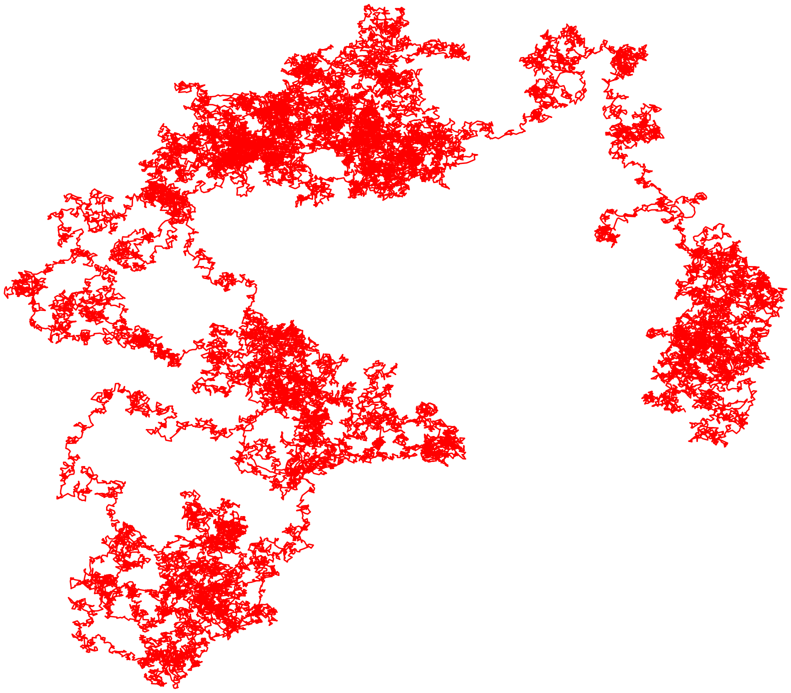
SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

École des Ponts
ParisTech

Master de Mathématiques et Applications

Spécialité *Mathématiques de la Modélisation*

Méthodes Numériques Probabilistes



Julien Reygner

Version finale — Année 2023 – 2024

Contents

Foreword

Contact. Comments and questions are welcome: julien.reygner@enpc.fr. Practical details and complementary material are provided on the webpage

https://cermics.enpc.fr/~reygnerj/anedp.html

Exercises. The exercises in these notes are labelled according to the following classification.

🗎 Can be made during class.
🏠 Can be made at home between two classes.
☕ Can be made in order to study before the exam.

Notebooks. Jupyter Notebooks are available on the course's webpage in order to implement some of the algorithms which are presented in the course. They are signaled with the icon </> in the notes.

Syllabus of the sessions. (update of December 19, 2023): this is the list of the notions which have been seen during the lectures. The programme of the exam will be based on this list.

- 17/10: Chapter 2 except Subsection 2.2.4.
- 24/10: Chapter 4.
- 07/11: Chapter 5, until the end of the proof of Theorem 5.3.3.
- 14/11: End of Section 5.3 in Chapter 5. Sections 6.1, 6.2 and 6.4 in Chapter 6, until the statement of Proposition 6.4.8.
- 21/11: Proof of Proposition 6.4.8, Subsection 6.4.3. Chapter 7.
- 28/11: Exercises (see webpage).
- 05/12: Crash course on Chapters 9 (except Subsection 9.3.5) and 10.
- 12/12: No lecture. The content of Chapter 11 is not examinable.
- 19/12: Exercises (see webpage).

Part I

Random number simulation and the Monte Carlo method

Chapter 1

Probability spaces and random variables

Contents

In this introductory Chapter we recall the basic notions of measure and probability theory with which will we work. Many results are stated without a proof; we refer to [6, 9] for details.

1.1 Probability space

1.1.1 Measure theory

We first recall the following basic definitions from measure theory and refer to [9] for complements. The complement of a set A is denoted by A^{c}. We call a set *countable* if it is either finite or in one-to-one correspondence with the set $\mathbb{N} = \{0, 1, \dots\}$ of nonnegative integers.

Definition 1.1.1 (σ-field)**.** *Let Ω be a set. A σ-*field *on Ω is a collection $\mathcal{A}$ of subsets of Ω such that:*

- $\Omega \in \mathcal{A}$*;*
- *for any $A \in \mathcal{A}$, the* complement $A^{\mathrm{c}} = \{\omega \in \Omega : \omega \notin A\}$ *belongs to $\mathcal{A}$;*
- *for any finite or countably infinite family $(A_n)_{n\geq 1}$ of elements of $\mathcal{A}$, the* union $\cup_{n\geq 1} A_n = \{\omega \in \Omega : \exists n \geq 1, \omega \in A_n\}$ *belongs to $\mathcal{A}$.*

If $\mathcal{A}$ is a σ-field on Ω, the pair $(\Omega, \mathcal{A})$ is called a *measurable space*. Elements of $\mathcal{A}$ are called *measurable sets*. Notice that by definition, σ-fields always contain the empty set $\varnothing$.

Definition 1.1.2 (Nonnegative measure)**.** *Let* $(\Omega, \mathcal{A})$ *be a measurable space. A* nonnegative measure *on* $(\Omega, \mathcal{A})$ *is a mapping* $\mu : \mathcal{A} \to [0, +\infty]$ *such that:*

- $\mu(\varnothing) = 0$*;*
- *for any finite or countably infinite family* $(A_n)_{n\geq 1}$ *of pairwise distinct elements of* $\mathcal{A}$*,* $\mu(\cup_{n\geq 1} A_n) = \sum_{n\geq 1} \mu(A_n)$*.*

The second property in Definition 1.1.2 is called *σ-additivity*. A nonnegative measure is called *finite* if $\mu(\Omega) < +\infty$, and *σ-finite* if there exists a nondecreasing sequence $(A_n)_{n\geq 1}$ of measurable sets such that $\mu(A_n) < +\infty$ for any $n \geq 1$, and $\cup_{n\geq 1} A_n = \Omega$.

A measurable set such that $\mu(A) = 0$ is called *μ-negligible*.

Definition 1.1.3 (σ-field generated by some family of sets)**.** *Let $\mathcal{C}$ be a family of subsets of Ω. The smallest σ-field which contains $\mathcal{C}$ is called the σ-field* generated *by $\mathcal{C}$.*

We shall often use the following result, which is a consequence of *Dynkin's System Theorem*.

Lemma 1.1.4 (Operational form of Dynkin's System Theorem)**.** *Let $\mathcal{C}$ be a family of subsets of Ω and $\mathcal{A}$ be the σ-field generated by $\mathcal{C}$. If $\mathcal{C}$ is stable by intersection, any two nonnegative and finite measures on $(\Omega, \mathcal{A})$ which coincide on $\mathcal{C}$ are necessarily equal.*

1.1.2 Probability space

Definition 1.1.5 (Probability space)**.** *A* probability space *is a triple* $(\Omega, \mathcal{A}, \mathbb{P})$ *such that:*

- Ω *is a set;*
- $\mathcal{A}$ *is a σ-field on Ω;*
- $\mathbb{P}$ *is a* probability measure *on* $(\Omega, \mathcal{A})$*, that is to say a nonnegative and finite measure such that* $\mathbb{P}(\Omega) = 1$.

Measurable sets $A \in \mathcal{A}$ are usually called *events*. An event A such that $\mathbb{P}(A) = 1$ is called *almost sure*. For any events A and B, we recall the elementary identity

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

from which one may for instance deduce that, denoting by A^{c} the complement of A,

$$\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A).$$

We shall sometimes refer to the inequality

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

as the *union bound*.

The following property relies on the σ-additivity property of the measure $\mathbb{P}$, its proof is omitted.

Proposition 1.1.6 (Monotonic continuity)**.** *Let $(B_n)_{n\geq 1}$ be a sequence of measurable sets.*

- *If $(B_n)_{n\geq 1}$ is nonincreasing, that is to say $B_{n+1} \subset B_n$ for any n, then*

$$\lim_{n\to+\infty} \mathbb{P}(B_n) = \mathbb{P}\left(\cap_{n\geq 1} B_n\right).$$

- *If $(B_n)_{n\geq 1}$ is nondecreasing, that is to say $B_n \subset B_{n+1}$ for any n, then*

$$\lim_{n\to+\infty} \mathbb{P}(B_n) = \mathbb{P}\left(\cup_{n\geq 1} B_n\right).$$

Proposition 1.1.6 has the following practical corollary.

Corollary 1.1.7 (Intersection of countable almost sure events)**.** *Let $(A_n)_{n\geq 1}$ be almost sure events. The event $\cap_{n\geq 1} A_n$ is almost sure.*

Proof. For any $n \geq 1$, set $B_n = \cap_{k=1}^n A_k$. By construction, the sequence $(B_n)_{n\geq 1}$ is nonincreasing, and satisfies $\cap_{n\geq 1} A_n = \cap_{n\geq 1} B_n$. Besides, we have, for any n,

$$\begin{aligned} \mathbb{P}(B_{n+1}^{\mathrm{c}}) &= \mathbb{P}((B_n \cap A_{n+1})^{\mathrm{c}}) \\ &= \mathbb{P}(B_n^{\mathrm{c}} \cup A_{n+1}^{\mathrm{c}}) \\ &\leq \mathbb{P}(B_n^{\mathrm{c}}) + \mathbb{P}(A_{n+1}^{\mathrm{c}}) \\ &= \mathbb{P}(B_n^{\mathrm{c}}), \end{aligned}$$

which by an immediate induction shows that $\mathbb{P}(B_n^{\mathrm{c}}) \leq \mathbb{P}(B_1^{\mathrm{c}}) = \mathbb{P}(A_1^{\mathrm{c}}) = 0$, and therefore $\mathbb{P}(B_n) = 1$ for any n. By Proposition 1.1.6, we conclude that

$$\mathbb{P}\left(\cap_{n\geq 1} A_n\right) = \mathbb{P}\left(\cap_{n\geq 1} B_n\right) = \lim_{n\to+\infty} \mathbb{P}(B_n) = 1,$$

which shows that the event $\cap_{n\geq 1} A_n$ is almost sure. □

1.1.3 Conditional probability

Definition 1.1.8 (Conditional probability)**.** *Let $B \in \mathcal{A}$ be such that $\mathbb{P}(B) > 0$. The* conditional probability *given B is the probability measure $\mathbb{P}(\cdot|B)$ defined on $(\Omega, \mathcal{A})$ by*

$$\forall A \in \mathcal{A}, \qquad \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Notice that $(\Omega, \mathcal{A}, \mathbb{P}(\cdot|B))$ is a probability space.

Lemma 1.1.9 (Total probability formula)**.** *Let us be given a partition of Ω into events $(B_n)_{n\geq 1}$. For any event A,*

$$\mathbb{P}(A) = \sum_{n\geq 1} \mathbb{P}(A \cap B_n) = \sum_{n\geq 1} \mathbb{P}(A|B_n)\mathbb{P}(B_n),$$

with the obvious convention that $\mathbb{P}(A|B_n)\mathbb{P}(B_n) = 0$ if $\mathbb{P}(B_n) = 0$.

1.2 Random variables

1.2.1 Definition

In this subsection, we consider measurable functions defined on $(\Omega, \mathcal{A})$ and taking their values in some measurable space $(E, \mathcal{E})$. If E is provided with a topology, we denote by $\mathcal{B}(E)$ the *Borel σ-field* on E, which is the smallest σ-field containing all open sets.

Definition 1.2.1 (Random variable)**.** *Let $(E, \mathcal{E})$ be a measurable space. A* random variable *in E is a measurable function $X : \Omega \to E$, that is to say a function such that*

$$\forall C \in \mathcal{E}, \qquad X^{-1}(C) := \{\omega \in \Omega : X(\omega) \in C\} \in \mathcal{A}.$$

The law*, or* distribution *of a random variable X is the probability measure P_X defined on $(E, \mathcal{E})$ by*

$$\forall C \in \mathcal{E}, \qquad P_X(C) = \mathbb{P}\left(X^{-1}(C)\right).$$

In other words, it is the pushforward $\mathbb{P} \circ X^{-1}$ *of $\mathbb{P}$ by X.*

The event $X^{-1}(C)$ is usually simply denoted by $\{X \in C\}$. Given a probability measure P on E, we shall also write $X \sim P$ to mean that X is distributed according to P, that is to say that $P_X = P$. Notice that the triple $(E, \mathcal{E}, P_X)$ is a probability space itself.

