7 Backward Martingales

A backward filtration is a sequence $(\mathcal{F}_n)_{n\in\mathbb{Z}_-}$ of sub-σ-fields of $\mathcal{A}$, which is indexed by the set $\mathbb{Z}_- = \{0, -1, -2, \dots\}$ of all *nonpositive* integers, and is such that $\mathcal{F}_n \subset \mathcal{F}_m$ whenever $n, m \in \mathbb{Z}_-$ and $n \le m$. We then set

$$\mathcal{F}_{-\infty} = \bigcap_{n\in\mathbb{Z}_-} \mathcal{F}_n$$

which is also a sub-σ-field of $\mathcal{A}$. It is important to observe that, in contrast with the "forward case" studied in the previous sections, the σ-field $\mathcal{F}_n$ becomes smaller and smaller when $n \downarrow -\infty$ (the larger $|n|$ is, the smaller the σ-field $\mathcal{F}_n$).

In what follows, we fix a backward filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_-}$. Let $X = (X_n)_{n\in\mathbb{Z}_-}$ be a sequence of real random variables indexed by $\mathbb{Z}_-$. We say that X is a backward martingale (resp. a backward supermartingale, a backward submartingale) if X_n is $\mathcal{F}_n$-measurable and $\mathbb{E}[|X_n|] < \infty$ for every $n \in \mathbb{Z}_-$, and if, for every $m, n \in \mathbb{Z}_-$ such that $n \le m$,

$$X_n = \mathbb{E}[X_m \,|\, \mathcal{F}_n] \quad (\text{resp. } X_n \ge \mathbb{E}[X_m \,|\, \mathcal{F}_n],\ X_n \le \mathbb{E}[X_m \,|\, \mathcal{F}_n]).$$

Theorem 12.30 *Let $(X_n)_{n\in\mathbb{Z}_-}$ be a backward supermartingale. Suppose that*

$$\sup_{n\in\mathbb{Z}_-} \mathbb{E}[|X_n|] < \infty. \tag{12.7}$$

Then the sequence $(X_n)_{n\in\mathbb{Z}_-}$ is uniformly integrable and converges a.s. and in L^1 to a random variable X_∞ as $n \to -\infty$. Moreover for every $n \in \mathbb{Z}_-$,

$$\mathbb{E}[X_n \,|\, \mathcal{F}_{-\infty}] \le X_\infty,$$

and equality holds in the last display if $(X_n)_{n\in\mathbb{Z}_-}$ is a backward martingale.

Remarks

(a) In the case of a backward martingale, condition (12.7) holds automatically since we have $X_n = \mathbb{E}[X_0 \mid \mathcal{F}_n]$ and thus $\mathbb{E}[|X_n|] \le \mathbb{E}[|X_0|]$ for every $n \in \mathbb{Z}_-$. By the same argument, the uniform integrability of the sequence $(X_n)_{n\in\mathbb{Z}_-}$ in the case of a martingale follows from Corollary 12.26.

(b) In the "forward" case studied in the previous sections, the fact that a supermartingale (or even a martingale) is bounded in L^1 does not imply its uniform integrability. In this sense, the backward case is very different from the forward case.

Proof We start by establishing the a.s. convergence of the sequence $(X_n)_{n\in\mathbb{Z}_-}$. This is again an application of Doob's upcrossing inequality. Let us fix an integer $K \ge 1$, and set, for every $n \in \{0, 1, \dots, K\}$

$$Y_n^K = X_{-K+n},$$
$$\mathcal{G}_n^K = \mathcal{F}_{-K+n}.$$

For $n > K$, set $Y_n^K = X_0$ and $\mathcal{G}_n^K = \mathcal{F}_0$. Then $(Y_n^K)_{n\in\mathbb{Z}_+}$ is a supermartingale with respect to the (forward) filtration $(\mathcal{G}_n^K)_{n\in\mathbb{Z}_+}$. By applying Lemma 12.14 to the submartingale $-Y_n^K$, we get, for every $a < b$,

$$(b-a)\,\mathbb{E}[N_K([a,b], -Y^K)] \le \mathbb{E}[(-Y_K^K - a)^+] = \mathbb{E}[(-X_0 - a)^+] \le |a| + \mathbb{E}[|X_0|].$$

When $K \uparrow \infty$, $N_K([a,b], -Y^K)$ increases to

$$N([a,b], -X) := \sup\{k \in \mathbb{Z}_+ : \exists m_1 < n_1 < \cdots < m_k < n_k \le 0,$$
$$- X_{m_1} \le a, -X_{n_1} \ge b, \dots, -X_{m_k} \le a, -X_{n_k} \ge b\}$$

which is the total upcrossing number of $(-X_n)_{n\in\mathbb{Z}_-}$ along $[a, b]$. The monotone convergence theorem thus implies that

$$(b-a)\,\mathbb{E}[N([a,b], -X)] \le |a| + \mathbb{E}[|X_0|] < \infty.$$

It follows that we have $N([a,b], -X) < \infty$ for every rationals $a < b$, a.s. By Lemma 12.13, this implies that the sequence $(X_n)_{n\in\mathbb{Z}_-}$ converges in $\overline{\mathbb{R}}$, a.s. as $n \to -\infty$. Furthermore, our assumption (12.7) and Fatou's lemma also show that the limit X_∞ verifies $\mathbb{E}[|X_\infty|] < \infty$.

Let us now show that the sequence $(X_n)_{n\in\mathbb{Z}_-}$ is uniformly integrable. We fix $\varepsilon > 0$ and prove that, if $a > 0$ is large enough, we have $\mathbb{E}[|X_n|\,\mathbf{1}_{\{|X_n|>a\}}] < \varepsilon$ for every $n \in \mathbb{Z}_-$.

Since the sequence $(\mathbb{E}[X_{-n}])_{n\in\mathbb{Z}_+}$ is increasing and bounded (by (12.7)) we may find an integer $K \le 0$ such that, for every $n \le K$,

$$\mathbb{E}[X_n] \le \mathbb{E}[X_K] + \frac{\varepsilon}{2}.$$

By Proposition 12.25 applied to $\{X_K\}$, we can fix $\delta > 0$ small enough so that, for every $A \in \mathcal{A}$ such that $\mathbb{P}(A) < \delta$ we have

$$\mathbb{E}[|X_K|\mathbf{1}_A] < \frac{\varepsilon}{2}.$$

The finite set $\{X_K, X_{K+1}, \dots, X_{-1}, X_0\}$ is uniformly integrable. Thus, we can find $a_0 > 0$ such that, for every $a \ge a_0$, we have

$$\mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>a\}}] < \varepsilon,$$

for every $n \in \{K, K+1, \dots, -1, 0\}$. Thanks to (12.7), we can also assume that, for every $a \geq a_0$,

$$\frac{1}{a}\,\sup\{\mathbb{E}[|X_n|] : n \in \mathbb{Z}_-\} < \delta. \tag{12.8}$$

Then, for every negative integer $n < K$,

$$\begin{aligned}
&\mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>a\}}] \\
&\quad = \mathbb{E}[-X_n\mathbf{1}_{\{X_n<-a\}}] + \mathbb{E}[X_n\mathbf{1}_{\{X_n>a\}}] \\
&\quad = -\mathbb{E}[X_n\mathbf{1}_{\{X_n<-a\}}] + \mathbb{E}[X_n] - \mathbb{E}[X_n\mathbf{1}_{\{X_n\leq a\}}] \\
&\quad \leq -\mathbb{E}[\mathbb{E}[X_K \mid \mathcal{F}_n]\mathbf{1}_{\{X_n<-a\}}] + \mathbb{E}[X_K] + \frac{\varepsilon}{2} - \mathbb{E}[\mathbb{E}[X_K \mid \mathcal{F}_n]\mathbf{1}_{\{X_n\leq a\}}] \\
&\quad = -\mathbb{E}[X_K\mathbf{1}_{\{X_n<-a\}}] + \mathbb{E}[X_K] + \frac{\varepsilon}{2} - \mathbb{E}[X_K\mathbf{1}_{\{X_n\leq a\}}] \\
&\quad = -\mathbb{E}[X_K\mathbf{1}_{\{X_n<-a\}}] + \mathbb{E}[X_K\mathbf{1}_{\{X_n>a\}}] + \frac{\varepsilon}{2} \\
&\quad \leq \mathbb{E}[|X_K|\mathbf{1}_{\{|X_n|>a\}}] + \frac{\varepsilon}{2}.
\end{aligned}$$

In the first inequality of the last display, we use the property $\mathbb{E}[X_n] \leq \mathbb{E}[X_K] + \varepsilon/2$ and the supermartingale inequality $X_n \geq \mathbb{E}[X_K \mid \mathcal{F}_n]$. The next equality uses the characteristic property of $\mathbb{E}[X_K \mid \mathcal{F}_n]$. We then observe that, for every $n < K$ and every $a \geq a_0$,

$$\mathbb{P}(|X_n| > a) \leq \frac{1}{a}\,\mathbb{E}[|X_n|] < \delta,$$

by (12.8). It follows from our choice of δ that we have

$$\mathbb{E}[|X_K|\mathbf{1}_{\{|X_n|>a\}}] < \frac{\varepsilon}{2},$$

for every $n < K$ and $a \geq a_0$. By combining this with our preceding bound on $\mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>a\}}]$, we obtain that, for every $a \geq a_0$ and $n < K$,

$$\mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>a\}}] < \varepsilon,$$

Since this bound also holds for $n \in \{K, K+1, \dots, -1, 0\}$, this completes the proof of the uniform integrability of the sequence $(X_n)_{n\in\mathbb{Z}_-}$.

The remaining part of the proof is easy. By Theorem 12.27, we get that the sequence $(X_n)_{n\in\mathbb{Z}_-}$ converges in L^1. Then, writing

$$\mathbb{E}[X_n\mathbf{1}_A] \leq \mathbb{E}[X_m\mathbf{1}_A]$$

for $m \leq n \leq 0$ and $A \in \mathcal{F}_{-\infty} \subset \mathcal{F}_m$, and passing to the limit $m \to -\infty$, we get

$$\mathbb{E}[X_n \mathbf{1}_A] \leq \mathbb{E}[X_\infty \mathbf{1}_A] , \quad \forall A \in \mathcal{F}_{-\infty}.$$

We have thus

$$\mathbb{E}[\mathbb{E}[X_n \,|\, \mathcal{F}_{-\infty}]\mathbf{1}_A] \leq \mathbb{E}[X_\infty \mathbf{1}_A] , \quad \forall A \in \mathcal{F}_{-\infty}.$$

and since X_∞ is $\mathcal{F}_{-\infty}$-measurable, this suffices to obtain $\mathbb{E}[X_n \mid \mathcal{F}_{-\infty}] \leq X_\infty$ (the same argument was used at the end of the proof of Theorem 12.29 (i)). If X is a martingale, the same inequality also holds for $-X$ and this gives the equality $\mathbb{E}[X_n \,|\, \mathcal{F}_{-\infty}] = X_\infty$. □

Corollary 12.31 *Let Z be a random variable in L^1, and let $(\mathcal{G}_n)_{n \in \mathbb{Z}_+}$ be a decreasing sequence of sub-σ-fields. Then,*

$$\mathbb{E}[Z \,|\, \mathcal{G}_n] \xrightarrow[n \to \infty]{a.s., L^1} \mathbb{E}[Z \,|\, \mathcal{G}_\infty]$$

where

$$\mathcal{G}_\infty = \bigcap_{n \in \mathbb{Z}_+} \mathcal{G}_n.$$

Proof For every $n \in \mathbb{Z}_-$, set $X_n = \mathbb{E}[Z \mid \mathcal{G}_{-n}]$ and $\mathcal{F}_n = \mathcal{G}_{-n}$. Thanks to Proposition 11.7, $(X_n)_{n \in \mathbb{Z}_-}$ is a martingale with respect to the backward filtration $(\mathcal{F}_n)_{n \in \mathbb{Z}_-}$. Theorem 12.30 thus ensures that X_n converges a.s. and in L^1 as $n \to -\infty$. Moreover, by the last assertion of Theorem 12.30,

$$X_\infty = \mathbb{E}[X_0 \,|\, \mathcal{F}_{-\infty}] = \mathbb{E}[\mathbb{E}[Z \,|\, \mathcal{F}_0] \,|\, \mathcal{F}_{-\infty}] = \mathbb{E}[Z \,|\, \mathcal{F}_{-\infty}].$$

□

Applications (A) *The strong law of large numbers*. Let us give a first application to an alternative proof of the strong law of large numbers (Theorem 10.8). We consider a sequence $(\xi_n)_{n \in \mathbb{N}}$ of independent and identically distributed random variables in L^1. We set $S_0 = 0$, and for every $n \geq 1$,

$$S_n = \xi_1 + \cdots + \xi_n.$$

We first observe that

$$\mathbb{E}[\xi_1 \,|\, S_n] = \frac{1}{n}\, S_n. \tag{12.9}$$

Let us justify this equality. We know that there exists a real measurable function g such that $\mathbb{E}[\xi_1 \,|\, S_n] = g(S_n)$. Let $k \in \{1, \dots, n\}$. Then the pair (ξ_k, S_n) has the same distribution as (ξ_1, S_n), so that, for every bounded measurable function h on $\mathbb{R}$,

$$\mathbb{E}[\xi_k h(S_n)] = \mathbb{E}[\xi_1 h(S_n)] = \mathbb{E}[g(S_n)h(S_n)]$$

which shows that we have also $\mathbb{E}[\xi_k \,|\, S_n] = g(S_n)$. It follows that

$$ng(S_n) = \mathbb{E}[\xi_1 + \cdots + \xi_n \,|\, S_n] = S_n$$

giving the desired formula (12.9).

For every $n \in \mathbb{N}$, set

$$\mathcal{G}_n = \sigma(S_n, \xi_{n+1}, \xi_{n+2}, \dots) = \sigma(S_n) \vee \sigma(\xi_{n+1}, \xi_{n+2}, \dots).$$

As a consequence of Proposition 11.10 applied with $Z = \xi_1$, $\mathcal{H}_1 = \sigma(S_n)$ and $\mathcal{H}_2 = \sigma(\xi_{n+1}, \xi_{n+2}, \dots)$, we have

$$\mathbb{E}[\xi_1 \,|\, \mathcal{G}_n] = \mathbb{E}[\xi_1 \,|\, S_n] = \frac{1}{n}\, S_n.$$

Note that the sequence $(\mathcal{G}_n)_{n\in\mathbb{N}}$ is decreasing (because $S_{n+1} = S_n + \xi_{n+1}$). We can thus apply Corollary 12.31 to get that $\frac{1}{n}S_n$ converges a.s. and in L^1. As we already noticed after the Kolmogorov zero-one law (Theorem 10.6), the limit must be constant and therefore equal also to $\lim \frac{1}{n}\mathbb{E}[S_n] = \mathbb{E}[\xi_1]$. In this way, we recover the strong law of large numbers, with the L^1-convergence that had not been established in the proof of Theorem 10.8. Note that the argument relies on Corollary 12.31, which uses only the (easier) martingale case of Theorem 12.30.

(B) *The Hewitt-Savage zero-one law*. Let $\xi_1, \xi_2, \dots$ be a sequence of independent and identically distributed random variables with values in a measurable space $(E, \mathcal{E})$. The mapping $\omega \mapsto (\xi_1(\omega), \xi_2(\omega), \dots)$ defines a random variable taking values in the product space $E^{\mathbb{N}}$, which is equipped with the smallest σ-field for which the coordinate mappings $(x_1, x_2, \dots) \longrightarrow x_i$ are measurable, for every $i \in \mathbb{N}$. Equivalently, this σ-field is generated by the “cylinder sets” of the form

$$\{(x_1, x_2, \dots) \in E^{\mathbb{N}} : x_1 \in B_1, x_2 \in B_2, \dots, x_k \in B_k\}$$

where $k \in \mathbb{N}$ and $B_1, \dots, B_k$ are measurable subsets of E.

A measurable function $F : E^{\mathbb{N}} \longrightarrow \mathbb{R}$ is said to be symmetric if

$$F(x_1, x_2, x_3, \dots) = F(x_{\pi(1)}, x_{\pi(2)}, x_{\pi(3)}, \dots)$$

for every $(x_1, x_2, \dots) \in E^{\mathbb{N}}$ and every permutation π of $\mathbb{N}$ with finite support (meaning that there exists an integer N such that $\pi(i) = i$ for every $i \geq N$).

Theorem 12.32 *Let F be a symmetric function on $E^{\mathbb{N}}$. Then, the random variable $F(\xi_1, \xi_2, \ldots)$ is almost surely constant.*

Example Suppose that the random variables $\xi_1, \xi_2, \ldots$ take values in $\mathbb{R}^d$, and consider the (d-dimensional) random walk

$$X_n = \xi_1 + \cdots + \xi_n.$$

If B is a Borel subset of $\mathbb{R}^d$,

$$\mathbf{1}_{\{\mathrm{card}\{n \geq 1 : X_n \in B\} = \infty\}}$$

is a symmetric function of $\xi_1, \xi_2, \ldots$. We have thus

$$\mathbb{P}(\mathrm{card}\{n \geq 1 : X_n \in B\} = \infty) = 0 \text{ or } 1.$$

Proof Without loss of generality, we may assume that F is bounded. Let us introduce the σ-fields

$$\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)\ ,\ \mathcal{F}_\infty = \bigvee_{n \in \mathbb{N}} \mathcal{F}_n = \sigma(\xi_1, \xi_2, \ldots),$$

and

$$\mathcal{G}_n = \sigma(\xi_{n+1}, \xi_{n+2}, \ldots)\ ,\ \mathcal{G}_\infty = \bigcap_{n \in \mathbb{N}} \mathcal{G}_n.$$

Also set $Y = F(\xi_1, \xi_2, \ldots)$ and, for every $n \in \mathbb{N}$,

$$X_n = \mathbb{E}[Y \mid \mathcal{F}_n]\ ,\ Z_n = \mathbb{E}[Y \mid \mathcal{G}_n].$$

Then Corollary 12.18 ensures that X_n converges to $\mathbb{E}[Y \mid \mathcal{F}_\infty] = Y$, a.s. and in L^1, whereas Corollary 12.31 shows that Z_n converge to $\mathbb{E}[Y \mid \mathcal{G}_\infty]$ a.s. and in L^1. Note that we must have $\mathbb{E}[Y \mid \mathcal{G}_\infty] = \mathbb{E}[Y]$ because the σ-field $\mathcal{G}_\infty$ contains only events of probability zero or one (Theorem 10.6).

Let us fix $\varepsilon > 0$. By the preceding considerations, we can fix a sufficiently large value of n such that

$$\mathbb{E}[|X_n - Y|] < \varepsilon\ ,\ \mathbb{E}[|Z_n - \mathbb{E}[Y]|] < \varepsilon. \tag{12.10}$$

On the other hand, by Proposition 8.9, there exists a measurable function $g_n : E^n \longrightarrow \mathbb{R}$ such that $X_n = g_n(\xi_1, \ldots, \xi_n)$, and the first bound in (12.10) can be rewritten as

$$\mathbb{E}[|F(\xi_1, \xi_2, \ldots) - g_n(\xi_1, \ldots, \xi_n)|] < \varepsilon.$$

Now note that the sequence $(\xi_{n+1}, \ldots, \xi_{2n}, \xi_1, \ldots, \xi_n, \xi_{2n+1}, \ldots)$ viewed as a random variable with values in $E^{\mathbb{N}}$ has the same law as $(\xi_1, \xi_2 \ldots)$ (just check that both laws coincide on cylinder sets, which form a class closed under finite intersections and generating the σ-field on $E^{\mathbb{N}}$). So the preceding bound also implies that

$$\mathbb{E}[|F(\xi_{n+1}, \ldots, \xi_{2n}, \xi_1, \ldots, \xi_n, \xi_{2n+1}, \ldots) - g_n(\xi_{n+1}, \ldots, \xi_{2n})|] < \varepsilon.$$

However,

$$F(\xi_{n+1}, \ldots, \xi_{2n}, \xi_1, \ldots, \xi_n, \xi_{2n+1}, \ldots) = F(\xi_1, \xi_2, \ldots) = Y$$

because F is symmetric, and we have obtained

$$\mathbb{E}[|Y - g_n(\xi_{n+1}, \ldots, \xi_{2n})|] < \varepsilon. \tag{12.11}$$

By conditioning on the σ-field $\mathcal{G}_n$, we have

$$\mathbb{E}[|\mathbb{E}[Y \mid \mathcal{G}_n] - \mathbb{E}[g_n(\xi_{n+1}, \ldots, \xi_{2n}) \mid \mathcal{G}_n]|] < \varepsilon,$$

and thus

$$\mathbb{E}[|Z_n - g_n(\xi_{n+1}, \ldots, \xi_{2n})|] < \varepsilon. \tag{12.12}$$

By combining (12.11) and (12.12) with the second bound in (12.10), we get

$$\mathbb{E}[|Y - \mathbb{E}[Y]|] < 3\varepsilon.$$

Since ε was arbitrary, we have proved that $Y = \mathbb{E}[Y]$ a.s. □

12.8 Exercises

Exercise 12.1 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent real random variables. Set $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$ for every $n \in \mathbb{N}$, and consider the canonical filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ of the process $(S_n)_{n\in\mathbb{Z}_+}$. Prove that:

(1) If $X_n \in L^1$ for every n, $\widetilde{S}_n = S_n - \mathbb{E}[S_n]$ is a martingale.
(2) If $X_n \in L^2$ for every n, $(\widetilde{S}_n)^2 - \mathbb{E}[(\widetilde{S}_n)^2]$ is a martingale.
(3) If, for some $\theta \in \mathbb{R}$, $\mathbb{E}[e^{\theta X_n}] < \infty$ for every $n \in \mathbb{N}$, then $e^{\theta S_n}/\mathbb{E}[e^{\theta S_n}]$ is a martingale.

Exercise 12.2 Let T be a stopping time.

(1) Prove that, for every $n \in \mathbb{Z}_+$ and every $A \in \mathcal{F}_n$, the set $A \cap \{T \geq n\}$ belongs to $\mathcal{F}_T$.
(2) For $B \in \mathcal{F}_\infty$, let $T^B : \Omega \longrightarrow [0, \infty]$ be defined by $T^B(\omega) = T(\omega)$ if $\omega \in B$, and $T^B(\omega) = \infty$ if $\omega \in B^c$. Prove that T^B is a stopping time if and only if $B \in \mathcal{F}_T$.

Exercise 12.3 Let T be a stopping time. Assume that there exists $\varepsilon \in (0, 1)$ and an integer $N \geq 1$ such that, for every $n \geq 0$,

$$\mathbb{P}(T \leq n + N \,|\, \mathcal{F}_n) \geq \varepsilon\,, \qquad \text{a.s.}$$

Prove that $T < \infty$ a.s. and $\mathbb{E}[T] < \infty$.

Exercise 12.4 Let $(S_n)_{n\in\mathbb{Z}_+}$ be a simple random walk on $\mathbb{Z}$, with $S_0 = k \in \mathbb{Z}$, and consider a function $\varphi : \mathbb{Z} \times \mathbb{Z}_+ \longrightarrow \mathbb{R}$. Prove that $\varphi(S_n, n)$ is a martingale if φ satisfies the functional relation

$$\varphi(s + 1, n + 1) + \varphi(s - 1, n + 1) = 2\,\varphi(s, n).$$

Infer that $S_n^2 - n$ and $S_n^3 - 3nS_n$ are martingales.

Exercise 12.5 Let $(X_n)_{n\in\mathbb{Z}_+}$ be an adapted random process such that $X_n \in L^1$ for every $n \in \mathbb{Z}_+$. Prove that this process is a martingale if and only if the property $\mathbb{E}[X_T] = \mathbb{E}[X_0]$ holds for every bounded stopping time T.

Exercise 12.6 Let $(X_n)_{n\in\mathbb{Z}_+}$ be a martingale, and let T be a stopping time such that

$$\mathbb{P}(T < \infty) = 1\,, \quad \mathbb{E}[|X_T|] < \infty\,, \quad \mathbb{E}[|X_T|\,\mathbf{1}_{\{T>n\}}] \underset{n\to\infty}{\longrightarrow} 0.$$

(1) Prove that $\mathbb{E}[|X_T - X_{T\wedge n}|] \underset{n\to\infty}{\longrightarrow} 0$.
(2) Conclude that $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

Exercise 12.7 Let $(X_n)_{n\in\mathbb{Z}_+}$ be a martingale with $X_0 = 0$. Assume that there exists a constant $M > 0$ such that $|X_{n+1} - X_n| \leq M$ for every $n \in \mathbb{Z}_+$.

(1) For $C > 0$ and $K > 0$, set $T_{C,K} = \inf\{n \geq 0 : X_n \geq K \text{ or } X_n \leq -C\}$. Prove that

$$\lim_{C\to+\infty} \mathbb{P}(T_{C,K} < \infty,\ X_{T_{C,K}} \leq -C) = 0.$$

(2) Prove that, $\mathbb{P}(d\omega)$ almost surely, (exactly) one of the following two properties holds:

- $X_n(\omega)$ has a finite limit as $n \to \infty$;
- $\sup_{n\geq 0} X_n(\omega) = +\infty$ and $\inf_{n\geq 0} X_n(\omega) = -\infty$.

Exercise 12.8 (Wald's Identity) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables in L^1. Set $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$ for every $n \geq 1$, and let $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ be the canonical filtration of $(S_n)_{n\in\mathbb{Z}_+}$.

(1) Let T be a stopping time such that $\mathbb{E}[T] < \infty$. Prove that the random process $M_n = S_{n\wedge T} - (n \wedge T)\mathbb{E}[X_1]$ is a uniformly integrable martingale.
(2) Prove that $S_T \in L^1$ and $\mathbb{E}[S_T] = \mathbb{E}[T]\,\mathbb{E}[X_1]$.

Exercise 12.9 (Another Proof of the Strong Law of Large Numbers) We consider a sequence $(X_n)_{n\in\mathbb{N}}$ of independent and identically distributed random variables in L^1, and assume that $\mathbb{E}[X_1] > 0$. For every $n \geq 0$, we set $S_n = X_1 + \cdots + X_n$ ($S_0 = 0$) and $I_n = \min\{S_k : 0 \leq k \leq n\}$. Finally, we set $T = \inf\{n \geq 0 : S_n > 0\} \leq +\infty$.

(1) Let $n \geq 0$. Verify that the two random vectors $(S_0, S_1, S_2, \ldots, S_n)$ and $(S_n - S_n, S_n - S_{n-1}, S_n - S_{n-2}, \ldots, S_n - S_0)$ have the same law and use this to obtain that $\mathbb{P}(T > n) = \mathbb{P}(S_n = I_n)$, and

$$\mathbb{E}[T] = \mathbb{E}\Big[\sum_{n=0}^{\infty} \mathbf{1}_{\{S_n = I_n\}}\Big].$$

(2) Verify that, for every $n \geq 0$, $\mathbb{E}[S_{n\wedge T}] = \mathbb{E}[X_1]\,\mathbb{E}[n \wedge T]$.
(3) In this question only, we assume that there is a constant $C > 0$ such that $X_1 \leq C$ a.s. Deduce from the preceding question that $\mathbb{E}[n \wedge T] \leq C/\mathbb{E}[X_1]$, for every integer $n \geq 0$. Using question (1), verify that

$$\sum_{n=0}^{\infty} \mathbf{1}_{\{S_n = I_n\}} < \infty, \qquad \text{a.s.}$$

and conclude that $\inf_{n\geq 0} S_n > -\infty$, a.s.
(4) Show that the conclusion of question (3) remains valid without the assumption that $X_1 \leq C$ a.s. (*Hint:* Choose $C > 0$ so that $\mathbb{E}[X_1\,\mathbf{1}_{\{X_1 \leq C\}}] > 0$.)
(5) Prove the strong law of large numbers (Theorem 10.8). (*This short proof of the strong law of large numbers is taken from* [5].)

Exercise 12.10 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. For every $n \geq 1$, set $\mathcal{G}_n = \sigma(X_n, X_{n+1}, \ldots)$ and

$$\mathcal{G}_\infty = \bigcap_{n=1}^{\infty} \mathcal{G}_n.$$

Use Corollary 12.18 to give a martingale proof of the fact that $\mathbb{P}(A) = 0$ or 1 for every $A \in \mathcal{G}_\infty$ (Theorem 10.6).

Exercise 12.11 The goal of this exercise is to prove that a Lipschitz function on $[0, 1]$ can be written as the integral of a bounded measurable function. We fix a Lipschitz function $f : [0, 1] \longrightarrow \mathbb{R}$ (there exists $L > 0$ such that $|f(x) - f(y)| \leq L|x - y|$ for every $x, y \in [0, 1]$). We also let X be a random variable with values in $[0, 1)$, which is uniformly distributed over $[0, 1)$. For every integer $n \geq 0$, we set

$$X_n = 2^{-n}\lfloor 2^n X\rfloor \quad \text{and} \quad Z_n = 2^n(f(X_n + 2^{-n}) - f(X_n)),$$

and we let $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$ be the canonical filtration of the process $(X_n)_{n\in\mathbb{Z}_+}$.

(1) Verify that $\sigma(X_0, X_1, \ldots) = \sigma(X)$, and $\mathcal{F}_n = \sigma(X_n)$ for every $n \geq 0$.
(2) Compute $\mathbb{E}[h(X_{n+1}) \,|\, \mathcal{F}_n]$ for any bounded Borel function $h : \mathbb{R} \longrightarrow \mathbb{R}$. Infer that $(Z_n)_{n\in\mathbb{Z}_+}$ is a bounded martingale with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$.
(3) Show that there exists a bounded Borel function $g : [0, 1) \longrightarrow \mathbb{R}$ such that $Z_n \longrightarrow g(X)$ as $n \to \infty$, a.s.
(4) Verify that a.s. for every $n \geq 0$,

$$Z_n = 2^n \int_{X_n}^{X_n+2^{-n}} g(u)\,\mathrm{d}u.$$

(5) Conclude that, for every $x \in [0, 1]$,

$$f(x) = f(0) + \int_0^x g(u)\,\mathrm{d}u.$$

Exercise 12.12 Consider a sequence $(X_n)_{n\in\mathbb{Z}_+}$ of random variables with values in $[0, 1]$, such that $X_0 = a$. For every $n \geq 0$, set $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$. We assume that, for every $n \geq 0$,

$$\mathbb{P}\Big(X_{n+1} = \frac{X_n}{2} \,\Big|\, \mathcal{F}_n\Big) = 1 - X_n\,, \qquad \mathbb{P}\Big(X_{n+1} = \frac{1 + X_n}{2} \,\Big|\, \mathcal{F}_n\Big) = X_n.$$

(1) Prove that $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}_+}$, which converges a.s. to a random variable Z.
(2) Prove that $\mathbb{E}[(X_{n+1} - X_n)^2] = \frac{1}{4}\mathbb{E}[X_n(1 - X_n)]$.
(3) Compute the distribution of Z.

Exercise 12.13 (Polya's Urn) At time 0, an urn contains a white balls and b red balls, where $a, b \in \mathbb{N}$. We draw one ball at random in the urn, and replace it by two balls of the same color to obtain the urn at time 1. We then proceed in the same manner to get the urn at time 2 from the urn at time 1, and so on. Thus, at time $n \geq 0$, the urn contains $a + b + n$ balls. To simplify notation, we set $N = a + b$.

(1) For every $n \geq 0$, let Y_n be the number of white balls in the urn at time n, and $X_n = Y_n/(N+n)$ (which is the proportion of white balls at time n). We consider the filtration $\mathcal{F}_n = \sigma(Y_0, Y_1, \dots, Y_n)$. Show that $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale that converges a.s. to a limiting random variable denoted by U.
(2) Consider the special case where $a = b = 1$. Prove by induction that, for every $n \geq 0$, Y_n is uniformly distributed over $\{1, 2, \dots, n + 1\}$. Give the distribution of U in that case.
(3) We come back to the general case. Fix $k \geq 1$, and, for every $n \geq 0$, set

$$Z_n = \frac{Y_n(Y_n + 1)\cdots(Y_n + k - 1)}{(N + n)(N + n + 1)\cdots(N + n + k - 1)}.$$

Prove that $(Z_n)_{n\in\mathbb{Z}_+}$ is a martingale, and then compute $\mathbb{E}[U^k]$.

Exercise 12.14 (Yet Another Proof of the Strong Law of Large Numbers)

(1) Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables in L^2, such that $\mathbb{E}[Z_n] = 0$ for every n and

$$\sum_{n=1}^{\infty} \frac{\mathrm{var}(Z_n)}{n^2} < \infty.$$

For every $n \in \mathbb{N}$, we set $S_n = \sum_{j=1}^{n} Z_j$ and $M_n = \sum_{j=1}^{n} \frac{Z_j}{j}$.
Prove that M_n converges a.s. as $n \to \infty$, and infer that S_n/n converges a.s. to 0 as $n \to \infty$. *Hint*: Verify that

$$\frac{S_n}{n} = M_n - \frac{1}{n}\sum_{j=1}^{n-1} M_j.$$

(2) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables in L^1. For every $n \in \mathbb{N}$, set

$$Y_n = X_n \mathbf{1}_{\{|X_n|\leq n\}}.$$

Verify that $\mathbb{E}[Y_n] \longrightarrow \mathbb{E}[X_1]$ as $n \to \infty$. Then prove that almost surely there exists an integer $n_0(\omega) \in \mathbb{N}$ such that $X_n = Y_n$ for every $n \geq n_0(\omega)$, and that

$$\sum_{n=1}^{\infty} \frac{\mathrm{var}(Y_n)}{n^2} < \infty.$$

(3) Conclude that $\frac{1}{n}(X_1 + \cdots + X_n) \longrightarrow \mathbb{E}[X_1]$ a.s. as $n \to \infty$.

Exercise 12.15 (Law of the Iterated Logarithm) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent Gaussian $\mathcal{N}(0, 1)$ random variables, and $S_n = X_1 + \ldots + X_n$ for every $n \in \mathbb{N}$.

(1) Prove that, for every $\theta > 0$ and $n \in \mathbb{N}$, we have for every $c > 0$,

$$\mathbb{P}\Big(\max_{1\leq k\leq n} S_k \geq c\Big) \leq e^{-c\theta}\, \mathbb{E}[e^{\theta S_n}]$$

and consequently

$$\mathbb{P}\Big(\max_{1\leq k\leq n} S_k \geq c\Big) \leq \exp(-\frac{c^2}{2n}).$$

(2) For every $x > \mathrm{e}$, set $h(x) = \sqrt{2x \log\log(x)}$. Prove that

$$\limsup_{n\to\infty} \frac{S_n}{h(n)} \leq 1\,, \qquad \text{a.s.}$$

Hint: For $K > 1$ fixed, bound the probabilities

$$\mathbb{P}\Big(\max_{1\leq k\leq K^n} S_k \geq K\, h(K^{n-1})\Big).$$

Exercise 12.16 (Kakutani's Theorem) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent positive random variables, such that $\mathbb{E}[X_n] = 1$ for every n. Set $M_0 = 1$ and, for every $n \in \mathbb{N}$,

$$M_n = \prod_{k=1}^{n} X_k.$$

(1) Prove that $(M_n)_{n\geq 0}$ is a martingale, which converges a.s. to a limit denoted by M_∞.
(2) For every $n \geq 1$, we set $a_n = \mathbb{E}[\sqrt{X_n}] \in (0, 1]$. Verify that the following three conditions are equivalent:

 (a) $\mathbb{E}[M_\infty] = 1$;
 (b) $M_n \longrightarrow M_\infty$ in L^1 as $n \to \infty$;
 (c) $\prod_{k=1}^{\infty} a_k > 0$.

 If these conditions do not hold, prove that $M_\infty = 0$ a.s.
 Hint: Use Scheffé's lemma (Proposition 10.5), and also consider the process

$$N_n = \prod_{k=1}^{n} \frac{\sqrt{X_k}}{a_k}.$$

Exercise 12.17 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent Bernoulli random variables with parameter $1/2$, and let $(\alpha_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers. For every $n \geq 1$, set

$$S_n = \sum_{j=1}^{n} \alpha_j X_j.$$

(1) Prove that the condition $\sum_{j=1}^{\infty} \alpha_j^2 < \infty$ implies that S_n converges a.s. as $n \to \infty$.
(2) Prove that if $\sum_{j=1}^{\infty} \alpha_j^2 = \infty$ then $\sup_{n\in\mathbb{N}} S_n = \infty$ and $\inf_{n\in\mathbb{N}} S_n = -\infty$, a.s. (*Hint:* Use Theorem 10.6 and consider the martingale $(S_n)^2 - \mathbb{E}[(S_n)^2]$).

Chapter 13
Markov Chains

This chapter is devoted to Markov chains with values in a finite or countable state space. In contrast with martingales, whose definition is based on conditional *means*, the definition of a Markov chain involves conditional *distributions*: it is required that the conditional law of X_{n+1} knowing the past of the process up to time n only depends on the value X_n at time n (and the dependence is always the same when n varies). This is the Markov property, which roughly speaking says that the knowledge of the past does not give more information than the present in order to predict the future. A reinforcement of this property, which is a very important tool for applications, is the strong Markov property, which replaces the deterministic time n by a stopping time. Using the strong Markov property, we investigate recurrent states (states at which the process returns infinitely many times) and transient states. The notion of invariant measure makes it possible to define the asymptotic frequency of visits of a recurrent state, via a form of the ergodic theorem. The last section discusses relations between Markov chains and martingales.

13.1 Definitions and First Properties

Throughout this chapter, E is a finite or countable space, which is equipped as usual with the σ-field $\mathcal{P}(E)$ (all functions defined on E are measurable!). A stochastic matrix on E is a collection $(Q(x, y))_{(x,y)\in E\times E}$ of real numbers that satisfies the following two properties:

(i) $0 \le Q(x, y) \le 1$, for every $x, y \in E$;
(ii) for every $x \in E$, $\sum_{y\in E} Q(x, y) = 1$.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_13

This notion is equivalent to the notion of a transition probability from E into E (see Definition 11.14). Indeed, if we set

$$\nu(x, A) = \sum_{y \in A} Q(x, y) \ , \qquad x \in E, \ A \subset E,$$

we get that ν is a transition probability from E into E, and conversely, starting from such a transition probability ν, the formula $Q(x, y) = \nu(x, \{y\})$ defines a stochastic matrix on E.

For every $n \geq 1$, we define $Q_n := (Q)^n$ by analogy with the usual matrix product: we set $Q_1 = Q$, and then, by induction,

$$Q_{n+1}(x, y) = \sum_{z \in E} Q_n(x, z) Q(z, y).$$

It is straightforward to check that Q_n is again a stochastic matrix on E. We also set $Q_0(x, y) = \mathbf{1}_{\{x=y\}}$. Then, for every integers $m, n \geq 0$, we have $Q_{m+n} = Q_m Q_n$, in the sense that

$$Q_{m+n}(x, y) = \sum_{z \in E} Q_m(x, z) Q_n(z, y).$$

Finally, for every integer $n \geq 2$, an easy induction shows that

$$Q_n(x, y) = \sum_{x_1, x_2, \ldots, x_{n-1} \in E} Q(x, x_1) Q(x_1, x_2) \cdots Q(x_{n-2}, x_{n-1}) Q(x_{n-1}, y). \tag{13.1}$$

For every function $f : E \longrightarrow \mathbb{R}_+$ (resp. for every bounded function $f : E \longrightarrow \mathbb{R}$), the notation Qf will stand for the function defined on E by

$$Qf(x) := \sum_{y \in E} Q(x, y) f(y) \ , \qquad x \in E,$$

which takes values in $[0, \infty]$ (resp. in $\mathbb{R}$). If ν is a measure on E, we also set, for every $y \in E$,

$$\nu Q(y) := \sum_{x \in E} \nu(x) Q(x, y).$$

Definition 13.1 Let Q be a stochastic matrix on E, and let $(X_n)_{n \in \mathbb{Z}_+}$ be a random process with values in E defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We say that $(X_n)_{n \in \mathbb{Z}_+}$ is a Markov chain with transition matrix Q if, for every integer $n \geq 0$, the conditional distribution of X_{n+1} knowing $(X_0, X_1, \ldots, X_n)$ is $Q(X_n, \cdot)$. This is equivalent to saying that

$$\mathbb{P}(X_{n+1} = y \,|\, X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = Q(x_n, y),$$

for every $x_0, \ldots, x_n, y \in E$ such that $\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) > 0$.

Let us comment on this definition and derive a few consequences. Saying that the conditional distribution of X_{n+1} knowing $(X_0, X_1, \ldots, X_n)$ is $Q(X_n, \cdot)$ means that, for every $y \in E$,

$$\mathbb{P}(X_{n+1} = y \,|\, X_0, X_1, \ldots, X_n) = Q(X_n, y)$$

(beware that the quantity in the left-hand side is a conditional expectation: we use the notation $\mathbb{P}(A \mid Z)$ for the conditional expectation $\mathbb{E}[\mathbf{1}_A \mid Z]$). The equivalence with the last sentence in the definition comes from the formula for the conditional expectation with respect to a discrete random variable, which we apply here to conditioning with respect to $(X_0, X_1, \ldots, X_n)$ (see Section 11.1).

Using the formula in the last display and the properties of conditioning with respect to nested sub-σ-fields (Proposition 11.7), we also get that, for every subset $\{i_1, \ldots, i_k\}$ of $\{0, 1, \ldots, n-1\}$ we have

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = y \,|\, X_{i_1}, \ldots, X_{i_k}, X_n) \\
&\quad = \mathbb{E}[\mathbb{P}(X_{n+1} = y \,|\, X_0, X_1, \ldots, X_n) \,|\, X_{i_1}, \ldots, X_{i_k}, X_n] \\
&\quad = \mathbb{E}[Q(X_n, y) \,|\, X_{i_1}, \ldots, X_{i_k}, X_n] \\
&\quad = Q(X_n, y).
\end{aligned}$$

In particular,

$$\mathbb{P}(X_{n+1} = y \,|\, X_n) = Q(X_n, y)$$

which is equivalent to saying that, for every $x \in E$ such that $\mathbb{P}(X_n = x) > 0$,

$$\mathbb{P}(X_{n+1} = y \,|\, X_n = x) = Q(x, y).$$

This last equality provides an intuitive explanation of the behavior of the Markov chain. If the chain sits at time n at the point x of E, it will "choose" the point visited at time $n + 1$ according to the probability measure $Q(x, \cdot)$.

Remarks

(i) For a general random process $(X_n)_{n \in \mathbb{Z}_+}$, the conditional distribution of X_{n+1} knowing $(X_0, X_1, \ldots, X_n)$ can be written in the form $\nu((X_0, X_1, \ldots, X_n), \cdot)$ where ν is a transition probability from E^{n+1} into E (cf. Definition 11.16): this conditional distribution depends a priori on all variables $X_0, X_1, \ldots, X_n$ and not only on the last one X_n. The fact that for a Markov chain the conditional distribution of X_{n+1} knowing $(X_0, X_1, \dots, X_n)$ only depends on X_n is the so-called **Markov property**: to predict the future X_{n+1}, the knowledge of the past $(X_0, X_1, \dots, X_n)$ does not give more information than the present X_n. We will later see other forms of the Markov property, which obey the same general principle.

(ii) The function $Q(x, \cdot)$ that gives the conditional law of X_{n+1} knowing that $X_n = x$ does not depend on the integer n. This is the time-homogeneity of the "transition mechanism". One could also consider inhomogeneous Markov chains for which the transition mechanism between times n and $n+1$ depends on n, but in the present course we consider only homogeneous Markov chains.

Proposition 13.2 *A random process $(X_n)_{n\in\mathbb{Z}_+}$ with values in E is a Markov chain with transition matrix Q if and only if, for every $n \geq 0$ and every $x_0, x_1, \dots, x_n \in E$, we have*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_0 = x_0)Q(x_0, x_1)Q(x_1, x_2)\cdots Q(x_{n-1}, x_n). \tag{13.2}$$

Proof If $(X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain with transition matrix Q, formula (13.2) is derived by induction on n, by writing

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n, X_{n+1} = x_{n+1}) = \\ &= \mathbb{P}(X_0 = x_0, \dots, X_n = x_n) \times \mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n),\end{aligned}$$

if $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) > 0$ (if $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = 0$, the desired formula at order $n+1$ is a trivial consequence of the one at order n).

Conversely, assuming that (13.2) holds, one immediately gets that, if $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) > 0$,

$$\begin{aligned}&\mathbb{P}(X_{n+1} = y \mid X_0 = x_0, \dots, X_n = x_n) \\ &\quad = \frac{\mathbb{P}(X_0 = x_0)Q(x_0, x_1)\cdots Q(x_{n-1}, x_n)Q(x_n, y)}{\mathbb{P}(X_0 = x_0)Q(x_0, x_1)\cdots Q(x_{n-1}, x_n)} = Q(x_n, y).\end{aligned}$$

□

Remark Formula (13.2) shows that, for a Markov chain $(X_n)_{n\in\mathbb{Z}_+}$, the law of $(X_0, X_1, \dots, X_n)$ is completely determined by the initial distribution (the law of X_0) and the transition matrix Q.

The next proposition gathers a few other simple properties of Markov chain. In particular, part (i) restates in a slightly different form some properties already stated after the definition.

Proposition 13.3 *Let $(X_n)_{n\in\mathbb{Z}_+}$ be a Markov chain with transition matrix Q.*

(i) For every integer $n \geq 0$ and every measurable function $f : E \longrightarrow \mathbb{R}_+$,

$$\mathbb{E}[f(X_{n+1}) \,|\, X_0, X_1, \ldots, X_n] = \mathbb{E}[f(X_{n+1}) \,|\, X_n] = Qf(X_n).$$

More generally, for every subset $\{i_1, \ldots, i_k\}$ of $\{0, 1, \ldots, n-1\}$, we have

$$\mathbb{E}[f(X_{n+1}) \,|\, X_{i_1}, \ldots, X_{i_k}, X_n] = Qf(X_n).$$

(ii) For every integers $n \geq 0$, $p \geq 1$ and for every $y_1, \ldots, y_p \in E$,

$$\mathbb{P}(X_{n+1} = y_1, \ldots, X_{n+p} = y_p \,|\, X_0, \ldots, X_n)$$
$$= Q(X_n, y_1)Q(y_1, y_2) \ldots Q(y_{p-1}, y_p),$$

and consequently,

$$\mathbb{P}(X_{n+p} = y_p \,|\, X_n) = \mathbb{P}(X_{n+p} = y_p \,|\, X_0, \ldots, X_n) = Q_p(X_n, y_p).$$

If we fix $n \geq 0$ and set $Y_p = X_{n+p}$ for every $p \in \mathbb{Z}_+$, the random process $(Y_p)_{p\in\mathbb{Z}_+}$ is also a Markov chain with transition matrix Q.

Proof

(i) Writing $f(X_{n+1}) = \sum_{y\in E} f(y)\, \mathbf{1}_{\{X_{n+1}=y\}}$, we get from the definition that

$$\begin{aligned}\mathbb{E}[f(X_{n+1}) \,|\, X_0, X_1, \ldots, X_n] &= \sum_{y\in E} f(y)\, \mathbb{P}(X_{n+1} = y \,|\, X_0, X_1, \ldots, X_n) \\ &= \sum_{y\in E} Q(X_n, y) f(y) \\ &= Qf(X_n).\end{aligned}$$

Then, if $\{i_1, \ldots, i_k\}$ is a subset of $\{0, 1, \ldots, n-1\}$,

$$\begin{aligned}&\mathbb{E}[f(X_{n+1}) \,|\, X_{i_1}, \ldots, X_{i_k}, X_n] \\ &\quad = \mathbb{E}[\mathbb{E}[f(X_{n+1}) \,|\, X_0, X_1, \ldots, X_n] \,|\, X_{i_1}, \ldots, X_{i_k}, X_n] \\ &\quad = \mathbb{E}[Qf(X_n) \,|\, X_{i_1}, \ldots, X_{i_k}, X_n] \\ &\quad = Qf(X_n).\end{aligned}$$

(ii) It immediately follows from (13.2) that

$$\mathbb{P}(X_{n+1} = y_1, \ldots, X_{n+p} = y_p \,|\, X_0 = x_0, \ldots, X_n = x_n)$$
$$= Q(x_n, y_1)Q(y_1, y_2) \cdots Q(y_{p-1}, y_p),$$

if $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0$. This gives the first assertion in (ii). The formula $\mathbb{P}(X_{n+p} = y_p \mid X_0, \ldots, X_n) = Q_p(X_n, y_p)$ is then obtained by summing over possible choices of $y_1, \ldots, y_{p-1}$ and using (13.1), and to get $\mathbb{P}(X_{n+p} = y_p \,|\, X_n)$ we just have to take conditional expectations with respect to X_n. To derive the last assertion, we use the preceding formulas to write

$$\begin{aligned}
&\mathbb{P}(Y_1 = y_1, \ldots, Y_p = y_p \,|\, Y_0) \\
&\quad = \mathbb{P}(X_{n+1} = y_1, \ldots, X_{n+p} = y_p \,|\, X_n) \\
&\quad = \mathbb{E}[\mathbb{P}(X_{n+1} = y_1, \ldots, X_{n+p} = y_p \,|\, X_0, \ldots, X_n) \,|\, X_n] \\
&\quad = Q(X_n, y_1)Q(y_1, y_2) \cdots Q(y_{p-1}, y_p)
\end{aligned}$$

and thus

$$\mathbb{P}(Y_0 = y_0, Y_1 = y_1, \ldots, Y_p = y_p) = \mathbb{P}(X_n = y_0)Q(y_0, y_1) \cdots Q(y_{p-1}, y_p).$$

The desired result then follows from the characterization in Proposition 13.2.

□

13.2 A Few Examples

13.2.1 Independent Random Variables

Let $(X_n)_{n\in\mathbb{Z}_+}$ be a sequence of independent and identically distributed random variables with values in E. If μ denotes the law of X_0, then $(X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain with transition matrix

$$Q(x, y) = \mu(y), \quad \forall x, y \in E.$$

The proof is immediate. This is not the most interesting example of a Markov chain!

13.2.2 Random Walks on $\mathbb{Z}^d$

Let $\eta, \xi_1, \xi_2, \ldots, \xi_n, \ldots$ be independent random variables with values in $\mathbb{Z}^d$. We assume that $\xi_1, \xi_2, \ldots$ have the same law denoted by μ, and we set, for every $n \geq 0$,

$$X_n = \eta + \xi_1 + \xi_2 + \cdots + \xi_n.$$

Then $(X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain in $E = \mathbb{Z}^d$ with transition matrix

$$Q(x, y) = \mu(y - x), \quad \forall x, y \in \mathbb{Z}^d.$$

To verify this, use the fact that ξ_{n+1} is independent of $(X_0, X_1, \ldots, X_n)$ to write

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = y \,|\, X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n)\\
&\quad = \mathbb{P}(\xi_{n+1} = y - x_n \,|\, X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n)\\
&\quad = \mathbb{P}(\xi_{n+1} = y - x_n)\\
&\quad = \mu(y - x_n).
\end{aligned}$$

We say that $(X_n)_{n\in\mathbb{Z}_+}$ is a random walk on $\mathbb{Z}^d$ with *jump distribution* μ.

Let $(e_1, \ldots, e_d)$ denote the canonical basis of $\mathbb{R}^d$. In the special case where $\mu(e_i) = \mu(-e_i) = \frac{1}{2d}$ for every $i \in \{1, \ldots, d\}$, the Markov chain $(X_n)_{n\in\mathbb{Z}_+}$ is called simple random walk on $\mathbb{Z}^d$. This generalizes the simple random walk on $\mathbb{Z}$ (coin-tossing process) already studied in Proposition 10.7 and in Chapter 12. Simple random walk on $\mathbb{Z}^d$ is a special case of simple random walk on a graph, which is discussed in the next section.

13.2.3 Simple Random Walk on a Graph

Let $\mathcal{P}_2(E)$ denote the set of all subsets of E with two elements, and let A be a subset of $\mathcal{P}_2(E)$. If $\{x, y\} \in A$, we say that the *vertices* x and y are linked by an *edge*. For every $x \in E$, set

$$A_x = \{y \in E : \{x, y\} \in A\},$$

which is the set of all vertices that are linked to x by an edge. We assume that A_x is finite and nonempty for every $x \in E$. We then define a transition matrix Q on E by setting, for every $x, y \in E$,

$$Q(x,y) := \begin{cases} \dfrac{1}{\operatorname{card} A_x} & \text{if } \{x,y\} \in A, \\ 0 & \text{otherwise.} \end{cases}$$

A Markov chain with transition matrix Q is called simple random walk on the graph (E, A). Informally, this Markov chain sitting at a point x of E chooses at the next step one of the vertices connected to x, uniformly at random.

13.2.4 Galton-Watson Branching Processes

Recall the definition of these processes, which were already studied in Chapter 12. Let μ be a probability measure on $\mathbb{Z}_+$ and let $\ell \in \mathbb{Z}_+$. We define by induction

$$X_0 = \ell$$

$$X_{n+1} = \sum_{j=1}^{X_n} \xi_{n,j} \ , \qquad \forall n \in \mathbb{Z}_+ \ ,$$

where the random variables $\xi_{n,j}$, $n \in \mathbb{Z}_+$, $j \in \mathbb{N}$ are independent and identically distributed with law μ. Then $(X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain in $E = \mathbb{Z}_+$ with transition matrix

$$Q(x,y) = \mu^{*x}(y), \quad \forall x, y \in \mathbb{Z}_+,$$

where the convolution μ^{*k} is defined for every integer $k \geq 1$ by induction, setting $\mu^{*1} = \mu$ and $\mu^{*(k+1)} = \mu^{*k} * \mu$, and by convention μ^{*0} is the Dirac measure at 0. By Proposition 9.12, μ^{*k} is the law of the sum of k independent random variables with law μ.

Indeed, by observing that the random variables $\xi_{n,j}$, $j \in \mathbb{N}$ are independent of $X_0, \ldots, X_n$, we have

$$\begin{aligned} &\mathbb{P}(X_{n+1} = y \,|\, X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\ &\quad = \mathbb{P}\Big(\sum_{j=1}^{x_n} \xi_{n,j} = y \,\Big|\, X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n\Big) \\ &\quad = \mathbb{P}\Big(\sum_{j=1}^{x_n} \xi_{n,j} = y\Big) \\ &\quad = \mu^{*x_n}(y). \end{aligned}$$

13.3 The Canonical Markov Chain

In this section, we explain how to make a canonical choice of a probability space (and a random process) to get a Markov chain with a given transition matrix. This is in contrast with the previous chapters, where the underlying probability space was not specified, except sometimes in examples. In the context of Markov chains, this canonical choice will have several advantages. In particular, it will enable us to consider simultaneously the given Markov chain with all possible initial points, which will be very useful in forthcoming developments.

We start with an existence result.

Proposition 13.4 *Let Q be a stochastic matrix on E. We can find a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which, for every $x \in E$, one can construct a random process $(X_n^x)_{n\in\mathbb{Z}_+}$ which is a Markov chain with transition matrix Q, such that $X_0^x = x$.*

Proof We consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which there exists a sequence $(U_n)_{n\in\mathbb{N}}$ of independent random variables uniformly distributed over $(0, 1)$. We saw in Section 9.4 that we can find such a probability space (we may indeed take $\Omega = [0, 1]$ equipped with its Borel σ-field and with Lebesgue measure). Then let $y_1, y_2, \ldots, y_k, \ldots$ be an enumeration of the elements of E. We assume here that E is infinite, but the finite case can be treated in exactly the same way.

Let us fix $x \in E$. We set $X_0^x = x$ and then, for every $k \in \mathbb{N}$,

$$X_1^x = y_k \quad \text{if} \quad \sum_{1\le j<k} Q(x, y_j) < U_1 \le \sum_{1\le j\le k} Q(x, y_j)$$

We see that X_1^x is well defined since $\sum_{j=1}^{\infty} Q(x, y_j) = 1$ and $0 < U_1 < 1$. Furthermore, we have $\mathbb{P}(X_1^x = y) = Q(x, y)$ for every $y \in E$ since U_1 is uniformly distributed over $(0, 1)$. We continue by induction and, for every $n \in \mathbb{N}$ we set

$$X_{n+1}^x = y_k \quad \text{if} \quad \sum_{1\le j<k} Q(X_n^x, y_j) < U_{n+1} \le \sum_{1\le j\le k} Q(X_n^x, y_j).$$

We observe that $(X_0^x, X_1^x, \ldots, X_n^x)$ is a (measurable) function of $(U_1, \ldots, U_n)$ and is therefore independent of U_{n+1}. We use this observation to write, for every $n \ge 1$ and $k \ge 1$,

$$\begin{aligned}
&\mathbb{P}(X_{n+1}^x = y_k \,|\, X_0^x = x_0, X_1^x = x_1, \ldots X_n^x = x_n) \\
&= \mathbb{P}\Big(\sum_{1\le j<k} Q(x_n, y_j) < U_{n+1} \le \sum_{1\le j\le k} Q(x_n, y_j) \,\Big|\, X_0^x = x_0, \ldots X_n^x = x_n \Big)
\end{aligned}$$

$$= \mathbb{P}\Big(\sum_{1\le j<k} Q(x_n, y_j) < U_{n+1} \le \sum_{1\le j\le k} Q(x_n, y_j)\Big)$$
$$= Q(x_n, y_k).$$

It follows that $(X_n^x)_{n\in\mathbb{Z}_+}$ is a Markov chain with transition matrix Q. □

Let us now explain our canonical choice of a probability space. We take

$$\mathbf{\Omega} = E^{\mathbb{Z}_+},$$

so that an element of $\mathbf{\Omega}$ is a sequence $\omega = (\omega_0, \omega_1, \omega_2, \ldots)$ of elements of E. The coordinate mappings $\mathbf{X}_n$, $n \in \mathbb{Z}_+$ are then defined by

$$\mathbf{X}_n(\omega) := \omega_n.$$

We equip $\mathbf{\Omega}$ with the smallest σ-field for which the coordinate mappings $\mathbf{X}_n$, $n \in \mathbb{Z}_+$ are measurable. We denote this σ-field by $\mathfrak{F}$. Equivalently, $\mathfrak{F}$ is generated by the "cylinder sets" of the form

$$\{(\omega_0, \omega_1, \omega_2, \ldots) \in E^{\mathbb{Z}_+} : \omega_0 = x_0, \ldots, \omega_n = x_n\}$$

where $n \in \mathbb{Z}_+$ and $x_0, \ldots, x_n \in E$.

Lemma 13.5 *Let $(G, \mathcal{G})$ be a measurable space, and let ψ be a function from G into $\mathbf{\Omega}$. Then ψ is measurable if and only if $\mathbf{X}_n \circ \psi$ is measurable, for every $n \in \mathbb{Z}_+$.*

Proof The "only if" part is immediate since a composition of measurable functions is measurable. So let us assume that $\mathbf{X}_n \circ \psi$ is measurable, for every $n \in \mathbb{Z}_+$, and prove that ψ is measurable. We observe that

$$\widetilde{\mathfrak{F}} := \{A \in \mathfrak{F} : \psi^{-1}(A) \in \mathcal{G}\}$$

is a σ-field on $\mathbf{\Omega}$, which by assumption contains all sets of the form $\mathbf{X}_n^{-1}(y)$, $y \in E$. It follows that all coordinate mappings $\mathbf{X}_n$ are measurable with respect to $\widetilde{\mathfrak{F}}$, which implies $\mathfrak{F} \subset \widetilde{\mathfrak{F}}$. □

Theorem 13.6 *Let Q be a stochastic matrix on E. For every $x \in E$, there exists a unique probability measure $\mathbf{P}_x$ on $(\mathbf{\Omega}, \mathfrak{F})$ such that, under $\mathbf{P}_x$, the coordinate process $(\mathbf{X}_n)_{n\in\mathbb{Z}_+}$ is a Markov chain with transition matrix Q, and $\mathbf{P}_x(\mathbf{X}_0 = x) = 1$.*

We will say that, on the probability space $(\mathbf{\Omega}, \mathfrak{F}, \mathbf{P}_x)$, the coordinate process $(\mathbf{X}_n)_{n\in\mathbb{Z}_+}$ is the *canonical Markov chain* with transition matrix Q and initial value x.

Proof Let $x \in E$. Proposition 13.4 allows us to construct a process $(X_n^x)_{n\in\mathbb{Z}_+}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, which is a Markov chain with transition matrix Q, such that $X_0^x = x$. We define $\mathbf{P}_x$ as the pushforward of $\mathbb{P}$ under the mapping

$$\omega \mapsto (X_n^x(\omega))_{n\in\mathbb{Z}_+}$$

from Ω into $\mathbf{\Omega}$. An application of Lemma 13.5 shows that this mapping is measurable (its composition with $\mathbf{X}_n$ is just the random variable X_n^x on Ω). We have $\mathbf{P}_x(\mathbf{X}_0 = x) = \mathbb{P}(X_0^x = x) = 1$ and moreover, for every $x_0, x_1, \dots, x_n \in E$,

$$\begin{aligned}\mathbf{P}_x(\mathbf{X}_0 = x_0, \mathbf{X}_1 = x_1, \dots, \mathbf{X}_n = x_n) &= \mathbb{P}(X_0^x = x_0, X_1^x = x_1, \dots, X_n^x = x_n)\\ &= \mathbb{P}(X_0^x = x_0) Q(x_0, x_1) \dots Q(x_{n-1}, x_n)\\ &= \mathbf{P}_x(\mathbf{X}_0 = x_0) Q(x_0, x_1) \dots Q(x_{n-1}, x_n)\end{aligned}$$

which shows that, under the probability measure $\mathbf{P}_x$, the coordinate process is a Markov chain with transition matrix Q (Proposition 13.2).

As for uniqueness, suppose that $\widetilde{\mathbf{P}}_x$ is another probability measure satisfying the properties stated in the theorem. Then $\mathbf{P}_x$ and $\widetilde{\mathbf{P}}_x$ assign the same value to any cylinder set. Since cylinder sets form a class closed under finite intersections, which generates the σ-field $\mathfrak{F}$, Corollary 1.19 implies that $\widetilde{\mathbf{P}}_x = \mathbf{P}_x$. □

Remarks

(a) By the last assertion of Proposition 13.2, we have, for every $n \ge 0$ and $x, y \in E$,

$$\mathbf{P}_x(\mathbf{X}_n = y) = Q_n(x, y).$$

(b) Let μ be a probability measure on E. We set

$$\mathbf{P}_\mu = \sum_{x\in E} \mu(x)\, \mathbf{P}_x$$

which defines a probability measure on $\mathbf{\Omega}$. By writing down the explicit formula for $\mathbf{P}_\mu(\mathbf{X}_0 = x_0, \dots, \mathbf{X}_n = x_n)$, and using Proposition 13.2, we get that, under $\mathbf{P}_\mu$, $(\mathbf{X}_n)_{n\in\mathbb{Z}_+}$ is a Markov chain with transition matrix Q, and the law of $\mathbf{X}_0$ is μ. Note in particular that $\mathbf{P}_{\delta_x} = \mathbf{P}_x$.

(c) Let $(Y_n)_{n\in\mathbb{Z}_+}$ be a Markov chain with transition matrix Q defined under the probability measure $\mathbb{P}$, such that $Y_0 = x$, $\mathbb{P}$ a.s. Then, for every measurable subset B of $\mathbf{\Omega} = E^{\mathbb{Z}_+}$, we have

$$\mathbb{P}((Y_n)_{n\in\mathbb{Z}_+} \in B) = \mathbf{P}_x(B).$$

Indeed, this equality holds when B is a cylinder set, and then a monotone class argument gives the general case. This equality shows that all the results that we will establish later for the canonical Markov chain (as given by Theorem 13.6) carry over to any Markov chain with the same transition matrix (no matter on which probability space it is defined).

In the remaining part of this section, we deal with the canonical Markov chain associated with a transition matrix Q. One of the main motivations for introducing this canonical Markov chain is the fact that it makes possible to define *shifts* on the space $\boldsymbol{\Omega}$. For every integer $k \geq 0$, we define the shift $\theta_k : \boldsymbol{\Omega} \longrightarrow \boldsymbol{\Omega}$ by

$$\theta_k((\omega_n)_{n\in\mathbb{Z}_+}) := (\omega_{k+n})_{n\in\mathbb{Z}_+}.$$

Lemma 13.5 shows that this mapping is measurable. By definition, we have $\mathbf{X}_n \circ \theta_k = \mathbf{X}_{k+n}$ for every $k, n \geq 0$. Equivalently, $\theta_k(\omega) = (X_k(\omega), X_{k+1}(\omega), \ldots)$ represents the future of the Markov chain after time k.

We will use the shifts θ_k to state a more precise version of the Markov property. We first introduce the canonical filtration on $\boldsymbol{\Omega}$ by setting $\mathfrak{F}_n = \sigma(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ for every $n \geq 0$. We also use the notation $\mathbf{E}_x$, resp. $\mathbf{E}_\mu$, to denote expected values under $\mathbf{P}_x$, resp. under $\mathbf{P}_\mu$.

Theorem 13.7 (Simple Markov Property) *Let F and G be two nonnegative measurable functions on $\boldsymbol{\Omega}$ and let $n \geq 0$. Suppose that F is $\mathfrak{F}_n$-measurable. Then, for every $x \in E$,*

$$\mathbf{E}_x[F\, G \circ \theta_n] = \mathbf{E}_x[F\, \mathbf{E}_{\mathbf{X}_n}[G]].$$

Equivalently,

$$\mathbf{E}_x[G \circ \theta_n \,|\, \mathfrak{F}_n] = \mathbf{E}_{\mathbf{X}_n}[G],$$

meaning that the conditional distribution of $\theta_n(\omega)$ knowing $(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ is $\mathbf{P}_{\mathbf{X}_n}$.

Note that the quantity $\mathbf{E}_{\mathbf{X}_n}[G]$ is the composition of $\mathbf{X}_n$ with the function $x \mapsto \mathbf{E}_x[G]$. In particular it is a function of $\mathbf{X}_n$ and thus an $\mathfrak{F}_n$-measurable random variable

Remark The formulas of Theorem 13.7 are immediately extended to the case where $\mathbf{E}_x$ is replaced by $\mathbf{E}_\mu$, for any probability measure μ on E. The same observation applies to Theorem 13.8 below.

Proof It is enough to prove the equality $\mathbf{E}_x[G \circ \theta_n \mid \mathfrak{F}_n] = \mathbf{E}_{\mathbf{X}_n}[G]$, as the first assertion then follows from the characteristic property of $\mathbf{E}_x[G \circ \theta_n \,|\, \mathfrak{F}_n]$. Assume first that G is of the form

$$G = \mathbf{1}_{\{\mathbf{X}_0=y_0,\mathbf{X}_1=y_1,\ldots,\mathbf{X}_p=y_p\}}$$

where $p \in \mathbb{Z}_+$ and $y_0, \ldots, y_p \in E$. In that case, for every $y \in E$,

$$\mathbf{E}_y[G] = \mathbf{1}_{\{y_0=y\}} Q(y_0, y_1) \ldots Q(y_{p-1}, y_p)$$

and the desired formula follows from Proposition 13.3 (ii). A monotone class argument then shows that the result still holds if $G = \mathbf{1}_A$, $A \in \mathfrak{F}$ (note that we use the version of the monotone convergence theorem for conditional expectations). Finally we get the desired result for nonnegative simple functions by linearity, and for nonnegative measurable functions by monotone convergence. □

Theorem 13.7 gives a general form of the Markov property: the conditional law of the future $\theta_n(\omega)$ given the past $(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ is given by $\mathbf{P}_{\mathbf{X}_n}$, and thus only depends on the present $\mathbf{X}_n$. It will be very important to extend this property to the case where the deterministic time n is replaced by a random time.

To motivate this extension, consider the problem of knowing whether the chain started from a point x returns to x infinitely many times. In other words, if

$$N_x = \sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{X}_n=x\}},$$

do we have $\mathbf{P}_x(N_x = \infty) = 1$? It is in fact sufficient to verify that the chain returns to x at least once. If $H_x = \inf\{n \geq 1 : \mathbf{X}_n = x\}$, with the usual convention $\inf \varnothing = +\infty$, then

$$\mathbf{P}_x(N_x = \infty) = 1 \Leftrightarrow \mathbf{P}_x(H_x < \infty) = 1.$$

The implication $\Rightarrow$ is trivial. Conversely, suppose that $\mathbf{P}_x(H_x < \infty) = 1$. Provided we can apply the extended Markov property mentioned above to the random time H_x, we get that the law of $\theta_{H_x}(\omega) = (\mathbf{X}_{H_x+n}(\omega))_{n\in\mathbb{Z}_+}$ is $\mathbf{P}_x$. But then, by writing

$$N_x(\omega) = 1 + N_x(\theta_{H_x}(\omega))$$

we get that N_x has the same law as $1 + N_x$ under $\mathbf{P}_x$, which is only possible if $N_x = \infty$, $\mathbf{P}_x$ a.s.

The next theorem will allow us to give a rigorous justification of the preceding argument (and the resulting property will be discussed in the next section). Recall from Section 12.2 the notion of a stopping time T in the filtration $(\mathfrak{F}_n)_{n\in\mathbb{Z}_+}$, and the definition of the σ-field $\mathfrak{F}_T$ representing the past up to time T (Definition 12.8). If T is a stopping time, then, provided that $T(\omega) < \infty$, we can define $\theta_T(\omega) = (\mathbf{X}_{T(\omega)+n}(\omega))_{n\in\mathbb{Z}_+}$, which represents the future after time T.

Theorem 13.8 (Strong Markov Property) *Let T be a stopping time of the filtration $(\mathfrak{F}_n)_{n\in\mathbb{Z}_+}$. Let F and G be two nonnegative measurable functions on $\mathbf{\Omega}$. Suppose that F is $\mathfrak{F}_T$-measurable. Then, for every $x \in E$,*

$$\mathbf{E}_x[\mathbf{1}_{\{T<\infty\}}\, F\;\, G\circ\theta_T] = \mathbf{E}_x[\mathbf{1}_{\{T<\infty\}}\, F\, \mathbf{E}_{\mathbf{X}_T}[G]].$$

Equivalently,

$$\mathbf{E}_x[\mathbf{1}_{\{T<\infty\}}\, G\circ\theta_T \,|\, \mathfrak{F}_T] = \mathbf{1}_{\{T<\infty\}}\, \mathbf{E}_{\mathbf{X}_T}[G].$$

Remark The random variable $\mathbf{X}_T$, which is defined on the $\mathfrak{F}_T$-measurable set $\{T < \infty\}$, is $\mathfrak{F}_T$-measurable (cf. Proposition 12.11, note that this proposition considered a real-valued random process, but the argument goes through without change in our setting). The random variable $\mathbf{E}_{\mathbf{X}_T}[G]$, which is defined also on $\{T < \infty\}$, is the composition of the mappings $\omega \mapsto \mathbf{X}_T(\omega)$ and $x \mapsto \mathbf{E}_x[G]$.

In the same way as Theorem 13.7, Theorem 13.8 can be interpreted (a little informally) by saying that the conditional distribution of the future after T (that is, $\theta_T(\omega) = (\mathbf{X}_{T+n})_{n\in\mathbb{Z}_+}$) knowing the past $(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_T)$, is the law $\mathbf{P}_{\mathbf{X}_T}$ of the Markov chain started at $\mathbf{X}_T$, and thus depends only on the "present" $\mathbf{X}_T$.

Proof It is enough to prove the first assertion. Since F is $\mathfrak{F}_T$-measurable, it is immediate to verify (from the definition of the σ-field $\mathfrak{F}_T$) that $\mathbf{1}_{\{T=n\}}F$ is $\mathfrak{F}_n$-measurable, for every integer $n \geq 0$. It follows that

$$\begin{aligned}\mathbf{E}_x[\mathbf{1}_{\{T=n\}}\, F\;\, G\circ\theta_T] = \mathbf{E}_x[\mathbf{1}_{\{T=n\}}\, F\;\, G\circ\theta_n] &= \mathbf{E}_x[\mathbf{1}_{\{T=n\}}\, F\, \mathbf{E}_{\mathbf{X}_n}[G]] \\ &= \mathbf{E}_x[\mathbf{1}_{\{T=n\}}\, F\, \mathbf{E}_{\mathbf{X}_T}[G]]\end{aligned}$$

by the simple Markov property (Theorem 13.7). We get the desired result by summing the formula of the last display over all $n \in \mathbb{Z}_+$.

□

Corollary 13.9 *Let $x \in E$ and let T be a stopping time. Assume that $\mathbf{P}_x(T < \infty) = 1$ and that there exists $y \in E$ such that $\mathbf{P}_x(\mathbf{X}_T = y) = 1$. Then, under the probability measure $\mathbf{P}_x$, $\theta_T(\omega)$ is independent of $\mathfrak{F}_T$ and distributed according to $\mathbf{P}_y$.*

Remark $\theta_T(\omega)$ is not defined on the event $\{T = \infty\}$, but this definition is irrelevant since we assume that $\mathbf{P}_x(T < \infty) = 1$.