1.2.2 Density

Definition 1.2.2 (Absolute continuity)**.** *Let μ be a nonnegative σ-finite measure on the measurable space $(E, \mathcal{E})$. A probability measure P on $(E, \mathcal{E})$ is called* absolutely continuous *with respect to μ if*

$$\forall C \in \mathcal{E}, \qquad \mu(C) = 0 \quad \Rightarrow \quad P(E) = 0.$$

In this case, we write $P \ll \mu$.

Theorem 1.2.3 (Radon–Nikodym Theorem)**.** *If $P \ll \mu$, then there exists a measurable function $p : E \to [0, +\infty)$ such that*

$$\forall C \in \mathcal{E}, \qquad P(C) = \int_{x \in E} \mathbb{1}_{\{x \in C\}} p(x) \mathrm{d}\mu(x).$$

The function p is unique up to a μ-negligible set, it is called the *density* of P with respect to μ and is also denoted by

$$p(x) = \frac{\mathrm{d}P}{\mathrm{d}\mu}(x),$$

so that we shall often write $\mathrm{d}P(x) = p(x)\mathrm{d}\mu(x)$ to mean that P has density p with respect to μ. Obviously, a probability density p with respect to μ necessarily satisfies

$$\int_{x \in E} p(x)\mathrm{d}\mu(x) = 1,$$

where the integral is understood in the sense of Lebesgue.

As far as densities are concerned, we shall essentially work in two particular frameworks:

- $E = \mathbb{R}^d$, $\mathcal{E}$ is the Borel σ-field $\mathcal{B}(\mathbb{R}^d)$ and μ is the Lebesgue measure;
- E is countable, $\mathcal{E}$ is the power set of E (called the *discrete* σ-field) and $\mu = \sum_{x \in E} \delta_x$ is the counting measure on E.

In particular, when a random variable in $\mathbb{R}^d$ is said 'to have density p' without more precision, it is implicitly understood that it is with respect to the Lebesgue measure. On the other hand, if a variable X takes its values in the countable space E endowed with the *discrete* σ-field, then its law is characterised by the family of numbers $(P_X(\{x\}))_{x \in E}$, which is called the *Probability Mass Function* of X.

1.3 Expectation

1.3.1 Definition

For all $p \in [1, +\infty)$, we denote by $\mathbf{L}^p(\Omega, \mathcal{A}, \mathbb{P})$, or simply $\mathbf{L}^p(\mathbb{P})$ when there is no ambiguity on the underlying measurable space $(\Omega, \mathcal{A})$, the set of random variables $X : \Omega \to \mathbb{R}$ such that $|X|^p$ is Lebesgue integrable on Ω. Random variables in $\mathbf{L}^1(\mathbb{P})$ are simply called *integrable*.

Definition 1.3.1 (Expectation). *Let $X \in \mathbf{L}^1(\mathbb{P})$. The* expectation *of X is the Lebesgue integral*

$$\mathbb{E}[X] := \int_{\omega \in \Omega} X(\omega) \mathrm{d}\mathbb{P}(\omega).$$

We recall that if X is a random variable in E then $(E, \mathcal{E}, P_X)$ is a probability space, so that the spaces $\mathbf{L}^p(P_X) = \mathbf{L}^p(E, \mathcal{E}, P_X)$ are defined similarly to $\mathbf{L}^p(\mathbb{P}) = \mathbf{L}^p(\Omega, \mathcal{A}, \mathbb{P})$.

Remark 1.3.2. *When X is nonnegative but not necessarily in $\mathbf{L}^1(\mathbb{P})$, the integral in Definition 1.3.1 still makes sense in $[0, +\infty]$. Therefore, in this case, we shall sometimes write $\mathbb{E}[X]$ as an element of $[0, +\infty]$, keeping in mind that $X \in \mathbf{L}^1(\mathbb{P})$ if and only if $\mathbb{E}[X] < +\infty$. This convention also includes the case of random variables with may take the value $+\infty$, such as series of nonnegative random variables. In the latter case, it is easily checked that if $\mathbb{E}[X] < +\infty$ then necessarily $X < +\infty$, almost surely*[1] *— but, of course, the converse statement does not hold in general.*

Theorem 1.3.3 (Transfer Theorem). *Let X be a random variable in E and $f : E \to \mathbb{R}$ be a measurable function. Then $f(X) \in \mathbf{L}^1(\mathbb{P})$ if and only if $f \in \mathbf{L}^1(P_X)$, and*

$$\mathbb{E}[f(X)] = \int_{\omega \in \Omega} f(X(\omega)) \mathrm{d}\mathbb{P}(\omega) = \int_{x \in E} f(x) P_X(\mathrm{d}x).$$

In addition, if X has density p with respect to some σ-finite measure μ on E, then

$$\mathbb{E}[f(X)] = \int_{x \in E} f(x) p(x) \mathrm{d}\mu(x).$$

1.3.2 Variance and moments

Lemma 1.3.4 (Jensen inequality). *Let $X \in \mathbf{L}^1(\mathbb{P})$ and $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then $\mathbb{E}[\phi(X)]$ is well-defined in $(-\infty, +\infty]$ and*

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

Exercise 1.3.5. *Prove Lemma 1.3.4.*

For any $p \in [1, +\infty)$, the quantity $\mathbb{E}[|X|^p]$ is called the *moment of order p* of the random variable X.

Exercise 1.3.6. *Check that if $1 \leq p \leq q$, then $\mathbf{L}^q(\mathbb{P}) \subset \mathbf{L}^p(\mathbb{P})$ and $\mathbb{E}[|X|^p]^{1/p} \leq \mathbb{E}[|X|^q]^{1/q}$.*

Definition 1.3.7 (Variance). *The* variance *of a random variable $X \in \mathbf{L}^2(\mathbb{P})$ is defined by*

$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Notice that, by Exercise 1.3.6, the assumption that $X \in \mathbf{L}^2(\mathbb{P})$ ensures that $\mathbb{E}[X]$ is well-defined.

Exercise 1.3.8. *Show that, for any $X \in \mathbf{L}^2(\mathbb{P})$, for any $a, b \in \mathbb{R}$, $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$.*

[1] See Exercise 1.4.1 for an illustrative example.

1.3.3 Independence

Definition 1.3.9 (Independence)**.** *Let $X_1, \dots, X_k$ be random variables taking their values in respective measurable spaces $(E_1, \mathcal{E}_1), \dots, (E_k, \mathcal{E}_k)$. These variables are called* independent *if, for any $C_1 \in \mathcal{E}_1, \dots, C_k \in \mathcal{E}_k$,*

$$\mathbb{P}(X_1 \in C_1, \dots, X_k \in C_k) = \mathbb{P}(X_1 \in C_1) \cdots \mathbb{P}(X_k \in C_k).$$

It is clear that, equivalently, the random variables $X_1, \dots, X_k$ are independent if the law of $(X_1, \dots, X_k) \in E_1 \times \cdots \times E_k$ is the product measure $P_{X_1} \otimes \cdots \otimes P_{X_k}$. When $\mathrm{d}P_{X_i} = p_i(x_i)\mathrm{d}\mu_i(x_i)$ for any i, the latter product measure has density $p_1(x_1) \cdots p_k(x_k)$ with respect to the product measure $\mu_1 \otimes \cdots \otimes \mu_k$. Besides, this characterisation shows that if the random variables $X_1, \dots, X_k$ are independent, then for any functions $f_1 \in \mathbf{L}^1(P_{X_1}), \dots, f_k \in \mathbf{L}^1(P_{X_k})$, the random variable $f_1(X_1) \cdots f_k(X_k)$ is integrable and satisfies

$$\mathbb{E}\left[f_1(X_1) \cdots f_k(X_k)\right] = \mathbb{E}\left[f_1(X_1)\right] \cdots \mathbb{E}\left[f_k(X_k)\right].$$

🗎 **Exercise 1.3.10.** *For any* independent *variables $X, Y \in \mathbf{L}^2(\mathbb{P})$, show that* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. *What is the value of* $\mathrm{Var}(X-Y)$*?*

The notion of independence can be extended to infinitely many random variables as follows: an arbitrary family $(X_i)_{i \in I}$ of random variables is called independent if, for any finite subset of indices $\{i_1, \dots, i_k\}$, the variables $X_{i_1}, \dots, X_{i_k}$ are independent. When, in addition, all spaces E_i are the same and all variables X_i have the same law, then the family is called *independent and identically distributed*, which we shall abbreviate to *iid*.

🏠 **Exercise 1.3.11** (Independence of events)**.** *A collection of events $(A_i)_{i \in I}$ is called independent if the random variables $(\mathbb{1}_{A_i})_{i \in I}$ are independent.*

1. *Show that two events A_1 and A_2 are independent if and only if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$.*
2. *Show that if $k \geq 3$, the identity $\mathbb{P}(A_1 \cap \cdots \cap A_k) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_k)$ is necessary but not sufficient for $A_1, \dots, A_k$ to be independent.*

Notice that the first question in Exercise 1.3.11 shows in particular that two events A and B, with $\mathbb{P}(B) > 0$, are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(B)$: the knowledge that the event B is realised does not affect the probability of A.

1.3.4 Transformation of random variables

A common problem in practical applications is the following: given a random variable $X \in E$ with law P_X and a measurable function $\phi : E \to F$, how to compute the law of $Y = \phi(X)$ (which, in technical terms, is the pushforward $P_X \circ \phi^{-1}$ of P_X by ϕ)? For example, if $E = \mathbb{R}^d$ and $F = \mathbb{R}^k$, and X has density p_X with respect to the Lebesgue measure on $\mathbb{R}^d$, does Y possess a density with respect to the Lebesgue measure on $\mathbb{R}^k$, and if so, can we get an explicit expression for this density?

The *dummy function method* provides a guideline to answer this question. Assume indeed that $E = \mathbb{R}^d$ and X has density p_X with respect to the Lebesgue measure on $\mathbb{R}^d$. Then, by Theorem 1.3.3, for any measurable and bounded function $f : F \to \mathbb{R}$, we have

$$\mathbb{E}\left[f(Y)\right] = \mathbb{E}\left[f(\phi(X))\right] = \int_{x \in \mathbb{R}^d} f(\phi(x)) p_X(x) \mathrm{d}x.$$

Assume that by a suitable change of variable $x \to y = \phi(x)$, one is able to rewrite the right-hand side under the form

$$\int_{y \in F} f(y) q(y) \mathrm{d}\mu(y)$$

for some σ-finite measure μ on F. Then necessarily q is a probability density with respect to μ, and it is the density of Y since we have written that

$$\mathbb{E}[f(Y)] = \int_{y \in F} f(y) q(y) \mathrm{d}\mu(y)$$

for all bounded and measurable functions f.

If p_X vanishes outside some open subset U of $\mathbb{R}^d$ and ϕ is a C^1-diffeomorphism between U and another open subset V of $\mathbb{R}^d$, then the change of variable is immediate and yields

$$\int_{x \in \mathbb{R}^d} f(\phi(x)) p_X(x) \mathrm{d}x = \int_{x \in U} f(\phi(x)) p_X(x) \mathrm{d}x = \int_{y \in V} f(y) p_X(\phi^{-1}(y)) |J_{\phi^{-1}}(y)| \mathrm{d}y,$$

where $J_{\phi^{-1}}$ is the Jacobian determinant of ϕ^{-1}. In this case, we deduce that Y has density $\mathbb{1}_{\{y \in U\}} p_X(\phi^{-1}(y)) |J_{\phi^{-1}}(y)|$ with respect to the Lebesgue measure on $\mathbb{R}^d$.

If ϕ is not bijective, then further manipulations of the integral in x are generally needed to reduce to a case where a bijective change of variable can be applied.

🏠 **Exercise 1.3.12.** *Let X and Y be two independent random variables in $\mathbb{R}^d$ with respective densities p and q. Show that $Z = X + Y$ has density*

$$p * q(z) = \int_{y \in \mathbb{R}^d} p(z - y) q(y) \mathrm{d}y.$$

1.4 Complement: the Borel–Cantelli Lemmas

☕ **Exercise 1.4.1** (Borel–Cantelli Lemma)**.** *Let $(A_n)_{n \geq 1}$ be a sequence of events which satisfies*

$$\sum_{n \geq 1} \mathbb{P}(A_n) < +\infty. \tag{1.1}$$

1. *Show that the $[0, +\infty]$-valued random variable*

   $$X = \sum_{n \geq 1} \mathbb{1}_{A_n}$$

   is almost surely finite.
2. *Deduce that the event*

   $$\limsup_{n \to +\infty} A_n = \{\omega \in \Omega : \forall N \geq 1, \exists n \geq N : \omega \in A_n\}$$

   has probability 0.