Proof Let F and G be as in Theorem 13.8. Then,

$$\mathbf{E}_x[F\;\, G(\theta_T(\omega))] = \mathbf{E}_x[F\, \mathbf{E}_{\mathbf{X}_T}[G]] = \mathbf{E}_x[F\, \mathbf{E}_y[G]] = \mathbf{E}_x[F]\, \mathbf{E}_y[G]$$

and the assertions of the corollary follow. □

13.4 The Classification of States

From now on, unless otherwise mentioned (in particular in examples), we will consider the canonical Markov chain introduced in the previous section for a given stochastic matrix Q on E. As explained above, all the results that we will derive carry over to Markov chains with the same transition matrix defined on any probability space.

Recall the notation

$$H_x = \inf\{n \geq 1 : \mathbf{X}_n = x\}$$

$$N_x = \sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{X}_n = x\}}.$$

Proposition 13.10 *Let $x \in E$. Then:*

- *either $\mathbf{P}_x(H_x < \infty) = 1$, and then*

$$N_x = \infty\,, \quad \mathbf{P}_x \text{ a.s.}$$

  in this case x is said to be recurrent;
- *or $\mathbf{P}_x(H_x < \infty) < 1$, and then*

$$N_x < \infty\,, \quad \mathbf{P}_x \text{ a.s.}$$

  and more precisely $\mathbf{P}_x(N_x = k) = \mathbf{P}_x(H_x = \infty)\,\mathbf{P}_x(H_x < \infty)^{k-1}$ for every $k \geq 1$, and $\mathbf{E}_x[N_x] = 1/\mathbf{P}_x(H_x = \infty) < \infty$; in this case x is said to be transient.

Proof For every $k \geq 1$, the strong Markov property shows that

$$\begin{aligned}
\mathbf{P}_x(N_x \geq k+1) &= \mathbf{E}_x[\mathbf{1}_{\{H_x<\infty\}}\,\mathbf{1}_{\{N_x \geq k\}} \circ \theta_{H_x}] \\
&= \mathbf{E}_x[\mathbf{1}_{\{H_x<\infty\}}\,\mathbf{E}_x[\mathbf{1}_{\{N_x \geq k\}}]] \\
&= \mathbf{P}_x(H_x < \infty)\,\mathbf{P}_x(N_x \geq k).
\end{aligned}$$

Since $\mathbf{P}_x(N_x \geq 1) = 1$, we get that $\mathbf{P}_x(N_x \geq k) = \mathbf{P}_x(H_x < \infty)^{k-1}$ for every $k \geq 1$ by induction. If $\mathbf{P}_x(H_x < \infty) = 1$ it immediately follows that $\mathbf{P}_x(N_x = \infty) = 1$. If $\mathbf{P}_x(H_x < \infty) < 1$, we obtain the law of N_x under $\mathbf{P}_x$, and in particular

$$\mathbf{E}_x[N_x] = \sum_{k=1}^{\infty} \mathbf{P}_x(N_x \geq k) = \frac{1}{\mathbf{P}_x(H_x = \infty)} < \infty.$$

□

Definition 13.11 The potential kernel of the Markov chain is the function defined on $E \times E$ by

$$U(x, y) := \mathbf{E}_x[N_y].$$

Note that $U(x, y) > 0$ if and only if the probability for the chain started from x to visit the point y is positive.

Proposition 13.12

(i) *For every* $x, y \in E$,

$$U(x, y) = \sum_{n=0}^{\infty} Q_n(x, y).$$

(ii) $U(x, x) = \infty$ *if and only if* x *is recurrent.*
(iii) *For every* $x, y \in E$, *with* $x \neq y$,

$$U(x, y) = \mathbf{P}_x(H_y < \infty)\, U(y, y).$$

Consequently, if y *is transient, we have* $\mathbf{P}_x(N_y < \infty) = 1$ *for every* $x \in E$.

Proof To get (i), we write

$$U(x, y) = \mathbf{E}_x\Big[\sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{X}_n = y\}}\Big] = \sum_{n=0}^{\infty} \mathbf{P}_x(\mathbf{X}_n = y) = \sum_{n=0}^{\infty} Q_n(x, y).$$

Property (ii) follows from Proposition 13.10 and the definition of U.
The first part of (iii) follows from the strong Markov property: if $x \neq y$,

$$\mathbf{E}_x[N_y] = \mathbf{E}_x[\mathbf{1}_{\{H_y<\infty\}} N_y \circ \theta_{H_y}] = \mathbf{E}_x[\mathbf{1}_{\{H_y<\infty\}} \mathbf{E}_y[N_y]] = \mathbf{P}_x(H_y < \infty)\, U(y, y).$$

The last assertion is then immediate since y is transient if and only if $U(y, y) < \infty$, which implies $\mathbf{E}_x[N_y] = U(x, y) < \infty$ for every $x \in E$. □

Example Consider the Markov chain on $\mathbb{Z}^d$ with transition kernel

$$Q((x_1, \ldots, x_d), (y_1, \ldots, y_d)) = \frac{1}{2^d} \prod_{i=1}^{d} \mathbf{1}_{\{|y_i - x_i| = 1\}}$$

(this is a special case of random walk on $\mathbb{Z}^d$). It is easy to verify that this Markov chain started from 0 can be constructed as $(Y_n^1, \ldots, Y_n^d)_{n\in\mathbb{Z}_+}$, where the random processes $Y^1, \ldots, Y^d$ are independent simple random walks on $\mathbb{Z}$, started from 0. It follows that

$$Q_n(0,0) = \mathbb{P}(Y_n^1 = 0, \dots, Y_n^d = 0) = \mathbb{P}(Y_n^1 = 0)^d.$$

However $\mathbb{P}(Y_n^1 = 0) = 0$ if n is odd, and if $n = 2k$ is even, a simple counting argument left to the reader shows that

$$\mathbb{P}(Y_{2k}^1 = 0) = 2^{-2k} \binom{2k}{k}.$$

Consequently,

$$U(0,0) = \sum_{k=0}^{\infty} Q_{2k}(0,0) = \sum_{k=0}^{\infty} \left(2^{-2k} \binom{2k}{k}\right)^d.$$

Stirling's formula shows that

$$2^{-2k} \binom{2k}{k} \underset{k\to\infty}{\sim} \frac{(\frac{2k}{e})^{2k}\sqrt{4\pi k}}{2^{2k}((\frac{k}{e})^k\sqrt{2\pi k})^2} = \sqrt{\frac{1}{\pi k}}.$$

We conclude that 0 is recurrent if $d = 1$ or $d = 2$, and transient if $d \geq 3$. When $d = 1$, the Markov chain with transition matrix Q is simple random walk on $\mathbb{Z}$, and the fact that 0 is recurrent also follows from Proposition 10.7, or from the martingale argument explained after Corollary 12.16. When $d = 2$, the Markov chain with transition matrix Q is *not* simple random walk on $\mathbb{Z}^2$ (as defined in Section 13.2.2), but a simple argument allows us to get the recurrence of 0 for simple random walk on $\mathbb{Z}^2$. Just observe that (Y_n^1, Y_n^2) takes values in the subgroup H of $\mathbb{Z}^2$ consisting of all pairs (i, j) such that $i + j$ is even, and note that this subgroup is isomorphic to $\mathbb{Z}^2$ through a group isomorphism Φ such that $\Phi((1,1)) = (0,1)$ and $\Phi((1,-1)) = (1,0)$. Then the claimed recurrence property follows from the fact that $\Phi((Y_n^1, Y_n^2))$ is simple random walk on $\mathbb{Z}^2$. This argument does not apply in dimension $d \geq 3$, but one can prove that 0 is transient for simple random walk on $\mathbb{Z}^d$ when $d \geq 3$.

Let us come back to the general setting. We denote the set of all recurrent points by R.

Lemma 13.13 *Assume that $x \in R$ and that y is a point of $E\backslash\{x\}$ such that $U(x,y) > 0$. Then, $y \in R$ and $\mathbf{P}_y(H_x < \infty) = 1$, hence in particular $U(y,x) > 0$.*

Proof Let us start by showing that $\mathbf{P}_y(H_x < \infty) = 1$. To this end, we write

$$\begin{aligned}
0 = \mathbf{P}_x(N_x < \infty) &\geq \mathbf{P}_x(H_y < \infty, H_x \circ \theta_{H_y} = \infty) \\
&= \mathbf{E}_x[\mathbf{1}_{\{H_y<\infty\}} \times \mathbf{1}_{\{H_x=\infty\}} \circ \theta_{H_y}] \\
&= \mathbf{E}_x[\mathbf{1}_{\{H_y<\infty\}} \mathbf{P}_y(H_x = \infty)] \\
&= \mathbf{P}_x(H_y < \infty)\, \mathbf{P}_y(H_x = \infty).
\end{aligned}$$

The assumption $U(x, y) > 0$ implies that $\mathbf{P}_x(H_y < \infty) > 0$, and it follows that $\mathbf{P}_y(H_x = \infty) = 0$, so that $\mathbf{P}_y(H_x < \infty) = 1$.

Then, we can find integers $n_1, n_2 \geq 1$ such that $Q_{n_1}(x, y) > 0$, and $Q_{n_2}(y, x) > 0$. For every integer $p \geq 0$, we have

$$\begin{aligned} Q_{n_2+p+n_1}(y, y) &\geq \mathbf{P}_y(\mathbf{X}_{n_2} = x, \mathbf{X}_{n_2+p} = x, \mathbf{X}_{n_2+p+n_1} = y) \\ &= Q_{n_2}(y, x) Q_p(x, x) Q_{n_1}(x, y) \end{aligned}$$

and therefore

$$U(y, y) \geq \sum_{p=0}^{\infty} Q_{n_2+p+n_1}(y, y) \geq Q_{n_2}(y, x) \Big(\sum_{p=0}^{\infty} Q_p(x, x) \Big) Q_{n_1}(x, y) = \infty$$

since $x \in R$ implies $\sum_{p=0}^{\infty} Q_p(x, x) = U(x, x) = \infty$. We conclude that $y \in R$. □

As a consequence of the lemma, if $x \in R$ and $y \in E \backslash R$, we have $U(x, y) = 0$: the chain started from a recurrent point cannot visit a transient point. This property plays an important role in the following statement.

Theorem 13.14 (Classification of States) *Let R be the set of all recurrent points, and $T = \inf\{n \geq 0 : \mathbf{X}_n \in R\}$. There exists a (unique) partition*

$$R = \bigcup_{i \in I} R_i$$

of R, such that the following properties hold:

(a) for every $i \in I$ and $x \in R_i$, we have $\mathbf{P}_x$ a.s.

- *$N_y = +\infty\,, \quad \forall y \in R_i$;*
- *$N_y = 0\,, \quad \forall y \in E \backslash R_i$;*

(b) for every $x \in E \backslash R$, we have $\mathbf{P}_x$ a.s.

- *either $T = \infty$ and $N_y < \infty$, $\forall y \in E$,*
- *or $T < \infty$, then there exists a (random) index $j \in I$ such that $\mathbf{X}_n \in R_j$ for every $n \geq T$, and $N_y = \infty$ for every $y \in R_j$.*

The sets R_i, $i \in I$, are called the recurrence classes of the Markov chain.

Proof For $x, y \in R$, set $x \sim y$ if and only if $U(x, y) > 0$. It follows from Lemma 13.13 that this defines an equivalence relation on R (note that $Q_n(x, y) > 0$ and $Q_m(y, z) > 0$ imply $Q_{n+m}(x, z) > 0$). The partition $(R_i)_{i \in I}$ then corresponds to the equivalence classes for this equivalence relation.

Let $i \in I$ and $x \in R_i$. We have then $U(x, y) = 0$ for every $y \in E\backslash R_i$ (in case $y \in E\backslash R$, we use Lemma 13.13) and thus $N_y = 0$, $\mathbf{P}_x$ a.s. for every $y \in E\backslash R_i$. On the other hand, if $y \in R_i$, we have $\mathbf{P}_x(H_y < \infty) = 1$ by Lemma 13.13, and the strong Markov property shows that

$$\mathbf{P}_x(N_y = \infty) = \mathbf{E}_x(\mathbf{1}_{\{H_y<\infty\}}\,\mathbf{1}_{\{N_y=\infty\}} \circ \theta_{H_y}] = \mathbf{P}_x(H_y < \infty)\,\mathbf{P}_y(N_y = \infty) = 1.$$

If $x \in E\backslash R$, then $\mathbf{P}_x$ a.s. on the event $\{T = \infty\}$, the chain does not visit R and furthermore $N_y < \infty$ for every $y \in E\backslash R$ by Proposition 13.12 (iii). On the event $\{T < \infty\}$, let j be the (random) index such that $\mathbf{X}_T \in R_j$. By applying the strong Markov property at T, and the first part of the statement, we easily get that, $\mathbf{P}_x$ a.s. on the event $\{T < \infty\}$, we have $\mathbf{X}_n \in R_j$ for every $n \geq T$, and $N_y = \infty$ for every $y \in R_j$. We leave the details to the reader. □

Definition 13.15 The chain is said to be irreducible if $U(x, y) > 0$ for every $x, y \in E$. We also say that Q is irreducible.

The chain is irreducible if, for every $x, y \in E$, the chain starting from x visits y with positive probability.

Corollary 13.16 *If the chain is irreducible,*

- *either all states are recurrent, there exists only one recurrence class, and we have for every $x \in E$,*

$$\mathbf{P}_x(N_y = \infty,\ \forall y \in E) = 1.$$

- *or all states are transient, and for every $x \in E$,*

$$\mathbf{P}_x(N_y < \infty,\ \forall y \in E) = 1.$$

When E is finite, only the first case occurs.

Proof If there exists a recurrent state, then Lemma 13.13 immediately shows that all states are recurrent, and since $U(x, y) > 0$ for every $x, y \in E$, we also see that there is just one recurrence class. The remaining part of the statement follows from Theorem 13.14, except for the last assertion, which we can prove by contradiction. If E is finite and all states are transient, then

$$\mathbf{P}_x \text{ a.s. },\quad \sum_{y\in E} N_y < \infty$$

which is absurd since

$$\sum_{y\in E} N_y = \sum_{y\in E}\sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{X}_n=y\}} = \sum_{n=0}^{\infty}\sum_{y\in E} \mathbf{1}_{\{\mathbf{X}_n=y\}} = \infty.$$

□

An irreducible Markov chain such that all states are recurrent will be called recurrent irreducible.

Examples We will now discuss the classification of states for each of the examples presented in Section 13.2. Before that, let us emphasize once again that all results obtained for the canonical Markov chain carry over to any Markov chain $(Y_n)_{n\in\mathbb{Z}_+}$ with the same transition matrix Q (and conversely). For instance, if $Y_0 = y$ a.s. then writing $N_x^Y = \sum_{n=0}^{\infty} \mathbf{1}_{\{Y_n=x\}}$, we have for every $k \in \mathbb{Z}_+ \cup \{\infty\}$,

$$\mathbb{P}(N_x^Y = k) = \mathbf{P}_y(N_x = k)),$$

since the left-hand side can be written as

$$\mathbb{P}((Y_n)_{n\in\mathbb{Z}_+} \in B),$$

with $B = \{\omega \in E^{\mathbb{Z}_+} : N_x(\omega) = k\}$, and it suffices to apply Remark (b) following Theorem 13.6.

(1) **Independent and identically distributed random variables with law** μ. In that case $Q(x, y) = \mu(y)$. The Borel-Cantelli lemma shows that y is recurrent if and only if $\mu(y) > 0$, and there is only one recurrence class. The chain is irreducible if and only if $\mu(y) > 0$ for every $y \in E$.
(2) **Random walk on** $\mathbb{Z}^d$. In that case,

$$Y_n = Y_0 + \sum_{i=1}^{n} \xi_i$$

where the random variables ξ_i take values in $\mathbb{Z}^d$, and are independent and distributed according to μ (and are also independent of Y_0). Then, since $Q(x, y) = \mu(y - x)$ only depends on $y - x$, it follows that $U(x, y)$ is also a function of $y - x$. Since x is recurrent if and only if $U(x, x) = \infty$, we get that all states are of the same type (recurrent or transient).

We assume that $\mathbb{E}[|\xi_1|] < \infty$, and we set $m = \mathbb{E}[\xi_1] \in \mathbb{R}^d$.

Case $m \neq 0$ In that case, the strong law of large numbers implies that

$$\lim_{n\to\infty} |Y_n| = \infty\ , \quad \text{a.s.}$$

It follows that all states are transient.

Case $m = 0$ In that case, things are more complicated. In the special case of simple random walk, the discussion following Proposition 13.12 shows that the chain is recurrent irreducible when $d = 1$ or $d = 2$ (one can also prove that all states are transient when $d \geq 3$).

We give a general result when $d = 1$.

Theorem 13.17 *Consider a random walk $(Y_n)_{n\in\mathbb{Z}_+}$ on $\mathbb{Z}$ whose jump distribution μ is such that $\sum_{k\in\mathbb{Z}} |k|\mu(k) < \infty$ and $\sum_{k\in\mathbb{Z}} k\mu(k) = 0$. Then all states are recurrent. Moreover the chain is irreducible if and only if the subgroup generated by $\{x \in \mathbb{Z} : \mu(x) > 0\}$ is $\mathbb{Z}$.*

Proof Let us assume that 0 is transient, so that $U(0, 0) < \infty$. We will see that this leads to a contradiction. Without loss of generality, we assume throughout the proof that $Y_0 = 0$. We observe that, for every $x \in \mathbb{Z}$,

$$U(0, x) \leq U(x, x) = U(0, 0)$$

where the first inequality follows from Proposition 13.12(iii). Consequently, for every $n \geq 1$,

$$\sum_{|x|\leq n} U(0, x) \leq (2n + 1)U(0, 0) \leq Cn \tag{13.3}$$

where $C = 3U(0, 0) < \infty$.

On the other hand, by the law of large numbers, $n^{-1}Y_n$ converges to 0 a.s., hence also in probability. Setting $\varepsilon = (4C)^{-1} > 0$, we can thus find N large enough, so that, for every $n \geq N$,

$$\mathbb{P}(|Y_n| \leq \varepsilon n) > \frac{1}{2},$$

or, equivalently,

$$\sum_{|x|\leq\varepsilon n} Q_n(0, x) > \frac{1}{2}.$$

If $n \geq p \geq N$, we have also

$$\sum_{|x|\leq\varepsilon n} Q_p(0, x) \geq \sum_{|x|\leq\varepsilon p} Q_p(0, x) > \frac{1}{2},$$

and then, by summing over $p \in \{N, \dots, n\}$, we get, for every $n \geq N$,

$$\sum_{|x| \leq \varepsilon n} U(0, x) \geq \sum_{p=N}^{n} \sum_{|x| \leq \varepsilon n} Q_p(0, x) > \frac{n - N}{2}.$$

However, by (13.3), if $\varepsilon n \geq 1$,

$$\sum_{|x| \leq \varepsilon n} U(0, x) \leq C\varepsilon n = \frac{n}{4}.$$

We get a contradiction as soon as n is large. This shows that 0 (and hence every $x \in \mathbb{Z}$) must be recurrent.

We still have to prove the last assertion. Let G be the subgroup generated by $\{x \in \mathbb{Z} : \mu(x) > 0\}$. It is immediate that

$$\mathbb{P}(Y_n \in G, \forall n \in \mathbb{Z}_+) = 1$$

(recall that we took $Y_0 = 0$). Hence, if $G \neq \mathbb{Z}$, the Markov chain is not irreducible. Conversely, assume that $G = \mathbb{Z}$. Then set

$$H = \{x \in \mathbb{Z} : U(0, x) > 0\}$$

and note that H is a subgroup of $\mathbb{Z}$:

- if $x, y \in H$, the lower bound

$$Q_{n+p}(0, x + y) \geq Q_n(0, x)\, Q_p(x, x + y) = Q_n(0, x)\, Q_p(0, y)$$

  shows that $x + y \in H$;
- if $x \in H$, the fact that 0 is recurrent and the property $U(0, x) > 0$ imply $U(x, 0) > 0$ (Lemma 13.13) and, since $U(x, 0) = U(0, -x)$, we get $-x \in H$.

Finally, since H clearly contains $\{x \in \mathbb{Z} : \mu(x) > 0\}$, we obtain that $H = \mathbb{Z}$. Recalling that $U(x, y) = U(0, y - x)$, we conclude that the chain is irreducible. □

For instance, if $\mu = \frac{1}{2}\delta_{-2} + \frac{1}{2}\delta_2$, all states are recurrent, but there are two recurrence classes, namely even integers and odd integers.

(3) **Random walk on a graph**. Consider the case of a finite graph: E is finite and A is a subset of $\mathcal{P}_2(E)$ such that $A_x := \{y \in E : \{x, y\} \in A\}$ is nonempty for every $x \in E$. The graph is said to be connected if, for every distinct $x, y \in E$, we can find an integer $p \geq 1$ and elements $x_0 = x, x_1, \dots, x_{p-1}, x_p = y$ of E such that $\{x_{i-1}, x_i\} \in A$ for every $i \in \{1, \dots, p\}$.

Proposition 13.18 *Simple random walk on a finite connected graph is recurrent irreducible.*

Proof Irreducibility immediately follows from the connectedness of the graph. Corollary 13.16 then gives the recurrence. □

(4) **Galton-Watson branching processes**. In that case, $E = \mathbb{Z}_+$ and $Q(x, y) = \mu^{*x}(y)$, where μ is a probability measure on $\mathbb{Z}_+$. We exclude the trivial case $\mu = \delta_1$.

We observe that 0 is an *absorbing* state, meaning that

$$\mathbf{P}_0(\forall n \in \mathbb{Z}_+ , \mathbf{X}_n = 0) = 1.$$

Hence 0 is (trivially) recurrent.

Proposition 13.19 *For a Galton-Watson branching process with offspring distribution $\mu \neq \delta_1$, 0 is the only recurrent state. Consequently, we have a.s.*

- *either there exists an integer N such that $\mathbf{X}_n = 0$ for every $n \geq N$;*
- *or $\mathbf{X}_n \longrightarrow +\infty$ as $n \to \infty$.*

Remark We saw in Chapter 12 that the first case (extinction of the population) occurs with probability 1 if $m = \sum k\mu(k) \leq 1$, and that the second case occurs with positive probability if $m > 1$ (we verified this under the additional assumption that $\sum k^2\mu(k) < \infty$, but this assumption can easily be removed, see Exercise 13.5).

Proof We first show that all states but 0 are transient. Consider the case $\mu(0) > 0$. If $x \geq 1$, $U(x, 0) \geq \mathbf{P}_x(\mathbf{X}_1 = 0) = \mu(0)^x > 0$ whereas $U(0, x) = 0$. This is only possible if x is transient. Suppose then that $\mu(0) = 0$. Since we exclude $\mu = \delta_1$, there exists $k \geq 2$ such that $\mu(k) > 0$, and, for every $x \geq 1$, $\mathbf{P}_x(\mathbf{X}_1 > x) > 0$, which implies that there exists $y > x$ such that $U(x, y) > 0$. Since it is clear that $U(y, x) = 0$ (in the case $\mu(0) = 0$ the population can only increase!), we conclude again that x is transient. The remaining part of the statement now follows from Theorem 13.14: if the population does not become extinct, the fact that transient states are visited only finitely many times ensures that, for every $m \geq 1$, states $1, \ldots, m$ are no longer visited after a sufficiently large time, which exactly means that $\mathbf{X}_n \to \infty$. □

By considering a branching process in the case where $\mu(0) > 0$ and $m > 1$ we see that the two possibilities in part (b) of Theorem 13.14 may both occur with positive probability: starting from a transient state, the chain may eventually reach the set of recurrent states (0 in the case of branching processes) or stay in the set of transient states.

13.5 Invariant Measures

A recurrent irreducible Markov chains visits any state infinitely many times. The notion of invariant measure will allow us to define a frequency of visit for each state, and thus to say that some states are visited more often than others.

Definition 13.20 Let μ be a (positive) measure on E, such that $\mu(x) < \infty$ for every $x \in E$ and $\mu(E) > 0$. We say that μ is invariant for the stochastic matrix Q (or simply invariant if there is no risk of confusion) if

$$\forall y \in E\,, \quad \mu(y) = \sum_{x \in E} \mu(x) Q(x, y).$$

In matrix form, the invariance property reads $\mu Q = \mu$. Since $Q_n = (Q)^n$ for every $n \geq 1$, we can iterate this relation and get that $\mu Q_n = \mu$ for every n.

Example For a random walk on $\mathbb{Z}^d$ with jump distribution γ, we know that $Q(x, y) = \gamma(y - x)$ is a function of $y - x$, and it immediately follows that the counting measure on $\mathbb{Z}^d$ is invariant.

Interpretation Assume that $\mu(E) < \infty$ (which holds automatically if E is finite). Up to replacing μ by $\mu(E)^{-1}\mu$, we can assume that $\mu(E) = 1$. Then, for every function $f : E \longrightarrow \mathbb{R}_+$,

$$\mathbf{E}_\mu[f(\mathbf{X}_1)] = \sum_{x \in E} \mu(x) \sum_{y \in E} Q(x, y) f(y) = \sum_{y \in E} f(y) \sum_{x \in E} \mu(x) Q(x, y) = \sum_{y \in E} \mu(y) f(y)$$

which shows that, under $\mathbf{P}_\mu$, $\mathbf{X}_1$ has the same law μ as $\mathbf{X}_0$. Using the relation $\mu Q_n = Q$, we get similarly that, for every $n \in \mathbb{Z}_+$ the law of $\mathbf{X}_n$ under $\mathbf{P}_\mu$ is μ. More precisely, for every nonnegative measurable function F on $\mathbf{\Omega} = E^{\mathbb{Z}_+}$,

$$\mathbf{E}_\mu[F \circ \theta_1] = \mathbf{E}_\mu[\mathbf{E}_{\mathbf{X}_1}[F]] = \sum_{x \in E} \mu(x)\, \mathbf{E}_x[F] = \mathbf{E}_\mu[F]$$

which shows that, under $\mathbf{P}_\mu$, $(\mathbf{X}_{1+n})_{n \in \mathbb{Z}_+}$ has the same law as $(\mathbf{X}_n)_{n \in \mathbb{Z}_+}$ (and similarly, for every integer $k \geq 0$, $(\mathbf{X}_{k+n})_{n \in \mathbb{Z}_+}$ has the same law as $(\mathbf{X}_n)_{n \in \mathbb{Z}_+}$).

Definition 13.21 Let μ be a measure on E such that $\mu(E) > 0$ and $\mu(x) < \infty$ for every $x \in E$. We say that μ is reversible (with respect to Q) if

$$\forall x, y \in E\,, \quad \mu(x) Q(x, y) = \mu(y) Q(y, x).$$

Proposition 13.22 *Any reversible measure is invariant.*

Proof If μ is reversible, then, for every $y \in E$,

$$\sum_{x \in E} \mu(x) Q(x, y) = \sum_{x \in E} \mu(y) Q(y, x) = \mu(y).$$

□

Conversely, there are invariant measures that are not reversible. For instance, the counting measure is invariant for any random walk on $\mathbb{Z}^d$, but it is reversible only if the jump distribution γ is symmetric ($\gamma(x) = \gamma(-x)$). In fact, reversibility is a much stronger condition than invariance. However, if a reversible measure exists, it is usually easy to find.

Examples

(a) **Biased coin-tossing**. This is the random walk on $\mathbb{Z}$ with transition matrix

$$Q(i, i+1) = p$$
$$Q(i, i-1) = q = 1 - p$$

where $p \in (0, 1)$. In that case, one immediately verifies that the measure

$$\mu(i) = \left(\frac{p}{q}\right)^i, \qquad i \in \mathbb{Z}$$

is reversible, hence invariant. Notice that μ is different from the counting measure (which is also invariant), except when $p = 1/2$.

(b) **Random walk on a graph**. The measure

$$\mu(x) = \mathrm{card}(A_x)$$

is reversible. Indeed, the property $\mu(x) Q(x, y) = \mu(y) Q(y, x)$ is trivial if $(x, y) \notin A$, and, if $\{x, y\} \in A$,

$$\mu(x) Q(x, y) = \mathrm{card}(A_x) \frac{1}{\mathrm{card}(A_x)} = 1 = \mu(y) Q(y, x).$$

(c) **Ehrenfest's urn model**. This is the Markov chain in $E = \{0, 1, \ldots, k\}$ with transition matrix

$$Q(j, j+1) = \frac{k-j}{k} \text{ if } 0 \le j \le k-1,$$
$$Q(j, j-1) = \frac{j}{k} \quad \text{ if } 1 \le j \le k.$$

This models the distribution of k particles in a box containing two compartments separated by a wall. At each integer time, a particle chosen at random crosses the wall. The Markov chain corresponds to the number of particles in the first compartment.

In that example, a measure μ is reversible if and only if

$$\mu(j)\,\frac{k-j}{k}=\mu(j+1)\,\frac{j+1}{k}$$

for $0\le j\le k-1$. This property holds for

$$\mu(j)=\binom{k}{j},\quad 0\le j\le k.$$

Let us come back to the general setting and state a first theorem.

Theorem 13.23 *Suppose that x is a recurrent point. The formula*

$$\nu(y)=\mathbf{E}_x\Big[\sum_{k=0}^{H_x-1}\mathbf{1}_{\{\mathbf{X}_k=y\}}\Big],\quad \forall y\in E, \tag{13.4}$$

defines an invariant measure. Moreover, $\nu(y)>0$ if and only if y belongs to the recurrence class of x.

Proof We first observe that, if y does not belong to the recurrence class of x, we have $\mathbf{E}_x[N_y]=U(x,y)=0$, and a fortiori $\nu(y)=0$.

We note that, in (13.4), we can replace the sum from $k=0$ to $k=H_x-1$ by a sum from $k=1$ to $k=H_x$. We use this simple observation to write, for every $y\in E$,

$$\begin{aligned}
\nu(y)&=\mathbf{E}_x\Big[\sum_{k=1}^{H_x}\mathbf{1}_{\{\mathbf{X}_k=y\}}\Big]\\
&=\sum_{z\in E}\mathbf{E}_x\Big[\sum_{k=1}^{H_x}\mathbf{1}_{\{\mathbf{X}_{k-1}=z,\,\mathbf{X}_k=y\}}\Big]\\
&=\sum_{z\in E}\sum_{k=1}^{\infty}\mathbf{E}_x\Big[\mathbf{1}_{\{k\le H_x,\,\mathbf{X}_{k-1}=z\}}\,\mathbf{1}_{\{\mathbf{X}_k=y\}}\Big]\\
&=\sum_{z\in E}\sum_{k=1}^{\infty}\mathbf{E}_x\Big[\mathbf{1}_{\{k\le H_x,\,\mathbf{X}_{k-1}=z\}}\Big]\,Q(z,y)\\
&=\sum_{z\in E}\mathbf{E}_x\Big[\sum_{k=1}^{H_x}\mathbf{1}_{\{\mathbf{X}_{k-1}=z\}}\Big]\,Q(z,y)\\
&=\sum_{z\in E}\nu(z)Q(z,y).
\end{aligned}$$

In the fourth equality, we used the fact that the event $\{k \leq H_x,\ \mathbf{X}_{k-1} = z\}$ is $\mathfrak{F}_{k-1}$-measurable to apply the Markov property at time $k - 1$.

We have thus obtained the equality $\nu Q = \nu$, which we can iterate to get $\nu Q_n = \nu$ for every integer $n \geq 0$. In particular, for every integer $n \geq 0$,

$$\sum_{z \in E} \nu(z) Q_n(z, x) = \nu(x) = 1.$$

Suppose that y belongs to the recurrence class of x. Then there exists $n \geq 0$ such that $Q_n(y, x) > 0$, and the preceding display shows that $\nu(y) < \infty$. We can also find m such that $Q_m(x, y) > 0$, and it follows that

$$\nu(y) = \sum_{z \in E} \nu(z) Q_m(z, y) \geq Q_m(x, y) > 0.$$

□

Remark If there is more than one recurrence class, then by picking a point in each class and using Theorem 13.23, we construct invariant measures with disjoint supports.

Theorem 13.24 *Assume that the Markov chain is recurrent irreducible. Then, there is only one invariant measure up to multiplication by positive constants.*

Proof Let μ be an invariant measure (Theorem 13.23 shows how to construct one). We prove by induction that, for every $p \geq 0$, for every $x, y \in E$,

$$\mu(y) \geq \mu(x)\, \mathbf{E}_x\Big[\sum_{k=0}^{p \wedge (H_x - 1)} \mathbf{1}_{\{\mathbf{X}_k = y\}}\Big]. \tag{13.5}$$

If $y = x$, this is immediate (equality even holds). So it is enough to prove (13.5) when $y \neq x$. If $p = 0$, this is trivial. Let $p \geq 0$, and assume that (13.5) holds at order p for every $x, y \in E$. Then, if $y \neq x$,

$$\begin{aligned}
\mu(y) &= \sum_{z \in E} \mu(z)\, Q(z, y) \\
&\geq \mu(x) \sum_{z \in E} \mathbf{E}_x\Big[\sum_{k=0}^{p \wedge (H_x - 1)} \mathbf{1}_{\{\mathbf{X}_k = z\}}\Big] Q(z, y) \\
&= \mu(x) \sum_{z \in E} \sum_{k=0}^{p} \mathbf{E}_x\Big[\mathbf{1}_{\{\mathbf{X}_k = z,\, k \leq H_x - 1\}}\Big] Q(z, y)
\end{aligned}$$

$$= \mu(x) \sum_{z \in E} \sum_{k=0}^{p} \mathbf{E}_x \Big[\mathbf{1}_{\{\mathbf{X}_k = z,\, k \le H_x - 1\}} \mathbf{1}_{\{\mathbf{X}_{k+1} = y\}} \Big]$$

$$= \mu(x) \mathbf{E}_x \Big[\sum_{k=0}^{p \wedge (H_x - 1)} \mathbf{1}_{\{\mathbf{X}_{k+1} = y\}} \Big]$$

$$= \mu(x) \mathbf{E}_x \Big[\sum_{k=1}^{(p+1) \wedge H_x} \mathbf{1}_{\{\mathbf{X}_k = y\}} \Big],$$

which gives the desired result at order $p + 1$. Note that, as in the proof of Theorem 13.23, we used the fact that $\{\mathbf{X}_k = z,\ k \le H_x - 1\}$ is $\mathfrak{F}_k$-measurable to apply the Markov property at time k.

By letting $p \to +\infty$ in (13.5) we get

$$\mu(y) \ge \mu(x)\, \mathbf{E}_x \Big[\sum_{k=0}^{H_x - 1} \mathbf{1}_{\{\mathbf{X}_k = y\}} \Big].$$

Fix $x \in E$. The measure

$$\nu_x(y) = \mathbf{E}_x \Big[\sum_{k=0}^{H_x - 1} \mathbf{1}_{\{\mathbf{X}_k = y\}} \Big]$$

is invariant (Theorem 13.23), and we have $\mu(y) \ge \mu(x)\nu_x(y)$ for every $y \in E$. Hence, for every $n \ge 1$,

$$\mu(x) = \sum_{z \in E} \mu(z) Q_n(z, x) \ge \sum_{z \in E} \mu(x) \nu_x(z) Q_n(z, x) = \mu(x)\nu_x(x) = \mu(x).$$

It follows that the inequality in the last display is an equality, which means that $\mu(z) = \mu(x)\nu_x(z)$ holds for every z such that $Q_n(z, x) > 0$. Irreducibility ensures that, for every $z \in E$, we can find an integer n such that $Q_n(z, x) > 0$, and we conclude that $\mu = \mu(x)\nu_x$, which completes the proof. □

Corollary 13.25 *Assume that the chain is recurrent irreducible. Then:*

(i) Either there exists an invariant probability measure μ, and we have, for every $x \in E$,

$$\mathbf{E}_x[H_x] = \frac{1}{\mu(x)}.$$

(ii) Or all invariant measures have infinite mass, and, for every $x \in E$,

$$\mathbf{E}_x[H_x] = \infty.$$

The chain is said to be positive recurrent in case (i) and null recurrent in case (ii).

Remark If E is finite, only case (i) can occur.