In other words, we have proved that under the condition (1.1), the set of ω which only belong to finitely many events A_n is almost sure. This statement is called the Borel–Cantelli Lemma*. To complete the exercise, we show that the converse statement does not hold true in general.*

3. *Consider $\Omega = [0, 1]$ provided with the Borel σ-field and $\mathbb{P}$ the Lebesgue measure. Let $A_n = [0, \epsilon_n]$ for some sequence ϵ_n which converges to 0. Show that $\mathbb{P}(\limsup_{n \to +\infty} A_n) = 0$, whether $\sum_{n \geq 1} \mathbb{P}(A_n)$ is finite or not.*

☕ **Exercise 1.4.2** (Second Borel–Cantelli Lemma)**.** *Let $(A_n)_{n\geq 1}$ be a sequence of independent events, which satisfies the condition that*

$$\sum_{n\geq 1} \mathbb{P}(A_n) = +\infty. \tag{1.2}$$

Our aim is to show that in this case, the event $\limsup_{n\to+\infty} A_n$ introduced in Exercise 1.4.1 is almost sure. Thus, this provides a partial converse to the Borel–Cantelli Lemma, in the case where the events are independent. In fact, this shows the stronger statement that in this case, the event $\limsup_{n\to+\infty} A_n$ has probability either 0 or 1, depending on whether (1.1) or (1.2) holds. This result is called the Borel Zero-One Law.

We denote by B the complement of $\limsup_{n\to+\infty} A_n$.

1. *For any $N \geq 1$, let $B_N = \cap_{n\geq N} A_n^{\mathrm{c}}$. Show that*

$$\mathbb{P}(B) \leq \sum_{N\geq 1} \mathbb{P}(B_N).$$

2. *For any $N \geq 1$, show that*

$$\mathbb{P}(B_N) \leq \liminf_{k\to+\infty} \prod_{n=N}^{N+k-1} (1 - \mathbb{P}(A_n)).$$

3. *Deduce that, for any $N \geq 1$, $\mathbb{P}(B_N) = 0$.*

Chapter 2

Random variables and their numerical simulation

Contents

In this Chapter, we introduce some classical laws (discrete or with density) and put an emphasis on the numerical simulation of associated random variables.

2.1 Random number simulation

2.1.1 Pseudo-random number generation

It is an obvious fact that a *deterministic* algorithm cannot generate a *truly random* sequence, as was written by von Neumann: 'Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.'[1]. Hence, *pseudo-random number generators* are deterministic algorithms which, starting from a *seed* x_0, return a sequence $x_1, x_2, \ldots$ of numbers which exhibits the same statistical properties as a sequence of *independent and identically distributed* random numbers.

Because of the finiteness of the memory of a computer, a pseudo-random number generator is necessarily *ultimately periodic*, that is to say that there exists $t \geq 0$, which may depend on x_0, such that for n large enough, $x_{n+t} = x_n$. In the sequel we call *maximal period* the largest value

[1] Quoted in D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd edition, Addison-Wesley, 1998.

of t over all possible values of x_0. Since 'truly random' sequences should not be periodic, it is an intuitive statement that a 'good' pseudo-random number generator should have a large maximal period.

We first present a class of pseudo-random generators which are relatively easy to describe. *Linear congruential generators* were introduced in 1948 and depend on the following integer parameters:

- a *modulus* $m > 0$;
- a *multiplier* $0 < a < m$;
- an *increment* $0 \leq c < m$.

The seed is an integer $x_0 \in \{0, \ldots, m-1\}$. The sequence $(x_n)_{n \geq 1}$ is then computed according to the recurrence relation

$$x_{n+1} = ax_n + c \mod m,$$

which produces integer numbers in $\{0, \ldots, m-1\}$. Typically, taking $m = 2^{32}$ allows to get integers encoded on 32 bits.

In general, the maximal period of linear congruential generators (which is at most m) can be computed. Yet, their quality remains very sensitive to the choice of a and m. More complex pseudo-random generators have thus been elaborated. The most widely used generator in current scientific computing languages is called *Mersenne Twister*. It was developed in 1997[2], it is based on the arithmetic properties of Mersenne numbers and its period is $2^{19937} - 1 \simeq 4.3 \cdot 10^{6001}$.

Whatever the chosen pseudo-random number generator, let us take as granted that given a seed $x_0 \in \{0, \ldots, m-1\}$, it returns a sequence $(x_n)_{n \geq 1}$ of numbers in $\{0, \ldots, m-1\}$, which has the following statistical properties:

(i) they look independent in the sense of Definition 1.3.9;

(ii) they look uniformly distributed in $\{0, \ldots, m-1\}$ in the sense that each integer $x \in \{0, \ldots, m-1\}$ appears in the sequence $(x_n)_{n \geq 1}$ with equal frequency $1/m$.

Defining $U_n = x_n/m \in [0, 1)$, we thus obtain a sequence of pseudo-random independent variables such that, for any $n \geq 1$, for any interval $C \subset [0, 1]$,

$$\mathbb{P}(U_n \in C) = \frac{1}{m} \sum_{x=0}^{m-1} \mathbb{1}_{\{x/m \in C\}} \simeq \int_{u=0}^{1} \mathbb{1}_{\{u \in C\}} \mathrm{d}u.$$

This motivates the following definition.

Definition 2.1.1 (Uniform distribution)**.** *A random variable U in $[0, 1]$ is called* uniformly distributed on $[0, 1]$ *if it has the density*

$$p(u) = \mathbb{1}_{\{u \in [0,1]\}}.$$

We denote $U \sim \mathcal{U}[0, 1]$.

🗎 **Exercise 2.1.2.** *Let $U \sim \mathcal{U}[0, 1]$. Compute $\mathbb{E}[U]$ and $\mathrm{Var}(U)$.*

🏠 **Exercise 2.1.3.** *Let $U \sim \mathcal{U}[0, 1]$. Show that the random variable $1-U$ has the same distribution as U.*

Remark 2.1.4 (Difference between variable and law)**.** *Exercise 2.1.3 allows to highlight the difference between the notions of* random variable *and their* law*: the random variables U and $1 - U$ are different, and in particular $U \neq 1 - U$, almost surely; however they have the same law.*

[2] Matsumoto, M. and Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator, *ACM Transactions on Modeling and Computer Simulations* (1998).

From now on, we shall thus work under the assumption that our computer is able to generate independent variables $(U_n)_{n\geq 1}$ which are uniformly distributed on $[0,1]$. In the sequel of this section, we study how to use this sequence in order to sample a random variable X with a given distribution.

Example 2.1.5 (Uniform distribution)**.** *The* uniform distribution on the interval $[a,b]$, *denoted by* $\mathcal{U}[a,b]$, *is the probability measure with density*

$$p(x) = \frac{1}{b-a}\mathbb{1}_{\{x\in[a,b]\}}.$$

If $U \sim \mathcal{U}[0,1]$, *then* $X := a + (b-a)U \sim \mathcal{U}[a,b]$.

In Example 2.1.5, the proof of the fact that $X \sim \mathcal{U}[a,b]$ relies on the change-of-variable technique explained in Subsection 1.3.4.

Remark 2.1.6. *Most scientific computing languages allow you to fix the seed of your pseudo-random number generator. This makes your code no longer random but this may prove very helpful for reproducibility, comparison of your code and experimental results with others, or simply debugging.*

2.1.2 Classical discrete distributions

We first introduce several discrete distributions.

Definition 2.1.7 (Bernoulli, binomial and geometric distributions)**.** *Let* $p \in [0,1]$.

(i) A random variable X *in* $\{0,1\}$ *such that* $\mathbb{P}(X=1)=p$ *and* $\mathbb{P}(X=0)=1-p$ *is called a* Bernoulli *random variable with parameter* p*. We denote* $X \sim \mathcal{B}(p)$.

(ii) Let $n \geq 1$ *and* $X_1,\ldots,X_n$ *be independent Bernoulli random variables with parameter* p*. The random variable* $S := X_1 + \cdots + X_n$ *is called a* binomial *random variable with parameters* n *and* p*. We denote* $S \sim \mathcal{B}(n,p)$.

(iii) Assume that $p \in (0,1]$ *and let* $(X_i)_{i\geq 1}$ *be a sequence of independent Bernoulli random variables with parameter* p*. The random variable* $T := \min\{i \geq 1 : X_i = 1\}$ *is called a* geometric *random variable with parameter* p*. We denote* $T \sim \mathcal{G}\mathrm{eo}(p)$.

🏠 **Exercise 2.1.8** (Properties of Bernoulli, binomial and geometric distributions)**.** *Let* X*,* S *and* T *be as in Definition 2.1.7.*

1. *Compute* $\mathbb{E}[X]$ *and* $\mathrm{Var}(X)$.
2. *Compute* $\mathbb{E}[S]$ *and* $\mathrm{Var}(S)$.
3. *Show that, for any* $k \in \{0,\ldots,n\}$, $\mathbb{P}(S=k) = \binom{n}{k}p^k(1-p)^{n-k}$.
4. *Show that, for any* $k \geq 1$, $\mathbb{P}(T=k) = p(1-p)^{k-1}$.
5. *Compute* $\mathbb{E}[T]$ *and* $\mathrm{Var}(T)$.

The numerical sampling of the Bernoulli, binomial and geometric distributions is addressed in the next exercise.

🗎 **Exercise 2.1.9.** *Let* $(U_n)_{n\geq 1}$ *be a sequence of independent uniform variables on* $[0,1]$.

1. *Using an* `if` *test, how to draw a random variable* $X \sim \mathcal{B}(p)$*?*
2. *Using a* `for` *loop, how to draw a random variable* $S \sim \mathcal{B}(n,p)$*?*
3. *Using a* `while` *loop, how to draw a random variable* $T \sim \mathcal{G}\mathrm{eo}(p)$*?*

The following exercise is attributed to Von Neumann.

☕ **Exercise 2.1.10** (Unbiasing a coin toss)**.** *Assume that you have a random number generator which returns independent Bernoulli variables with an* unknown *parameter* $p \in (0,1)$*. How to use it to draw a Bernoulli random variable with parameter* $1/2$*?*

2.1.3 The inverse CDF method

Let X be a random variable taking its values in some finite set E, and let $(p_x)_{x \in E}$ be its probability mass function (that is to say, $p_x = \mathbb{P}(X = x)$). A somehow intuitive algorithm allowing to sample X from a $\mathcal{U}[0,1]$ random variable U is the following:

1. label the elements of E in some arbitrary order $x_1, \ldots, x_{\mathsf{m}}$;
2. select the unique index $i \in \{1, \ldots, \mathsf{m}\}$ such that $p_{x_1} + \cdots + p_{x_{i-1}} < U \leq p_{x_1} + \cdots + p_{x_i}$;
3. return $X = x_i$.

It is clear that we have

$$\mathbb{P}(X = x_i) = \mathbb{P}(p_{x_1} + \cdots + p_{x_{i-1}} < U \leq p_{x_1} + \cdots + p_{x_i}) = p_{x_i},$$

so that X has the correct law.

The generalisation of this approach to arbitrary, real-valued random variables, is based on the introduction of the *Cumulative Distribution Function* of such variables.

Definition 2.1.11 (Cumulative Distribution Function)**.** *Let X be a real-valued random variable. The* Cumulative Distribution Function *(CDF) of X is the function $F_X : \mathbb{R} \to [0,1]$ defined by*

$$\forall x \in \mathbb{R}, \qquad F_X(x) := \mathbb{P}(X \leq x).$$

Remark 2.1.12. *Since the Borel σ-field on $\mathbb{R}$ is generated by the intervals of the form $(-\infty, x]$, by Lemma 1.1.4, two random variables have the same CDF if and only if they have the same law.*

🏠 **Exercise 2.1.13** (Properties of CDFs)**.** *Let F_X be the CDF of a random variable X. Show that:*

1. *F_X is nondecreasing;*
2. $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F_X(x) = 1$*;*
3. *F_X is right continuous and has left limits.*

When X has a density p, Definition 2.1.11 yields the identity

$$\forall x \in \mathbb{R}, \qquad F_X(x) = \int_{y=-\infty}^{x} p(y)\mathrm{d}y,$$

which shows that F_X is continuous and $\mathrm{d}x$-almost everywhere differentiable, with $F_X' = p$.