Proof By Theorem 13.24, all invariant measures are proportional. Hence, either they all have infinite mass (case (ii)) or they are all finite, and one of them is a probability measure (case (i)). In case (i), let μ be the unique invariant probability measure, and let $x \in E$. Then, if ν_x denotes the invariant measure constructed in Theorem 13.23,

$$\nu_x(y) = \mathbf{E}_x\Big[\sum_{k=0}^{H_x-1} \mathbf{1}_{\{\mathbf{X}_k=y\}}\Big],$$

we know that μ is proportional to ν_x: $\mu = C\nu_x$ with $C > 0$. Writing $1 = \mu(E) = C\,\nu_x(E)$, we find $C = (\nu_x(E))^{-1}$, and thus

$$\mu(x) = \frac{\nu_x(x)}{\nu_x(E)} = \frac{1}{\nu_x(E)}.$$

However,

$$\nu_x(E) = \sum_{y\in E} \mathbf{E}_x\Big[\sum_{k=0}^{H_x-1} \mathbf{1}_{\{\mathbf{X}_k=y\}}\Big] = \mathbf{E}_x\Big[\sum_{k=0}^{H_x-1}\Big(\sum_{y\in E} \mathbf{1}_{\{\mathbf{X}_k=y\}}\Big)\Big] = \mathbf{E}_x[H_x],$$

which gives the desired formula for $\mathbf{E}_x[H_x]$. In case (ii), ν_x is infinite, and the calculation in the last display gives $\mathbf{E}_x[H_x] = \nu_x(E) = \infty$. □

The next proposition gives a useful criterion to prove the recurrence of a Markov chain.

Proposition 13.26 *Assume that the chain is irreducible. If there exists a finite invariant measure, then the chain is recurrent (and thus positive recurrent).*

Proof Let γ be a finite invariant measure, and let $y \in E$ such that $\gamma(y) > 0$. For every $x \in E$, Proposition 13.12(iii) gives the bound

$$\sum_{n=0}^{\infty} Q_n(x, y) = U(x, y) \leq U(y, y).$$

We multiply both sides of the preceding display by $\gamma(x)$ and sum over all $x \in E$. We get

$$\sum_{n=0}^{\infty} \gamma Q_n(y) \le \gamma(E)\, U(y,y).$$

Since γ is invariant, we have $\gamma Q_n(y) = \gamma(y) > 0$ for every $n \ge 0$. We conclude that

$$\gamma(E)\, U(y,y) = \infty.$$

Since $\gamma(E) < \infty$ by assumption, it follows that $U(y,y) = \infty$. Hence y is recurrent. The desired result then follows from Corollary 13.16. □

Remark The existence of an infinite invariant measure does not allow one to prove recurrence. For instance, we have seen that the counting measure is invariant for any random walk on $\mathbb{Z}^d$, but many of these random walks are irreducible without being recurrent.

Example Let $p \in (0,1)$. Consider the Markov chain on $E = \mathbb{Z}_+$ with transition matrix

$$Q(k,k+1) = p\ ,\ \ Q(k,k-1) = 1-p\ , \qquad \text{if } k \ge 1,$$
$$Q(0,1) = 1.$$

This chain is irreducible. Furthermore, one immediately verifies that the measure μ defined by

$$\mu(k) = \Big(\frac{p}{1-p}\Big)^{k-1}\ , \qquad \text{if } k \ge 1,$$
$$\mu(0) = 1-p\ ,$$

is reversible hence invariant.

If $p < 1/2$, μ is a finite measure, and Proposition 13.26 implies that the chain is positive recurrent (we will see later that it is null recurrent when $p = 1/2$ and transient when $p > 1/2$).

13.6 Ergodic Theorems

We keep on considering the canonical Markov chain associated with a transition matrix Q.

Theorem 13.27 *Assume that the chain is recurrent irreducible, and let μ be an invariant measure. Let f and g be two nonnegative measurable functions on E such that $\int f\,\mathrm{d}\mu < \infty$ and $0 < \int g\,\mathrm{d}\mu < \infty$. Then, for every $x \in E$, we have, $\mathbf{P}_x$ a.s.,*

$$\frac{\sum_{k=0}^{n} f(\mathbf{X}_k)}{\sum_{k=0}^{n} g(\mathbf{X}_k)} \underset{n\to\infty}{\longrightarrow} \frac{\int f\,\mathrm{d}\mu}{\int g\,\mathrm{d}\mu}.$$

Remark The result still holds when $\int f\mathrm{d}\mu = \infty$: just use a comparison argument by writing $f = \lim \uparrow f_k$, with an increasing sequence $(f_k)_{k\in\mathbb{N}}$ of nonnegative functions such that $\int f_k\mathrm{d}\mu < \infty$ for every k.

Corollary 13.28 *If the chain is irreducible and positive recurrent, and if μ denotes the unique invariant probability measure, then, for any nonnegative measurable function f on E and for every $x \in E$, we have, $\mathbf{P}_x$ a.s.,*

$$\frac{1}{n}\sum_{k=0}^{n} f(\mathbf{X}_k) \underset{n\to\infty}{\longrightarrow} \int f\,\mathrm{d}\mu.$$

This corollary immediately follows from Theorem 13.27 by taking $g = 1$ in this statement.

Proof of Theorem 13.27 We fix $x \in E$. We set

$$T_0 = 0\ ,\ T_1 = H_x$$

and then by induction,

$$T_{n+1} = \inf\{k > T_n : \mathbf{X}_k = x\}.$$

It is straightforward to verify that T_n is a stopping time for every $n \geq 0$. Since x is recurrent, we also know that $T_n < \infty$ for every $n \geq 0$, $\mathbf{P}_x$ a.s. Let us set, for every $k \geq 0$,

$$Z_k(f) = \sum_{n=T_k}^{T_{k+1}-1} f(\mathbf{X}_n).$$

□

Lemma 13.29 *The random variables $Z_k(f)$, $k = 0, 1, 2, \ldots$, are independent and identically distributed under $\mathbf{P}_x$.*

Proof Let $g_0, g_1, g_2, \dots$ be bounded measurable functions on $\mathbb{R}_+$. It is enough to prove that, for every integer $k \geq 0$, we have

$$\mathbf{E}_x\Big[\prod_{i=0}^{k} g_i(Z_i(f))\Big] = \prod_{i=0}^{k} \mathbf{E}_x[g_i(Z_0(f))].$$

We prove this by induction on k. For $k = 0$ there is nothing to prove. Suppose that the preceding display holds at order $k-1$, for some $k \geq 1$. Then we observe that:

- the random variables $Z_0(f), Z_1(f), \dots, Z_{k-1}(f)$ are $\mathfrak{F}_{T_k}$-measurable (we leave this as an exercise for the reader);
- $\theta_{T_k}(\omega)$ is independent of $\mathfrak{F}_{T_k}$ and distributed according to $\mathbf{P}_x$, by Corollary 13.9;
- we have $Z_k(f) = Z_0(f) \circ \theta_{T_k}$, by construction.

It follows from these observations that

$$\begin{aligned}\mathbf{E}_x\Big[\prod_{i=0}^{k} g_i(Z_i(f))\Big] &= \mathbf{E}_x\Big[\Big(\prod_{i=0}^{k-1} g_i(Z_i(f))\Big) g_k(Z_0(f) \circ \theta_{T_k})\Big] \\ &= \mathbf{E}_x\Big[\prod_{i=0}^{k-1} g_i(Z_i(f))\Big] \mathbf{E}_x[g_k(Z_0(f))],\end{aligned}$$

which gives the desired result at order k. □

Let us return to the proof of the theorem. As previously we write ν_x for the invariant measure defined in formula (13.4). Then $\mu = \mu(x)\nu_x$ since $\nu_x(x) = 1$ and we know from Theorem 13.24 that all invariant measures are proportional. We then observe that

$$\mathbf{E}_x[Z_0(f)] = \mathbf{E}_x\Big[\sum_{k=0}^{H_x-1} \sum_{y \in E} f(y)\, \mathbf{1}_{\{\mathbf{X}_k = y\}}\Big] = \sum_{y \in E} f(y)\, \nu_x(y) = \frac{\int f \,\mathrm{d}\mu}{\mu(x)}.$$

Lemma 13.29 and the strong law of large numbers then show that, $\mathbf{P}_x$ a.s.

$$\frac{1}{n} \sum_{k=0}^{n-1} Z_k(f) \underset{n\to\infty}{\longrightarrow} \frac{\int f \,\mathrm{d}\mu}{\mu(x)}. \tag{13.6}$$

For every integer $n \geq 0$, write $N_x(n)$ for the number of returns to x for the chain up to time n, $N_x(n) = \sum_{k=1}^{n} \mathbf{1}_{\{\mathbf{X}_k = n\}}$. We have then $T_{N_x(n)} \leq n < T_{N_x(n)+1}$. Writing

$$\frac{\sum_{k=0}^{T_{N_x(n)}-1} f(\mathbf{X}_k)}{N_x(n)} \leq \frac{\sum_{k=0}^{n} f(\mathbf{X}_k)}{N_x(n)} \leq \frac{\sum_{k=0}^{T_{N_x(n)+1}-1} f(\mathbf{X}_k)}{N_x(n)},$$

or equivalently

$$\frac{\sum_{j=0}^{N_x(n)-1} Z_j(f)}{N_x(n)} \leq \frac{\sum_{k=0}^{n} f(\mathbf{X}_k)}{N_x(n)} \leq \frac{\sum_{j=0}^{N_x(n)} Z_j(f)}{N_x(n)},$$

we deduce from the convergence (13.6) and the property $N_x(n) \longrightarrow \infty$ as $n \to \infty$ that, $\mathbf{P}_x$ a.s.,

$$\frac{1}{N_x(n)} \sum_{k=0}^{n} f(\mathbf{X}_k) \underset{n\to\infty}{\longrightarrow} \frac{\int f \, \mathrm{d}\mu}{\mu(x)}.$$

The proof is then completed by using the same result with f replaced by g. □

Corollary 13.30 *Assume that the chain is recurrent irreducible. Then, for every $y \in E$,*

(i) in the positive recurrent case,

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{\mathbf{X}_k = y\}} \underset{n\to\infty}{\longrightarrow} \mu(y), \quad a.s.$$

where μ is the unique invariant probability measure;

(ii) in the null recurrent case,

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{\mathbf{X}_k = y\}} \underset{n\to\infty}{\longrightarrow} 0, \quad a.s.$$

In both cases, the convergence holds a.s. under $\mathbf{P}_x$ for any $x \in E$.

The corollary is an immediate consequence of Theorem 13.27. We just apply this result with $g = \mathbf{1}_{\{y\}}$ and $f = 1$, using the remark after the statement of the theorem in the null recurrent case.

Definition 13.31 Let x be a recurrent point, and

$$L_x = \{n \geq 0 : Q_n(x, x) > 0\}.$$

The period of x (relative to Q) is the largest common divisor to L_x. It is denoted by $d(x)$.

Remark Since L_x is closed under addition ($Q_{n+m}(x, x) \geq Q_n(x, x)Q_m(x, x)$), the subgroup generated by L_x can be written in the form $d(x)\mathbb{Z} = L_x - L_x (= \{n - m : n, m \in L_x\})$.

Proposition 13.32 *Assume that the chain is recurrent irreducible.*

(i) All points of E have the same period, which is called the period of the chain and denoted by d.
(ii) The chain is called aperiodic if $d = 1$. In that case, for every $x, y \in E$, there exists an integer $n_0 \geq 0$ such that $Q_n(x, y) > 0$ for every $n \geq n_0$.

Proof

(i) Let $x, y \in E$. Since the chain is irreducible, there exist two integers n_1 and n_2 such that $Q_{n_1}(x, y) > 0$ and $Q_{n_2}(y, x) > 0$. But then, if $n \in L_x$, we have $Q_{n_2+n+n_1}(y, y) \geq Q_{n_2}(y, x)Q_n(x, x)Q_{n_1}(x, y) > 0$ and thus $n_1 + n + n_2 \in L_y$, which implies that $L_x - L_x \subset L_y - L_y$ and thus $d(y)$ is a divisor of $d(x)$. By symmetry, we get $d(x) = d(y)$.
(ii) It is clearly enough to treat the case where $y = x$. Since $d(x) = 1$, we can find two integers $n_1, m_1 \geq 0$ such that $1 = n_1 - m_1$ and

$$Q_{n_1}(x, x) > 0, \;\; Q_{m_1}(x, x) > 0.$$

If $m_1 = 0$, then $n_1 = 1$ and the result is obvious with $n_0 = 0$. If $m_1 \geq 1$, then, for every $j \in \{0, 1, \dots, m_1 - 1\}$, we have

$$Q_{m_1^2+j}(x, x) = Q_{jn_1+(m_1-j)m_1}(x, x) > 0.$$

Since $Q_{m_1}(x, x) > 0$, we also get $Q_{m_1^2+km_1+j}(x, x) > 0$ for every integers $k \geq 0$ and $j \in \{0, 1, \dots, m_1 - 1\}$, and finally we have for every $j \geq 0$,

$$Q_{m_1^2+j}(x, x) > 0.$$

□

Theorem 13.33 *Assume that the chain is irreducible, positive recurrent and aperiodic, and let μ denote the unique invariant probability measure. Then, for every $x \in E$,*

$$\sum_{y \in E} |\mathbf{P}_x(\mathbf{X}_n = y) - \mu(y)| \underset{n\to\infty}{\longrightarrow} 0.$$

In particular, we have $\mathbf{P}_x(\mathbf{X}_n = y) \longrightarrow \mu(y)$ as $n \to \infty$ for every $x, y \in E$.

Proof The formula

$$\overline{Q}((x_1, x_2), (y_1, y_2)) = Q(x_1, y_1)Q(x_2, y_2)$$

defines a stochastic matrix on $E \times E$, and we have

$$\overline{Q}_n((x_1, x_2), (y_1, y_2)) = Q_n(x_1, y_1)Q_n(x_2, y_2)$$

for every integer $n \geq 1$. Let $((\mathbf{X}^1_n, \mathbf{X}^2_n)_{n\in\mathbb{Z}_+}, (\overline{\mathbf{P}}_{(x_1,x_2)})_{(x_1,x_2)\in E\times E})$ denote the canonical Markov chain associated with $\overline{Q}$. It is straightforward to verify that, under $\overline{\mathbf{P}}_{(x_1,x_2)}$, $\mathbf{X}^1_n$ and $\mathbf{X}^2_n$ are two independent Markov chains with transition matrix Q started at x_1 and x_2 respectively.

We observe that $\overline{Q}$ is irreducible. Indeed, if $(x_1, x_2), (y_1, y_2) \in E \times E$, Proposition 13.32(ii) gives the existence of two integers n_1 and n_2 such that $Q_n(x_1, y_1) > 0$ for every $n \geq n_1$, and $Q_n(x_2, y_2) > 0$ for every $n \geq n_2$. If $n \geq n_1 \vee n_2$, we have thus $\overline{Q}_n((x_1, x_2), (y_1, y_2)) > 0$.

Moreover, the product measure $\mu \otimes \mu$ is invariant for $\overline{Q}$:

$$\begin{aligned}\sum_{(x_1,x_2)\in E\times E} &\mu(x_1)\mu(x_2)Q(x_1, y_1)Q(x_2, y_2)\\ &= \sum_{x_1\in E} \mu(x_1)Q(x_1, y_1) \sum_{x_2\in E} \mu(x_2)Q(x_2, y_2)\\ &= \mu(y_1)\mu(y_2).\end{aligned}$$

By Proposition 13.26, we get that the Markov chain $(\mathbf{X}^1_n, \mathbf{X}^2_n)$ is positive recurrent.

Let us fix $x \in E$. Then, for every $y \in E$,

$$\begin{aligned}\mathbf{P}_x(\mathbf{X}_n = y) - \mu(y) &= \overline{\mathbf{P}}_{\mu\otimes\delta_x}(\mathbf{X}^2_n = y) - \overline{\mathbf{P}}_{\mu\otimes\delta_x}(\mathbf{X}^1_n = y)\\ &= \overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{\mathbf{X}^2_n=y\}} - \mathbf{1}_{\{\mathbf{X}^1_n=y\}}].\end{aligned}$$

Consider the stopping time $T = \inf\{n \geq 0 : \mathbf{X}^1_n = \mathbf{X}^2_n\}$. The recurrence of the chain $(\mathbf{X}^1_n, \mathbf{X}^2_n)$ ensures that $T < \infty$, $\overline{\mathbf{P}}_{\mu\otimes\delta_x}$ a.s. Then the preceding equality implies that, for every $y \in E$,

$$\begin{aligned}\mathbf{P}_x(\mathbf{X}_n = y) - \mu(y) &= \overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T>n\}}(\mathbf{1}_{\{\mathbf{X}^2_n=y\}} - \mathbf{1}_{\{\mathbf{X}^1_n=y\}})]\\ &\quad + \sum_{k=0}^{n} \sum_{z\in E} \overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T=k,\mathbf{X}^1_k=\mathbf{X}^2_k=z\}}(\mathbf{1}_{\{\mathbf{X}^2_n=y\}} - \mathbf{1}_{\{\mathbf{X}^1_n=y\}})].\end{aligned} \tag{13.7}$$

However, for every $k \in \{0, 1, \ldots, n\}$ and $z \in E$, the Markov property at time k gives

$$\begin{aligned}\overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T=k,\mathbf{X}^1_k=\mathbf{X}^2_k=z\}}\mathbf{1}_{\{\mathbf{X}^2_n=y\}}] &= \overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T=k,\mathbf{X}^1_k=\mathbf{X}^2_k=z\}}]\, Q_{n-k}(z, y)\\ &= \overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T=k,\mathbf{X}^1_k=\mathbf{X}^2_k=z\}}\mathbf{1}_{\{\mathbf{X}^1_n=y\}}],\end{aligned}$$

and thus the second term in the right-hand side of (13.7) is equal to 0. In this way, we obtain that

$$\begin{aligned}\sum_{y\in E} |\mathbf{P}_x(\mathbf{X}_n = y) - \mu(y)| &= \sum_{y\in E} |\overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T>n\}}(\mathbf{1}_{\{\mathbf{X}^2_n=y\}} - \mathbf{1}_{\{\mathbf{X}^1_n=y\}})]| \\ &\leq \sum_{y\in E} \overline{\mathbf{E}}_{\mu\otimes\delta_x}[\mathbf{1}_{\{T>n\}}(\mathbf{1}_{\{\mathbf{X}^2_n=y\}} + \mathbf{1}_{\{\mathbf{X}^1_n=y\}})] \\ &= 2\,\overline{\mathbf{P}}_{\mu\otimes\delta_x}(T > n),\end{aligned}$$

which tends to 0 as $n \to \infty$, since $T < \infty$, $\overline{\mathbf{P}}_{\mu\otimes\delta_x}$ a.s. □

Remark The preceding proof is a good example of the so-called *coupling method*. Roughly speaking, this method involves constructing two random variables, or two random processes, with prescribed distributions, in such a way that they are almost surely close, in some appropriate sense. Although this is not made explicit in the argument, the underlying idea of the preceding proof is to construct the Markov chain started from x and the same Markov chain started from its invariant probability measure, in such a way that these two processes coincide after the (finite) stopping time T.

13.7 Martingales and Markov Chains

In this last section, we discuss some relations between Markov chains and martingale theory. We consider again the canonical Markov chain with transition matrix Q.

Definition 13.34 A function $f : E \longrightarrow \mathbb{R}_+$ is said to be Q-harmonic (resp. Q-superharmonic) if, for every $x \in E$,

$$f(x) = Qf(x) \qquad \text{(resp. } f(x) \geq Qf(x)\text{)}.$$

More generally, if $F \subset E$, we say that f is Q-harmonic on F (resp. Q-superharmonic on F) if the property $f(x) = Qf(x)$ (resp. $f(x) \geq Qf(x)$) holds for every $x \in F$.

Remark We could more generally consider harmonic and superharmonic functions of arbitrary sign, but in this book we limit ourselves to nonnegative functions.

Proposition 13.35

(i) *A function $f : E \longrightarrow \mathbb{R}_+$ is Q-harmonic (resp. Q-superharmonic) if and only if, for every $x \in E$, the process $(f(\mathbf{X}_n))_{n\in\mathbb{Z}_+}$ is a martingale (resp. a supermartingale) under $\mathbf{P}_x$, with respect to the filtration $(\mathfrak{F}_n)_{n\in\mathbb{Z}_+}$.*

(ii) *Let $F \subset E$ and $G = E\backslash F$. Define the stopping time*

$$T_G := \inf\{n \geq 0 : \mathbf{X}_n \in G\}.$$

Then, if f is Q-harmonic (resp. if f is Q-superharmonic) on F, the process $(f(\mathbf{X}_{n\wedge T_G}))_{n\in\mathbb{Z}_+}$ is a martingale (resp. a supermartingale) under $\mathbf{P}_x$, for every $x \in F$.

Proof

(i) Suppose that f is Q-harmonic. Then, by Proposition 13.3(i), for every $n \geq 0$,

$$\mathbf{E}_x[f(\mathbf{X}_{n+1})\,|\,\mathfrak{F}_n] = Qf(\mathbf{X}_n) = f(\mathbf{X}_n).$$

In particular, we have $\mathbf{E}_x[f(\mathbf{X}_n)] = \mathbf{E}_x[f(\mathbf{X}_0)] = f(x)$, hence $f(\mathbf{X}_n) \in L^1(\mathbf{P}_x)$, for every $n \in \mathbb{Z}_+$, and the preceding equality shows that $(f(\mathbf{X}_n))_{n\in\mathbb{Z}_+}$ is a martingale under $\mathbf{P}_x$ (with respect to the filtration $(\mathfrak{F}_n)_{n\in\mathbb{Z}_+}$).

Conversely, suppose that $f(\mathbf{X}_n)$ is a martingale under $\mathbf{P}_x$. We immediately get that

$$f(x) = \mathbf{E}_x[f(\mathbf{X}_0)] = \mathbf{E}_x[f(\mathbf{X}_1)] = Qf(x).$$

The case of a superharmonic function is treated in a similar manner.

(ii) Suppose that f is Q-harmonic on F. Then, for every $x \in F$ and $n \geq 0$,

$$\begin{aligned}
\mathbf{E}_x[f(\mathbf{X}_{(n+1)\wedge T_G})\,|\,\mathfrak{F}_n] &= \mathbf{E}_x[f(\mathbf{X}_{n+1})\mathbf{1}_{\{T_G>n\}}\,|\,\mathfrak{F}_n] + \mathbf{E}_x[f(\mathbf{X}_{T_G})\mathbf{1}_{\{T_G\leq n\}}\,|\,\mathfrak{F}_n] \\
&= \mathbf{1}_{\{T_G>n\}}\,\mathbf{E}_x[f(\mathbf{X}_{n+1})\,|\,\mathfrak{F}_n] + f(\mathbf{X}_{T_G})\,\mathbf{1}_{\{T_G\leq n\}} \\
&= \mathbf{1}_{\{T_G>n\}}\,Qf(\mathbf{X}_n) + f(\mathbf{X}_{T_G})\mathbf{1}_{\{T_G\leq n\}} \\
&= \mathbf{1}_{\{T_G>n\}}\,f(\mathbf{X}_n) + f(\mathbf{X}_{T_G})\mathbf{1}_{\{T_G\leq n\}} \\
&= f(\mathbf{X}_{n\wedge T_G}).
\end{aligned}$$

In the second equality, we noticed that $f(\mathbf{X}_{T_G})\,\mathbf{1}_{\{T_G\leq n\}} = f(\mathbf{X}_{T_G\wedge n})\,\mathbf{1}_{\{T_G\leq n\}}$ is $\mathfrak{F}_n$-measurable. As in part (i), it follows from the last display that $\mathbf{E}_x[f(\mathbf{X}_{n\wedge T_G})] = f(x) < \infty$ and that $f(\mathbf{X}_{n\wedge T_G})$ is a martingale. The case where f is Q-superharmonic on F is treated similarly.

□

Theorem 13.36 *Let F be a nonempty proper subset of E and $G = E\backslash F$. Let $g : G \longrightarrow \mathbb{R}_+$ be a bounded function.*

(i) The function

$$h(x) = \mathbf{E}_x[g(\mathbf{X}_{T_G})\,\mathbf{1}_{\{T_G<\infty\}}], \qquad x \in E$$

is Q-harmonic on F.

(ii) Assume that $T_G < \infty$, $\mathbf{P}_x$ a.s. for every $x \in F$. Then the function h is the unique bounded function on E that

- *is Q-harmonic on F,*
- *coincides with g on G.*

The case $g = 1$ is especially useful.

Corollary 13.37 *Let G be a nonempty proper subset of E. The function $x \mapsto \mathbf{P}_x(T_G < \infty)$ is Q-harmonic on $E\backslash G$.*

Proof of Theorem 13.36

(i) We first notice that, if $x \in F$, we have $\mathbf{P}_x$ a.s.

$$g(\mathbf{X}_{T_G})\,\mathbf{1}_{\{T_G<\infty\}} = g(\mathbf{X}_{T_G} \circ \theta_1)\,\mathbf{1}_{\{T_G\circ\theta_1<\infty\}},$$

because, if $\omega_0 \notin G$, the first point in the sequence $(\omega_1, \omega_2, \ldots)$ that belongs to G is the same as in the sequence $(\omega_0, \omega_1, \ldots)$. Put differently, if $U(\omega) = g(\mathbf{X}_{T_G}(\omega))\,\mathbf{1}_{\{T_G(\omega)<\infty\}}$, we have $U = U \circ \theta_1$, $\mathbf{P}_x$ a.s. Hence, for every $x \in F$, Theorem 13.7 gives

$$h(x) = \mathbf{E}_x[U] = \mathbf{E}_x[U \circ \theta_1] = \mathbf{E}_x[\mathbf{E}_{\mathbf{X}_1}[U]] = \mathbf{E}_x[h(\mathbf{X}_1)] = Qh(x),$$

which shows that h is Q-harmonic on F.

(ii) It is obvious that $h(x) = g(x)$ if $x \in G$, and, by part (i), we know that h is Q-harmonic on F. Let h' be a bounded function that is Q-harmonic on F and coincides with g on G. Let $x \in F$. By Proposition 13.35, $Y_n = h'(\mathbf{X}_{n\wedge T_G})$ is a martingale under $\mathbf{P}_x$. This martingale is bounded and therefore uniformly integrable, and converges to $h'(\mathbf{X}_{T_G}) = g(\mathbf{X}_{T_G})$, $\mathbf{P}_x$ a.s. By Theorem 12.28, we have thus

$$h'(x) = \mathbf{E}_x[Y_0] = \mathbf{E}_x[Y_\infty] = \mathbf{E}_x[g(\mathbf{X}_{T_G})] = h(x).$$

□

Examples

(a) Let us consider the gambler's ruin problem already discussed in Section 12.6. We consider simple random walk on $\mathbb{Z}$ and an integer $m \geq 2$. We set

$$T = \inf\{n \geq 0 : \mathbf{X}_n \leq 0 \text{ or } \mathbf{X}_n \geq m\} = \inf\{n \geq 0 : \mathbf{X}_n \notin F\},$$

where $F = \{1, 2, \ldots, m-1\}$. Write $G = \mathbb{Z} \setminus F$ as above. Define $g(i) = 1$ if $i \geq m$ and $g(i) = 0$ if $i \leq 0$. Then Theorem 13.36 shows that the function

$$h(k) = \mathbf{E}_k[g(\mathbf{X}_T)] = \mathbf{P}_k(\mathbf{X}_T \geq m), \quad k \in \mathbb{Z}$$

is Q-harmonic on F. The property $Qh(k) = h(k)$ for $k \in F$ holds if $h(k) = \frac{1}{2}(h(k-1) + h(k+1))$, or equivalently $h(k) - h(k-1) = h(k+1) - h(k)$, for every $k \in F$. Since $h(k) = 1$ if $k \geq m$ and $h(k) = 0$ if $k \leq 0$, we easily get that, for every $k \in F = \{1, \ldots, m-1\}$,

$$h(k) = \frac{k}{m}.$$

This formula was already obtained in Section 12.6 via a different method.

(b) *Discrete Dirichlet problem*. Let F be a finite subset of $\mathbb{Z}^d$. Define the boundary ∂F by

$$\partial F := \{y \in \mathbb{Z}^d \setminus F : \exists x \in F,\ |y - x| = 1\},$$

and set $\overline{F} = F \cup \partial F$.

A function h defined on $\overline{F}$ is called discrete harmonic on F if, for every $x \in F$, the value $h(x)$ of h at x is equal to its mean over the $2d$ nearest neighbors of x. This is equivalent to saying that h is Q-harmonic on F, as defined in Definition 13.34, if Q is the transition matrix of simple random walk on $\mathbb{Z}^d$: $Q(x, x \pm e_j) = \frac{1}{2d}$ for $j = 1, \ldots, d$, where $(e_1, \ldots, e_d)$ is the canonical basis of $\mathbb{R}^d$. Note that Definition 13.34 requires h to be defined on $\mathbb{Z}^d$ and not only on $\overline{F}$, but the values of h on $\mathbb{Z}^d \setminus \overline{F}$ are irrelevant to verify that h is Q-harmonic on F.

Theorem 13.36 then yields the following result. For any nonnegative function g defined on ∂F, the unique function $h : \overline{F} \longrightarrow \mathbb{R}_+$ such that:

- h is discrete harmonic on F,
- $h(y) = g(y), \forall y \in \partial F$,

is given by

$$h(x) = \mathbf{E}_x[g(\mathbf{X}_{T_{\partial F}})], \qquad x \in F,$$

where $\mathbf{P}_x$ refers to simple random walk on $\mathbb{Z}^d$ started from x, and

$$T_{\partial F} = \inf\{n \geq 0 : \mathbf{X}_n \in \partial F\}.$$

Harmonic and superharmonic functions provide criteria allowing one to decide whether an irreducible Markov chain is recurrent or transient. We give one result in this direction. Recall that superharmonic functions are always nonnegative for us.

Theorem 13.38 *Assume that the chain is irreducible. Then it is transient if and only if there exists a nonconstant Q-superharmonic function.*

Proof Assume that the chain is recurrent, and let h be a (nonnegative) Q-superharmonic function. Also let $x, y \in E$. By Proposition 13.35, the process $(h(\mathbf{X}_n))_{n\in\mathbb{Z}_+}$ is a nonnegative supermartingale under $\mathbf{P}_x$, and thus converges $\mathbf{P}_x$ a.s. Since the chain visits both x and y infinitely many times, $h(\mathbf{X}_n)$ takes both values $h(x)$ and $h(y)$ infinitely many times, and it follows that $h(x) = h(y)$. So, if the chain is recurrent, any Q-superharmonic function is constant.

Conversely, assume that the chain is transient and fix $x \in E$. Consider the function $h(y) = \mathbf{P}_y(T_{\{x\}} < \infty)$. By Corollary 13.37, h is harmonic on $E\backslash\{x\}$. But since $h \leq 1$ and $h(x) = 1$, we have also $h(x) \geq Qh(x)$, and thus h is superharmonic on E. Finally it is clear that h is not constant (if we had $h(y) = 1$ for every $y \in E$ an application of the Markov property at time 1 would show that $\mathbf{P}_x(H_x < \infty) = 1$ and x would be recurrent). □

Example Let us consider the example given at the end of Section 13.5. For this Markov chain, one immediately checks that the function

$$h(k) = \Big(\frac{1-p}{p}\Big)^k\ ,\ k \in \mathbb{Z}_+$$

is Q-superharmonic when $p \geq 1/2$. The function h is not constant when $p > 1/2$, and so the chain is transient if $p > 1/2$. This argument complements the one given at the end of Section 13.5, which showed positive recurrence when $p < 1/2$. When $p = 1/2$, the chain is (null) recurrent: recurrence can be obtained by observing that $Q_n(0,0) = \widetilde{Q}_n(0,0)$, where $\widetilde{Q}$ denotes the transition matrix of simple random walk on $\mathbb{Z}$.

13.8 Exercises

Exercise 13.1 Let E and F be two countable sets and let $f : E \longrightarrow F$ be a surjective map. Let $(X_n)_{n\in\mathbb{Z}_+}$ be a Markov chain on E with transition matrix Q. Assume that, for every $a \in F$, and every $x, x' \in E$, the equality

$$\sum_{y\in f^{-1}(a)} Q(x, y) = \sum_{y\in f^{-1}(a)} Q(x', y)$$

holds whenever $f(x) = f(x')$. Show that $(f(X_n))_{n\in\mathbb{Z}_+}$ is a Markov chain on F and give its transition matrix.

Exercise 13.2 (Random Walk on the Binary Tree) Consider the countable set

$$E = \bigcup_{k=0}^{\infty} \{1, 2\}^k,$$

where we make the convention that $\{1, 2\}^0 = \{\dagger\}$ consists of a single element $\dagger$. An element of E other than $\dagger$ is thus a k-tuple $(i_1, \ldots, i_k)$, where $i_1, \ldots, i_k \in \{1, 2\}$, and we define $\pi((i_1, \ldots, i_k)) = (i_1, \ldots, i_{k-1})$ ($= \dagger$ when $k = 1$). We view E as a graph whose edge set is $A = \{\{x, \pi(x)\} : x \in E\backslash\{\dagger\}\}$. Prove that simple random walk on this graph is transient. (*Hint:* Use the preceding exercise.)

Exercise 13.3 Let $(X_n)_{n\in\mathbb{Z}_+}$ be a simple random walk on $\mathbb{Z}$, with $X_0 = 0$. For every $k \in \mathbb{Z}_+$, set $T_k = \inf\{n \geq 0 : X_n = k\}$. Prove that the random variables $T_k - T_{k-1}$, $k \in \mathbb{N}$, are independent and identically distributed.

Exercise 13.4 Let S be a countable set, and let $(G, \mathcal{G})$ be a measurable space. Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in G, and let $\phi : S \times G \longrightarrow S$ be a measurable function. Also fix $x \in S$. We define a random process $(X_n)_{n\in\mathbb{Z}_+}$ with values in S by induction, by setting $X_0 = x$ and then, for every $n \geq 0$, $X_{n+1} = \phi(X_n, Z_{n+1})$. Prove that $(X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain, and determine its transition matrix in terms of the distribution of the random variables Z_n.

Exercise 13.5 Consider the Galton-Watson process $(X_n)_{n\in\mathbb{N}}$ of Section 13.2.4, and assume that $X_0 = 1$ and that the offspring distribution μ satisfies $\mu(0) + \mu(1) < 1$. Write $m = \sum_{k=0}^{\infty} k\mu(k)$ for the mean of μ, and g for the generating function of μ,

$$g(r) = \sum_{k=0}^{\infty} \mu(k)\, r^k\,, \quad \forall r \in [0, 1].$$

(1) Verify that, for every $n \in \mathbb{N}$ and $r \in [0, 1]$, $\mathbb{E}[r^{X_n}] = g_n(r)$, where g_n is the n-th iterate of g ($g_1 = g$ and $g_{n+1} = g \circ g_n$ for every $n \in \mathbb{N}$).
(2) Set $T_0 = \inf\{n \geq 0 : X_n = 0\}$. Verify that $\mathbb{P}(T_0 < \infty) = \lim_{n\to\infty} g_n(0)$.
(3) Prove that $\mathbb{P}(T_0 < \infty)$ is the smallest solution of the equation $g(t) = t$ in the interval $[0, 1]$. Verify that $\mathbb{P}(T_0 < \infty) < 1$ if and only $m > 1$.

Exercise 13.6 (Birth and Death Process) Let Q be the transition matrix on $\mathbb{Z}_+$ that is determined by

$$Q(0, 0) = r_0\,,\ \ Q(0, 1) = p_0\,,$$

and, for every $k \geq 1$,

$$Q(k, k-1) = q_k\,,\ \ Q(k, k) = r_k\,,\ \ Q(k, k+1) = p_k\,,$$

where $r_0 + p_0 = 1$ and $q_k + r_k + p_k = 1$ for every $k \geq 1$. We assume that $p_j > 0$ for every $j \geq 0$ and $q_j > 0$ for every $j \geq 1$. We consider the canonical Markov chain associated with Q.

(1) Verify that the chain is irreducible and has a reversible measure unique up to a multiplicative constant.
(2) Under the condition

$$\sum_{i=1}^{\infty} \frac{p_0 p_1 \cdots p_{i-1}}{q_1 q_2 \cdots q_i} < \infty,$$

prove that the chain is positive recurrent.
(3) We suppose that $p_k = p$ for every $k \geq 0$ and $q_k = q$ for every $k \geq 1$, where $q > p > 0$. Compute the quantities $\mathbf{E}_k[H_k]$ for every $k \geq 0$.

Exercise 13.7 Let $N \in \mathbb{N}$ and let Q be the transition matrix on $E = \{0, 1, \ldots, N\}$ defined by

$$Q(i, j) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}, \quad \forall i, j \in \{0, 1, \ldots, N\}.$$

In other words, $Q(i, \cdot)$ is the binomial $\mathcal{B}(N, i/N)$ distribution. Let $k \in \{0, 1, \ldots, N\}$ and let $(X_n)_{n\in\mathbb{Z}_+}$ be a Markov chain with transition matrix Q started at $X_0 = k$.