Definition 2.1.14. *Let F_X be the CDF of a random variable X. The* pseudo-inverse *of F_X is the function $F_X^{-1} : [0,1] \to [-\infty, +\infty]$ defined by*

$$\forall u \in [0,1], \qquad F_X^{-1}(u) := \inf\{x \in \mathbb{R} : F_X(x) \geq u\},$$

with the conventions that $\inf \mathbb{R} = -\infty$ *and* $\inf \varnothing = +\infty$.

The pseudo-inverse of a CDF is nondecreasing, left continuous with right limits. When F_X is continuous and increasing, then F_X^{-1} is the usual inverse bijection of F_X. In general, it need not hold that $F_X(F_X^{-1}(u)) = u$ or $F_X^{-1}(F_X(x)) = x$, but the following weaker statement remains true.

Lemma 2.1.15 (CDF and pseudo-inverse)**.** *Let F_X be the CDF of a random variable X. For all $x \in \mathbb{R}$, $u \in (0,1)$, we have $F_X^{-1}(u) \leq x$ if and only if $u \leq F_X(x)$.*

Proof. Since F_X is right continuous, for any $u \in (0,1)$ the set $\{x \in \mathbb{R} : F_X(x) \geq u\}$ is closed, therefore $F_X(F_X^{-1}(u)) \geq u$. Since F_X is nondecreasing, we deduce that if $F_X^{-1}(u) \leq x$ then $u \leq F_X(x)$. Conversely, if $u \leq F_X(x)$, then by the definition of F_X^{-1}, $F_X^{-1}(u) \leq x$. □

Corollary 2.1.16 (The inverse CDF method)**.** *Let F_X be the CDF of a random variable X, and let $U \sim \mathcal{U}[0,1]$. The random variables X and $F_X^{-1}(U)$ have the same distribution.*

Proof. By Lemma 2.1.15 and Definition 2.1.1, for all $x \in \mathbb{R}$,

$$\mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = \int_{u=0}^{F_X(x)} \mathrm{d}u = F_X(x),$$

so that the random variables X and $F_X^{-1}(U)$ have the same CDF. From Remark 2.1.12 we conclude that they have the same distribution. □

We illustrate this method on the exponential distribution.

Definition 2.1.17 (Exponential distribution)**.** *Let $\lambda > 0$. A random variable X in $[0, +\infty)$ is called* exponential *with parameter λ if it has the density*

$$p(x) = \mathbb{1}_{\{x>0\}} \lambda \mathrm{e}^{-\lambda x}.$$

We denote $X \sim \mathcal{E}(\lambda)$.

🏠 **Exercise 2.1.18** (Properties of exponential distributions)**.** *Let $X \sim \mathcal{E}(\lambda)$.*

1. *Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.*
2. *If $a > 0$, what is the law of aX?*

An immediate computation shows that the CDF of X writes

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - \mathrm{e}^{-\lambda x} & \text{otherwise.} \end{cases}$$

As a consequence, for all $u \in [0, 1]$,

$$F_X^{-1}(u) = -\frac{1}{\lambda} \ln(1-u),$$

with the obvious convention that $\ln 0 = -\infty$. Therefore, to draw a random variable $X \sim \mathcal{E}(\lambda)$, one may take a uniform variable U on $[0, 1]$ and return $-\frac{1}{\lambda}\ln(1-U)$. Notice that, by Exercise 2.1.3, it is also equivalent to return $-\frac{1}{\lambda}\ln(U)$.

🏠 **Exercise 2.1.19** (Other standard densities)**.** *Apply the inverse CDF method to the following standard probability densities.*

1. *The Pareto distribution with parameter $\alpha > 0$, with density $\mathbb{1}_{\{x>1\}} \alpha x^{-(\alpha+1)}$.*
2. *The Cauchy distribution with parameter $a > 0$, with density $\frac{a}{\pi} \frac{1}{a^2+x^2}$.*
3. *The Weibull distribution with parameter $m > 0$, with density $\mathbb{1}_{\{x>0\}} m x^{m-1} \exp(-x^m)$.*
4. *The Rayleigh distribution with parameter $\sigma^2 > 0$, with density $\mathbb{1}_{\{x>0\}} \frac{x}{\sigma^2} \exp(-\frac{x^2}{2\sigma^2})$.*

🏠 **Exercise 2.1.20** (Back to geometric distribution)**.** *Let $X \sim \mathcal{E}(\lambda)$.*

1. *What is the law of $\lceil X \rceil$?*[3]
2. *Deduce an algorithm which returns a $\mathcal{G}\mathrm{eo}(p)$ random variable with a single uniform random variable U.*

[3] For any $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the unique integer such that $\lceil x \rceil - 1 < x \leq \lceil x \rceil$.

🏠 **Exercise 2.1.21** (Poisson distribution)**.** *A random variable $N \in \mathbb{N}$ is distributed according to the* Poisson distribution *with parameter $\lambda > 0$ if, for any $k \in \mathbb{N}$,*

$$\mathbb{P}(N = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}.$$

We denote $N \sim \mathcal{P}(\lambda)$.

1. *Compute $\mathbb{E}[N]$ and $\mathrm{Var}(N)$.*
2. *Show that if $(S_n)_{n\geq 1}$ is a sequence of binomial random variables such that $S_n \sim \mathcal{B}(n, p_n)$ with $np_n \to \lambda$, then for all $k \geq 0$, $\mathbb{P}(S_n = k) \to \mathbb{P}(N = k)$. Which interpretation of the Poisson distribution can you deduce?*
3. *Let $(X_i)_{i\geq 1}$ be a sequence of independent exponential random variables with parameter λ. Show that $\inf\{n \geq 0 : X_1 + \cdots + X_{n+1} \geq 1\} \sim \mathcal{P}(\lambda)$.*
4. *Deduce an algorithm to draw a random variable $N \sim \mathcal{P}(\lambda)$ using a sequence $(U_i)_{i\geq 1}$ of independent uniform variables on $[0, 1]$.*

🗎 **Exercise 2.1.22.** *Show that if the CDF F_X of X is continuous, then $F_X(X) \sim \mathcal{U}[0, 1]$.*

2.1.4 Gaussian random variables

We recall that the *Gauss integral* is equal to[4]

$$\int_{x\in\mathbb{R}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \sqrt{2\pi}.$$

Definition 2.1.23 (Standard Gaussian variables)**.** *A random variable G in $\mathbb{R}$ is a* standard Gaussian variable *if it has the density*

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

🗎 **Exercise 2.1.24.** *If G is a standard Gaussian variable, show that $G \in \mathbf{L}^p(\mathbb{P})$ for any $p \in [1, +\infty)$ and compute $\mathbb{E}[G]$ and $\mathrm{Var}(G)$.*

It follows from this exercise that for any $\mu, \sigma \in \mathbb{R}$, the random variable $X = \mu + \sigma G$ satisfies $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. This remark is used in the next definition.

Definition 2.1.25 (Gaussian variable)**.** *If G is a standard Gaussian variable, then for any $\mu, \sigma \in \mathbb{R}$, the random variable*

$$X = \mu + \sigma G$$

is called a Gaussian random variable *with mean μ and variance σ^2. Its law is denoted by $\mathcal{N}(\mu, \sigma^2)$.*

Gaussian variables are also called *normal*. The fact that the law of X only depends on σ through σ^2 is justified by the following result.

🗎 **Exercise 2.1.26.** *Show that if $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 > 0$, then X has density*

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

[4] Do not hesitate to redo the computation just to be sure that you still know how to!

We insist on the fact that the definition of Gaussian random variables also includes the case where $\sigma = 0$, in which case X is the almost surely constant random variable equal to μ. In this case, the law of X is the Dirac measure δ_μ and therefore it does not have a density.

By definition, the problem of sampling from Gaussian distributions reduces to the case of the standard Gaussian distribution. Let $\Phi : \mathbb{R} \to [0, 1]$ denote its CDF, given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{y=-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y.$$

It is known that Φ cannot be expressed in terms of usual functions, such as polynomials, exponentials or logarithms. Hence the inverse CDF method cannot be applied in the present case. We shall present an *ad hoc* approach, called the *Box–Muller method*.

Proposition 2.1.27 (Box–Muller method)**.** *Let $R \sim \mathcal{E}(1/2)$ and $\Theta \sim \mathcal{U}[0, 2\pi]$ be independent random variables. The random variables*

$$X := \sqrt{R}\cos\Theta, \qquad Y := \sqrt{R}\sin\Theta,$$

are independent and follow the standard Gaussian distribution.

Proof. We use the dummy function method introduced in Subsection 1.3.4 and let $f : \mathbb{R}^2 \to \mathbb{R}$ be measurable and bounded. Since R and Θ are independent, the law of the pair (R, Θ) is the product of the marginal densities, and therefore

$$\begin{aligned}
\mathbb{E}[f(X,Y)] &= \mathbb{E}\left[f\left(\sqrt{R}\cos\Theta, \sqrt{R}\sin\Theta\right)\right] \\
&= \int_{\omega\in\Omega} f(\sqrt{R(\omega)}\cos(\Theta(\omega)), \sqrt{R(\omega)}\sin(\Theta(\omega)))\mathrm{d}\mathbb{P}(\omega) \\
&= \int_{r=0}^{+\infty}\int_{\theta=0}^{2\pi} f(\sqrt{r}\cos\theta, \sqrt{r}\sin\theta)\frac{\mathrm{d}\theta}{2\pi}\frac{1}{2}\mathrm{e}^{-r/2}\mathrm{d}r.
\end{aligned}$$

Using the polar change of coordinates $x = \sqrt{r}\cos\theta$, $y = \sqrt{r}\sin\theta$ in the right-hand side, we get

$$\mathbb{E}[f(X,Y)] = \int_{x,y\in\mathbb{R}} f(x,y)\frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)\mathrm{d}x\mathrm{d}y,$$

which shows that the pair (X, Y) has density

$$\frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right),$$

which implies that X and Y are independent standard Gaussian variables. □

Since both R and Θ can be sampled using the inverse CDF method, Proposition 2.1.27 provides a method to sample X and Y from two independent uniform random variables on $[0, 1]$.

Remark 2.1.28 (What does my computer really do?[5])**.** *The Box–Muller method is used by NumPy's random.standard_normal function to generate Gaussian variables. Its newer random number generator class, called Generator, uses another method called the Ziggurat algorithm, which is based on the rejection method described in the next Subsection. In contrast, the statistical software R uses the inverse CDF method to generate Gaussian samples, with a numerical approximation of the function Φ^{-1}.*

</> You may implement both the Box–Muller method and the Ziggurat algorithm in the Notebook `Ziggurat.ipynb` available on the course's webpage.

[5] According to the blog post https://medium.com/mti-technology/how-to-generate-gaussian-samples-3951f2203ab0.

2.1.5 Rejection sampling

We start with the following simple question: given a bounded subset D of $\mathbb{R}^d$ with positive Lebesgue measure, how to draw a point X *uniformly* in D, that is to say according to the density

$$p(x) = \frac{1}{|D|}\mathbb{1}_{\{x \in D\}},$$

where $|D|$ denotes the Lebesgue measure of D?

If D is a *rectangle*, that is to say a Cartesian product $\prod_{i=1}^d [a_i, b_i]$ of intervals, in which case it is actually more convenient to denote it by R, then it is easily checked that the vector $(X_1, \ldots, X_d)$ of independent coordinates, such that each X_i is uniformly distributed on $[a_i, b_i]$, is uniformly distributed on R.

In the general case, an intuitive procedure can be formulated as follows (see also Figure 2.1):

(i) start to 'frame' D into a rectangle $R \supset D$;
(ii) draw X uniformly in R;
(iii) if $X \in D$ then return it, otherwise restart at Step (ii).

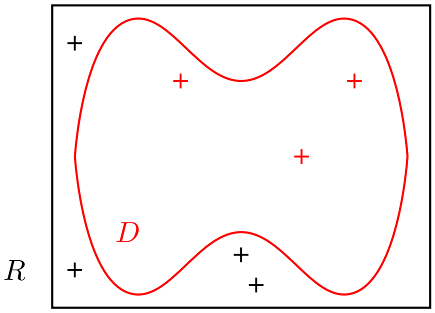


Figure 2.1: The domain D framed into a rectangle R. Random points are drawn in R, only those falling into D are kept.

Let us prove that this procedure produces a correct result. Let $X_1, X_2, \ldots$ be independent random variables uniformly distributed in R, and $N := \inf\{n \geq 1 : X_n \in D\}$, so that the algorithm returns the random variable X_N. We may already remark that the law of N is easy to compute.

Exercise 2.1.29. *Show that* $N \sim \mathcal{G}\mathrm{eo}(|D|/|R|)$.

In particular, $\mathbb{E}[N] = |R|/|D|$ so the smaller R, the faster the algorithm, which is a reasonable statement. As far as the law of X_N is concerned, let us take $C \in \mathcal{B}(\mathbb{R}^d)$ and compute

$$\begin{aligned}\mathbb{P}(X_N \in C) &= \sum_{n=1}^{+\infty} \mathbb{P}(X_n \in C, N = n)\\ &= \sum_{n=1}^{+\infty} \mathbb{P}(X_1 \notin D, \ldots, X_{n-1} \notin D, X_n \in C \cap D).\end{aligned}$$

Since the random variables $X_1, \dots, X_n$ are independent, each term of the sum rewrites

$$\begin{aligned}\mathbb{P}(X_1 \notin D, \dots, X_{n-1} \notin D, X_n \in C \cap D) &= \mathbb{P}(X_1 \notin D) \cdots \mathbb{P}(X_{n-1} \notin D)\mathbb{P}(X_n \in C \cap D) \\ &= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \int_{x \in R} \mathbb{1}_{\{x \in C \cap D\}} \frac{\mathrm{d}x}{|R|} \\ &= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \frac{|D|}{|R|} \int_{x \in \mathbb{R}^d} \mathbb{1}_{\{x \in C\}} p(x) \mathrm{d}x,\end{aligned}$$

where p denotes the uniform density on D. Summing over n, we deduce that

$$\mathbb{P}(X_N \in C) = \int_{x \in \mathbb{R}^d} \mathbb{1}_{\{x \in C\}} p(x) \mathrm{d}x,$$

which shows that X_N has density p.

🏠 **Exercise 2.1.30.** *Show that the random variables X_N and N are independent.*

This *rejection* method can be generalised to non-uniform densities as follows.

Theorem 2.1.31 (Rejection sampling)**.** *Let $p : \mathbb{R}^d \to [0, +\infty)$ be a probability density. Assume that there exist a probability density $q : \mathbb{R}^d \to [0, +\infty)$ and $k \geq 1$ such that, $\mathrm{d}x$-almost everywhere, $p(x) \leq kq(x)$. Let $(X_n)_{n \geq 1}$ be a sequence of independent random variables in $\mathbb{R}^d$ with density q, and $(U_n)_{n \geq 1}$ be a sequence of independent random variables uniformly distributed in $[0, 1]$, independent from $(X_n)_{n \geq 1}$. Let*

$$N := \inf\{n \geq 1 : kq(X_n)U_n \leq p(X_n)\}.$$

We have the following results:
(i) $N \sim \mathcal{G}\mathrm{eo}(1/k)$;
(ii) X_N has density p;
(iii) N and X_N are independent.

The proof of Theorem 2.1.31 follows from the same computation as in the example of uniform distributions, for which q is the uniform distribution on the rectangle, and $k = |R|/|D|$. In fact, at a slightly more conceptual level, it can be seen as a consequence of this example. Indeed, set

$$D = \{(x, y) \in \mathbb{R}^d \times \mathbb{R} : 0 \leq y \leq p(x)\},$$

and

$$R = \{(x, y) \in \mathbb{R}^d \times \mathbb{R} : 0 \leq y \leq kq(x)\}.$$

Since these sets have respective $d+1$-dimensional Lebesgue measure 1 and k, the uniform density thereon is well-defined. Moreover, the following two points are easy to check:

- if X has density q and $U \sim \mathcal{U}[0, 1]$ is independent from X, then $(X, kq(X)U)$ is uniformly distributed in R;
- if (X, Y) is uniformly distributed in D then X has marginal density p.

As a consequence, the rejection algorithm described in Theorem 2.1.31 is *exactly* the rejection algorithm for uniform densities, where one generates uniform samples $(X_n, kq(X_n)U_n)$ in R and keeps the first which is in D, that is to say such that $kq(X_n)U_n \leq p(X_n)$.

Remark 2.1.32. *Theorem 2.1.31 can easily be generalised to the case where one wants to draw X from a probability measure P on some abstract space E, and has access to samples under $Q \gg P$, with $\frac{\mathrm{d}P}{\mathrm{d}Q} \leq k$, Q-almost everywhere. Then the statement of Theorem 2.1.31 remains in force, with N defined as the first index for which $kU_n \leq \frac{\mathrm{d}P}{\mathrm{d}Q}(X_n)$.*

Rejection sampling is useful when one is not able to sample directly from p, but can find q such that $p \leq kq$ and sampling from q is easier. Just like in the example of uniform distributions, the smaller k, the faster the algorithm, therefore from a computational point of view it is of interest to take q as a ‘good approximation’ of p.

🏠 **Exercise 2.1.33** (Gamma distribution)**.** *The* Gamma distribution *with (shape) parameter* $a > 0$ *is the probability measure on* $\mathbb{R}$ *with density*

$$p(x) = \mathbb{1}_{\{x>0\}} \frac{1}{\Gamma(a)} x^{a-1} \mathrm{e}^{-x},$$

where Γ *is Euler’s function*

$$\Gamma(a) := \int_{x=0}^{+\infty} x^{a-1} \mathrm{e}^{-x} \mathrm{d}x.$$

We assume that $a > 1$ *and want to implement the rejection sampling method with q the density of the exponential distribution with parameter* λ*. Which value of* λ *should we take? What will be the resulting value of k?*

2.2 Random vector simulation

In this section, we consider the issue of simulating random vectors, and in particular Gaussian vectors. For any $p \geq 1$, we denote by $\mathbf{L}^p(\mathbb{P}; \mathbb{R}^d)$ the set of random vectors whose coordinates are random variables in $\mathbf{L}^p(\mathbb{P})$. If $X = (X_1, \ldots, X_n) \in \mathbf{L}^1(\mathbb{P}; \mathbb{R}^d)$, we denote by $\mathbb{E}[X]$ the vector $(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$.

2.2.1 Covariance

Definition 2.2.1 (Covariance between two random variables)**.** *Let* $X, Y \in \mathbf{L}^2(\mathbb{P})$*. The* covariance *between* X *and* Y *is defined by*

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

It is clear that the covariance is symmetric and bilinear on $\mathbf{L}^2(\mathbb{P})$, and that

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X).$$

Besides, we have the formula

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

and the Cauchy–Schwarz inequality yields

$$|\,\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}.$$

The latter inequality shows that the *correlation coefficient* between X and Y, defined by

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

is always between -1 and 1.

📄 **Exercise 2.2.2.** 1. *Show that, if* X *and* Y *are independent, then* $\mathrm{Cov}(X, Y) = 0$.

2. *Let $X \sim \mathcal{N}(0,1)$ and $Y = X^2$. Compute* $\mathrm{Cov}(X,Y)$ *on the one hand, and determine whether X and Y are independent on the other hand.*

Definition 2.2.3 (Covariance matrix)**.** *Let $X = (X_1, \ldots, X_d) \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$. The* covariance matrix *of X is the $d \times d$ matrix* $\mathrm{Cov}[X]$ *with coefficients* $\mathrm{Cov}(X_i, X_j)$*.*

Clearly, a covariance matrix is symmetric. Exercise 2.2.4 below shows that it is also nonnegative. We shall see in Proposition 2.2.18 that, conversely, any symmetric and nonnegative matrix is the covariance matrix of a (Gaussian) random vector.

🗎 **Exercise 2.2.4.** *Show that if $X \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$ has covariance matrix K, then for any $u \in \mathbb{R}^d$,*

$$\mathrm{Var}(\langle u, X\rangle) = \langle u, Ku\rangle.$$

🗎 **Exercise 2.2.5.** *Let $X \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$ with covariance matrix K, and let $b \in \mathbb{R}^k$, $A \in \mathbb{R}^{k\times d}$. Show that*

$$\mathrm{Cov}[b + AX] = AKA^\top.$$

2.2.2 Characteristic function

The characteristic function is a useful tool to study random vectors.

Definition 2.2.6 (Characteristic function)**.** *Let $X \in \mathbb{R}^d$ be a random vector. Its* characteristic function *is the function $\Psi_X : \mathbb{R}^d \to \mathbb{C}$ defined by*

$$\forall u \in \mathbb{R}^d, \qquad \Psi_X(u) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle u,X\rangle}\right] = \mathbb{E}\left[\cos(\langle u, X\rangle)\right] + \mathrm{i}\mathbb{E}\left[\sin(\langle u, X\rangle)\right].$$

By Theorem 1.3.3, we get that

$$\forall u \in \mathbb{R}^d, \qquad \Psi_X(u) = \int_{x\in\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle u,x\rangle} \mathrm{d}P_X(x),$$

so that up to sign change and dilation, the characteristic function of X coincides with the Fourier transform of the measure P_X. Since the latter is injective, we deduce the following important property.

Proposition 2.2.7 (Characterisation of the law)**.** *Two random vectors X and Y in $\mathbb{R}^d$ have the same law if and only if*

$$\forall u \in \mathbb{R}^d, \qquad \Psi_X(u) = \Psi_Y(u).$$

🗎 **Exercise 2.2.8.** *Let X, Y be two independent random vectors. Show that, for any $u \in \mathbb{R}^d$, $\Phi_{X+Y}(u) = \Phi_X(u)\Phi_Y(u)$.*

🏠 **Exercise 2.2.9.** *Compute the characteristic function of X when $X \sim \mathcal{B}(p)$, $\mathcal{B}(n,p)$, $\mathcal{G}\mathrm{eo}(p)$, $\mathcal{P}(\lambda)$, $\mathcal{U}[a,b]$, $\mathcal{E}(\lambda)$.*

The characteristic function of Gaussian variables plays a central role in the sequel of this section, but its direct computation requires to compute an integral along a complex-valued line, which is beyond the scope of these notes. A more elementary approach is proposed in Exercise 2.2.11 below. It first requires the following technical statement, which will also be used to prove the Central Limit Theorem in Chapter 3.

Lemma 2.2.10 (Derivatives of the characteristic function)**.** *If $X \in \mathbf{L}^p(\mathbb{P};\mathbb{R}^d)$ for some integer $p \geq 1$, then Ψ_X is of class C^p on $\mathbb{R}^d$ and, for any multi-index $q = (q_1, \ldots, q_d)$ with $|q| = q_1 + \cdots + q_d \leq p$,*

$$\forall u \in \mathbb{R}^d, \qquad \frac{\partial^{|q|}\Psi_X}{\partial u_1^{q_1} \cdots \partial u_d^{q_d}}(u) = \mathrm{i}^{|q|}\mathbb{E}\left[X_1^{q_1} \cdots X_d^{q_d} \mathrm{e}^{\mathrm{i}\langle u, X\rangle}\right].$$

Proof. The proof consists in the application of a standard derivative-under-the-integral argument. To proceed, we note that almost surely, the function $u \mapsto \mathrm{e}^{\mathrm{i}\langle u,X\rangle}$ is C^∞ on $\mathbb{R}$, and for any multi-index $q = (q_1, \ldots, q_d)$,

$$\frac{\partial^{|q|}}{\partial u_1^{q_1} \cdots \partial u_d^{q_d}} \mathrm{e}^{\mathrm{i}\langle u,X\rangle} = \mathrm{i}^{|q|} X_1^{q_1} \cdots X_d^{q_d} \mathrm{e}^{\mathrm{i}\langle u,X\rangle}.$$

The right-hand side satisfies

$$\left|\mathrm{i}^{|q|} X_1^{q_1} \cdots X_d^{q_d} \mathrm{e}^{\mathrm{i}\langle u,X\rangle}\right| = |X_1|^{q_1} \cdots |X_d|^{q_d}.$$

If $p \geq |q|$ is such that $X_1, \ldots, X_d \in \mathbf{L}^p(\mathbb{P})$, then one may set $\alpha_i = p/q_i \in [1, +\infty]$ for all i, and then deduce from Hölder's inequality that

$$\begin{aligned}\mathbb{E}\left[|X_1|^{q_1} \cdots |X_d|^{q_d}\right] &\leq \mathbb{E}\left[(|X_1|^{q_1})^{\alpha_1}\right]^{1/\alpha_1} \cdots \mathbb{E}\left[(|X_d|^{q_d})^{\alpha_d}\right]^{1/\alpha_d} \\ &= \mathbb{E}\left[|X_1|^p\right]^{1/\alpha_1} \cdots \mathbb{E}\left[|X_d|^p\right]^{1/\alpha_d} < +\infty.\end{aligned}$$

This shows that the considered partial derivative is dominated by an integrable random variable, uniformly in u, which allows to conclude by Lebesgue's Differentiation Theorem. □

Exercise 2.2.11 (Characteristic function of Gaussian random variables)**.** *Let $G \sim \mathcal{N}(0,1)$.*

1. *Show that Ψ_G is C^1 on $\mathbb{R}$, and that for all $u \in \mathbb{R}$, $\Psi_G'(u) + u\Psi_G(u) = 0$.*
2. *Deduce that $\Psi_G(u) = \exp(-u^2/2)$.*
3. *If $X \sim \mathcal{N}(\mu, \sigma^2)$, what is the expression of $\Psi_X(u)$?*
4. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \tau^2)$ be independent. Compute the law of $X + Y$.*

We conclude this subsection with a characterisation of independence by characteristic functions.