(1) Classify the states of $(X_n)_{n\in\mathbb{Z}_+}$.
(2) Verify that $(X_n)_{n\in\mathbb{Z}_+}$ is a martingale, which converges to a limit X_∞, a.s. as $n \to \infty$. Determine the law of X_∞.

Exercise 13.8 Let $(S_n)_{n\in\mathbb{Z}_+}$ be a simple random walk on $\mathbb{Z}$, with $S_0 = 0$. For $k \in \mathbb{Z}\backslash\{0\}$, show that the expected value of the number of visits of k before the first return to 0 is equal to 1.

Exercise 13.9 (Kolmogorov's Criterion for Reversibility) Consider an irreducible Markov chain on E with transition matrix Q. Prove that the chain has a reversible measure if and only if the following two conditions hold.

- For every $x, y \in E$, the property $Q(x, y) > 0$ implies $Q(y, x) > 0$.
- For every finite sequence $x_0, x_1, \ldots, x_n$ in E such that $x_n = x_0$ and $Q(x_{i-1}, x_i) > 0$ for every $i \in \{1, \ldots, n\}$,

$$\prod_{i=1}^{n} \frac{Q(x_i, x_{i-1})}{Q(x_{i-1}, x_i)} = 1.$$

Exercise 13.10 (h-Transform) Let Q be a transition matrix on a countable space E, and let $h : E \longrightarrow \mathbb{R}_+$ be a nonnegative function on E. We assume that the set $F = \{x \in E : h(x) > 0\}$ is not empty and that h is Q-harmonic on F.

(1) For every $x, y \in F$, set

$$Q'(x, y) = \frac{h(y)}{h(x)}\, Q(x, y).$$

Verify that Q' is a transition matrix on F.

(2) Fix $a \in F$, and suppose that $(X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain in E with transition matrix Q, such that $X_0 = a$, and $(Y_n)_{n\in\mathbb{Z}_+}$ is a Markov chain in F with transition matrix Q', such that $Y_0 = a$. Prove that, for every integer $n \geq 1$ and every function $F : E^{n+1} \longrightarrow \mathbb{R}_+$,

$$\mathbb{E}[F(Y_0, Y_1, \ldots, Y_n)] = \frac{1}{h(a)}\, \mathbb{E}[\mathbf{1}_{\{T>n\}}\, h(X_n)\, F(X_0, X_1, \ldots, X_n)],$$

where $T = \inf\{k \geq 0 : X_k \notin F\}$.

(3) We now assume that Q is the transition matrix of simple random walk on $\mathbb{Z}$. Verify that the assumptions are satisfied when $h(i) = i \vee 0$ and $F = \mathbb{N}$, and compute Q'. Retain the assumptions and notation of question (2) (in particular $a \in \mathbb{N}$ is fixed), and for every integer $N > a$ set

$$\tau_N = \inf\{n \geq 0 : X_n = N\}\,,\ \sigma_N = \inf\{n \geq 0 : Y_n = N\}.$$

Verify that $\sigma_N < \infty$ a.s., and that the law of $(Y_0, Y_1, \ldots, Y_{\sigma_N})$ coincides with the conditional distribution of $(X_0, X_1, \ldots, X_{\tau_N})$ under $\mathbb{P}(\cdot \mid \tau_N < T)$.

Exercise 13.11 Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{N}$. We assume that:

- $a = \mathbb{E}[Y_1] < \infty$;
- the greatest common divisor of $\{k \geq 1 : \mathbb{P}(Y_1 = k) > 0\}$ is 1.

We then define $Z_1 = Y_1, Z_2 = Y_1 + Y_2, \ldots, Z_k = Y_1 + \cdots + Y_k, \ldots$, and we set $X_0 = 0$ and, for every integer $n \geq 0$,

$$X_n = Z_{k(n)} - n, \text{ where } k(n) = \min\{k \geq 1 : Z_k \geq n\}.$$

(1) Verify that $(X_n)_{n\in\mathbb{Z}_+}$ is an irreducible Markov chain with values in a subset of $\mathbb{Z}_+$ to be determined. Prove that this Markov chain is positive recurrent and aperiodic.
(2) Consider the random set of integers $\mathcal{Z} = \{Z_1, Z_2, Z_3, \ldots\}$. Prove that

$$\lim_{n\to\infty} \mathbb{P}(n \in \mathcal{Z}) = \frac{1}{a}.$$

Exercise 13.12 We consider a random walk $(S_n)_{n\in\mathbb{Z}_+}$ on $\mathbb{Z}$ starting from 0, with jump distribution μ satisfying the following two properties:

- $\mu(0) < 1$ and $\mu(k) = 0$ for every $k < -1$;
- $\sum_{k\in\mathbb{Z}} |k|\, \mu(k) < \infty$ and $\sum_{k\in\mathbb{Z}} k\mu(k) = 0$.

(1) Verify that the Markov chain $(S_n)_{n\in\mathbb{Z}_+}$ is recurrent and irreducible.
(2) Let $H = \inf\{n \geq 1 : S_n = 0\}$ and $R = \inf\{n \geq 1 : S_n \geq 0\}$. Prove that, for every $k \in \mathbb{Z}$,

$$\mathbb{E}\Big[\sum_{n=0}^{H-1} \mathbf{1}_{\{S_n=k\}}\Big] = 1$$

and infer that we have also for every integer $k \leq 0$,

$$\mathbb{E}\Big[\sum_{n=0}^{R-1} \mathbf{1}_{\{S_n=k\}}\Big] = 1.$$

(3) Let $p \in \mathbb{Z}_+$. Verify that

$$\mathbb{P}(S_R = p) = \sum_{n=0}^{\infty} \mathbb{E}[\mathbf{1}_{\{n<R\}}\, \mathbf{1}_{\{S_{n+1}=p\}}]$$

and using question (2) conclude that

$$\mathbb{P}(S_R = p) = \sum_{k=p}^{\infty} \mu(k).$$

Exercise 13.13 A student owns three books numbered 1,2,3, which are stored on a shelf. Each morning, the student chooses at random one of the books, in such a way that the probability that the book i is chosen is $\alpha_i > 0$, and the choices are made independently every day. At the end of the day, the student places the chosen book back on the shelf, to the left of the other two. Suppose that on the morning of the first day the books stand in the order 1,2,3 from left to right on the shelf, and for every $n \geq 1$, let p_n be the probability that the books are in the same order on the morning of the n-th day. Compute the limit of p_n as $n \to \infty$.

Exercise 13.14 Suppose that we move a knight randomly on a chess board according to the following rules. Initially, the knight stands on one of the corners of the chess board. Then, at each step, the knight chooses one of the possible moves at random with equal probabilities, independently of what happened before. (Recall that a chess board is a square board of eight rows and eight columns, resulting in 64 squares, and that the knight's moves form an "L" shape, two squares horizontally and one square vertically, or two squares vertically and one square horizontally.) Compute the expected value of the number of steps before the knight comes back to its starting point. (Answer: 168).

Chapter 14
Brownian Motion

This chapter is devoted to the study of Brownian motion, which, together with the Poisson process studied in Chapter 9, is one of the most important continuous-time random processes. We motivate the definition of Brownian motion from an approximation by (discrete-time) random walks, which is reminiscent of the physical interpretation of the phenomenon first observed by Brown. We then provide a detailed construction of a Brownian motion $(B_t)_{t\geq 0}$, where a key step is the continuity of the sample paths $t \mapsto B_t(\omega)$ for every $\omega \in \Omega$. This construction leads to the definition of the Wiener measure or law of Brownian motion, which is a probability measure on the space of all continuous functions from $\mathbb{R}_+$ into $\mathbb{R}$ (or into $\mathbb{R}^d$ in the d-dimensional case). Simple considerations relying on a zero-one law for Brownian motion yield several remarkable almost sure properties of Brownian sample paths. In a way similar to Markov chains, Brownian motion satisfies a strong Markov property, which is very useful in explicit calculations of distributions. In the last part of the chapter, we discuss the close links between Brownian motion and harmonic functions on $\mathbb{R}^d$. Brownian motion provides an explicit formula for the solution of the classical Dirichlet problem, which consists in finding a harmonic function in a domain with a prescribed boundary condition. As a consequence, the Poisson kernel of a ball exactly corresponds to the distribution of the first exit point from the ball by Brownian motion. The relations with harmonic functions are also useful to study various properties of multidimensional Brownian motion, such as recurrence in dimension two or transience in higher dimensions.

14.1 Brownian Motion as a Limit of Random Walks

The physical explanation of Brownian motion justifies the irregular and unpredictable displacement of a Brownian particle by the numerous collisions of this particle with the molecules of the ambient medium, which produce continual

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5_14

changes of direction. From a mathematical perspective, this suggests that we consider the evolution in discrete time of a point particle that moves on the lattice $\mathbb{Z}^d$ and at each time $n \in \mathbb{Z}_+$ makes a step in a direction chosen at random independently of the past. In other words, we consider the model of a random walk on $\mathbb{Z}^d$, which we already discussed in Chapter 13. Our goal in this section is to introduce the definition of Brownian motion as the limit of random walks suitably rescaled in time and space. The results of this section are not used in the remaining part of the chapter, but serve as a strong motivation for our study.

Let $(S_n)_{n\in\mathbb{Z}_+}$ be a random walk on $\mathbb{Z}^d$ started from 0. This means that $S_0 = 0$ and, for every $n \in \mathbb{N}$,

$$S_n = Y_1 + \cdots + Y_n$$

where the random variables $Y_1, Y_2, \dots$ take values in $\mathbb{Z}^d$ and are independent and identically distributed with law μ. We assume that μ satisfies the following two properties:

- $\displaystyle\sum_{k\in\mathbb{Z}^d} |k|^2 \mu(k) < \infty$, and, for every $i, j \in \{1, \dots, d\}$,

$$\sum_{k\in\mathbb{Z}^d} k_i k_j\, \mu(k) = \begin{cases} \sigma^2 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{14.1}$$

  where $\sigma > 0$ is a constant and we write $k = (k_1, \dots, k_d)$ for $k \in \mathbb{Z}^d$;
- $\displaystyle\sum_{k\in\mathbb{Z}^d} k\mu(k) = 0$ (μ is centered).

Simple random walk on $\mathbb{Z}^d$ satisfies these assumptions, with $\sigma^2 = 1/d$, and there are many other examples. Assumption (14.1) is equivalent to $\mathbb{E}[(\xi \cdot Y_1)^2] = \sigma^2|\xi|^2$, for every $\xi \in \mathbb{R}^d$, which means informally that the random walk moves at the same speed in all directions (isotropy condition).

We are interested in the long-time behavior of the function $n \mapsto S_n$. By the multidimensional central limit theorem (Theorem 10.18), we already know that $n^{-1/2}S_n$ converges in distribution to a centered Gaussian vector with covariance matrix $\sigma^2\mathrm{Id}$, where Id denotes the identity matrix. In view of deriving more information, we set, for every $n \in \mathbb{N}$ and every real $t \geq 0$,

$$S_t^{(n)} := \frac{1}{\sqrt{n}}\, S_{\lfloor nt\rfloor},$$

where we recall that $\lfloor x\rfloor$ denotes the integer part of x.

Proposition 14.1 *For every choice of the integer $p \geq 1$ and of the real numbers $0 = t_0 < t_1 < \cdots < t_p$, we have*

$$(S^{(n)}_{t_1}, S^{(n)}_{t_2}, \ldots, S^{(n)}_{t_p}) \xrightarrow[n\to\infty]{\text{(d)}} (U_1, U_2, \ldots, U_p)$$

where the distribution of the limiting random vector $(U_1, U_2, \ldots, U_p)$ can be characterized as follows:

- *the random variables $U_1, U_2 - U_1, \ldots, U_p - U_{p-1}$ are independent;*
- *for every $j \in \{1, \ldots, p\}$, $U_j - U_{j-1}$ is a centered Gaussian vector with covariance matrix $\sigma^2(t_j - t_{j-1})\mathrm{Id}$ (we write $U_0 = 0$ by convention).*

Remark It is easy to write down an explicit formula for the density of the limiting distribution. The density of $U_j - U_{j-1}$ is $p_{\sigma^2(t_j - t_{j-1})}(x)$, where, for every $a > 0$,

$$p_a(x) = \frac{1}{(2\pi a)^{d/2}} \exp\Big(-\frac{|x|^2}{2a}\Big), \quad \forall x \in \mathbb{R}^d, \tag{14.2}$$

is the density of the centered Gaussian vector with covariance matrix $a\,\mathrm{Id}$ (recall from Proposition 11.12 that the components of this vector are independent Gaussian $\mathcal{N}(0, a)$ random variables, so that the preceding formula for p_a follows from Corollary 9.6). Using the independence of $U_1, U_2 - U_1, \ldots, U_p - U_{p-1}$, we get that the density of $(U_1, U_2 - U_1, \ldots, U_p - U_{p-1})$ is the function $g : (\mathbb{R}^d)^p \longrightarrow \mathbb{R}_+$ defined by

$$g(x_1, \ldots, x_p) = p_{\sigma^2 t_1}(x_1) p_{\sigma^2(t_2 - t_1)}(x_2) \cdots p_{\sigma^2(t_p - t_{p-1})}(x_p), \quad x_1, \ldots, x_p \in \mathbb{R}^d.$$

By a straightforward change of variables, the density of $(U_1, U_2, \ldots, U_p)$ is

$$\begin{aligned} f(y_1, \ldots, y_p) &= g(y_1, y_2 - y_1, \ldots, y_p - y_{p-1}) \\ &= p_{\sigma^2 t_1}(y_1) p_{\sigma^2(t_2 - t_1)}(y_2 - y_1) \cdots p_{\sigma^2(t_p - t_{p-1})}(y_p - y_{p-1}). \end{aligned}$$

Proof By Theorem 10.13, it is enough to verify that, for every $\xi_1, \ldots, \xi_p \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}\sum_{j=1}^{p} \xi_j \cdot S^{(n)}_{t_j}\Big)\Big] \xrightarrow[n\to\infty]{} \mathbb{E}\Big[\exp\Big(\mathrm{i}\sum_{j=1}^{p} \xi_j \cdot U_j\Big)\Big].$$

Via a simple linear transformation this is equivalent to verifying that, for every $\eta_1, \ldots, \eta_p \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}\sum_{j=1}^{p} \eta_j \cdot (S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}})\Big)\Big] \xrightarrow[n\to\infty]{} \mathbb{E}\Big[\exp\Big(\mathrm{i}\sum_{j=1}^{p} \eta_j \cdot (U_j - U_{j-1})\Big)\Big]. \tag{14.3}$$

Let us prove (14.3). Thanks to the independence of $U_1, U_2 - U_1, \ldots, U_p - U_{p-1}$ and to formula (10.5), we know that

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}\sum_{j=1}^{p}\eta_j\cdot(U_j - U_{j-1})\Big)\Big] = \prod_{i=1}^{p}\mathbb{E}\Big[\exp\Big(\mathrm{i}\,\eta_j\cdot(U_j - U_{j-1})\Big)\Big]$$
$$= \exp\Big(-\sum_{j=1}^{p}\frac{\sigma^2|\eta_j|^2(t_j - t_{j-1})}{2}\Big).$$

On the other hand,

$$S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}} = \frac{1}{\sqrt{n}}\sum_{k=\lfloor nt_{j-1}\rfloor+1}^{\lfloor nt_j\rfloor} Y_k,$$

and, using the grouping by blocks principle of Section 9.2, it follows that, for every fixed n, the random variables $S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}}$, $1 \le j \le p$, are independent. Moreover, for every j, $S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}}$ has the same distribution as

$$\frac{1}{\sqrt{n}}S_{\lfloor nt_j\rfloor - \lfloor nt_{j-1}\rfloor} = \frac{\sqrt{\lfloor nt_j\rfloor - \lfloor nt_{j-1}\rfloor}}{\sqrt{n}}\,\frac{1}{\sqrt{\lfloor nt_j\rfloor - \lfloor nt_{j-1}\rfloor}}\,S_{\lfloor nt_j\rfloor - \lfloor nt_{j-1}\rfloor}.$$

Thanks to the multidimensional central limit theorem (Theorem 10.18), and using the easy fact stated in Exercise 10.9, we get that $S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}}$ converges in distribution to $\sqrt{t_j - t_{j-1}}\,N$ as $n \to \infty$, where N is a centered d-dimensional Gaussian vector with covariance matrix $\sigma^2\mathrm{Id}$. Consequently, for every fixed $j \in \{1, \ldots, p\}$,

$$\mathbb{E}\Big[\exp\Big(\mathrm{i}\,\eta_j\cdot(S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}})\Big)\Big] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[\exp(\mathrm{i}\sqrt{t_j - t_{j-1}}\,\eta_j\cdot N)]$$
$$= \exp\Big(-\frac{\sigma^2|\eta_j|^2(t_j - t_{j-1})}{2}\Big).$$

Together with the independence of the random variables $S^{(n)}_{t_j} - S^{(n)}_{t_{j-1}}$, $1 \le j \le p$, this leads to the desired convergence (14.3). □

We now turn to the definition of Brownian motion. In this definition, Brownian motion is a (continuous-time) random process with values in $\mathbb{R}^d$, that is, a collection $(B_t)_{t\ge 0}$ of random variables with values in $\mathbb{R}^d$ indexed by $t \in \mathbb{R}_+$.

Definition 14.2 A random process $(B_t)_{t\ge 0}$ with values in $\mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a d-dimensional Brownian motion started from 0 if:

(P1) We have $B_0 = 0$ a.s. and, for every choice of the integer $p \ge 1$ and of $0 = t_0 < t_1 < \cdots < t_p$, the random variables $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_p} - B_{t_{p-1}}$ are independent and, for every $j \in \{1, \ldots, p\}$, $B_{t_j} - B_{t_{j-1}}$ is a centered Gaussian vector with covariance matrix $(t_j - t_{j-1})\mathrm{Id}$.

(P2) For every $\omega \in \Omega$, the function $t \mapsto B_t(\omega)$ is continuous.

We emphasize that the existence of a random process satisfying both properties (P1) and (P2) is not obvious, but will be established below.

The functions $t \mapsto B_t(\omega)$ (indexed by the parameter $\omega \in \Omega$) are called the *sample paths* of B, and property (P2) is often stated by saying that Brownian motion has continuous sample paths.

Assuming the existence of Brownian motion, we can reformulate Proposition 14.1 by saying that, for every choice of $t_1 < \cdots < t_p$,

$$(S_{t_1}^{(n)}, S_{t_2}^{(n)}, \ldots, S_{t_p}^{(n)}) \xrightarrow[n\to\infty]{\text{(d)}} (\sigma B_{t_1}, \sigma B_{t_2}, \ldots, \sigma B_{t_p}).$$

Brownian motion thus appears as a continuous limit of suitably rescaled discrete random walks. As it was explained in the remark following Proposition 14.1, the distribution of the vector $(B_{t_1}, B_{t_2}, \ldots, B_{t_p})$ is given by

$$\begin{aligned}&\mathbb{P}\Big((B_{t_1}, B_{t_2}, \ldots, B_{t_p}) \in A\Big)\\ &\quad = \int_A p_{t_1}(y_1)p_{t_2-t_1}(y_2-y_1)\cdots p_{t_p-t_{p-1}}(y_p-y_{p-1})\,\mathrm{d}y_1\ldots\mathrm{d}y_p, \qquad (14.4)\end{aligned}$$

for every Borel subset A of $(\mathbb{R}^d)^p$, where the functions $p_a(x)$ are defined in (14.2). Formula (14.4) gives the *finite-dimensional marginal distributions* of Brownian motion. This should be compared with the analogous result for the Poisson process derived at the end of Chapter 9.

14.2 The Construction of Brownian Motion

We now state the basic existence result for Brownian motion.

Theorem 14.3 *Brownian motion exists. In other words, on a suitable probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can construct a collection $(B_t)_{t\geq 0}$ of random variables satisfying both properties (P1) and (P2) of Definition 14.2.*

Proof We first consider the case $d = 1$ (corresponding to linear Brownian motion), and in a first step we will construct the collection $(B_t)_{t\in[0,1]}$. We choose the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ so that we can construct a sequence of independent Gaussian $\mathcal{N}(0, 1)$ random variables on this space (see Section 9.4).

Let us start by introducing the Haar functions, which are real functions defined on the interval $[0,1)$. We first set

$$h_0(t) = 1, \quad \forall t \in [0,1)$$

and then, for every integer $n \geq 0$ and for every $k \in \{0, 1, \ldots, 2^n - 1\}$,

$$h_n^k(t) = 2^{n/2}\, \mathbf{1}_{[\frac{2k}{2^{n+1}}, \frac{2k+1}{2^{n+1}})}(t) - 2^{n/2}\, \mathbf{1}_{[\frac{2k+1}{2^{n+1}}, \frac{2k+2}{2^{n+1}})}(t), \quad \forall t \in [0,1).$$

These functions belong to the Hilbert space $L^2([0,1), \mathcal{B}([0,1)), \lambda)$ where we recall that λ denotes Lebesgue measure. Furthermore, one easily verifies that $\|h_0\|_2 = \|h_n^k\|_2 = 1$, where $\|h\|_2$ refers to the L^2-norm, and the functions h_0, h_n^k are orthogonal in L^2. Hence the collection of Haar functions,

$$\Big(h_0,\, (h_n^k)_{n \geq 0, 0 \leq k \leq 2^n - 1}\Big) \tag{14.5}$$

forms an orthonormal system in $L^2([0,1), \mathcal{B}([0,1)), \lambda)$ (see the Appendix below for definitions). This orthonormal system is in fact an orthonormal basis. Indeed, it is easy to verify that a step function f on $[0,1)$ such that there exists an integer $n \in \mathbb{N}$ with the property that f is constant on every interval of the form $[(i-1)2^{-n}, i2^{-n})$, $1 \leq i \leq 2^n$, must be a linear combination of the Haar functions, and on the other hand, the set of all such step functions is dense in $L^2([0,1), \mathcal{B}([0,1)), \lambda)$ (note that any continuous function with compact support on $[0,1)$ can be approximated uniformly on $[0,1)$ by step functions of this type, and use Theorem 4.8).

Write $\langle f, g \rangle = \int_0^1 f(t) g(t) \mathrm{d}t$ for the scalar product in $L^2([0,1), \mathcal{B}([0,1)), \lambda)$. Then, for every $f \in L^2([0,1), \mathcal{B}([0,1)), \lambda)$ we have

$$f = \langle f, h_0 \rangle h_0 + \sum_{n=0}^{\infty} \sum_{k=0}^{2^n - 1} \langle f, h_n^k \rangle h_n^k,$$

where the series converges in L^2 (Theorem A.5).

On the other hand, we may consider a countable collection

$$\Big(N_0,\ (N_n^k)_{n \geq 0, 0 \leq k \leq 2^n - 1}\Big)$$

of independent Gaussian $\mathcal{N}(0,1)$ random variables defined on our probability space $(\Omega, \mathcal{A}, \mathbb{P})$. One immediately verifies that this collection forms an orthonormal system in the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$ (orthogonality follows from independence). By standard Hilbert space theory (Proposition A.6), there exists a linear isometric mapping $\mathcal{I}$ from $L^2([0,1), \mathcal{B}([0,1)), \lambda)$ into $L^2(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathcal{I}(h_0) = N_0$ and $\mathcal{I}(h_n^k) = N_n^k$ for every $n \geq 0$ and $0 \leq k \leq 2^n - 1$. More precisely, for every

$f \in L^2([0,1), \mathcal{B}([0,1)), \lambda)$,

$$\mathcal{I}(f) = \langle f, h_0 \rangle N_0 + \sum_{n=0}^{\infty} \sum_{k=0}^{2^n-1} \langle f, h_n^k \rangle N_n^k,$$

where the series converges in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. We note that

$$\mathbb{E}[\mathcal{I}(f)^2] = \|f\|_2^2$$

by the isometry property, and that $\mathbb{E}[\mathcal{I}(f)] = 0$ since all random variables N_0, N_n^k are centered. Moreover the next lemma will show that $\mathcal{I}(f)$ is a (centered) Gaussian random variable.

Lemma 14.4 *Let $(U_n)_{n\in\mathbb{N}}$ be a sequence of Gaussian (real) random variables that converges to U in L^2. Then U is also Gaussian.*

Proof Let $m_n = \mathbb{E}[U_n]$ and $\sigma_n^2 = \text{var}(U_n)$. The L^2-convergence ensures that $m_n \longrightarrow m = \mathbb{E}[U]$ and $\sigma_n^2 \longrightarrow \sigma^2 = \text{var}(U)$ as $n \to \infty$. However, since the L^2-convergence implies the convergence in distribution, we have also, for every $\xi \in \mathbb{R}$,

$$e^{\mathrm{i} m_n \xi - \sigma_n^2 \xi^2/2} = \mathbb{E}[e^{\mathrm{i}\xi U_n}] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[e^{\mathrm{i}\xi U}]$$

which shows that the characteristic function of U is

$$\mathbb{E}[e^{\mathrm{i}\xi U}] = e^{\mathrm{i} m \xi - \sigma^2 \xi^2/2}$$

and thus U is distributed according to $\mathcal{N}(m, \sigma^2)$. □

Let $f \in L^2([0,1), \mathcal{B}([0,1)), \lambda)$. Writing

$$\mathcal{I}(f) = \lim_{m\to\infty} \Big(\langle f, h_0 \rangle N_0 + \sum_{n=0}^{m} \sum_{k=0}^{2^n-1} \langle f, h_n^k \rangle N_n^k \Big), \quad \text{in } L^2,$$

and using the fact that a linear combination of independent Gaussian variables is still Gaussian, we deduce from Lemma 14.4 that $\mathcal{I}(f)$ is Gaussian $\mathcal{N}(0, \|f\|_2^2)$. We also note that, if $f, f' \in L^2([0,1), \mathcal{B}([0,1)), \lambda)$,

$$\text{cov}(\mathcal{I}(f), \mathcal{I}(f')) = \mathbb{E}[\mathcal{I}(f)\mathcal{I}(f')] = \langle f, f' \rangle$$

by the isometry property.

We then set, for every $t \in [0,1]$,

$$B_t = \mathcal{I}(\mathbf{1}_{[0,t)}).$$

In particular, $B_0 = \mathcal{I}(0) = 0$ a.s.

Let us first verify that $(B_t)_{t\in[0,1]}$ satisfies property (P1) (restricted to times belonging to $[0,1]$). Fix $0 = t_0 < t_1 < \cdots < t_p \leq 1$. By linearity, we have

$$B_{t_i} - B_{t_{i-1}} = \mathcal{I}(\mathbf{1}_{[t_{i-1},t_i)})$$

which is $\mathcal{N}(0, t_i - t_{i-1})$ distributed by the previous observations. Moreover, if $i \neq j$,

$$\begin{aligned}\mathrm{cov}(B_{t_i} - B_{t_{i-1}}, B_{t_j} - B_{t_{j-1}}) &= \mathbb{E}[(B_{t_i} - B_{t_{i-1}})(B_{t_j} - B_{t_{j-1}})] \\ &= \langle \mathbf{1}_{[t_{i-1},t_i)}, \mathbf{1}_{[t_{j-1},t_j)} \rangle \\ &= 0.\end{aligned}$$

It is easy to verify that $(B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_p} - B_{t_{p-1}})$ is a Gaussian vector: if $\lambda_1, \ldots, \lambda_p \in \mathbb{R}$,

$$\sum_{j=1}^{p} \lambda_j (B_{t_j} - B_{t_{j-1}}) = \mathcal{I}\Big(\sum_{j=1}^{p} \lambda_j \mathbf{1}_{[t_{j-1},t_j)}\Big)$$

is a Gaussian random variable By Proposition 11.12, the fact that the covariance matrix $(\mathrm{cov}(B_{t_i} - B_{t_{i-1}}, B_{t_j} - B_{t_{j-1}}))_{i,j=1,\ldots,p}$ is diagonal implies that the random variables $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_p} - B_{t_{p-1}}$, are independent, which completes the proof of (P1).

We now need to establish the continuity property (P2) (recall that, for the moment, we restrict our attention to $t \in [0,1]$). At the present stage, for every $t \in [0,1]$, $B_t = \mathcal{I}(\mathbf{1}_{[0,t)})$ is defined as an element of $L^2(\Omega, \mathcal{A}, \mathbb{P})$ and thus as an equivalence class of random variables which are almost surely equal. For (P2) to be meaningful, it is necessary to select a representative in this equivalence class, for each $t \in [0,1]$ (the selection of such a representative did not matter for (P1) but it does for (P2)). To this end, we will study the series defining B_t in a more precise manner. We first introduce the functions defined on the interval $[0,1]$ by

$$g_0(t) := \langle \mathbf{1}_{[0,t)}, h_0 \rangle = t$$

$$g_n^k(t) := \langle \mathbf{1}_{[0,t)}, h_n^k \rangle = \int_0^t h_n^k(s)\mathrm{d}s\,, \text{ for } n \geq 0,\ 0 \leq k \leq 2^n - 1.$$

By construction, for every $t \in [0, 1]$,

$$B_t = \mathcal{I}(\mathbf{1}_{[0,t)}) = tN_0 + \sum_{n=0}^{\infty} \sum_{k=0}^{2^n-1} g_n^k(t) N_n^k \tag{14.6}$$

where the series converges in $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

For every integer $m \geq 0$ and every $\omega \in \Omega$, set

$$B_t^{(m)}(\omega) = tN_0(\omega) + \sum_{n=0}^{m} \sum_{k=0}^{2^n-1} g_n^k(t) N_n^k(\omega), \quad \forall t \in [0, 1]. \tag{14.7}$$

Note that $t \mapsto B_t^{(m)}(\omega)$ is continuous for every $\omega \in \Omega$.

Claim There exists a measurable set A with $\mathbb{P}(A) = 0$, such that, for every $\omega \in A^c$, the function $t \mapsto B_t^{(m)}(\omega)$ converges uniformly on $[0, 1]$ as $m \to \infty$ to a limiting function denoted by $t \mapsto B_t^\circ(\omega)$.

Assuming that the claim is proved, we also define $B_t^\circ(\omega) = 0$ for every $t \in [0, 1]$ when $\omega \in A$. In this way, $B_t^\circ(\omega)$ is obtained as an almost sure limit, which must coincide with the L^2-limit in (14.6). Thus our definition of $B_t^\circ(\omega)$ will give an element of the equivalence class of $\mathcal{I}(\mathbf{1}_{[0,t)})$ as desired (in other words, $B_t^\circ = B_t$ a.s.), but we have also obtained that $t \mapsto B_t^\circ(\omega)$ is continuous since a uniform limit of continuous functions is continuous (of course if $\omega \in A$ the desired continuity is trivial).

To prove our claim, we observe that $0 \leq g_n^k \leq 2^{-n/2}$ and that, for every fixed n, the functions g_n^k, $0 \leq k \leq 2^n - 1$ have disjoint supports ($g_n^k(t) > 0$ only if $k2^{-n} < t < (k+1)2^{-n}$). Hence, for every $n \geq 0$,

$$\sup_{t\in[0,1]} \left| \sum_{k=0}^{2^n-1} g_n^k(t) N_n^k \right| \leq 2^{-n/2} \sup_{0\leq k\leq 2^n-1} |N_n^k|. \tag{14.8}$$

Lemma 14.5 *If N is $\mathcal{N}(0, 1)$ distributed, we have for every $a \geq 1$,*

$$\mathbb{P}(|N| \geq a) \leq e^{-a^2/2}.$$

Proof We have

$$\mathbb{P}(|N| \geq a) = \frac{2}{\sqrt{2\pi}} \int_a^\infty \mathrm{d}x\, e^{-x^2/2} \leq \frac{2}{\sqrt{2\pi}} \int_a^\infty \mathrm{d}x\, \frac{x}{a}\, e^{-x^2/2} = \frac{2}{a\sqrt{2\pi}} e^{-a^2/2}.$$

□

Since all variables N_n^k are $\mathcal{N}(0,1)$ distributed, we can use the lemma to bound

$$\mathbb{P}\Big(\sup_{0\le k\le 2^n-1}|N_n^k|>2^{n/4}\Big)\le\sum_{k=0}^{2^n-1}\mathbb{P}(|N_n^k|>2^{n/4})\le 2^n\exp(-2^{(n/2)-1}).$$

Setting

$$A_n=\Big\{\sup_{0\le k\le 2^n-1}|N_n^k|>2^{n/4}\Big\}$$

We deduce from the Borel-Cantelli lemma (Lemma 9.3) and the preceding bound that

$$\mathbb{P}(\limsup A_n)=0.$$

We set $A=\limsup A_n$, so that $\mathbb{P}(A)=0$ and, on the other hand, if $\omega\notin A$, we have for every large enough n,

$$\sup_{0\le k\le 2^n-1}|N_n^k|\le 2^{n/4}$$

and thus, by (14.8),

$$\sup_{t\in[0,1]}\Big|\sum_{k=0}^{2^n-1}g_n^k(t)N_n^k\Big|\le 2^{-n/4}$$

which implies the desired uniform convergence of the functions $t\mapsto B_t^{(m)}$ defined in (14.7). This completes the verification of our claim, and we conclude that the collection $(B_t^\circ)_{t\in[0,1]}$ satisfies both (P1) and (P2) (restricted to the time interval $[0,1]$).

It remains to get rid of the restriction to $t\in[0,1]$, and to generalize the construction in any dimension d. In a first step, we consider independent collections $(B_t^{(1)})_{t\in[0,1]}$, $(B_t^{(2)})_{t\in[0,1]}$, etc. constructed as above: for each of these collections, we use a sequence of independent Gaussian $\mathcal{N}(0,1)$ random variables which is independent of the sequences used previously (to do this we need only countably many independent Gaussian $\mathcal{N}(0,1)$ random variables). We then set, for every $t\ge 0$,

$$B_t=B_1^{(1)}+B_1^{(2)}+\cdots+B_1^{(k)}+B_{t-k}^{(k+1)}\qquad\text{if } t\in[k,k+1).$$

It is straightforward to verify that $(B_t)_{t\ge 0}$ satisfies properties (P1) and (P2) in dimension one.

Finally, in dimension $d \geq 2$, we just have to consider d independent Brownian motions in dimension one $(B_t^1)_{t\geq 0}, \ldots, (B_t^d)_{t\geq 0}$ (again, it is enough to have countably many independent $\mathcal{N}(0, 1)$ random variables on our probability space) and we set

$$B_t = (B_t^1, B_t^2, \ldots, B_t^d)$$

for every $t \in \mathbb{R}_+$. The verification of (P1) and (P2) is again easy, and this completes the proof of the theorem. □

If $x \in \mathbb{R}^d$, a random process $(B_t)_{t\geq 0}$ is called a (d-dimensional) Brownian motion started from x if $(B_t - x)_{t\geq 0}$ is a (d-dimensional) Brownian motion started from 0.

14.3 The Wiener Measure

Let $C(\mathbb{R}_+, \mathbb{R}^d)$ denote the space of all continuous functions from $\mathbb{R}_+$ into $\mathbb{R}^d$. We equip $C(\mathbb{R}_+, \mathbb{R}^d)$ with the σ-field $\mathcal{C}$ defined as the smallest σ-field for which the coordinate mappings $\mathrm{w} \mapsto \mathrm{w}(t)$, from $C(\mathbb{R}_+, \mathbb{R})$ into $\mathbb{R}$, are measurable, for every $t \in \mathbb{R}_+$.

Lemma 14.6 *The σ-field $\mathcal{C}$ coincides with the Borel σ-field when $C(\mathbb{R}_+, \mathbb{R}^d)$ is endowed with the topology of uniform convergence on every compact set.*

Proof Write $\mathcal{B}$ for the Borel σ-field associated with the topology of uniform convergence on every compact set. The fact that $\mathcal{C} \subset \mathcal{B}$ is immediate since the coordinate mappings are continuous hence $\mathcal{B}$-measurable. Conversely, a distance defining the topology on $C(\mathbb{R}_+, \mathbb{R}^d)$ is given by

$$D(\mathrm{w}, \mathrm{w}') = \sum_{n=1}^{\infty} 2^{-n} \sup_{0\leq t\leq n} (|\mathrm{w}(t) - \mathrm{w}'(t)| \wedge 1).$$

Let us fix $\mathrm{w}_0 \in C(\mathbb{R}_+, \mathbb{R}^d)$. Since, for every $\mathrm{w} \in C(\mathbb{R}_+, \mathbb{R}^d)$,

$$\sup_{t\in[0,n]} (|\mathrm{w}(t) - \mathrm{w}_0(t)| \wedge 1) = \sup_{t\in[0,n]\cap\mathbb{Q}} (|\mathrm{w}(t) - \mathrm{w}_0(t)| \wedge 1),$$

and a supremum of countably many measurable functions is measurable, we get that the mapping $\mathrm{w} \mapsto d(\mathrm{w}_0, \mathrm{w})$ is $\mathcal{C}$-measurable. It follows that balls of $C(\mathbb{R}_+, \mathbb{R}^d)$ are $\mathcal{C}$-measurable. Finally, since $C(\mathbb{R}_+, \mathbb{R}^d)$ equipped with the distance D is separable, any open set is the union of a countable collection of balls and is thus $\mathcal{C}$-measurable. This implies that $\mathcal{B} \subset \mathcal{C}$. □

Definition 14.7 Let $(B_t)_{t\geq 0}$ be a d-dimensional Brownian motion started at 0, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The ($d$-dimensional) Wiener measure is the probability measure $\mathbf{P}_0$ on $(C(\mathbb{R}_+, \mathbb{R}^d), \mathcal{C})$, which is defined as the pushforward of $\mathbb{P}(\mathrm{d}\omega)$ under the mapping

$$\omega \mapsto (t \mapsto B_t(\omega))$$

from Ω into $C(\mathbb{R}_+, \mathbb{R}^d)$.