Proposition 2.2.12 (Characterisation of independence)**.** *Two random vectors $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^k$ are independent if and only if*

$$\forall u \in \mathbb{R}^d, \quad \forall v \in \mathbb{R}^k, \qquad \Psi_{(X,Y)}(u,v) = \Psi_X(u)\Psi_Y(v).$$

2.2.3 Gaussian vectors

Definition 2.2.13 (Gaussian vector)**.** *A random vector $X \in \mathbb{R}^d$ is* Gaussian *if, for any $u \in \mathbb{R}^d$, the random variable $\langle u, X\rangle$ is Gaussian in the sense of Definition 2.1.25.*

Let $X \in \mathbf{L}^2(\mathbb{P};\mathbb{R}^d)$. Set $m = \mathbb{E}[X] \in \mathbb{R}^d$ and $K = \mathrm{Cov}[X] \in \mathbb{R}^{d\times d}$. For any $u \in \mathbb{R}^d$, it is immediate that

$$\mathbb{E}[\langle u, X\rangle] = \langle u, m\rangle,$$

and by Exercise 2.2.4,

$$\mathrm{Var}(\langle u, X\rangle) = \langle u, Ku\rangle.$$

Therefore, if X is Gaussian, then necessarily, $\langle u, X\rangle \sim \mathcal{N}(\langle u, m\rangle, \langle u, Ku\rangle)$, and thus by Exercise 2.2.11,

$$\Psi_X(u) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle u,X\rangle}\right] = \exp\left(\mathrm{i}\langle u, m\rangle - \frac{1}{2}\langle u, Ku\rangle\right).$$

We deduce the following statement.

Proposition 2.2.14 (Characteristic function of Gaussian vectors)**.** *The random vector X is Gaussian if and only if there exist $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d\times d}$ such that, for any $u \in \mathbb{R}^d$,*

$$\Psi_X(u) = \exp\left(\mathrm{i}\langle u, m\rangle - \frac{1}{2}\langle u, Ku\rangle\right).$$

In this case, we have $m = \mathbb{E}[X]$ and $K = \mathrm{Cov}[X]$, and we denote by $\mathcal{N}_d(m, K)$ the law of X.

Exercise 2.2.15 (Stability of Gaussian vectors by affine transform)**.** *Show that if $X \sim \mathcal{N}_d(m, K)$ and $b \in \mathbb{R}^k$, $A \in \mathbb{R}^{k\times d}$, then $b + AX \sim \mathcal{N}_k(b + Am, AKA^\top)$.*

Exercise 2.2.16 (Gaussian vectors and Gaussian coordinates)**.** *The following results should clarify the links between Gaussian vectors and Gaussian coordinates.*

1. *Let $(X_1, \ldots, X_d)$ be a Gaussian vector. Show that the coordinates $X_1, \ldots, X_d$ are Gaussian random variables.*
2. *Construct an example of a vector $(X_1, \ldots, X_d)$ such that each coordinate X_i is a Gaussian random variable but the vector is not a Gaussian vector.*
3. *Let $X_1, \ldots, X_d$ be* independent *Gaussian variables. Show that the vector $(X_1, \ldots, X_d)$ is Gaussian.*

In the sequel of the course, we will use the following characterisation of independence for Gaussian vectors.

Proposition 2.2.17 (Independence in Gaussian vectors)**.** *Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^k$ such that $(X, Y) \in \mathbb{R}^{d+k}$ is a Gaussian vector. The vectors X and Y are independent if and only if*

$$\forall i \in \{1, \ldots, d\}, \quad \forall j \in \{1, \ldots, k\}, \qquad \mathrm{Cov}(X_i, Y_j) = 0.$$

Proof. Write the covariance matrix K of (X, Y) under the block form

$$K = \begin{pmatrix} K_X & K_{X,Y} \\ K_{X,Y}^\top & K_Y \end{pmatrix},$$

so that the claim to prove is that X and Y are independent if and only if $K_{X,Y} = 0$. The direct implication is straightforward by Exercise 2.2.2. Conversely, assume that $K_{X,Y} = 0$, and set $m_X = \mathbb{E}[X]$, $m_Y = \mathbb{E}[Y]$. Then by Exercise 2.2.2 again, (X, Y) has the same expectation and covariance matrix as the vector (X', Y'), with $X' \sim \mathcal{N}_d(m_X, K_X)$ and $Y' \sim \mathcal{N}_k(m_Y, K_Y)$ independent from each other. Since both (X, Y) and (X', Y') are Gaussian, this assertion is enough to imply that they have the same law, and as a consequence X and Y are independent. □

To complete this subsection, we address the question of how to simulate a random vector drawn from the Gaussian measure $\mathcal{N}_d(m, K)$ for some given $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d\times d}$. To proceed, we first remark that the Box–Muller method described in Proposition 2.1.27 allows to simulate independent realisations $G_1, \ldots, G_d$ of the standard Gaussian distribution. We next recall that, by the Spectral Theorem, for any symmetric nonnegative matrix $K \in \mathbb{R}^{d\times d}$, there exists $\lambda_1, \ldots, \lambda_d \geq 0$ and an orthonormal basis $(e_1, \ldots, e_d)$ of $\mathbb{R}^d$ such that for any i, $Ke_i = \lambda_i e_i$.

Proposition 2.2.18 (Simulation of Gaussian vectors)**.** *Let $m \in \mathbb{R}^d$ and $K \in \mathbb{R}^{d\times d}$ be a symmetric and nonnegative matrix, with associated eigenvalues $\lambda_1, \ldots, \lambda_d \geq 0$ and eigenvectors $(e_1, \ldots, e_d)$. Let $G_1, \ldots, G_d$ be independent standard Gaussian variables. Then*

$$X = m + \sum_{i=1}^{d} G_i \sqrt{\lambda_i} e_i \sim \mathcal{N}_d(m, K).$$

Proof. For any $u \in \mathbb{R}^d$,

$$\langle u, X\rangle = \langle u, m\rangle + \sum_{i=1}^{d} G_i \sqrt{\lambda_i} \langle u, e_i\rangle$$

is a sum of independent Gaussian variables, therefore by Exercise 2.2.11, it is a Gaussian variable. Hence, X is a Gaussian vector. Besides, it is immediate that $\mathbb{E}[\langle u, X\rangle] = \langle u, m\rangle$, and by independence,

$$\mathrm{Var}(\langle u, X\rangle) = \sum_{i=1}^{d} \lambda_i \langle u, e_i\rangle^2 = \langle u, Ku\rangle,$$

which shows that $\mathbb{E}[X] = m$ and $\mathrm{Cov}[X] = K$. □

Proposition 2.2.18 has the practical interest to show that, up to diagonalising the covariance matrix, it is possible to sample from the Gaussian measure $\mathcal{N}_d(m, K)$ as soon as independent standard Gaussian random variables are available. It may also be useful for theoretical purposes, as in the next exercise.

☕ **Exercise 2.2.19.** *Show that, if K is invertible, $X \sim \mathcal{N}_d(m, K)$ has density*

$$\frac{1}{\sqrt{(2\pi)^d \det(K)}} \exp\left(-\frac{\langle x - m, K^{-1}(x - m)\rangle}{2}\right)$$

with respect to the Lebesgue measure on $\mathbb{R}^d$. If K is not invertible, can you find a similar density with respect to another measure?

Remark 2.2.20. *The decomposition of X as a sum of uncorrelated variables can be performed far beyond both the Gaussian and finite-dimensional case. In general, it is called the* Karhunen–Loeve *expansion of X, and may be performed as soon as X can be written as a (possibly infinite) collection of random variables $(X_t)_{t\in I}$, where the set of indices I is endowed with a σ-field and a measure making $t \mapsto K(s, t) := \mathrm{Cov}(X_s, X_t)$ a square-integrable, measurable function for any $s \in I$.*

2.2.4 Copulas

Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$. In general, the collection of the *marginal* laws of $X_1, \ldots, X_d$ does not characterise the *joint* law of the vector, and a supplementary information is needed to describe how these variables depend on each other. For Gaussian vectors, this information is contained in the *correlation matrix* $R = (\rho(X_i, X_j))_{1\leq i,j\leq d}$ of the vector, since it is easily checked that if one knows the parameters (μ_i, σ_i^2) of each X_i on the one hand, and the correlation matrix R on the other hand, then one can reconstruct the law of the full vector $\mathcal{N}_d(m, K)$ by letting $m_i = \mu_i$ and $K_{i,j} = \sigma_i \sigma_j R_{i,j}$. Beyond the case of Gaussian vectors, the notion of *copula* allows to characterise the dependency between the coordinates of a random vector.

Definition 2.2.21 (Copula)**.** *A function $C : [0,1]^d \to [0,1]$ is called a* copula *if there exists a random vector $(U_1, \ldots, U_d) \in [0,1]^d$ such that:*
(i) for any $i \in \{1, \ldots, d\}$, $U_i \sim \mathcal{U}[0,1]$;
(ii) for any $(u_1, \ldots, u_d) \in [0,1]^d$, $C(u_1, \ldots, u_d) = \mathbb{P}(U_1 \leq u_1, \ldots, U_d \leq u_d)$.

As a consequence of Definition 2.2.21, a copula has the following properties:

- it is nondecreasing in each coordinate;
- for any $u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_d$, $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$;
- for any u_i, $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$.

Some elementary examples of copulas are given by the *independent* copula

$$C(u_1, \ldots, u_d) = u_1 \cdots u_d,$$

and the *independent* copula

$$C(u_1, \ldots, u_d) = \min(u_1, \ldots, u_d).$$

🗎 **Exercise 2.2.22.** *Describe the law of the random vectors $(U_1, \ldots, U_d)$ respectively associated with the independent and comonotonic copulas.*

The main result about copulas is the following statement, in which we generalise Definition 2.1.11 to random vectors by letting $F_X(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d)$. It remains true that the CDF of X characterises its law.

Theorem 2.2.23 (Sklar's Theorem)**.** *Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector with CDF F_X.*
(i) There exists a copula C_X such that for any $(x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$F_X(x_1, \ldots, x_d) = C_X\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right).$$

(ii) If the marginal CDFs $F_{X_1}, \ldots, F_{X_d}$ are continuous, then the copula is unique and given by, for any $(u_1, \ldots, u_d) \in [0,1]^d$,

$$C_X(u_1, \ldots, u_d) = F_X\left(F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(x_d)\right).$$

The copula of a random vector therefore allows to isolate the dependency structure of its components, apart from their marginal distributions.

🏠 **Exercise 2.2.24** (The Gaussian copula)**.** *Let $X \sim \mathcal{N}(m, K)$ and R the associated correlation matrix. Show that the copula of X is given by*

$$C_X(u_1, \ldots, u_d) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

where Φ is the CDF of the standard Gaussian distribution on $\mathbb{R}$, and Φ_R is the CDF of the Gaussian measure $\mathcal{N}_d(0, R)$.