Write Φ for the mapping considered in the definition. For this definition to be meaningful, we of course need Φ to be measurable. However, the same argument as in the proof of Lemma 13.5 shows that this measurability property holds if (and only if) the composition of Φ with any of the coordinate mappings $\mathrm{w} \mapsto \mathrm{w}(t)$ is measurable, and this is immediate since these compositions are the random variables B_t.

The previous definition makes sense because it does not depend on the choice of the Brownian motion B (nor on the underlying probability space on which this Brownian motion is defined). Indeed, for any choice of $0 = t_0 < t_1 < \cdots < t_p$, we have, for every Borel subsets $A_0, A_1, \ldots, A_p$ of $\mathbb{R}^d$,

$$\begin{aligned}
&\mathbf{P}_0(\{\mathrm{w} \in C(\mathbb{R}_+, \mathbb{R}^d) : \mathrm{w}(t_0) \in A_0, \mathrm{w}(t_1) \in A_1, \ldots, \mathrm{w}(t_p) \in A_p\}) \\
&= \mathbb{P}(B_{t_0} \in A_0, B_{t_1} \in A_1, \ldots, B_{t_p} \in A_p) \\
&= \mathbf{1}_{A_0}(0) \int_{A_1\times\cdots\times A_p} p_{t_1}(y_1) p_{t_2-t_1}(y_2 - y_1) \cdots p_{t_p-t_{p-1}}(y_p - y_{p-1})\, \mathrm{d}y_1 \ldots \mathrm{d}y_p,
\end{aligned}$$

by formula (14.4), which holds for any Brownian motion B (this is just a reformulation of (P1)). An application of Corollary 1.19 shows that a probability measure on $C(\mathbb{R}_+, \mathbb{R}^d)$ is characterized by its values on the "cylinder sets" of the form

$$\{\mathrm{w} \in C(\mathbb{R}_+, \mathbb{R}^d) : \mathrm{w}(t_0) \in A_0, \mathrm{w}(t_1) \in A_1, \ldots, \mathrm{w}(t_p) \in A_p\}.$$

It follows that $\mathbf{P}_0$ is uniquely determined independently of the choice of the Brownian motion B. We may and will often see a Brownian motion B as the random variable $\omega \mapsto (t \mapsto B_t(\omega))$ with values in $C(\mathbb{R}_+, \mathbb{R}^d)$, which is considered in Definition 14.7. From this point of view, the Wiener measure is just the law of Brownian motion (started at 0). The fact that this law is uniquely defined also means that, if B' is another Brownian motion (started at 0) then, for any $A \in \mathcal{C}$,

$$\mathbb{P}\Big((B'_t)_{t\geq 0} \in A\Big) = \mathbf{P}_0(A) = \mathbb{P}\Big((B_t)_{t\geq 0} \in A\Big), \tag{14.9}$$

so that, in particular, any almost sure property that holds for a given Brownian motion holds automatically for any other Brownian motion.

If $x \in \mathbb{R}^d$, we also write $\mathbf{P}_x(\mathrm{dw})$ for the pushforward of $\mathbf{P}_0(\mathrm{dw})$ under the translation $\mathrm{w} \to x + \mathrm{w}$. This is the law of Brownian motion started at x. In a way similar to Chapter 13, $\mathbf{E}_x$ will be used to denote the expectation under $\mathbf{P}_x$.

The Canonical Construction of Brownian Motion In a way very similar to the case of Markov chains studied in the previous chapter, we will now present a canonical construction of Brownian motion involving a special choice of the probability space and of the process $(B_t)_{t\geq 0}$. This canonical construction allows us to deal simultaneously with all possible starting points just by changing the underlying probability measure. This will also enable us to write the strong Markov property in a form that is particularly convenient for our applications to harmonic functions (see Theorem 14.19 below).

We set $\mathbf{\Omega} = C(\mathbb{R}_+, \mathbb{R}^d)$ (without risk of confusion, we use the same notation as in the chapter on Markov chains for a different space!). We equip $\mathbf{\Omega}$ with the σ-field $\mathcal{C}$ and the probability measure $\mathbf{P}_0$. We then set, for every $t \geq 0$,

$$B_t(\mathrm{w}) = \mathrm{w}(t), \quad \forall \mathrm{w} \in \mathbf{\Omega}.$$

The collection $(B_t)_{t\geq 0}$, which is defined on the probability space $(\mathbf{\Omega}, \mathcal{C}, \mathbf{P}_0)$, is a Brownian motion started from 0. Notice that property (P2) is here obvious. Property (P1) follows from the formula given above for

$$\mathbf{P}_0(\{\mathrm{w} \in C(\mathbb{R}_+, \mathbb{R}^d) : \mathrm{w}(t_0) \in A_0, \mathrm{w}(t_1) \in A_1, \ldots, \mathrm{w}(t_p) \in A_p\}).$$

Similarly, under $\mathbf{P}_x$, $(B_t)_{t\geq 0}$ is a Brownian motion started from x.

14.4 First Properties of Brownian Motion

In this section, we consider a d-dimensional Brownian motion B started at 0. For every $t \geq 0$, we write $\mathcal{F}_t = \sigma(B_t, 0 \leq r \leq t)$ for the σ-field generated by the random variables $B_r, 0 \leq r \leq t$. We also write $\mathcal{F}_\infty = \sigma(B_t, t \geq 0)$. The collection $(\mathcal{F}_t)_{t\geq 0}$ is a (continuous-time) filtration—meaning that it is an increasing collection of σ-fields indexed by $\mathbb{R}_+$.

Proposition 14.8

(i) *Let φ be a vector isometry of $\mathbb{R}^d$, then $(\varphi(B_t))_{t\geq 0}$ is also a Brownian motion (in particular, $-B$ is a Brownian motion).*

(ii) *For every $\gamma > 0$, the process $(B_t^\gamma)_{t\geq 0}$ defined by $B_t^\gamma := \frac{1}{\gamma} B_{\gamma^2 t}$ is also a Brownian motion* (scaling invariance).

(iii) *For every $s \geq 0$, the process $(B_t^{(s)})_{t\geq 0}$ defined by $B_t^{(s)} := B_{s+t} - B_s$ is also a Brownian motion, and the collection $(B_t^{(s)})_{t\geq 0}$ is independent of $\mathcal{F}_s$ (simple Markov property).*

Remarks

(i) All these properties remain trivially valid if B is a Brownian motion started at $x \neq 0$.
(ii) One may compare (iii) with the analogous property for the Poisson process (Theorem 9.21).

Proof (i) and (ii) are very easy, by checking that both $\varphi(B_t)$ and B^γ_t satisfy property (P1) — property (P2) is obvious. For (i), note that, if X is a d-dimensional Gaussian vector with covariance matrix $a\,\mathrm{Id}$, then $\varphi(X)$ has the same distribution as X, as it can be seen by writing down the characteristic function of $\varphi(X)$. As for (iii), the fact that $B^{(s)}$ is a Brownian motion is also immediate and the independence property can be derived as follows. For every choice of $t_1 < t_2 < \cdots < t_p$ and $r_1 < r_2 < \cdots < r_q \leq s$, property (P1) implies that the vector $(B^{(s)}_{t_1}, \dots, B^{(s)}_{t_p})$ is independent of $(B_{r_1}, \dots, B_{r_q})$. Using Proposition 9.7 (or the remark at the end of Section 9.2), we infer that the collection $(B^{(s)}_t)_{t\geq 0}$ is independent of $(B_r)_{0\leq r\leq s}$. □

Theorem 14.9 (Blumenthal's Zero-One Law) *Let*

$$\mathcal{F}_{0+} = \bigcap_{s>0} \mathcal{F}_s.$$

Then, for every $A \in \mathcal{F}_{0+}$, we have $\mathbb{P}(A) = 0$ or 1.

Proof Let $A \in \mathcal{F}_{0+}$ and $t_1, \dots, t_p > 0$. For $\varepsilon > 0$ small enough (smaller than $\min\{t_i : 1 \leq i \leq p\}$) the simple Markov property in Proposition 14.8 (iii) implies that $(B_{t_1} - B_\varepsilon, \dots, B_{t_p} - B_\varepsilon)$ is independent of $\mathcal{F}_\varepsilon$, and a fortiori of $\mathcal{F}_{0+}$. Consequently, for every bounded continuous function f on $(\mathbb{R}^d)^p$,

$$\mathbb{E}[\mathbf{1}_A\, f(B_{t_1} - B_\varepsilon, \dots, B_{t_p} - B_\varepsilon)] = \mathbb{P}(A)\,\mathbb{E}[f(B_{t_1} - B_\varepsilon, \dots, B_{t_p} - B_\varepsilon)].$$

By letting $\varepsilon \to 0$, we get

$$\mathbb{E}[\mathbf{1}_A\, f(B_{t_1}, \dots, B_{t_p})] = \mathbb{P}(A)\,\mathbb{E}[f(B_{t_1}, \dots, B_{t_p})],$$

and thus $(B_{t_1}, \dots, B_{t_p})$ is independent of $\mathcal{F}_{0+}$. Thanks to Proposition 9.7, it follows that $\mathcal{F}_\infty$ is independent of $\mathcal{F}_{0+}$. In particular, $\mathcal{F}_{0+} \subset \mathcal{F}_\infty$ is independent of itself, so that $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$ for every $A \in \mathcal{F}_{0+}$. □

In the next two corollaries, we consider the case $d = 1$.

Corollary 14.10 $(d = 1)$ *We have, a.s. for every $\varepsilon > 0$,*

$$\sup_{0\leq s\leq\varepsilon} B_s > 0, \qquad \inf_{0\leq s\leq\varepsilon} B_s < 0.$$

For every $a \in \mathbb{R}$, set $T_a = \inf\{t \geq 0 : B_t = a\}$ ($\inf \varnothing = \infty$). Then

$$a.s., \ \forall a \in \mathbb{R}, \quad T_a < \infty.$$

Consequently, we have a.s.

$$\limsup_{t\to\infty} B_t = +\infty, \quad \liminf_{t\to\infty} B_t = -\infty.$$

Remark It is not a priori obvious that $\sup_{0\leq s\leq \varepsilon} B_s$ is measurable, since this is an uncountable supremum of measurable functions. However, since we know that the functions $t \mapsto B_t(\omega)$ are continuous, the latter supremum is also equal to $\sup_{s\in[0,\varepsilon]\cap\mathbb{Q}} B_s$, and is therefore measurable as the supremum of a countable collection of measurable functions. Alternatively we could observe that we are considering the composition of the two functions $\omega \mapsto (t \mapsto B_t(\omega))$ and $\mathrm{w} \mapsto \sup_{0\leq s\leq \varepsilon} \mathrm{w}(t)$, which are both measurable, the second one thanks to Lemma 14.6.

Proof Fix a decreasing sequence $(\varepsilon_p)_{p\in\mathbb{N}}$ of positive reals converging to 0, and set

$$A = \bigcap_{p\in\mathbb{N}} \Big\{ \sup_{0\leq s\leq \varepsilon_p} B_s > 0 \Big\}.$$

Let $t > 0$ and $p_0 \in \mathbb{N}$ such that $\varepsilon_{p_0} \leq t$. Since the intersection defining A is decreasing, we have also

$$A = \bigcap_{p\geq p_0} \Big\{ \sup_{0\leq s\leq \varepsilon_p} B_s > 0 \Big\},$$

which is $\mathcal{F}_t$-measurable. We conclude that A is $\mathcal{F}_t$-measurable for every $t > 0$ and thus A is $\mathcal{F}_{0+}$-measurable. On the other hand,

$$\mathbb{P}(A) = \lim_{p\to\infty} \downarrow \mathbb{P}\Big(\sup_{0\leq s\leq \varepsilon_p} B_s > 0 \Big),$$

and

$$\mathbb{P}\Big(\sup_{0\leq s\leq \varepsilon_p} B_s > 0 \Big) \geq \mathbb{P}(B_{\varepsilon_p} > 0) = \frac{1}{2},$$

since B_{ε_p} is $\mathcal{N}(0, \varepsilon_p)$-distributed and thus has a symmetric density. This shows that $\mathbb{P}(A) \geq 1/2$. From Theorem 14.9, we get that $\mathbb{P}(A) = 1$, and it follows that

$$\text{a.s.}, \ \forall \varepsilon > 0, \quad \sup_{0\leq s\leq \varepsilon} B_s > 0.$$

The assertion about $\inf_{0\le s\le\varepsilon} B_s$ is obtained by replacing B by $-B$.

We then write

$$1 = \mathbb{P}\Big(\sup_{0\le s\le 1} B_s > 0\Big) = \lim_{\delta\downarrow 0} \uparrow \mathbb{P}\Big(\sup_{0\le s\le 1} B_s > \delta\Big),$$

and we observe that, thanks to the scaling invariance property (Proposition 14.8 (ii)) with $\gamma = \delta$,

$$\mathbb{P}\Big(\sup_{0\le s\le 1} B_s > \delta\Big) = \mathbb{P}\Big(\sup_{0\le s\le 1/\delta^2} B^\delta_s > 1\Big) = \mathbb{P}\Big(\sup_{0\le s\le 1/\delta^2} B_s > 1\Big)$$

where the last equality holds because the law of Brownian motion is uniquely defined, see the remarks following Definition 14.7 and in particular formula (14.9). By letting $\delta \to 0$, we get

$$\mathbb{P}\Big(\sup_{s\ge 0} B_s > 1\Big) = \lim_{\delta\downarrow 0} \uparrow \mathbb{P}\Big(\sup_{0\le s\le 1/\delta^2} B_s > 1\Big) = 1.$$

Another scaling argument then shows that, for every $A > 0$,

$$\mathbb{P}\Big(\sup_{s\ge 0} B_s > A\Big) = 1$$

and replacing B by $-B$ we also get

$$\mathbb{P}\Big(\inf_{s\ge 0} B_s < -A\Big) = 1.$$

The last assertions of the corollary easily follow. For the last one, we observe that a continuous function $f : \mathbb{R}_+ \longrightarrow \mathbb{R}$ takes all real values only if $\limsup_{t\to+\infty} f(t) = +\infty$ and $\liminf_{t\to+\infty} f(t) = -\infty$. □

Corollary 14.11 ($d = 1$) *Almost surely, the function $t \mapsto B_t$ is not monotone on any non-trivial interval.*

Proof From the first assertion of Corollary 14.10 and the simple Markov property in Proposition 14.8 (iii), we immediately get that a.s. for every rational $q \in \mathbb{Q}_+$, for every $\varepsilon > 0$,

$$\sup_{q\le t\le q+\varepsilon} B_t > B_q, \qquad \inf_{q\le t\le q+\varepsilon} B_t < B_q.$$

The desired result follows. Notice that we restricted ourselves to rational values of q in order to discard a *countable* union of sets of zero probability (and in fact the stated result would fail if we would consider all real values of q). □

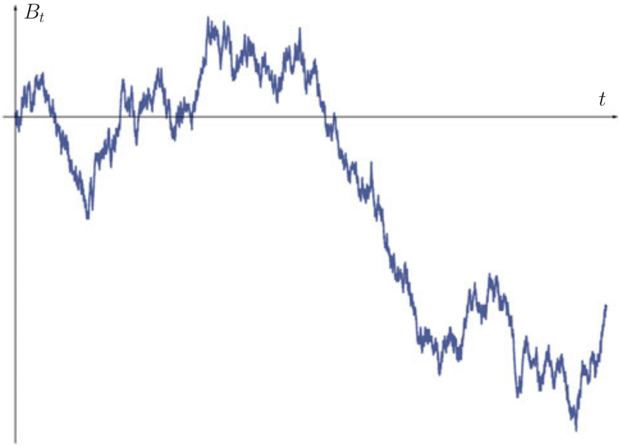

Fig. 14.1 Simulation of the graph of a linear Brownian motion

Other properties related to the irregularity of Brownian sample paths will be found in the exercises below. Figure 14.1 gives a rough idea of what the graph of a linear Brownian motion may look like.

14.5 The Strong Markov Property

Our next goal is to extend the simple Markov property (Proposition 14.8 (iii)) to the case where the deterministic time s is replaced by a random time T. In a way very similar to the case of Markov chains (cf. Section 13.3), the admissible random times are the stopping times, which we need to define in the continuous-time setting of this chapter.

In this section, B is a d-dimensional Brownian motion started at $x \in \mathbb{R}^d$. We use the same notation $\mathcal{F}_t = \sigma(B_r, 0 \le r \le t)$ and $\mathcal{F}_\infty = \sigma(B_t, t \ge 0)$ as in the previous section.

Definition 14.12 A random variable T with values in $[0, \infty]$ is a stopping time (of the filtration $(\mathcal{F}_t)_{t \ge 0}$) if $\{T \le t\} \in \mathcal{F}_t$, for every $t \ge 0$.

If T is a stopping time, then, for every $t > 0$,

$$\{T < t\} = \bigcup_{q \in \mathbb{Q} \cap [0,t)} \{T \le q\}$$

is also in $\mathcal{F}_t$, and consequently $\{T = t\} = \{T \le t\} \backslash \{T < t\} \in \mathcal{F}_t$. It also follows from the definition that a stopping time is $\mathcal{F}_\infty$-measurable. As in the discrete-time setting of Chapter 12, one easily verifies that $S \wedge T$ is a stopping time if S and T

are both stopping times. Since constant times are obviously stopping times, we get that $T \wedge t$ is a stopping time, for any stopping time T and $t \geq 0$. Similarly, it is straightforward to verify that $T + t$ is a stopping time.

Example In dimension $d = 1$, $T_a = \inf\{t \geq 0 : B_t = a\}$ is a stopping time. Indeed,

$$\{T_a \leq t\} = \left\{ \inf_{r \in \mathbb{Q} \cap [0,t]} |B_r - a| = 0 \right\} \in \mathcal{F}_t.$$

Definition 14.13 Let T be a stopping time. The σ-field of the past before T is defined by

$$\mathcal{F}_T := \{A \in \mathcal{F}_\infty : \forall t \geq 0,\ A \cap \{T \leq t\} \in \mathcal{F}_t\}.$$

We leave it as an exercise to check that $\mathfrak{F}_T$ is a σ-field. If S and T are two stopping times such that $S \leq T$, then, for every $A \in \mathcal{F}_S$ and $t \geq 0$, we have

$$A \cap \{T \leq t\} = (A \cap \{S \leq t\}) \cap \{T \leq t\} \in \mathcal{F}_t,$$

which shows that $\mathcal{F}_S \subset \mathcal{F}_T$.

Lemma 14.14 *Let T be a stopping time. The random variable T is $\mathcal{F}_T$-measurable. For every $\omega \in \{T < \infty\}$, set*

$$B_T(\omega) = B_{T(\omega)}(\omega).$$

Then B_T is $\mathcal{F}_T$-measurable.

Remark As in Proposition 12.11, we use the notion of measurability for a function defined only on a measurable subset of the underlying space.

Proof The first assertion is very easy. It is enough to verify that $\{T \leq s\} \in \mathcal{F}_T$, for every $s \geq 0$. However, we have $\{T \leq s\} \cap \{T \leq t\} = \{T \leq s \wedge t\} \in \mathcal{F}_{s \wedge t} \subset \mathcal{F}_t$ which gives the desired result.

To prove the second assertion, we introduce the following notation. For $n \in \mathbb{N}$ and for any $t \geq 0$, $\lfloor t \rfloor_n$ stands for the largest real of the form $k2^{-n}$, $k \in \mathbb{Z}_+$, smaller than or equal to t. We then observe that, for $\omega \in \{T < \infty\}$, we have $B_T(\omega) = \lim B_{\lfloor T \rfloor_n}(\omega)$ as $n \to \infty$, and so it is sufficient to prove that $B_{\lfloor T \rfloor_n}$ is $\mathcal{F}_T$-measurable. If $A \in \mathcal{B}(\mathbb{R}^d)$, we have

$$\{B_{\lfloor T \rfloor_n} \in A\} = \bigcup_{k=0}^{\infty} \Big(\{k2^{-n} \leq T < (k+1)2^{-n}\} \cap \{B_{k2^{-n}} \in A\} \Big).$$

We claim that $\{k2^{-n} \leq T < (k+1)2^{-n}\} \cap \{B_{k2^{-n}} \in A\}$ is in $\mathcal{F}_T$, for every integer $k \geq 0$. From the last display, this will show that $\{B_{\lfloor T \rfloor_n} \in A\} \in \mathcal{F}_T$ and thus $B_{\lfloor T \rfloor_n}$ is $\mathcal{F}_T$-measurable, which is the desired result.

To get our claim, we have to verify that, for every $t \geq 0$, the event

$$\Big(\{k2^{-n} \leq T < (k+1)2^{-n}\} \cap \{B_{k2^{-n}} \in A\}\Big) \cap \{T \leq t\}$$

is in $\mathcal{F}_t$. However, this event is empty if $t < k2^{-n}$, and, if $t \geq k2^{-n}$, it is equal to $\{k2^{-n} \leq T \leq t \wedge (k+1)2^{-n}\} \cap \{B_{k2^{-n}} \in A\}$, which belongs to $\mathcal{F}_t$, because, on one hand, $\{B_{k2^{-n}} \in A\} \in \mathcal{F}_{k2^{-n}} \subset \mathcal{F}_t$ and, on the other hand, $\{k2^{-n} \leq T \leq t \wedge (k+1)2^{-n}\} = \{T \leq t \wedge (k+1)2^{-n}\} \backslash \{T < k2^{-n}\}$. This completes the proof of our claim. □

Theorem 14.15 (Strong Markov Property) *Let T be a stopping time such that $\mathbb{P}(T < \infty) > 0$. For every $t \geq 0$, define a random variable $B_t^{(T)}$ by setting*

$$B_t^{(T)} := \begin{cases} B_{T+t} - B_T & \text{if } T < \infty, \\ 0 & \text{if } T = \infty. \end{cases}$$

Then, under the conditional probability $\mathbb{P}(\cdot \mid T < \infty)$, the process $(B_t^{(T)})_{t \geq 0}$ is a Brownian motion started at 0 and independent of $\mathcal{F}_T$.

Remark The fact that $B_t^{(T)}$ is a random variable (and is in fact $\mathcal{F}_{T+t}$-measurable) follows from Lemma 14.14.

Proof Clearly it is enough to consider the case when $B_0 = 0$ (otherwise write $B_t = x + B'_t$ and note that $\sigma(B'_s, s \leq t) = \sigma(B_s, s \leq t)$).

We first assume that $T < \infty$ a.s. We claim that, for every $A \in \mathcal{F}_T$ and $0 \leq t_1 < \cdots < t_p$, and every bounded continuous function F from $(\mathbb{R}^d)^p$ into $\mathbb{R}_+$, we have

$$\mathbb{E}[\mathbf{1}_A\, F(B_{t_1}^{(T)}, \ldots, B_{t_p}^{(T)})] = \mathbb{P}(A)\, \mathbb{E}[F(B_{t_1}, \ldots, B_{t_p})]. \tag{14.10}$$

The statement of the theorem (when $T < \infty$ a.s.) follows from (14.10). Indeed, by taking $A = \Omega$, we get that $(B^{(T)})_{t \geq 0}$ satisfies property (P1), and since property (P2) is obvious by construction, we obtain that $(B_t^{(T)})_{t \geq 0}$ is a Brownian motion started at 0. Furthermore, (14.10) also shows that, for every choice of $0 \leq t_1 < \cdots < t_p$, the random vector $(B_{t_1}^{(T)}, \ldots, B_{t_p}^{(T)})$ is independent of $\mathcal{F}_T$, which implies that $B^{(T)}$ is independent of $\mathcal{F}_T$.

Let us verify our claim (14.10). For every integer $n \geq 1$ and every $s \geq 0$, let $\lceil s \rceil_n$ be the smallest real of the form $k2^{-n}$, $k \in \mathbb{Z}_+$ greater than or equal to s. Clearly, $0 \leq \lceil s \rceil_n - s < 2^{-n}$. We have thus a.s.

$$\begin{aligned} &F(B_{t_1}^{(T)}, \ldots, B_{t_p}^{(T)}) \\ &\quad = F(B_{T+t_1} - B_T, \ldots, B_{T+t_p} - B_T) \end{aligned}$$

$$= \lim_{n\to\infty} F(B_{\lceil T\rceil_n+t_1} - B_{\lceil T\rceil_n}, \dots, B_{\lceil T\rceil_n+t_p} - B_{\lceil T\rceil_n})$$

$$= \lim_{n\to\infty} \sum_{k=0}^{\infty} \mathbf{1}_{\{(k-1)2^{-n}<T\le k2^{-n}\}} F(B_{k2^{-n}+t_1} - B_{k2^{-n}}, \dots, B_{k2^{-n}+t_p} - B_{k2^{-n}}),$$

where we notice that the sum over k contains at most one non-zero term. By dominated convergence, it follows that

$$\mathbb{E}[\mathbf{1}_A\, F(B^{(T)}_{t_1}, \dots, B^{(T)}_{t_p})]$$
$$= \lim_{n\to\infty} \sum_{k=0}^{\infty} \mathbb{E}[\mathbf{1}_A \mathbf{1}_{\{(k-1)2^{-n}<T\le k2^{-n}\}} F(B^{(k2^{-n})}_{t_1}, \dots, B^{(k2^{-n})}_{t_p})], \tag{14.11}$$

where we use the same notation $B^{(k2^{-n})}_t = B_{k2^{-n}+t} - B_{k2^{-n}}$ as in Proposition 14.8 (iii). Since $A \in \mathcal{F}_T$, the event $A \cap \{(k-1)2^{-n} < T \le k2^{-n}\}$ is $\mathcal{F}_{k2^{-n}}$-measurable. By the simple Markov property (Proposition 14.8 (iii)), the latter event is independent of $(B^{(k2^{-n})}_t)_{t\ge 0}$, and we have thus

$$\mathbb{E}[\mathbf{1}_{A\cap\{(k-1)2^{-n}<T\le k2^{-n}\}} F(B^{(k2^{-n})}_{t_1}, \dots, B^{(k2^{-n})}_{t_p})]$$
$$= \mathbb{P}(A \cap \{(k-1)2^{-n} < T \le k2^{-n}\})\, \mathbb{E}[F(B^{(k2^{-n})}_{t_1}, \dots, B^{(k2^{-n})}_{t_p})]$$
$$= \mathbb{P}(A \cap \{(k-1)2^{-n} < T \le k2^{-n}\})\, \mathbb{E}[F(B_{t_1}, \dots, B_{t_p})],$$

since $B^{(k2^{-n})}$ is a Brownian motion started at 0. Our claim (14.10) follows since summing the latter equality over $k \in \mathbb{Z}_+$ shows that the series in the right-hand side of (14.11) is equal to $\mathbb{P}(A)\,\mathbb{E}[F(B_{t_1}, \dots, B_{t_p})]$ (independently of n!).

When $\mathbb{P}(T = \infty) > 0$, the very same argument gives

$$\mathbb{E}[\mathbf{1}_{A\cap\{T<\infty\}}\, F(B^{(T)}_{t_1}, \dots, B^{(T)}_{t_p})] = \mathbb{P}(A \cap \{T < \infty\})\, \mathbb{E}[F(B_{t_1}, \dots, B_{t_p})]$$

and the statement of the theorem follows in the same way as in the case $T < \infty$ a.s. □

The following theorem is an important application of the strong Markov property, which is known as the *reflection principle* (Fig. 14.2).

Theorem 14.16 *Suppose that $(B_t)_{t\ge 0}$ is a one-dimensional Brownian motion started at* 0*, and, for every $t \ge 0$, set $S_t = \sup_{s\le t} B_s$. Then, for every $t > 0$, $a \ge 0$ and $b \in (-\infty, a]$, we have*

$$\mathbb{P}(S_t \ge a,\ B_t \le b) = \mathbb{P}(B_t \ge 2a - b).$$

Consequently S_t has the same law as $|B_t|$.

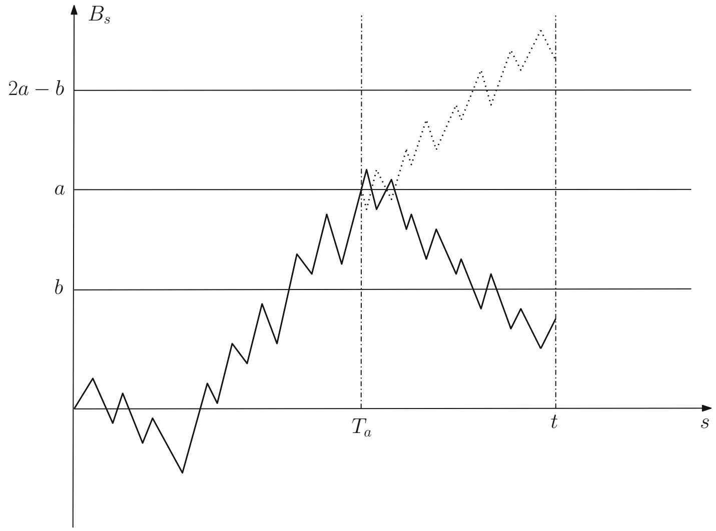

Fig. 14.2 Illustration of the reflection principle. Knowing that Brownian motion has hit a before time t, the probability that it lies below b at time t is the same as the probability that it lies above $2a - b$ at the same time. This corresponds to replacing the part of the graph of Brownian motion between times T_a and t by its reflection across the horizontal line of ordinate a (this reflection appears in dotted lines in the figure)

Proof We apply the strong Markov property to the stopping time

$$T_a = \inf\{t \geq 0 : B_t = a\}.$$

We know from Corollary 14.10 that $T_a < \infty$ a.s. We have $\{S_t \geq a\} = \{T_a \leq t\}$ a.s. (by the equality $B_0 = 0$ a.s. and the continuity of sample paths). Hence,

$$\mathbb{P}(S_t \geq a,\ B_t \leq b) = \mathbb{P}(T_a \leq t,\ B_t \leq b)$$

Now notice that, with the notation of Theorem 14.15, we have

$$\{T_a \leq t, B_t \leq b\} = \{T_a \leq t, B^{(T_a)}_{t-T_a} \leq b - a\}$$

because, on the event $\{T_a \leq t\}$, we have $B^{(T_a)}_{t-T_a} = B_t - B_{T_a} = B_t - a$. Define $H = \{(s, \mathrm{w}) \in \mathbb{R}_+ \times C(\mathbb{R}_+, \mathbb{R});\ s \leq t,\ \mathrm{w}(t-s) \leq b - a\}$ and observe that H is a measurable subset of $\mathbb{R}_+ \times C(\mathbb{R}_+, \mathbb{R})$ (we leave the proof to the reader). Then, the event in the last display can also be written as

$$\{(T_a, B^{(T_a)}) \in H\}$$

where we view $B^{(T_a)} = (B^{(T_a)}_t)_{t\geq 0}$ as a random variable taking values in $C(\mathbb{R}_+, \mathbb{R})$, as explained after Definition 14.7.

By Theorem 14.15, $B^{(T_a)}$ is a Brownian motion independent of $\mathcal{F}_{T_a}$, and in particular of T_a (by Lemma 14.14, T_a is $\mathcal{F}_{T_a}$-measurable). Since $B^{(T_a)}$ has the same law as $-B^{(T_a)}$, the law of the pair $(T_a, B^{(T_a)})$, which by independence is the product of the laws of T_a and of $B^{(T_a)}$, is also the same as the law of $(T_a, -B^{(T_a)})$. Hence,

$$\begin{aligned}
\mathbb{P}(S_t \geq a,\ B_t \leq b) &= \mathbb{P}((T_a, B^{(T_a)}) \in H] \\
&= \mathbb{P}[(T_a, -B^{(T_a)}) \in H) \\
&= \mathbb{P}(T_a \leq t,\ -B^{(T_a)}_{t-T_a} \leq b - a) \\
&= \mathbb{P}(T_a \leq t,\ B_t \geq 2a - b) \\
&= \mathbb{P}(B_t \geq 2a - b)
\end{aligned}$$

because the fact that $2a - b \geq a$ shows that the event $\{B_t \geq 2a - b\}$ is contained in $\{T_a \leq t\}$.

As for the second assertion, by noting that $\mathbb{P}(B_t = a) = 0$ and using the case $b = a$ of the first assertion, we get

$$\mathbb{P}(S_t \geq a) = \mathbb{P}(S_t \geq a, B_t \geq a) + \mathbb{P}(S_t \geq a, B_t \leq a) = 2\mathbb{P}(B_t \geq a) = \mathbb{P}(|B_t| \geq a),$$

which gives the desired result. □

In the next two corollaries, we keep the notation and assumptions of Theorem 14.16 ($(B_t)_{t\geq 0}$ is a one-dimensional Brownian motion started at 0).

Corollary 14.17 *Let $t > 0$. The law of the pair (S_t, B_t) has a density given by*

$$g_t(a, b) = \frac{2(2a - b)}{\sqrt{2\pi t^3}} \exp\left(-\frac{(2a - b)^2}{2t}\right) \mathbf{1}_{\{a>0,b<a\}}.$$

Proof The previous theorem shows that, for every $a \geq 0$ and $b \in (-\infty, a]$,

$$\mathbb{P}(S_t \geq a,\ B_t \leq b) = \mathbb{P}(B_t \geq 2a - b) = \frac{1}{\sqrt{2\pi t}} \int_{2a-b}^{\infty} e^{-x^2/(2t)}\,\mathrm{d}x.$$

since B_t is Gaussian $\mathcal{N}(0, t)$. It is then a simple matter to verify that

$$\mathbb{P}(S_t \geq a,\ B_t \leq b) = \int_{[a,\infty)\times(-\infty,b]} g_t(x, y)\,\mathrm{d}x\mathrm{d}y,$$

and this suffices to obtain that g_t is the density of (S_t, B_t). □

Corollary 14.18 *Let $a > 0$ and $T_a = \inf\{t \geq 0 : B_t = a\}$. Let N be a Gaussian $\mathcal{N}(0, 1)$ random variable. Then T_a has the same law as a^2/N^2 and has a density given by*

$$f_a(t) = \frac{a}{\sqrt{2\pi t^3}} \exp\Big(-\frac{a^2}{2t}\Big) \mathbf{1}_{\{t>0\}}.$$

Proof For every $t > 0$,

$$\begin{aligned}
\mathbb{P}(T_a \leq t) &= \mathbb{P}(S_t \geq a) \\
&= \mathbb{P}(|B_t| \geq a) \qquad \text{(Theorem 14.16)} \\
&= \mathbb{P}(B_t^2 \geq a^2) \\
&= \mathbb{P}(tN^2 \geq a^2) \qquad (B_t \text{ has the same law as } \sqrt{t}N) \\
&= \mathbb{P}(\frac{a^2}{N^2} \leq t).
\end{aligned}$$

Then an easy calculation gives the density of a^2/N^2 (Exercise 8.7). □

Reformulation on the Canonical Space In view of forthcoming developments, it will be convenient to reformulate the strong Markov property in the framework of the canonical construction of Brownian motion given at the end of Section 14.3. **From now on**, we thus consider the canonical space $\mathbf{\Omega} = C(\mathbb{R}_+, \mathbb{R}^d)$ equipped with the σ-field $\mathcal{C}$, on which we define $B_t(\mathrm{w}) := \mathrm{w}(t)$, and the σ-fields $\mathfrak{F}_t := \sigma(B_s, 0 \leq s \leq t)$. Recall that, for every $x \in \mathbb{R}^d$, we have a (unique) probability measure $\mathbf{P}_x$ on $(\mathbf{\Omega}, \mathcal{C})$ such that $(B_t)_{t\geq 0}$ is a Brownian motion started at x under $\mathbf{P}_x$. We will use w rather than ω for the generic element of $\mathbf{\Omega}$.

As in the case of Markov chains, one of the reasons for using the canonical construction is the fact that it allows the definition of *translation operators*. For every $s \geq 0$, the mapping $\theta_s : \mathbf{\Omega} \longrightarrow \mathbf{\Omega}$ is defined by

$$(\theta_s \mathrm{w})(t) := \mathrm{w}(s + t) \ , \ \forall t \geq 0.$$

Equivalently, $B_t \circ \theta_s = B_{s+t}$, for every $t \geq 0$.

Let T be a stopping time of the filtration $(\mathfrak{F}_t)_{t\geq 0}$. Then we may consider the mapping

$$\theta_T : \{T < \infty\} \longrightarrow \mathbf{\Omega}$$

which is (obviously) defined by $\theta_T(\mathrm{w}) = \theta_{T(\mathrm{w})}(\mathrm{w})$. It is not hard to prove that this mapping is measurable, using the same method as in the proof of Lemma 14.14. In fact, with the notation introduced in this proof, we have for every w such that $T(\mathrm{w}) < \infty$, $\theta_T(\mathrm{w}) = \lim_{n\to\infty} \theta_{\lfloor T \rfloor_n}(\mathrm{w})$, and it therefore suffices to prove that $\theta_{\lfloor T \rfloor_n}$

is measurable for every n. This is easy since $\lfloor T \rfloor_n$ takes only countably many values, and $\theta_{\lfloor T \rfloor_n}$ coincides with $\theta_{k2^{-n}}$ on the (measurable) event where $\lfloor T \rfloor_n = k2^{-n}$.

Theorem 14.19 *Let T be a stopping time, and let F and G be two nonnegative measurable functions on $\mathbf{\Omega}$. We assume that F is $\mathfrak{F}_T$-measurable. Then, for every $x \in \mathbb{R}^d$,*

$$\mathbf{E}_x[\mathbf{1}_{\{T<\infty\}} F\; G \circ \theta_T] = \mathbf{E}_x[\mathbf{1}_{\{T<\infty\}} F\; \mathbf{E}_{B_T}[G]].$$

Remark This statement is obviously similar to Theorem 13.8.

Proof We may assume that $\mathbf{P}_x(T < \infty) > 0$ (otherwise the result is trivial). To simplify notation, we write $\mathbf{P}_x^{(T)}$ for the probability $\mathbf{P}_x^{(T)} = \mathbf{P}_x(\cdot \mid T < \infty)$ (and $\mathbf{E}_x^{(T)}$ for the expectation under $\mathbf{P}_x^{(T)}$). We first observe that, if $T(\mathrm{w}) < \infty$,

$$(\theta_T \mathrm{w})(t) = \mathrm{w}(T+t) = \mathrm{w}(T) + (\mathrm{w}(T+t) - \mathrm{w}(T)) = B_T(\mathrm{w}) + B_t^{(T)}(\mathrm{w}),$$

with the notation of Theorem 14.15. Then,

$$\mathbf{E}_x^{(T)}[F\; G \circ \theta_T] = \mathbf{E}_x^{(T)}[F\; G(B_T + B^{(T)})] = \mathbf{E}_x^{(T)}[F\, \mathbf{E}_x^{(T)}[G(B_T + B^{(T)}) \,|\, \mathfrak{F}_T]],$$

where as previously $B^{(T)}$ denotes the random continuous function $(B_t^{(T)})_{t \geq 0}$. Then, on one hand, B_T is $\mathfrak{F}_T$-measurable, and, on the other hand, Theorem 14.15 shows that $B^{(T)}$ is independent of $\mathfrak{F}_T$ under $\mathbf{P}_x^{(T)}$, and its law under $\mathbf{P}_x^{(T)}$ is $\mathbf{P}_0$. Using Theorem 11.9, we get

$$\mathbf{E}_x^{(T)}[G(B_T + B^{(T)}) \,|\, \mathfrak{F}_T] = \int \mathbf{P}_0(d\mathrm{w})\, G(B_T + \mathrm{w}) = \mathbf{E}_{B_T}[G]$$

and the desired result follows. □

14.6 Harmonic Functions and the Dirichlet Problem

Throughout this section and the next one, we use the canonical construction of Brownian motion, and in particular $(B_t)_{t \geq 0}$ is a Brownian motion started at x under the probability measure $\mathbf{P}_x$, for every $x \in \mathbb{R}^d$. To avoid trivialities, we assume that $d \geq 2$.

In Section 7.3, we introduced Lebesgue measure on the unit sphere $\mathbb{S}^{d-1}$, which is denoted by ω_d. The *uniform probability measure* on $\mathbb{S}^{d-1}$ is the probability measure σ_d obtained by normalizing ω_d: $\sigma_d(A) = (\omega_d(\mathbb{S}^{d-1}))^{-1}\omega_d(A)$ for every Borel subset A of $\mathbb{S}^{d-1}$. By Theorem 7.4, σ_d is related to Lebesgue measure λ_d on

$\mathbb{R}^d$ via the explicit formula

$$\sigma_d(A) = \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}} \lambda_d(\{rx : 0 \le r \le 1,\ x \in A\}).$$

Similarly as ω_d, the measure σ_d is invariant under vector isometries. Furthermore, by Theorem 7.4, we have for every Borel measurable function $f : \mathbb{R}^d \longrightarrow \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = c_d \int_0^\infty \int_{\mathbb{S}^{d-1}} f(rz)\, r^{d-1}\,\mathrm{d}r\, \sigma_d(\mathrm{d}z). \tag{14.12}$$

where $c_d = \dfrac{2\pi^{d/2}}{\Gamma(d/2)}$.

Lemma 14.20 *The measure σ_d is the only probability measure on $\mathbb{S}^{d-1}$ that is invariant under vector isometries.*

Proof Let μ be a probability measure on $\mathbb{S}^{d-1}$ and assume that μ is invariant under vector isometries. Then, for every $\xi \in \mathbb{R}^d$ and every vector isometry Φ,

$$\widehat{\mu}(\xi) = \int e^{\mathrm{i}\xi\cdot x}\mu(\mathrm{d}x) = \int e^{\mathrm{i}\xi\cdot\Phi^{-1}(x)}\mu(\mathrm{d}x) = \int e^{\mathrm{i}\Phi(\xi)\cdot x}\mu(\mathrm{d}x) = \widehat{\mu}(\Phi(\xi)).$$

It follows that $\widehat{\mu}(\xi)$ only depends on $|\xi|$, and therefore there is a bounded measurable function $f : \mathbb{R}_+ \longrightarrow \mathbb{C}$ such that, for every $\xi \in \mathbb{R}^d$,

$$\widehat{\mu}(\xi) = f(|\xi|).$$

By the same argument, there is a bounded measurable function $g : \mathbb{R}_+ \longrightarrow \mathbb{C}$ such that

$$\widehat{\sigma}_d(\xi) = g(|\xi|).$$

Then, for every $r \ge 0$,

$$\int_{\mathbb{S}^{d-1}} \Big(\int_{\mathbb{S}^{d-1}} e^{\mathrm{i}r\xi\cdot x}\mu(\mathrm{d}x)\Big)\sigma_d(\mathrm{d}\xi) = \int_{\mathbb{S}^{d-1}} f(r)\,\sigma_d(\mathrm{d}\xi) = f(r)$$

and, by the Fubini theorem, this is also equal to

$$\int_{\mathbb{S}^{d-1}} \Big(\int_{\mathbb{S}^{d-1}} e^{\mathrm{i}rx\cdot\xi}\sigma_d(\mathrm{d}\xi)\Big)\mu(\mathrm{d}x) = \int_{\mathbb{S}^{d-1}} g(r)\,\mu(\mathrm{d}x) = g(r).$$

Hence $f = g$, so that $\widehat{\mu} = \widehat{\sigma}_d$ and $\mu = \sigma_d$ by Theorem 8.16. □

If $x \in \mathbb{R}^d$ and $r > 0$, we denote the open ball of radius r centered at x by $B(x,r)$, and $\bar{B}(x,r)$ stands for the closed ball. The uniform probability measure on the sphere of radius r centered at x is by definition the pushforward of $\sigma_d(\mathrm{d}y)$ under the mapping $y \to x + ry$, and is denoted by $\sigma_{x,r}$.

Recall that, until the end of this chapter, we consider the canonical construction of Brownian motion.

Proposition 14.21 *Let $x \in \mathbb{R}^d$ and $r > 0$. Let S be the stopping time*

$$S = \inf\{t \geq 0 : |B_t - x| \geq r\}.$$

The law of B_S under $\mathbf{P}_x$ is the probability measure $\sigma_{x,r}$.

Proof By elementary translation and scaling arguments, it is enough to consider the case $x = 0$, $r = 1$, and then $\sigma_{x,r} = \sigma_d$. In that case, Proposition 14.8 (i) shows that the law of B_S is invariant under vector isometries. Thanks to Lemma 14.20, the law of B_S must be σ_d. □

Recall that a domain D is a connected open subset of $\mathbb{R}^d$. A function $h : D \longrightarrow \mathbb{R}$ is said to be locally bounded if it is bounded on every compact subset of D.

Definition 14.22 Let D be a domain of $\mathbb{R}^d$. A locally bounded measurable function $h : D \longrightarrow \mathbb{R}$ is called harmonic if, for every $x \in D$ and every $r > 0$ such that the closed ball $\bar{B}(x, r)$ is contained in D, we have

$$h(x) = \int h(y)\, \sigma_{x,r}(\mathrm{d}y). \tag{14.13}$$

In other words, the value of h at x coincides with its mean over the sphere of radius r centered at x, provided that the closed ball $\bar{B}(x, r)$ is contained in D.

The Classical Dirichlet Problem Let D be a domain of $\mathbb{R}^d$ with $D \neq \mathbb{R}^d$, and let $g : \partial D \longrightarrow \mathbb{R}$ be a continuous function. A function $h : D \longrightarrow \mathbb{R}$ is said to satisfy the Dirichlet problem in D with boundary condition g if

- $h_{|\partial D} = g$ in the sense that, for every $y \in \partial D$,

$$g(y) = \lim_{x \to y, x \in D} h(x)\ ;$$

- h is harmonic on D.

The next theorem provides a candidate for the solution of the Dirichlet problem.

Theorem 14.23 *Let D be a bounded domain, and let g be a bounded measurable function on ∂D. Set*

$$T = \inf\{t \geq 0 : B_t \notin D\}.$$

Then the function

$$h(x) := \mathbf{E}_x[g(B_T)], \qquad x \in D$$

is harmonic on D.

We may view this theorem as a continuous analog of Theorem 13.36 (i).

Proof To see that T is a stopping time, we write

$$\{T \le t\} = \Big\{ \inf_{0 \le s \le t, s \in \mathbb{Q}} \mathrm{dist}(B_s, D^c) = 0 \Big\},$$

using the notation $\mathrm{dist}(x, A) = \inf\{|x - y| : y \in A\}$. Since D is bounded, properties of one-dimensional Brownian motion (Corollary 14.10) imply that $T < \infty$, $\mathbf{P}_x$ a.s. Obviously, $B_T \in \partial D$, $\mathbf{P}_x$ a.s. if $x \in D$. We know that B_T is a random variable (it is even $\mathfrak{F}_T$-measurable) and thus $h(x) := \mathbf{E}_x[g(B_T)]$ is well-defined for $x \in D$ and bounded by $\sup\{|g(y)| : y \in \partial D\}$.

Let us explain why h is measurable on D. Recall the notation $\mathcal{C}$ for the σ-field on $C(\mathbb{R}_+, \mathbb{R}^d)$. Then, for every $A \in \mathcal{C}$, the mapping $x \to \mathbf{P}_x(A)$ is measurable: this property holds for cylinder sets of the form $A = \{\mathrm{w} : \mathrm{w}(t_1) \in A_1, \ldots, \mathrm{w}(t_p) \in A_p\}$, since there is an explicit formula in that case, and it then suffices to use a monotone class argument. It follows that the mapping $x \to \mathbf{E}_x[F]$ is measurable, for every nonnegative measurable function F on $C(\mathbb{R}_+, \mathbb{R}^d)$. By applying this property to the function

$$F(\mathrm{w}) = \mathbf{1}_{\{T(\mathrm{w})<\infty\}} g(B_T(\mathrm{w}))$$

which is measurable by Lemma 14.14, we get that h is measurable.

Let us now fix $x \in D$ and $r > 0$ such that $\bar{B}(x, r) \subset D$. Set

$$S = \inf\{t \ge 0 : B_t \notin B(x, r)\} = \inf\{t \ge 0 : |B_t - x| \ge r\}.$$

Clearly, $S \le T$, $\mathbf{P}_x$ a.s. (in fact $S(\mathrm{w}) \le T(\mathrm{w})$ for every $\mathrm{w} \in \boldsymbol{\Omega} = C(\mathbb{R}_+, \mathbb{R}^d)$). Moreover

$$B_T = B_T \circ \theta_S\,, \quad \mathbf{P}_x \text{ a.s.}$$

Indeed, this is just saying that, if $t \to \mathrm{w}(t)$ is a continuous path starting from x, the exit point from D for w is the same as the exit point from D for the path derived from w by "erasing" the initial portion of w between time 0 and the first time when w exits $B(x, r)$.

We can thus use the strong Markov property in the form given by Theorem 14.19 and get

$$h(x) = \mathbf{E}_x[g(B_T)] = \mathbf{E}_x[g(B_T) \circ \theta_S] = \mathbf{E}_x[\mathbf{E}_{B_S}[g(B_T)]] = \mathbf{E}_x[h(B_S)]$$
$$= \int h(y)\, \sigma_{x,r}(\mathrm{d}y)$$

where the last equality is Proposition 14.21. This completes the proof. □

In order to show that the function h in the preceding theorem solves the Dirichlet problem (this would at least require g to be continuous), we would need to show that, for every $y \in \partial D$,

$$g(y) = \lim_{x \to y, x \in D} \mathbf{E}_x[g(B_T)].$$

Intuitively, if $x \in D$ is close to $y \in \partial D$, one might expect that Brownian motion starting from x will quickly exit D, so that the exit point B_T will be close to x, and thus also close to y, and the fact that g is continuous will ensure that $g(B_T)$ is close to $g(y)$. To make this argument rigorous, we will need certain additional assumptions. Before that, we start by discussing the uniqueness to the Dirichlet problem.

The next proposition shows that harmonic functions are automatically smooth.

Proposition 14.24 *If h is harmonic on the domain D, then h is infinitely differentiable on D. Moreover, if $x \in D$ and $r > 0$ are such that $\bar{B}(x, r) \subset D$, we have*

$$h(x) = \frac{1}{\lambda_d(B(x,r))} \int_{B(x,r)} h(y)\, \mathrm{d}y. \tag{14.14}$$

Proof We may assume that D is bounded and h is bounded on D (we may replace D by an open ball with closure contained in D). Fix $r_0 > 0$, and set

$$D_0 = \{x \in D : \mathrm{dist}(x, D^c) > r_0\}.$$

It is enough to show that h is infinitely differentiable on D_0—we assume implicitly that r_0 is small enough so that D_0 is not empty. To this end, consider a C^∞ function $\phi : \mathbb{R} \to \mathbb{R}_+$ with compact support contained in $(0, r_0)$, and not identically zero. For every $x \in D_0$ and every $r \in (0, r_0)$, we have

$$h(x) = \int \sigma_{x,r}(\mathrm{d}z)\, h(z) = \int \sigma_d(\mathrm{d}y)\, h(x + ry).$$

We multiply both the left-hand side and the right-hand side by $r^{d-1}\phi(r)$ and then integrate with respect to Lebesgue measure $\mathrm{d}r$ between 0 and r_0. Using formula (14.12) we get, with a constant $c > 0$ depending only on ϕ, for every

$x \in D_0$,

$$\begin{aligned} c\,h(x) &= c_d \int_0^{r_0} \mathrm{d}r\, r^{d-1}\phi(r) \int \sigma_d(\mathrm{d}y)\, h(x+ry) \\ &= \int_{B(0,r_0)} \mathrm{d}z\, \phi(|z|) h(x+z) \\ &= \int_{B(x,r_0)} \mathrm{d}z\, \phi(|z-x|) h(z) \\ &= \int_{\mathbb{R}^d} \mathrm{d}z\, \phi(|z-x|) \widetilde{h}(z) \end{aligned}$$

where in the last equality we wrote $\widetilde{h}$ for the function defined on $\mathbb{R}^d$ such that $\widetilde{h}(z) = h(z)$ if $z \in D$ and $\widetilde{h}(z) = 0$ if $z \notin D$.

Hence, on D_0, h is equal to the convolution of the function $z \to \phi(|z|)$, which is infinitely differentiable and has compact support, with the function $\widetilde{h}$, which is bounded and measurable. An application of Theorem 2.13 easily shows that such a convolution is infinitely differentiable.

We still have to prove the last assertion. We replace ϕ by $\mathbf{1}_{[0,r_0)}$ in the preceding calculation, and we get that, for every $x \in D_0$,

$$h(x) = c' \int_{B(x,r_0)} \mathrm{d}z\, h(z)$$

with a constant c' independent of h. Taking $h = 1$ (which is harmonic!), we get that $c' = (\lambda_d(B(x, r_0)))^{-1}$ giving the desired result. □

Corollary 14.25 *Let D be a bounded domain, and let g be a continuous function on ∂D. If a solution to the Dirichlet problem in D with boundary condition g exists, it is unique.*

Proof Suppose that h_1 and h_2 are two solutions, and set $f = h_1 - h_2$. We argue by contradiction and assume that f is not identically zero. Up to interchanging the roles of h_1 and h_2, we may assume that there exists $x \in D$ such that $f(x) > 0$. If we extend f by the value 0 on ∂D, the resulting function is continuous on $\bar{D}$, and thus attains its maximum $M > 0$ in D. Let $x_0 \in D$ such that $f(x_0) = M$. By the preceding proposition, for every $r < \mathrm{dist}(x_0, D^c)$,

$$f(x_0) = \frac{1}{\lambda_d(B(x_0, r))} \int_{B(x_0,r)} \mathrm{d}y\, f(y),$$

so that

$$\int_{B(x_0,r)} \mathrm{d}y\, (f(x_0) - f(y)) = 0.$$

Since $f(x_0) \geq f(y)$ for every $y \in D$, this is only possible if $f(x_0) = f(y)$, $\lambda_d(dy)$ a.e. on $B(x_0, r)$. Since f is continuous (again by Proposition 14.24) we have thus $f(y) = f(x_0) = M$ for every $y \in B(x_0, r)$. We get that $\{x \in D : f(x) = M\}$ is an open set. On the other hand, this set is also closed and since D is connected we have necessarily $\{x \in D : f(x) = M\} = D$, which is absurd since f tends to 0 at the boundary of D. □

Definition 14.26 We say that a domain D of $\mathbb{R}^d$ satisfies the exterior cone condition if, for every $y \in \partial D$, there exists $r > 0$ and a circular cone C with apex y such that $\mathsf{C} \cap B(y, r) \subset D^c$.

Theorem 14.27 *Assume that D is a bounded domain that satisfies the exterior cone condition, and let g be a continuous function on ∂D. Let T be as in Theorem 14.23. Then the function*

$$h(x) = \mathbf{E}_x[g(B_T)], \qquad x \in D$$

is the unique solution of the Dirichlet problem in D with boundary condition g.

Proof Thanks to Theorem 14.23 and Corollary 14.25, it is enough to verify that, for every fixed $y \in \partial D$,

$$\lim_{x \to y, x \in D} h(x) = g(y). \tag{14.15}$$

Let us fix $\varepsilon > 0$. To prove (14.15), we have to verify that we can choose $\alpha > 0$ small enough so that the conditions $x \in D$ and $|x - y| < \alpha$ imply $|h(x) - g(y)| < \varepsilon$.

By the continuity of g, we may first choose $\delta > 0$ such that the conditions $z \in \partial D$ and $|z - y| < \delta$ imply

$$|g(z) - g(y)| < \frac{\varepsilon}{3}.$$

Then let $M > 0$ such that $|g(z)| < M$ for every $z \in \partial D$. We have, for every $x \in D$ and $\eta > 0$,

$$\begin{aligned}
|\mathbf{E}_x[g(B_T)] - g(y)| &\leq \mathbf{E}_x[|g(B_T) - g(y)|\mathbf{1}_{\{T \leq \eta\}}] + \mathbf{E}_x[|g(B_T) - g(y)|\mathbf{1}_{\{T > \eta\}}] \\
&\leq \mathbf{E}_x[|g(B_T) - g(y)|\mathbf{1}_{\{T \leq \eta\}}\mathbf{1}_{\{\sup_{t \leq \eta} |B_t - x| \leq \delta/2\}}] \\
&\quad + 2M\mathbf{P}_x\Big(\sup_{t \leq \eta} |B_t - x| > \frac{\delta}{2}\Big) + 2M\,\mathbf{P}_x(T > \eta) \\
&= I + II + III.
\end{aligned}$$

We will now bound separately the three terms I, II, III.

Consider first the term II. From translation invariance, we have

$$II = 2M\mathbf{P}_0\Big(\sup_{t\leq\eta}|B_t| > \frac{\delta}{2}\Big)$$

so that the term II does not depend on x. Clearly, II tends to 0 when $\eta \to 0$ (this is saying that $\sup_{t\leq\eta}|B_t| \longrightarrow 0$ in probability under $\mathbf{P}_0$, which holds since the continuity of sample paths even gives a.s. convergence). We can therefore fix $\eta > 0$ small enough in such a way that $II < \varepsilon/3$.

Then, if $x \in D$ and $|x-y| < \frac{\delta}{2}$, we have, on the event $\{T \leq \eta\} \cap \{\sup_{t\leq\eta}|B_t - x| \leq \delta/2\}$,

$$|B_T - y| \leq |B_T - x| + |x - y| < \delta$$

and our choice of δ ensures that the term I is bounded above by $\varepsilon/3$.

Since ε was arbitrary, the desired result (14.15) will follow if we can choose $\alpha \in (0, \delta/2)$ sufficiently small so that the condition $|x-y| < \alpha$ implies that the term $III = 2M\,\mathbf{P}_x(T > \eta)$ is also bounded above by $\varepsilon/3$. This is a consequence of the next lemma, which therefore completes the proof of the theorem. □

Lemma 14.28 *Under the exterior cone condition, we have for every $y \in \partial D$ and every $\eta > 0$,*

$$\lim_{x\to y, x\in D} \mathbf{P}_x(T > \eta) = 0.$$

Remark As it was already suggested after the proof of Theorem 14.23, the key point to verify the boundary condition (14.15) is to check that Brownian motion starting near the boundary of D will quickly exit D with high probability. This is precisely what the previous lemma tells us. The exterior cone condition is not the best possible one, but it already yields certain interesting applications.

Proof We start by making the exterior cone condition more explicit. Fix $y \in \partial D$. For every $u \in \mathbb{S}^{d-1}$ and $\gamma \in (0, 1)$, consider the circular cone

$$C(u, \gamma) = \{z \in \mathbb{R}^d : z \cdot u > (1-\gamma)|z|\}.$$

Then, we can choose $r > 0$, $u \in \mathbb{S}^{d-1}$ and $\gamma \in (0, 1)$ such that

$$y + (C(u, \gamma) \cap B(0, r)) \subset D^c.$$

To simplify notation, set $C = C(u, \gamma) \cap B(0, r)$, and

$$\widetilde{C} = \{z \in \mathbb{R}^d : z \cdot u > (1 - \frac{\gamma}{2})|z|\} \cap B(0, \frac{r}{2})$$

which is the intersection with $B(0, \frac{r}{2})$ of a cone "slightly smaller" than $C(u, r)$.

If O is an open subset of $\mathbb{R}^d$, write $T_O = \inf\{t \geq 0 : B_t \in O\}$. Note that T_O is a random variable because $\{T_O < t\} = \cup_{s\in[0,t)\cap\mathbb{Q}}\{B_s \in O\}$. As an easy consequence of the zero-one law (Theorem 14.9), we have

$$T_{\widetilde{C}} = 0\ ,\ \ \mathbf{P}_0 \text{ a.s..}$$

Indeed, if $(\varepsilon_n)_{n\in\mathbb{N}}$ is a sequence of positive reals decreasing to 0, the event $\limsup\{B_{\varepsilon_n} \in \widetilde{C}\}$ belongs to the σ-field $\mathfrak{F}_{0+}$, and an argument analogous to the proof of Corollary 14.10 shows that this event has positive probability.

For $a \in (0, r/2)$, set

$$\widetilde{C}_a = \widetilde{C} \cap \bar{B}(0, a)^c.$$

Since the open sets $\widetilde{C}_a$ increase to $\widetilde{C}$ when $a \downarrow 0$, we have $T_{\widetilde{C}_a} \downarrow T_{\widetilde{C}} = 0$ as $a \downarrow 0$, $\mathbf{P}_0$ a.s.

Let us fix $\beta > 0$. By the last observation, we can choose $a > 0$ small enough so that

$$\mathbf{P}_0(T_{\widetilde{C}_a} \leq \eta) > 1 - \beta.$$

Using the fact that $y + C \subset D^c$, we have for $x \in D$,

$$\mathbf{P}_x(T \leq \eta) \geq \mathbf{P}_x(T_{y+C} \leq \eta) = \mathbf{P}_0(T_{y-x+C} \leq \eta).$$

However simple geometric considerations (left to the reader!) show that, as soon as $|y - x|$ is small enough, the shifted cone $y - x + C$ contains $\widetilde{C}_a$ (see Fig. 14.3), and then

$$\mathbf{P}_x(T \leq \eta) \geq \mathbf{P}_0(T_{\widetilde{C}_a} \leq \eta) > 1 - \beta$$

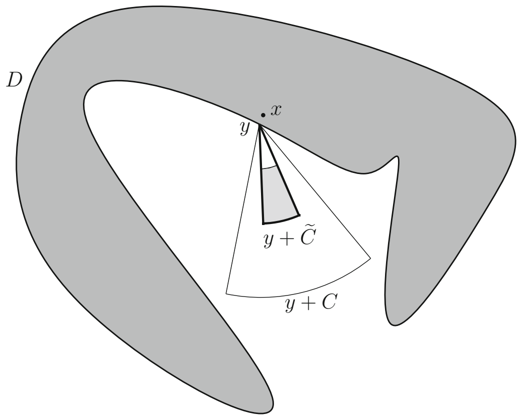

Fig. 14.3 Illustration of the proof. The set $y + C$ is delimited by thin lines, and the set $y + \widetilde{C}$ by thick lines. Then $y + \widetilde{C}_a$ appears as the shaded part inside these thick lines. One observes that, if $x \in D$ and x is close to y, the set $x + \widetilde{C}_a$ will still be contained in $y + C$

by our choice of a. Since β was arbitrary, this completes the proof of the lemma and of Theorem 14.27. □

We now derive an analytic characterization of harmonic functions, which is often used as the definition of these functions. Recall that, if f is twice continuously differentiable on D, the Laplacian Δf is defined by

$$\Delta f(x) = \sum_{j=1}^{d} \frac{\partial^2 f}{\partial y_j^2}(x), \quad x \in D.$$

Proposition 14.29 *Let h be a real function defined on D. Then h is harmonic on D if and only if h is twice continuously differentiable on D and $\Delta h = 0$.*

Proof Let h be harmonic on D. By Proposition 14.24, h is infinitely differentiable on D. Let $x \in D$ and let $r_0 > 0$ be such that the ball $\bar{B}(x, r_0)$ is contained in D. By Proposition 14.24 again, we have, for every $r \in (0, r_0]$,

$$h(x) = \frac{1}{\lambda_d(B(x,r))} \int_{B(x,r)} h(y)\,\mathrm{d}y. \tag{14.16}$$

On the other hand, Taylor's formula at order 2 shows that, if $y = (y_1, \ldots, y_d) \in B(x, r)$,

$$h(y) = h(x) + \sum_{i=1}^{d} \frac{\partial h}{\partial y_i}(x)\,(y_i - x_i) + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 h}{\partial y_i \partial y_j}(x)\,(y_i - x_i)(y_j - x_j) + o(r^2)$$

where the remainder $o(r^2)$ is uniform when y varies in $B(x, r)$. By integrating the equality of the last display over $B(x, r)$, and using obvious symmetries, we get

$$\int_{B(x,r)} h(y)\,\mathrm{d}y = \lambda_d(B(x,r))\,h(x) + \frac{1}{2} \sum_{i=1}^{d} \frac{\partial^2 h}{\partial y_i^2}(x) \int_{B(x,r)} (y_i - x_i)^2 \mathrm{d}y + o(r^{d+2}).$$

Set $C_1 = \int_{B(0,1)} y_1^2 \mathrm{d}y > 0$. The last display and (14.16) give

$$\frac{C_1}{2}\,\Delta h(x)\,r^{d+2} + o(r^{d+2}) = 0$$

which is only possible if $\Delta h(x) = 0$.

Conversely, assume that h is twice continuously differentiable on D and $\Delta h = 0$. It is enough to prove that, if U is an open ball whose closure $\bar{U}$ is contained in D, then h is harmonic on U. By Theorem 14.27, there exists a (unique) function $\widetilde{h}$ that is continuous on $\bar{U}$, harmonic on U, and such that $\widetilde{h}(x) = h(x)$ for $x \in \partial U$. The first part of the proof shows that $\Delta \widetilde{h} = 0$ sur U. By applying the next lemma to the two functions $h - \widetilde{h}$ and $\widetilde{h} - h$ (viewed as defined on $\bar{U}$) we get that $h = \widetilde{h}$ on $\bar{U}$, which completes the proof. □

Lemma 14.30 (Maximum Principle) *Let V be a bounded open subset of $\mathbb{R}^d$, and let u be a continuous function on the closure $\bar{V}$ of V. Assume that u is is twice continuously differentiable on V and $\Delta u \geq 0$ on V. Then,*

$$\sup_{x \in \bar{V}} u(x) = \sup_{x \in \partial V} u(x).$$

Proof First assume that we have the stronger property $\Delta u > 0$ on V. We argue by contradiction and assume that

$$\sup_{x \in \bar{V}} u(x) > \sup_{x \in \partial V} u(x).$$

Then, we can find $x_0 \in V$ such that

$$u(x_0) = \sup_{x \in V} u(x).$$

It follows that

$$\frac{\partial u}{\partial y_j}(x_0) = 0\ ,\ \forall j \in \{1, \ldots, d\}$$

and moreover Taylor's formula at order 2 ensures that the symmetric matrix

$$M_{x_0} = \Big(\frac{\partial^2 u}{\partial y_i \partial y_j}(x_0)\Big)_{i,j \in \{1,\ldots,d\}}$$

is nonpositive, meaning that the associated quadratic form takes nonpositive values. In particular, all eigenvalues of M_{x_0} are nonpositive and so is the trace of M_{x_0}. This is a contradiction since the trace of M_{x_0} is $\Delta u(x_0) > 0$.

Under the weaker assumption $\Delta u \geq 0$ on D, we set, for every $\varepsilon > 0$, and every $x = (x_1, \ldots, x_d) \in \bar{V}$,

$$u_\varepsilon(x) = u(x) + \varepsilon x_1^2,$$

in such a way that $\Delta u_\varepsilon = \Delta u + 2\varepsilon > 0$. The first part of the proof ensures that

$$\sup_{x \in \bar{V}} u_\varepsilon(x) = \sup_{x \in \partial V} u_\varepsilon(x),$$

and the desired result follows by letting ε tend to 0. □

14.7 Harmonic Functions and Brownian Motion

We start with an important result connecting Brownian motion, martingales and harmonic functions. To state this result, we first need to define continuous-time martingales (this is a direct generalization of the discrete-time martingales studied in Chapter 12). Recall that we assume $d \geq 2$ and that we argue on the canonical space of Brownian motion as defined at the end of Section 14.3. As previously, for every $t \geq 0$, $\mathfrak{F}_t$ denotes the σ-field generated by $(B_s, s \leq t)$. A collection $(M_t)_{t \geq 0}$ of integrable random variables indexed by the nonnegative reals is a (continuous-time) martingale under the probability measure $\mathbf{P}_x$ if M_t is $\mathfrak{F}_t$-measurable, for every $t \geq 0$, and the relation $M_s = \mathbf{E}_x[M_t \,|\, \mathfrak{F}_s]$ holds for every $0 \leq s \leq t$.

For every open subset U of $\mathbb{R}^d$, we set $H_U = \inf\{t \geq 0 : B_t \notin U\}$. It is easy to verify that H_U is a stopping time (just write $\{H_U \leq t\} = \{\inf\{\text{dist}(B_s, U^c) : s \in [0, t] \cap \mathbb{Q}\} = 0\}$).

Theorem 14.31 *Let D be a domain in $\mathbb{R}^d$. A continuous function $h : D \longrightarrow \mathbb{R}$ is harmonic if and only if, for every bounded open set U with closure $\bar{U} \subset D$, the random process*

$$(h(B_{t \wedge H_U}))_{t \geq 0}$$

is a martingale under $\mathbf{P}_x$, for every $x \in U$.

Informally, a function is harmonic if and only if its composition with Brownian motion is a martingale.

Proof Let us first assume that h is harmonic, and let U be an open set satisfying the conditions of the statement. To simplify notation, set $H = H_U$. Also fix $x \in U$ and note that the random variables $h(B_{t \wedge H})$ are bounded above, $\mathbf{P}_x$ a.s., by $\sup\{|h(y)| : y \in \bar{U}\} < \infty$.

Let $s \leq t$. The random variable $B_{s \wedge H}$ is $\mathfrak{F}_{s \wedge H}$-measurable (Lemma 14.14) hence also $\mathfrak{F}_s$-measurable (by the remark before Lemma 14.14). To get the desired equality $E[h(B_{t \wedge H}) \,|\, \mathfrak{F}_s] = h(B_{s \wedge H})$, it is therefore enough to show that, for every bounded $\mathfrak{F}_s$-measurable random variable F, one has

$$\mathbf{E}_x[F\, h(B_{s \wedge H})] = \mathbf{E}_x[F\, h(B_{t \wedge H})].$$

We may view the restriction of h to U as the (unique) solution to the Dirichlet problem in U with a boundary condition given by the restriction of h to ∂U. If U satisfies the exterior cone condition, Theorem 14.27 shows that necessarily for every $y \in U$,

$$h(y) = \mathbf{E}_y[h(B_H)]. \tag{14.17}$$

In fact this equality holds even if U does not satisfy the exterior cone condition. To see this, set $U_\varepsilon := \{y \in U : \text{dist}(y, U^c) > \varepsilon\}$ for every $\varepsilon > 0$ and notice that (if ε is small) U_ε is nonempty and satisfies the exterior cone condition. So for every $y \in U$, for ε small enough so that $y \in U_\varepsilon$, the preceding considerations give $h(y) = \mathbf{E}_y[h(B_{H_{U_\varepsilon}})]$, and we just have to let $\varepsilon \to 0$ using dominated convergence (note that $H_{U_\varepsilon} \uparrow H_U$ as $\varepsilon \downarrow 0$).

It follows from (14.17) that

$$\mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}h(B_s)] = \mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}\,\mathbf{E}_{B_s}[h(B_H)]].$$

However, since $F\,\mathbf{1}_{\{s<H\}}$ is $\mathfrak{F}_s$-measurable, the Markov property, in the form stated in Theorem 14.19 with $T = s$, implies that

$$\mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}\,\mathbf{E}_{B_s}[h(B_H)]] = \mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}h(B_H)\circ\theta_s] = \mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}h(B_H)],$$

since $B_H \circ \theta_s = B_H$ on the event $\{s < H\}$. We thus get

$$\begin{aligned}
\mathbf{E}_x[F\,h(B_{s\wedge H})] &= \mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}h(B_s)] + \mathbf{E}_x[F\,\mathbf{1}_{\{s\geq H\}}h(B_H)]\\
&= \mathbf{E}_x[F\,\mathbf{1}_{\{s<H\}}h(B_H)] + \mathbf{E}_x[F\,\mathbf{1}_{\{s\geq H\}}h(B_H)]\\
&= \mathbf{E}_x[F\,h(B_H)].
\end{aligned}$$

The same argument gives

$$\mathbf{E}_x[F\,h(B_{t\wedge H})] = \mathbf{E}_x[F\,h(B_H)]$$

and we arrive at the desired equality.