Given the system of marginal distributions and the copula of a random vector X, we now ask how to generate samples of X. This is done with the following two-step procedure.

Lemma 2.2.25 (Sampling vectors with given marginal distributions and copulas)**.** *Let C be a copula and $F_1, \ldots, F_d$ be CDFs on $\mathbb{R}$. Consider the following algorithm:*
1. Generate $(U_1, \ldots, U_d)$ with CDF C.
2. Return $X = (F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d))$.

The vector X has copula C and each component X_i has CDF F_i.

The proof of Lemma 2.2.25 is straightforward. We now focus on the first step, namely: given a copula C, how to sample $(U_1, \dots, U_d) \in [0,1]^d$ with CDF C?

Lemma 2.2.26 (Sampling from a given copula)**.** *Let C be a copula and $(U_1, \dots, U_d) \in [0,1]^d$ with CDF C. Assume that all derivatives*

$$\frac{\partial^k C}{\partial u_1 \cdots \partial u_k}, \qquad k = 1, \dots, d-1,$$

exist. Then for any $k \in \{1, \dots, d-1\}$, the conditional CDF of U_{k+1} given $U_1, \dots, U_k$ writes

$$\forall u_{k+1} \in [0,1], \qquad \mathbb{P}(U_{k+1} \le u_{k+1} | U_1, \dots, U_k) = \frac{\partial^k C}{\partial u_1 \cdots \partial u_k}(U_1, \dots, U_k, u_{k+1}, 1, \dots, 1).$$

Lemma 2.2.26 thus provides the following algorithm to sample $(U_1, \dots, U_d)$:

1. Draw $U_1 \sim \mathcal{U}[0,1]$.
2. For $k = 1, \dots, d-1$, draw U_{k+1} conditionally on $U_1, \dots, U_k$ by using the inverse CDF method with the conditional CDF given by Lemma 2.2.26.

Chapter 3

Convergence of random variables and limit theorems

Contents

3.1 Convergence of random variables

Throughout this section, we consider random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking their values in $\mathbb{R}^d$, on which we fix an arbitrary norm $|\cdot|$. Most of the material of this section could easily be generalised to the case of random variable taking their values in a metric space[1].

3.1.1 Convergence of random variables: definitions and basic results

Definition 3.1.1 (Convergences)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables in $\mathbb{R}^d$ and X be a random variable in $\mathbb{R}^d$.*

(i) X_n converges to X almost surely if there exists an event $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 1$ and

$$\forall \omega \in A, \qquad \lim_{n\to+\infty} X_n(\omega) = X(\omega).$$

(ii) X_n converges to X in probability if, for any $\epsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

[1] To the notable exception of results involving characteristic functions, which would require a linear structure with duality properties.

(iii) For any $p \in [1, +\infty)$, X_n converges to X in $\mathbf{L}^p$ if[2]

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

In measure theoretic terms, the almost sure convergence of random variables corresponds to the $\mathbb{P}$-almost everywhere convergence of measurable functions. Likewise, the notion of convergence in $\mathbf{L}^p$ is the standard strong convergence in the linear space $\mathbf{L}^p(\mathbb{P}; \mathbb{R}^d)$. Convergence in probability is a bit more unusual from this point of view[3], however it plays a pivotal role in the articulation of the various modes of convergence.

Proposition 3.1.2 (Hierarchy of convergences)**.** *Let $(X_n)_{n \geq 1}$ be a sequence of random variables in $\mathbb{R}^d$ and X be a random variable in $\mathbb{R}^d$.*

(i) If $X_n \to X$ almost surely, then $X_n \to X$ in probability.

(ii) For any $1 \leq p \leq q$, if $X_n \to X$ in $\mathbf{L}^q$, then $X_n \to X$ in $\mathbf{L}^p$.

(iii) If $X_n \to X$ in $\mathbf{L}^1$, then $X_n \to X$ in probability.

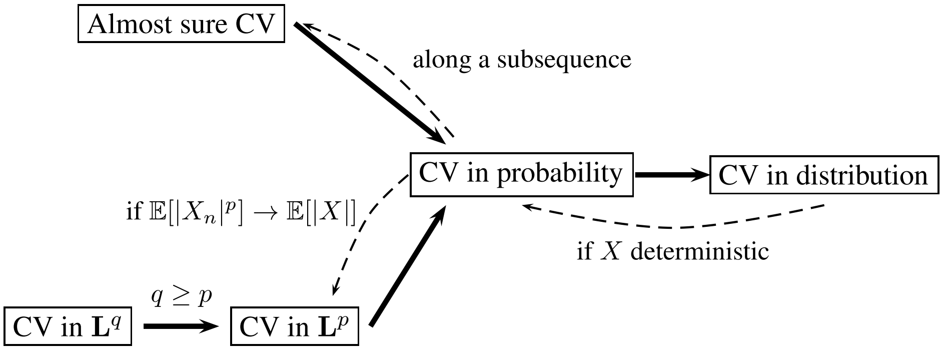


Figure 3.1: Hierarchy of various modes of convergence (including the convergence in distribution which will be seen in Subsection 3.1.3) and partial converse statements.

The hierarchy between these modes of convergence is summarised on Figure 3.1. In the proof of Proposition 3.1.2, we shall need the following two results.

Lemma 3.1.3 (Dominated Convergence Theorem for random variables)**.** *Assume that $X_n \to X$ almost surely and that there exists $Y \in \mathbf{L}^1(\mathbb{P})$ such that $|X_n| \leq Y$ for any n. Then $\mathbb{E}[X_n]$ converges to $\mathbb{E}[X]$.*

The statement of Lemma 3.1.3 is nothing but a reformulation of Lebesgue's Dominated Convergence Theorem, therefore we omit its proof. We however point out the important remark that it applies in particular if the sequence X_n is bounded by a deterministic constant y.

Lemma 3.1.4 (Markov's inequality)**.** *Let $Y \in \mathbf{L}^1(\mathbb{P})$ be such that $Y \geq 0$, almost surely. For any $a > 0$, we have*

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

Proof. Observe that, for any $y \geq 0$, $\mathbb{1}_{\{y \geq a\}} \leq y/a$ and take the expectation of this inequality evaluated in $y = Y$. □

[2] We should rather write in $\mathbf{L}^p(\mathbb{P}; \mathbb{R}^d)$ to be completely consistent with the notation of Chapter 2, however in order not to overweight the exposition we shall adopt this shorthand notation throughout the chapter.

[3] It is referred to as *convergence in measure* in analysis.

We are now in position to prove Proposition 3.1.2.

Proof of Proposition 3.1.2. We first assume that $X_n \to X$ almost surely and let A be the associated almost sure event on which $X_n(\omega) \to X(\omega)$ for any ω. For any $\epsilon > 0$, for any $\omega \in A$, we have $|X_n(\omega) - X(\omega)| < \epsilon$ for n large enough, and therefore

$$\lim_{n\to+\infty} \mathbb{1}_{\{|X_n(\omega)-X(\omega)|\geq\epsilon\}} = 0.$$

As a consequence, the random variable $\mathbb{1}_{\{|X_n-X|\geq\epsilon\}}$ converges to 0, almost surely, and therefore by Lemma 3.1.3,

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \mathbb{E}\left[\mathbb{1}_{\{|X_n-X|\geq\epsilon\}}\right] \to 0.$$

This proves the first point.

The second point is an immediate consequence of Exercise 1.3.6.

To prove the third point, we assume that $X_n \to X$ in $\mathbf{L}^1$ and fix $\epsilon > 0$. By Lemma 3.1.4, we then have

$$\mathbb{P}(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\epsilon} \to 0,$$

and the proof is completed. □

Proposition 3.1.5 (Properties of convergence in probability). *(i) If $X_n \to X$ in probability, then for any continuous function $f : \mathbb{R}^d \to \mathbb{R}^k$, $f(X_n) \to f(X)$ in probability.*

(ii) If $X_n \to X$ in probability and $Y_n \to Y$ in probability then $(X_n, Y_n) \to (X, Y)$ in probability.

(iii) If $X_n \to X$ in probability and $X_n \to Y$ in probability then $X = Y$ almost surely.

Before detailing the proof of Proposition 3.1.5, we point out several remarks.

- You should first convince yourself that the three statements of Proposition 3.1.5 become trivial if convergence in probability is replaced with almost sure convergence.
- If convergence in probability is replaced with convergence in $\mathbf{L}^p$ then the points (ii) and (iii) also remain true, however the first point no longer holds: it may depend on the growth of f.
- A straightforward application of the points (i) and (ii) is that if $X_n \to X$ and $Y_n \to Y$ in probability, then $X_n + Y_n \to X + Y$, $X_nY_n \to XY$ (if $d = 1$), and so on.

Proof of Proposition 3.1.5. If the function f is assumed to be uniformly continuous, then the proof of (i) is an easy exercise. To reduce the proof to this case, we use a localisation argument. Let $\epsilon > 0$. Since f is continuous on $\mathbb{R}^d$, for any $M \geq 0$, this function is uniformly continuous on the closed ball $\overline{B}(0, M+1)$, so that there exists $\delta_{M,\epsilon} \in (0,1]$ such that for any $x, x' \in \overline{B}(0, M+1)$, if $|x - x'| \leq \delta_{M,\epsilon}$ then $|f(x) - f(x')| \leq \epsilon$. We fix $M \geq 0$ and first write

$$\begin{aligned}\mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon\right) &= \mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| > M\right)\\ &\quad + \mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M\right)\\ &\leq \mathbb{P}\left(|X| > M\right) + \mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M\right).\end{aligned}$$

The second term in the right-hand side rewrites

$$\begin{aligned}\mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M\right) &= \mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| \leq \delta_{M,\epsilon}\right)\\ &\quad + \mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| > \delta_{M,\epsilon}\right).\end{aligned}$$

By definition of $\delta_{M,\epsilon}$, if $|X| \leq M$ and $|X - X_n| \leq \delta_{M,\epsilon}$ then $X, X_n \in \overline{B}(0, M+1)$, therefore

$$\mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| \leq \delta_{M,\epsilon}\right) = 0.$$

On the other hand, it is immediate that

$$\mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| > \delta_{M,\epsilon}\right) \leq \mathbb{P}\left(|X_n - X| > \delta_{M,\epsilon}\right).$$

Since $X_n \to X$ in probability, the right-hand side above goes to 0 when $n \to +\infty$. Overall, we deduce that

$$\limsup_{n\to+\infty} \mathbb{P}\left(|f(X_n) - f(X)| \geq \epsilon\right) \leq \mathbb{P}\left(|X| > M\right).$$

By the Monotone Convergence Theorem, the right-hand side goes to 0 when $M \to +\infty$, which completes the proof of (i).

To prove the point (ii), we endow the space of pairs (x, y) with the norm $|x| + |y|$ and use the union bound to write, for any $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| + |Y_n - Y| \geq \epsilon) \leq \mathbb{P}(|X_n - X| \geq \epsilon/2) + \mathbb{P}(|Y_n - Y| \geq \epsilon/2),$$

which easily leads to the claimed statement.

To prove the last point, we use the triangle inequality and the same application of the union bound to write, for any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(|X - Y| \geq \epsilon\right) &\leq \mathbb{P}\left(|X_n - X| + |X_n - Y| \geq \epsilon\right) \\ &\leq \mathbb{P}\left(|X_n - X| \geq \epsilon/2\right) + \mathbb{P}\left(|X_n - Y| \geq \epsilon/2\right),\end{aligned}$$

which shows that for any $\epsilon > 0$, the event $\{|X - Y| < \epsilon\}$ is almost sure. Taking a countable sequence $(\epsilon_M)_{M\geq 1}$ decreasing to 0, we deduce from Corollary 1.1.7 that almost surely, $|X - Y|$ is smaller than any ϵ_M, and therefore $X = Y$. □

3.1.2 Convergence of random variables: complements

In general, none of the converse statements to those of Proposition 3.1.2 hold true: counter-examples are provided in Exercise 3.1.6. However, some partial converse statements are gathered in Proposition 3.1.7 and Exercise 3.1.9.

🏠 **Exercise 3.1.6** (Counter-examples to converse statements to Proposition 3.1.2). *1. Let $(X_n)_{n\geq 1}$ be a sequence of independent random variables such that $X_n \sim \mathcal{B}(1/n)$ for any $n \geq 1$. Show that $X_n \to 0$ in probability but not almost surely.* You may use the Borel Zero-One Law from Exercise 1.4.2.