The converse is easier. Assume that h satisfies the property in the statement, and let U be an open ball with closure contained in D. By the martingale property, for every $x \in U$,

$$h(x) = \mathbf{E}_x[h(B_{0\wedge H_U})] = \mathbf{E}_x[h(B_{t\wedge H_U})].$$

By letting $t \to \infty$ and using dominated convergence, we have $h(x) = \mathbf{E}_x[h(B_{H_U})]$, and Theorem 14.23 shows that h is harmonic on U. □

Proposition 14.32 *Let $0 \leq a < b$ and let $D_{a,b}$ be the domain*

$$D_{a,b} = B(0,b)\backslash\bar{B}(0,a).$$

Let $f : D_{a,b} \longrightarrow \mathbb{R}$ be a radial function, meaning that $f(x)$ only depends on $|x|$. Then f is harmonic if and only if there exist two constants $C, C' \in \mathbb{R}$ such that

$$f(x) = \begin{cases} C + C'\log|x| & \text{if } d = 2,\\ C + C'|x|^{2-d} & \text{if } d \geq 3.\end{cases}$$

Proof Suppose that f is harmonic. By Proposition 14.24, f is infinitely differentiable. Let $g : (a, b) \longrightarrow \mathbb{R}$ be the function such that $f(x) = g(|x|)$. The formula for the Laplacian of a radial function reads

$$\Delta f(x) = g''(|x|) + \frac{d-1}{|x|}\, g'(|x|).$$

By Proposition 14.29, g solves the differential equation

$$g''(r) + \frac{d-1}{r}\, g'(r) = 0.$$

Solving this equation shows that f is of the form given in the proposition. Conversely, if f is of this form, Proposition 14.29 shows that f is harmonic. □

In the next two statements, we use the notation $T_A = \inf\{t \geq 0 : B_t \in A\}$ for any closed subset A of $\mathbb{R}^d$.

Proposition 14.33 *Let $x \in \mathbb{R}^d \backslash \{0\}$, and let $\varepsilon, R > 0$ such that $\varepsilon < |x| < R$. Then,*

$$\mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < T_{B(0,R)^c}) = \begin{cases} \dfrac{\log R - \log|x|}{\log R - \log \varepsilon} & \text{if } d = 2, \\ \dfrac{|x|^{2-d} - R^{2-d}}{\varepsilon^{2-d} - R^{2-d}} & \text{if } d \geq 3. \end{cases} \tag{14.18}$$

Remark In dimension $d = 1$, the corresponding formula is

$$\mathbf{P}_x(T_a < T_b) = \frac{b-x}{b-a}$$

for $a < x < b$. This can be proved in exactly the same manner, noting that harmonic functions are affine functions in dimension one. This may be compared with the analogous formula for simple random walk on $\mathbb{Z}$ in the gambler's ruin example of Section 12.6.

Proof Consider the domain $D = D_{\varepsilon,R}$, with the notation of Proposition 14.32. This domain satisfies the exterior cone condition. Let g be the continuous function on ∂D defined by

$$\begin{cases} g(y) = 1 & \text{if } |y| = \varepsilon, \\ g(y) = 0 & \text{if } |y| = R. \end{cases}$$

Then Theorem 14.27 shows that

$$h(x) := \mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < T_{B(0,R)^c}) \,, \quad \varepsilon < |x| < R$$

is the unique solution to the Dirichlet problem with boundary condition g. Using Proposition 14.32, we immediately see that the right-hand side of (14.18) solves the same Dirichlet problem. The desired result follows. □

The preceding proposition yields interesting information on the almost sure behavior of the Brownian sample paths.

Corollary 14.34

(i) *If $d \geq 3$, for every $\varepsilon > 0$ and $x \in \mathbb{R}^d$ such that $\varepsilon < |x|$,*

$$\mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < \infty) = \Big(\frac{\varepsilon}{|x|}\Big)^{d-2}.$$

Moreover, for every $x \in \mathbb{R}^d$,

$$\lim_{t\to\infty} |B_t| = \infty \ , \qquad \mathbf{P}_x \textit{ a.s.}$$

(ii) *If $d = 2$, for every $\varepsilon > 0$ and $x \in \mathbb{R}^2$ such that $\varepsilon < |x|$,*

$$\mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < \infty) = 1$$

but

$$\mathbf{P}_x(T_{\{0\}} < \infty) = 0.$$

Moreover, $\mathbf{P}_x$ a.s., for every open subset U of $\mathbb{R}^2$, the set $\{t \geq 0 : B_t \in U\}$ is unbounded.

By analogy with the case of Markov chains, we say that Brownian motion is transient in dimension $d \geq 3$ and recurrent in dimension $d = 2$. Note that this recurrence property does not mean that all points are visited: on the contrary, a fixed point of $\mathbb{R}^2$ other than the starting point is not visited with probability one.

Proof

(i) The first assertion is easy since

$$\mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < \infty) = \lim_{n\uparrow\infty} \mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < T_{B(0,n)^c})$$

and we just have to apply (14.18).

Then, for every integer $n \geq 1$, set

$$T_{(n)} = T_{B(0,2^n)^c}.$$

By applying the strong Markov property at $T_{(n)}$ (in the form given by Theorem 14.19), and using the first assertion, we get for $|x| \leq 2^n$,

$$\mathbf{P}_x\Big(\inf_{t \geq T_{(n)}} |B_t| \leq n\Big) = \mathbf{E}_x\Big[\mathbf{P}_{B_{T_{(n)}}}(T_{\bar{B}(0,n)} < \infty)\Big] = (\frac{n}{2^n})^{d-2}.$$

The Borel-Cantelli lemma then implies that, $\mathbf{P}_x$ a.s. for every large enough integer n, we have

$$\inf_{t \geq T_{(n)}} |B_t| > n.$$

It follows that the function $t \to |B_t|$ converges to ∞ as $t \to \infty$, a.s.

(ii) By formula (14.18) we have

$$\mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < T_{B(0,R)^c}) = \frac{\log R - \log |x|}{\log R - \log \varepsilon}$$

if $\varepsilon < |x| < R$. By letting R tend to ∞ in this formula, we get for $\varepsilon < |x|$,

$$\mathbf{P}_x(T_{\bar{B}(0,\varepsilon)} < \infty) = 1.$$

Letting ε tend to 0 in the same formula gives for $|x| < R$,

$$\mathbf{P}_x(T_{\{0\}} < T_{B(0,R)^c}) = 0.$$

Since $T_{B(0,R)^c} \uparrow \infty$ as $R \uparrow \infty$, this implies that, for $x \neq 0$,

$$\mathbf{P}_x(T_{\{0\}} < \infty) = 0.$$

If $x \neq 0$, we have thus both

$$\mathbf{P}_x \text{ a.s. } \forall \varepsilon > 0,\ T_{\bar{B}(0,\varepsilon)} < \infty$$

and

$$\mathbf{P}_x \text{ a.s. } 0 \notin \{B_t : t \geq 0\}.$$

These two properties imply that, $\mathbf{P}_x$ a.s., 0 is an accumulation point of the function $t \to B_t$ when $t \to \infty$. Hence, a.s. for every open set U containing 0, the set $\{t \geq 0 : B_t \in U\}$ is $\mathbf{P}_x$ a.s. unbounded. A translation invariance argument then gives the last assertion (note that it is enough to consider a countable collection of open sets U).

□

We conclude this chapter with a discussion of the Poisson kernel and its interpretation as the exit distribution of Brownian motion from a ball. The Poisson kernel (of the unit ball) is the function defined on $B(0,1) \times \mathbb{S}^{d-1}$ by

$$K(x,y) := \frac{1-|x|^2}{|x-y|^d}\,, \quad x \in B(0,1),\ y \in \mathbb{S}^{d-1}.$$

Lemma 14.35 *For every fixed $y \in \mathbb{S}^{d-1}$, the function $x \to K(x,y)$ is harmonic on $B(0,1)$.*

Proof Set $K_y(x) = K(x,y)$ for $x \in B(0,1)$. A (tedious but straightforward) calculation shows that $\Delta K_y = 0$ on $B(0,1)$, and the desired result follows from Proposition 14.29. □

Lemma 14.36 *For every $x \in B(0,1)$,*

$$\int_{\mathbb{S}^{d-1}} K(x,y)\,\sigma_d(\mathrm{d}y) = 1.$$

Proof For $x \in B(0,1)$, set

$$F(x) = \int_{\mathbb{S}^{d-1}} K(x,y)\,\sigma_d(\mathrm{d}y).$$

One easily infers from the preceding lemma that the function F is harmonic on $B(0,1)$: one way to see this is to apply Fubini's theorem to check that the mean value property (14.13) holds for F since it holds for the functions K_y (or alternatively one may differentiate under the integral sign to verify that $\Delta F = 0$). On the other hand, using the invariance properties of σ_d and K under vector isometries, one gets that F is a radial function. The restriction of F to $B(0,1)\backslash\{0\}$ must then be of the form given in Proposition 14.32. Since F is also continuous at 0, the constant C' in the formulas of this proposition must be zero and F is constant. Finally, for every $x \in B(0,1)$, $F(x) = F(0) = 1$. □

Theorem 14.37 *Let g be a continuous function on $\mathbb{S}^{d-1}$. The solution of the Dirichlet problem in $B(0,1)$ with boundary condition g is*

$$h(x) = \int_{\mathbb{S}^{d-1}} K(x,y)\,g(y)\,\sigma_d(\mathrm{d}y)\,,\ x \in B(0,1).$$

Moreover, for every $x \in B(0,1)$, the law of the exit point of Brownian motion started at x from the ball $B(0,1)$ has a continuous density with respect to $\sigma_d(\mathrm{d}y)$, and this density is the function $y \mapsto K(x,y)$.

Proof The same application of Fubini's theorem as outlined in the proof of Lemma 14.36 shows that h is harmonic in $B(0,1)$. To verify the boundary condition, fix $y_0 \in \mathbb{S}^{d-1}$. For every $\delta > 0$, the explicit form of the Poisson kernel shows that, if $x \in B(0,1)$ and $y \in \mathbb{S}^{d-1}$ are such that $|x - y_0| < \delta/2$ and $|y - y_0| > \delta$, we have

$$K(x,y) \le \left(\frac{2}{\delta}\right)^d (1 - |x|^2).$$

It follows from this bound that, for every $\delta > 0$,

$$\lim_{x \to y_0, x \in B(0,1)} \int_{\{|y-y_0|>\delta\}} K(x,y)\,\sigma(\mathrm{d}y) = 0. \tag{14.19}$$

Then, given $\varepsilon > 0$, we choose $\delta > 0$ small enough so that $|g(y) - g(y_0)| \le \varepsilon$ whenever $y \in \mathbb{S}^{d-1}$ and $|y - y_0| \le \delta$. If $M = \sup\{|g(y)| : y \in \mathbb{S}^{d-1}\}$, we have

$$\begin{aligned} |h(x) - g(y_0)| &= \left| \int_{\mathbb{S}^{d-1}} K(x,y)\,(g(y) - g(y_0))\,\sigma_d(\mathrm{d}y) \right| \\ &\le 2M \int_{\{|y-y_0|>\delta\}} K(x,y)\,\sigma(\mathrm{d}y) + \varepsilon, \end{aligned}$$

using Lemma 14.36 in the first equality, and then our choice of δ. Thanks to (14.19), we now get

$$\limsup_{x \to y_0, x \in B(0,1)} |h(x) - g(y_0)| \le \varepsilon.$$

Since ε was arbitrary, we obtain the desired boundary condition.

Finally, for the last assertion, we use Theorem 14.27, which asserts that the unique solution to the same Dirichlet problem is also given by

$$h(x) = \mathbf{E}_x[g(B_T)],$$

where $T = \inf\{t \ge 0 : B_t \notin D\}$. By comparing the two formulas for h we obtain precisely that the law of B_T under $\mathbf{P}_x$ is the measure $K(x,y)\sigma_d(\mathrm{d}y)$. □

14.8 Exercises

In Exercises 14.1 to 14.11, $(B_t)_{t \ge 0}$ is a one-dimensional Brownian motion with $B_0 = 0$, and $S_t = \sup\{B_s : 0 \le s \le t\}$.

Exercise 14.1 For every $a \ge 0$, set $T_a = \inf\{t \ge 0 : B_t = a\}$.

(1) Prove that, for every $0 \le a < b$, the random variable $T_b - T_a$ is independent of $\sigma(T_c, 0 \le c \le a)$ and has the same distribution as T_{b-a}.

(2) Prove that, for every $a_1, \dots, a_n \in \mathbb{R}_+$ and $\lambda > 0$, the vector $(T_{\lambda a_1}, \dots, T_{\lambda a_n})$ has the same law as $(\lambda^2 T_{a_1}, \dots, \lambda^2 T_{a_n})$.
(3) Let $n \in \mathbb{N}$ and let $T^{(1)}, T^{(2)}, \dots, T^{(n)}$ be n independent random variables distributed as T_1. Verify that $T^{(1)} + \cdots + T^{(n)}$ has the same distribution as $n^2 T_1$. Comment on the relation between this result and the strong law of large numbers.

Exercise 14.2 Let $a > 0$ and $T_a = \inf\{t \geq 0 : B_t = a\}$. Prove that we have almost surely

$$T_a = \inf\{t \geq 0 : B_t > a\}.$$

Exercise 14.3 Prove that

$$\left(\int_0^t e^{B_s}\,\mathrm{d}s\right)^{1/\sqrt{t}} \xrightarrow[t\to\infty]{\text{(d)}} e^{|N|},$$

where N is a Gaussian $\mathcal{N}(0,1)$ random variable.

Exercise 14.4

(1) Prove that a.s.,

$$\limsup_{t\downarrow 0} \frac{B_t}{\sqrt{t}} = +\infty\,, \quad \liminf_{t\downarrow 0} \frac{B_t}{\sqrt{t}} = -\infty.$$

(2) Let $s > 0$. Prove that a.s. the function $t \mapsto B_t$ is not differentiable at s.

Exercise 14.5

(1) For every integer $n \geq 1$, set

$$\Sigma_n = \sum_{k=1}^{2^n} \left(B_{k2^{-n}} - B_{(k-1)2^{-n}}\right)^2.$$

Compute $\mathbb{E}[\Sigma_n]$ and $\mathrm{var}(\Sigma_n)$, and prove that Σ_n converges in L^2 and a.s. to a constant as $n \to \infty$.
(2) Prove that a.s. the function $t \mapsto B_t(\omega)$ is not of bounded variation on the interval $[0,1]$ (see Exercise 6.1 for the definition of functions of bounded variation).

Exercise 14.6

(1) For every $t \in [0, 1]$, set $B'_t = B_{1-t} - B_1$. Prove that the two random processes $(B_t)_{t\in[0,1]}$ and $(B'_t)_{t\in[0,1]}$ have the same law (as in the definition of the Wiener measure, this law is a probability measure on the space $C([0, 1], \mathbb{R})$).
(2) Let $t > 0$. Prove that $S_t - B_t$ and S_t have the same law without using Corollary 14.17.

Exercise 14.7 Let $\tau = \inf\{t \geq 0 : B_t = S_1\}$.

(1) Prove that $0 < \tau < 1$ a.s. (one may use the preceding exercise) and then that τ is **not** a stopping time.
(2) Using question (2) of the preceding exercise, verify (without making any calculation) that, for every $a \in (0, 1)$,

$$\mathbb{P}(\tau > a) = \mathbb{P}(\sqrt{1-a}|N| > \sqrt{a}|N'|)$$

where N and N' are two independent Gaussian $\mathcal{N}(0, 1)$ random variables.
(3) Conclude that the law of τ is the arcsine distribution of Exercise 9.3.

Exercise 14.8 Show the local maxima of B are almost surely distinct. In other words, a.s. for any rationals $0 \leq a < b < c < d$, we have

$$\sup_{a\leq t\leq b} B_t \neq \sup_{c\leq t\leq d} B_t.$$

Exercise 14.9 Let $H = \{t \in [0, 1] : B_t = 0\}$. Using Corollary 14.10 and the strong Markov property, prove that H is a.s. a compact subset of $[0, 1]$ with no isolated points and zero Lebesgue measure.

Exercise 14.10

(1) For every $a > 0$, set $\sigma_a = \inf\{t \geq 0 : |B_t| \geq a\}$. Show that there is a constant $\gamma \in (0, 1)$ depending on a such that, for every integer $N \geq 1$,

$$\mathbb{P}(\sigma_a > N) \leq \gamma^N.$$

(2) For every $n \geq 1$, define a sequence $T_0^n, T_1^n, \ldots$ by induction by setting

$$T_0^n = 0,\ T_1^n = \sigma_{2^{-n}},\ T_{k+1}^n = \inf\{t > T_k^n : |B_t - B_{T_k^n}| = 2^{-n}\}.$$

Verify that the random times T_k^n are almost surely finite and are stopping times. Prove that the random variables $T_k^n - T_{k-1}^n$, $k = 1, 2, \ldots$, are independent and identically distributed, and similarly the random variables $B_{T_k^n} - B_{T_{k-1}^n}$, $k = 1, 2, \ldots$, are independent and identically distributed.

(3) We set $X^n_k = 2^n\, B_{T^n_k}$ for every $k \in \mathbb{Z}_+$. Verify that $(X^n_k)_{k\in\mathbb{Z}_+}$ is a simple random walk on $\mathbb{Z}$.
(4) Show that there exists a constant $c > 0$ such that, for every $t \geq 0$,

$$\lim_{n\to\infty} T^n_{\lfloor 2^{2n}t\rfloor} = c\,t\,, \quad \text{a.s.}$$

(5) Infer that, for every $t > 0$,

$$\lim_{n\to\infty} \sup_{0\leq s\leq t} \left| \frac{1}{2^n} X^n_{\lfloor 2^{2n}s\rfloor} - B_{cs} \right| = 0\,, \quad \text{a.s.}$$

and finally that $c = 1$.

Exercise 14.11 For $\alpha \in (0, 1]$, a continuous function $f : [0, 1] \longrightarrow \mathbb{R}$ is said to be α-Hölder if there exists a constant C such that $|f(s) - f(t)| \leq C\,|s - t|^\alpha$ for every $s, t \in [0, 1]$.

(1) Prove that the function $[0, 1] \ni t \mapsto B_t(\omega)$ is a.s. not $\frac{1}{2}$-Hölder.
(2) Let $\delta \in (0, \frac{1}{2})$. Prove that a.s. there exists an integer $n_0(\omega)$ such the bound $|B_{k2^{-n}} - B_{(k-1)2^{-n}}| \leq 2^{-n\delta}$ holds for every $n \geq n_0(\omega)$ and every $k \in \{1, \ldots, 2^n\}$.
(3) Prove that the function $[0, 1] \ni t \mapsto B_t(\omega)$ is a.s. δ-Hölder.

Exercise 14.12 Let $d \geq 3$ and let B be a d-dimensional Brownian motion started from 0. Fix $A > 1$ and $\delta \in (0, 1)$, and set

$$\mathcal{R}_{\delta,A} = \{x \in \mathbb{R}^d : \delta \leq |x| \leq A\}.$$

For every $n \geq 1$ and every $k_1, \ldots, k_d \in \mathbb{Z}$, define the cube

$$C^{(n)}_{k_1,\ldots,k_d} = [k_1 2^{-n}, (k_1+1)2^{-n}]\times[k_2 2^{-n}, (k_2+1)2^{-n}]\times\cdots\times[k_d 2^{-n}, (k_d+1)2^{-n}].$$

Set

$$N_n = \mathrm{card}\{(k_1, \ldots, k_d) \in \mathbb{Z}^d : C^{(n)}_{k_1,\ldots,k_d} \cap \mathcal{R}_{\delta,A} \cap \{B_t, t \geq 0\} \neq \varnothing\}.$$

(1) Prove that there is a constant K depending on δ and A such that, for every $n \geq 1$,

$$\mathbb{E}[N_n] \leq K\, 2^{2n}.$$

(2) Recall from Exercise 3.4 the definition of the Hausdorff dimension $\dim(A)$ of a subset A of $\mathbb{R}^d$. Prove that $\dim(\{B_t, t \geq 0\}) \leq 2$ a.s.

Exercise 14.13 Let D be a bounded domain in $\mathbb{R}^d$, $d \geq 2$, and let g be a continuous function on ∂D. Suppose that h solves the Dirichlet problem with boundary condition g. Show that, for every $x \in D$,

$$h(x) = \mathbf{E}_x[g(B_T)],$$

with the notation of Theorem 14.23 (in particular $T = \inf\{t \geq 0 : B_t \notin D\}$).

Exercise 14.14 Let $d \geq 2$, and let $D = \{x \in \mathbb{R}^d : 0 < |x| < 1\}$ be the punctured open unit ball. Define $g : \partial D \longrightarrow \mathbb{R}$ by setting $g(x) = 1$ if $|x| = 1$ and $g(0) = 0$. Prove that the Dirichlet problem in D with boundary condition g has no solution. (*Hint:* Use the result of the preceding exercise).

Exercise 14.15 Let $d \geq 3$. Let K be a compact subset of the closed unit ball, and $D = \mathbb{R}^d \backslash K$. We assume that D is connected and satisfies the exterior cone condition. Let $g : \partial D \longrightarrow \mathbb{R}$ be a continuous function. We consider a function u that satisfies the Dirichlet problem in D with boundary condition g, and assume that u is bounded.

We use the canonical representation of Brownian motion in $\mathbb{R}^d$, and set $T_K = \inf\{t \geq 0 : B_t \in K\} \in [0, \infty]$.

(1) Let $(R_n)_{n \in \mathbb{N}}$ be a sequence of real numbers in $(1, \infty)$ such that $R_n \uparrow \infty$ as $n \to \infty$. For every n, set $T_{(n)} = \inf\{t \geq 0 : |B_t| \geq R_{(n)}\}$. Prove that, for every $n \geq 1$ and every $x \in D$ such that $|x| < R_{(n)}$,

$$u(x) = \mathbf{E}_x[g(B_{T_K})\,\mathbf{1}_{\{T_K < T_{(n)}\}}] + \mathbf{E}_x[u(B_{T_{(n)}})\,\mathbf{1}_{\{T_{(n)} < T_K\}}].$$

(2) Prove that, up to replacing the sequence $(R_n)_{n \in \mathbb{N}}$ by a subsequence, we can assume that there exists a constant $\alpha \in \mathbb{R}$ such that, for every $x \in \mathbb{R}$,

$$\lim_{n \to \infty} \mathbf{E}_x[u(B_{T_{(n)}})] = \alpha.$$

(3) Deduce from questions (1) and (2) that

$$\lim_{|x| \to \infty} u(x) = \alpha$$

and then prove that, for every $x \in D$,

$$u(x) = \mathbf{E}_x[g(B_{T_K})\,\mathbf{1}_{\{T_K < \infty\}}] + \alpha\,\mathbf{P}_x(T_K = \infty).$$

(4) Conversely, verify that, for any $\alpha \in \mathbb{R}$, the right-hand side of the last display gives a solution of the Dirichlet problem in D with boundary condition g.

Appendix A
A Few Facts from Functional Analysis

In this appendix, we recall some basic results about Hilbert spaces and Banach spaces, which are used in the text. Proofs of these results can be found in [22] or in most textbooks on functional analysis.

Normed Linear Spaces and Banach Spaces

Let E be a linear space over the field $\mathbb{R}$. A norm on E is a mapping $x \mapsto \|x\|$ from E into $\mathbb{R}_+$ that satisfies the following properties:

- $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in E$ (triangle inequality);
- $\|ax\| = |a| \, \|x\|$ for every $x \in E$ and $a \in \mathbb{R}$;
- $\|x\| = 0$ if and only if $x = 0$.

One says that $(E, \|\cdot\|)$ is a *normed linear space*. The norm induces a distance d on E, which is defined by the formula

$$d(x, y) = \|x - y\|.$$

The normed linear space $(E, \|\cdot\|)$ is called a *Banach space* if it is complete for the distance induced by the norm. In other words, if $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in E, meaning that

$$\lim_{m,n\to\infty} \|x_m - x_n\| = 0,$$

there exists $x \in E$ such that the sequence $(x_n)_{n\in\mathbb{N}}$ converges to x.

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5

A linear form on E is a linear mapping from E into $\mathbb{R}$. A linear form Φ is continuous if and only if

$$|||\Phi||| := \sup_{\|x\|\le 1} |\Phi(x)| < \infty,$$

and we have then

$$|\Phi(x)| \le |||\Phi|||\, \|x\|$$

for every $x \in E$. The space of all continuous linear forms on E is denoted by E', and $(E', |||\cdot|||)$ is a normed linear space called the (topological) dual of E. The quantity $|||\Phi|||$ is called the operator norm of Φ.

The dual space $(E', |||\cdot|||)$ of a normed linear space $(E, \|\cdot\|)$ is always a Banach space (even if $(E, \|\cdot\|)$ is not a Banach space).

The following special case of the Hahn-Banach theorem is used in this book only to construct certain counterexamples.

Theorem A.1 (Hahn-Banach) *Let $(E, \|\cdot\|)$ be a normed linear space, and let F be a linear subspace of E. Let ϕ be a linear mapping from F into $\mathbb{R}$ such that $|\phi(y)| \le \|y\|$ for every $y \in F$. Then, there exists a continuous linear form Φ on E such that we have both $\Phi(y) = \phi(y)$ for every $y \in F$ and $|||\Phi||| \le 1$.*

Hilbert Spaces

Let H be a linear space over the field $\mathbb{R}$. A scalar product on H is a mapping $(x, y) \mapsto \langle x, y\rangle$ from $H \times H$ into $\mathbb{R}$ which satisfies the following properties:

- $\langle x, y\rangle = \langle y, x\rangle$ for every $x, y \in H$;
- $\langle x, y + z\rangle = \langle x, y\rangle + \langle x, z\rangle$ and $\langle ax, y\rangle = a\langle x, y\rangle$ for every $x, y, z \in H$ and $a \in \mathbb{R}$;
- $\langle x, x\rangle \ge 0$ for every $x \in H$, and $\langle x, x\rangle = 0$ if and only $x = 0$.

Given a scalar product on H, the formula

$$\|x\| := \sqrt{\langle x, x\rangle}$$

defines a norm on H. The triangle inequality for this norm is deduced from the Cauchy-Schwarz inequality

$$|\langle x, y\rangle| \le \|x\|\, \|y\|$$

for every $x, y \in H$.

If the normed linear space $(H, \|\cdot\|)$ is complete (for the distance induced by the norm), we say that $(H, \langle\cdot,\cdot\rangle)$ is a (real) *Hilbert space*.

In the following statements, we consider a Hilbert space $(H, \langle\cdot,\cdot\rangle)$. The following theorem applied with $H = L^2(E, \mathcal{A}, \mu)$ plays a key role in the proof of the Radon-Nikodym theorem (Theorem 4.11).

Theorem A.2 (Riesz) *For every $y \in H$, define a continuous linear form Φ_y on H by setting*

$$\Phi_y(x) := \langle x, y\rangle$$

for every $x \in H$. Then the mapping $y \mapsto \Phi_y$ is a linear isometry from H onto the dual H', meaning that it is linear and bijective, and $|||\Phi_y||| = \|y\|$ for every $y \in H$.

We now state (a special case of) the projection theorem in a Hilbert space, which is used in Chapter 11 to give an interpretation of conditional expectation.

Theorem A.3 *Let F be a closed linear subspace of H, and let $x \in H$. Then there is unique element of F denoted by $p_F(x)$ such that*

$$\|x - p_F(x)\| = \min\{\|x - y\| : y \in F\}.$$

Furthermore, $p_F(x)$ is characterized by the two properties

- *$p_F(x) \in F$;*
- *$\langle x - p_F(x), y\rangle = 0$, for every $y \in F$.*

$p_F(x)$ is called the orthogonal projection of x on F.

Two elements x and y of H are said to be orthogonal if $\langle x, y\rangle = 0$.

Proposition A.4 *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of pairwise orthogonal elements in a Hilbert space H. Then the limit*

$$\lim_{k\to\infty} \sum_{n=1}^{k} x_n$$

exists in H if and only if $\sum_{n=1}^{\infty} \|x_n\|^2 < \infty$. This limit is denoted by $\sum_{n=1}^{\infty} x_n$, and we have

$$\Big\| \sum_{n=1}^{\infty} x_n \Big\|^2 = \sum_{n=1}^{\infty} \|x_n\|^2.$$

We finally give two statements that are used in our construction of Brownian motion in Chapter 14. We say that a sequence $(e_n)_{n\in\mathbb{N}}$ of elements of H is an orthonormal system if we have

- $\langle e_m, e_n\rangle = 0$ for every $m, n \in \mathbb{N}$, $m \neq n$;
- $\|e_n\| = 1$ for every $n \in \mathbb{N}$.

We say that $(e_n)_{n\in\mathbb{N}}$ is an orthonormal basis of H if it is an orthonormal system and moreover the linear space spanned by $\{e_n : n \in \mathbb{N}\}$ is dense in H.

Theorem A.5 *Suppose that $(e_n)_{n\in\mathbb{N}}$ is an orthonormal basis of H. Then, for every $x \in H$, we have*

$$\sum_{n=1}^{\infty} \langle x, e_n\rangle^2 = \|x\|^2$$

and

$$x = \sum_{n=1}^{\infty} \langle x, e_n\rangle\, e_n.$$

The series in the second display of the theorem converges in the sense of Proposition A.4.

The last statement shows that an isometric mapping between two Hilbert spaces can be constructed by mapping an orthonormal basis of the first space to an orthonormal system in the second one.

Proposition A.6 *Let H and K be two Hilbert spaces. Without risk of confusion, we use the same notation $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ for the norm and the scalar product on both H and K. Suppose that $(e_n)_{n\in\mathbb{N}}$ is an orthonormal basis of H and $(f_n)_{n\in\mathbb{N}}$ is an orthonormal system of K. Then there is a unique mapping Ψ from H into K such that:*

- *Ψ is linear and isometric, in the sense that $\|\Psi(x)\| = \|x\|$ for every $x \in H$;*
- *$\Psi(e_n) = f_n$ for every $n \in \mathbb{N}$.*

Moreover, for every $x \in H$,

$$\Psi(x) = \sum_{n=1}^{\infty} \langle x, e_n\rangle\, f_n.$$

Notes and Suggestions for Further Reading

We provide below some suggestions for further reading. We emphasize that the list of references is very far from being exhaustive. There is a huge number of books dealing with the same topics at a comparable or more advanced level.

Chapters 1–7 The classic book of Rudin [22] is still a good source for measure theory and its applications in analysis. The books by Billingsley [2] and Dudley [7] give a detailed treatment of measure theory with a view to applications in probability. Both [22] and [7] also provide a glimpse of the functional analytic approach to measure theory, where, roughly speaking, measures are constructed from linear forms defined on an appropriate space of functions. Stroock [24] presents measure theory with an emphasis toward analytic applications.

Chapter 8 We have made no attempt to trace back the history of probability theory. Much information about this history (and the history of measure theory) can be found in the notes of the books by Dudley [7] and Kallenberg [10], which are also excellent sources for more advanced study. General references on probability theory at a level comparable to the present book include Chung [4], Durrett [8] and Grimmett and Stirzaker [9].

Chapter 9 Our elementary approach for constructing sequences of independent real random variables in Section 9.4 is sufficient for our needs in the present book, but a more elegant (and more general) method would be to extend the construction of product measures in Chapter 5 to the case of infinite products, see for instance Chapter 8 in [7]. More generally, the Kolmogorov extension theorem (see, in particular, Chapter 7 of Billingsley [1]) allows one to construct random processes with prescribed finite-dimensional marginals, and this applies to independent sequences as a very special case. A more elegant approach to the Poisson process calculations in Section 9.7 can be given via the concept of Poisson measures, see Kingman [11].

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5

Chapter 10 I learnt of the proof of the strong law of large numbers that is presented in this chapter from Neveu [15]. Other proofs are given in Chapter 12 as applications of martingale theory. The weak convergence of probability measures can be extended to measures on abstract metric spaces, see Billingsley [1] for a comprehensive account. The proof of the central limit theorem (Theorem 10.15) via characteristic functions may appear mysterious: see, in particular, Chapter 2 of Stroock [23] for proofs of more general statements that avoid using characteristic functions.

Chapter 11 Conditional distributions as defined in Definition 11.16 are usually called *regular* conditional distributions. This notion can be extended to conditioning with respect to a sub-σ-field rather than with respect to a random variable X as in Definition 11.16. See Chapter 6 of [10], which also gives a proof of Theorem 11.17.

Chapter 12 Both Williams [25] and Neveu [16] are good sources for more advanced study and applications of discrete-time martingales. Continuous-time martingales, which are briefly mentioned at the end of Chapter 14, are studied in the reference book [6] by Dellacherie and Meyer (see also Revuz and Yor [20], Rogers and Williams [21], and Le Gall [12]).

Chapter 13 See Norris [17] for more about (discrete-time) Markov chains in a countable state space, as well as for continuous-time Markov chains. Both [17] and Brémaud [3] also discuss the use of Markov chains in various areas of applications. Revuz [19] deals with Markov chains with values in general state spaces. In the setting of Theorem 13.33 giving the convergence of the law of process at time n to the invariant probability measure, a crucial question is to get information on the speed of this convergence: see Levin et al. [13] for a thorough discussion of this problem.

Chapter 14 The construction of Brownian motion in the proof of Theorem 14.3 is known as Lévy's construction. The approximation by random walks that is outlined in the first section holds in a strong form known as Donsker's theorem, see in particular Billingsley [1]. The continuity of sample paths (property (P2) of the definition of Brownian motion) is usually obtained via the Kolmogorov lemma, which applies to more general random processes (see e.g. [12] or [20]). Mörters and Peres [14] is an excellent source for various properties of Brownian motion and its sample paths. The connection between Brownian motion and harmonic functions is more commonly derived via the Itô formula of stochastic calculus, see, in particular, [12] and [20] for a presentation and various applications of stochastic calculus. Much more about connections between Brownian motion, harmonic functions and classical potential theory can be found in Port and Stone [18].

References

1. Billingsley, P., *Convergence of probability measures*. Wiley and Sons, 1968.
2. Billingsley, P., *Probability and measure*. Wiley and Sons, 1979.
3. Brémaud, P., *Markov chains, Gibbs Fields, Monte-Carlo simulations and queues*. Texts in Applied Mathematics 31. Springer, 1999.
4. Chung, K.L., *A course in probability theory*. 3rd ed. Academic Press, 2001.
5. Curien, N., Yet another proof of the strong law of large numbers. *The American Mathematical Monthly*, 2022. https://doi.org/10.1080/00029890.2022.2115820.
6. Dellacherie, C., Meyer, P.-A., *Probabilities and potential. B. Theory of martingales*. North-Holland Publishing Co., 1982.
7. Dudley, R.M., *Real analysis and probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.
8. Durrett, R., *Probability: theory and examples*. 5th ed. Cambridge University Press, 2019.
9. Grimmett, G., Stirzaker, D., *Probability and random processes*. 3rd ed. Oxford University Press, 2001.
10. Kallenberg, O., *Foundations of modern probability*. Probability and its Applications. 2nd ed. Springer, 2002.
11. Kingman, J.F.C., *Poisson processes*. Oxford University Press, 1993.
12. Le Gall, J.-F., *Brownian motion, martingales and stochastic calculus*. Graduate Texts in Mathematics 274. Springer, 2016.
13. Levin, D.A., Peres, Y., Wilmer, E.L., *Markov chains and mixing times*. American Mathematical Society, 2009.
14. Mörters, P., Peres, Y., *Brownian motion*. Cambridge University Press, 2010.
15. Neveu, J., *Probabilités*. Cours de l'Ecole polytechnique (1997). Unpublished lecture notes.
16. Neveu, J., *Discrete-parameter martingales*. North-Holland Publishing Co., 1975.
17. Norris, J., *Markov chains*. Cambridge University Press, 1997.
18. Port, S.C., Stone, C.J., *Brownian motion and classical potential theory*. Academic Press, 1978.
19. Revuz, D., *Markov chains*. North-Holland Publishing Co., 1975.
20. Revuz, D., Yor, M., *Continuous martingales and Brownian motion*. Grundlehren der mathematischen Wissenschaften 293. 3rd ed. Springer, 1999.
21. Rogers, L.C.G., Williams, D., *Diffusions, Markov processes and martingales, vol. I: Foundations*. 2nd ed. Wiley, 1994

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5

22. Rudin, W., *Real and complex analysis*. McGraw-Hill, 1974.
23. Stroock, D.W., *Probability theory, an analytic view*. Cambridge University Press, 1993.
24. Stroock, D.W., *Essentials of integration theory for analysis*. Graduate Texts in Mathematics 262. 2nd ed. Springer 2020.
25. Williams, D., *Probability with martingales*. Cambridge University Press, 1991.

Index

J.-F. Le Gall, *Measure Theory, Probability, and Stochastic Processes*,
Graduate Texts in Mathematics 295, https://doi.org/10.1007/978-3-031-14205-5

J

K

L

M

N

O

P