2. For $a > 0$, $b > 0$, let $(X_n)_{n\geq 1}$ be a sequence of random variables such that

$$X_n = \begin{cases} n^b & \textit{with probability } 1/n^a, \\ 0 & \textit{with probability } 1 - 1/n^a. \end{cases}$$

(a) For any $p \geq 1$, compute $\mathbb{E}[|X_n|^p]$.
(b) For given $1 \leq p < q$, choose a and b so that $X_n \to 0$ in $\mathbf{L}^p$ but not in $\mathbf{L}^q$.
(c) Choose a and b so that $X_n \to 0$ in probability but not in $\mathbf{L}^1$.

Proposition 3.1.7 (Almost sure convergence up to a subsequence). *If $X_n \to X$ in probability, then there is a (deterministic) increasing sequence of integers $(n_k)_{k\geq 1}$ such that the subsequence $(X_{n_k})_{k\geq 1}$ converges almost surely to X.*

Proof. Fix an increasing sequence of integers $(n_k)_{k\geq 1}$ and notice that the almost sure convergence of X_{n_k} to X is equivalent to the statement that

$$\mathbb{P}\left(\forall M \geq 1, \exists K \geq 1 : \forall k \geq K, |X_{n_k} - X| \leq \epsilon_M\right) = 1,$$

where $(\epsilon_M)_{M\geq 1}$ is a deterministic sequence of positive numbers which converges to 0. Since, by Corollary 1.1.7, a countable intersection of almost sure events remains almost sure, we deduce that it suffices to construct $(n_k)_{k\geq 1}$ such that

$$\forall M \geq 1, \qquad \mathbb{P}\left(\exists K \geq 1 : \forall k \geq K, |X_{n_k} - X| \leq \epsilon_M\right) = 1.$$

By the Borel–Cantelli Lemma (see Exercise 1.4.1), for any $M \geq 1$, the event $\{\exists K \geq 1 : \forall k \geq K, |X_{n_k} - X| \leq \epsilon_M\}$ is almost sure if

$$\sum_{k=1}^{+\infty} \mathbb{P}(|X_{n_k} - X| > \epsilon_M) < +\infty.$$

For a fixed value of M, since $X_n \to X$ in probability, it is easy to construct a sequence $(n_{k,M})_{k\geq 1}$ such that

$$\forall k \geq 1, \qquad \mathbb{P}\left(|X_{n_{k,M}} - X| > \epsilon_M\right) \leq \frac{1}{k^2},$$

and therefore the associated series is finite. To complete the proof, we need to remove the dependency upon M of this sequence. To this aim we use a diagonal argument and set $n_k = n_{k,k}$. Then, using the fact that ϵ_M is assumed to decrease, we have, for any $M \geq 1$,

$$\begin{aligned}\sum_{k=1}^{+\infty} \mathbb{P}\left(|X_{n_k} - X| > \epsilon_M\right) &\leq M + \sum_{k=M+1}^{+\infty} \mathbb{P}\left(|X_{n_{k,k}} - X| > \epsilon_M\right)\\ &\leq M + \sum_{k=M+1}^{+\infty} \mathbb{P}\left(|X_{n_{k,k}} - X| > \epsilon_k\right)\\ &\leq M + \sum_{k=M+1}^{+\infty} \frac{1}{k^2},\end{aligned}$$

which completes the proof. □

Proposition 3.1.7 allows to prove the following generalisation of Lemma 3.1.3, in which the almost sure convergence requirement is relaxed to convergence in probability, and which will be useful in the sequel.

🏠 **Exercise 3.1.8** (Dominated Convergence Theorem with convergence in probability)**.** *Assume that $X_n \to X$ in probability, and that there exists $Y \in \mathbf{L}^1(\mathbb{P})$ such that, almost surely, $|X_n| \leq Y$ for any n. The purpose of this exercise is to prove that $\mathbb{E}[X_n]$ converges to $\mathbb{E}[X]$.*

1. *Using Proposition 3.1.7, show that $|X| \leq Y$, almost surely.*
2. *Complete the proof of the claimed statement.* Hint: you may remark that, for any $\epsilon > 0$, there exists $M \geq 1$ such that $\mathbb{E}[Y \mathbb{1}_{\{Y>M\}}] \leq \epsilon$.
3. *Deduce that if $X_n \to X$ in probability and there is a random variable $Y \in \mathbf{L}^p(\mathbb{P})$ such that $|X_n - X| \leq Y$, then $X_n \to X$ in $\mathbf{L}^p$.*

☕ **Exercise 3.1.9** (Riesz–Scheffé's Lemma)**.** *Assume that $X_n \to X$ almost surely, and that $X_n, X \in \mathbf{L}^p(\mathbb{P})$ with $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$.*

1. *Show that $2^{p-1}(|X_n|^p + |X|^p) - |X_n - X|^p \geq 0$, almost surely.*
2. *Using Fatou's Lemma, deduce that $X_n \to X$ in $\mathbf{L}^p$.*
3. *Show that this conclusion still holds if X_n is only assumed to converge to X in probability.*

3.1.3 Convergence in distribution

Convergence in distribution is a bit different from the modes of convergence introduced in Definition 3.1.1, because it concerns the law of X_n rather than the variable X_n itself.

Definition 3.1.10 (Weak convergence of probability measures)**.** *A sequence $(P_n)_{n\geq 1}$ of probability measures on $\mathbb{R}^d$ converges weakly to P if, for any continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\lim_{n\to+\infty} \int_{x\in\mathbb{R}^d} f(x)\mathrm{d}P_n(x) = \int_{x\in\mathbb{R}^d} f(x)\mathrm{d}P(x).$$

The following statement, which goes by the name of Portmanteau's Theorem, provides equivalent characterisations of the weak convergence.

Theorem 3.1.11 (Portmanteau's Theorem)**.** *The following statements are equivalent.*

(i) *P_n converges weakly to P.*

(ii) *For any uniformly continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$, $\lim_{n\to+\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$.*

(iii) *For any closed set F, $\limsup_{n\to+\infty} P_n(F) \leq P(F)$.*

(iv) *For any open set G, $\liminf_{n\to+\infty} P_n(G) \geq P(G)$.*

(v) *For any measurable set A such that $P(\partial A) = 0$, $\lim_{n\to+\infty} P_n(A) = P(A)$.*

Definition 3.1.12 (Convergence in distribution)**.** *A sequence of random variables $(X_n)_{n\geq 1}$ converges in distribution to X if the law of X_n converges weakly to the law of X; in other words, if for any continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$,*

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

In order to compare convergence in distribution with the modes of convergence introduced in Definition 3.1.1, one should compare the following three statements with the contents of Proposition 3.1.5.

Proposition 3.1.13 (Properties of convergence in distribution)**.** *(i)* *If $X_n \to X$ in distribution, then for any continuous function $f : \mathbb{R}^d \to \mathbb{R}^k$, $f(X_n)$ converges in distribution to $f(X)$.*

(ii) *If $X_n \to X$ in distribution and $Y_n \to Y$ in distribution, then* nothing can be said *about the convergence in distribution of the pair (X_n, Y_n).*

(iii) *If $X_n \to X$ in distribution and $X_n \to Y$ in distribution, then the random variables X and Y have the same law.*

The first and third points of Proposition 3.1.13 are straightforward consequences of Definition 3.1.12. As far as the second point is concerned, the assertion that $X_n \to X$ and $Y_n \to Y$ is a statement on the marginal distributions of X_n and Y_n, which is not sufficient to characterise the joint distribution of the pair (X_n, Y_n), and therefore does not allow to describe the asymptotic behaviour of the law of (X_n, Y_n) in general. There are however particular cases in which the convergence in distribution of the pair (X_n, Y_n) can be deduced from the marginal convergence in distribution of X_n and Y_n.

Lemma 3.1.14 (Convergence in distribution of (X_n, Y_n))**.** *(i)* *Assume that, for any n, X_n and Y_n are independent, and that $X_n \to X$, $Y_n \to Y$ in distribution. Then the pair (X_n, Y_n) converges in distribution to (X', Y'), where X' and Y' are independent, and X' (resp. Y') has the same law as X (resp. Y).*

(ii) *(Slutsky's Lemma) Assume that $X_n \to X$ in distribution and $Y_n \to y$ in probability, where y is a constant. Then $(X_n, Y_n) \to (X, y)$ in distribution.*

The proof of Lemma 3.1.14 is postponed to Exercise 3.1.18. We first detail important properties and characterisations of the convergence in distribution.

Proposition 3.1.15 (Convergence in probability and convergence in distribution)**.** *If $X_n \to X$ in probability, then $X_n \to X$ in distribution.*

Conversely, for any $x \in \mathbb{R}^d$, if $X_n \to x$ in distribution then $X_n \to x$ in probability.

Proof. Let us assume that $X_n \to X$ in probability and take $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded. By the triangle inequality, for any $\epsilon > 0$,

$$\begin{aligned} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| &\leq \mathbb{E}\left[|f(X_n) - f(X)|\right] \\ &= \mathbb{E}\left[|f(X_n) - f(X)|\mathbb{1}_{\{|f(X_n)-f(X)|<\epsilon\}}\right] \\ &\quad + \mathbb{E}\left[|f(X_n) - f(X)|\mathbb{1}_{\{|f(X_n)-f(X)|\geq\epsilon\}}\right] \\ &\leq \epsilon + 2\|f\|_\infty \mathbb{P}(|f(X_n) - f(X)| \geq \epsilon). \end{aligned}$$

By Proposition 3.1.5 (i), $\mathbb{P}(|f(X_n) - f(X)| \geq \epsilon)$ goes to 0 when $n \to \infty$. Therefore

$$\limsup_{n\to+\infty} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leq \epsilon$$

for any ϵ, which shows that $X_n \to X$ in distribution.

For the converse statement, let us assume that $X_n \to x$ in distribution, fix $\epsilon > 0$, and consider a continuous function $\psi_\epsilon : [0, +\infty) \to [0, 1]$ such that $\psi_\epsilon(0) = 0$ and $\psi_\epsilon(r) \geq \mathbb{1}_{\{r\geq\epsilon\}}$ for any $r \geq 0$. Then the function $f : x' \in \mathbb{R}^d \mapsto \psi_\epsilon(|x' - x|)$ is continuous and bounded, and it satisfies

$$\forall x' \in \mathbb{R}^d, \quad \mathbb{1}_{\{|x'-x|\geq\epsilon\}} \leq f(x'), \qquad \text{and} \quad f(x) = 0.$$

Applying Definition 3.1.12 with this function, we get

$$\mathbb{P}(|X_n - X| \geq \epsilon) \leq \mathbb{E}[f(X_n)] \to \mathbb{E}[f(x)] = 0,$$

which shows that $X_n \to x$ in probability. □

Proposition 3.1.16 (Lévy's Theorem)**.** *$X_n \to X$ in distribution if and only if, for any $u \in \mathbb{R}^d$, $\Psi_{X_n}(u) \to \Psi_X(u)$.*

Proof. The direct implication is straightforward since for any $u \in \mathbb{R}^d$, the functions $x \mapsto \cos(\langle u, x\rangle)$ and $x \mapsto \sin(\langle u, x\rangle)$ are continuous and bounded.

We admit the converse implication. From an analytic perspective, the main idea is to write the mapping $P_{X_n} \to \Psi_{X_n}$ as a (slighlty modified) Fourier transform $\mathcal{F}$, so that the claim to prove reduces to showing some continuity property of the inverse transform $\mathcal{F}^{-1}$. □

🏠 **Exercise 3.1.17** (Gaussian vectors and convergence in distribution)**.** *Let $(X^n)_{n\geq 1}$ be a sequence of Gaussian vectors in $\mathbb{R}^d$. For all $n \geq 1$, let $m^n = \mathbb{E}[X^n]$ and $K^n = \mathrm{Cov}[X^n]$.*

1. *Show that if $m^n \to m$ and $K^n \to K$, then X^n converges in distribution to $\mathcal{N}_d(m, K)$.*
2. *Conversely, show that if X^n converges in distribution to some random vector X, then there exist m and K such that $m^n \to m$ and $K^n \to K$, and $X \sim \mathcal{N}_d(m, K)$.*

📄 **Exercise 3.1.18** (Proof of Lemma 3.1.14)**.** *Prove the two statements of Lemma 3.1.14 using Proposition 3.1.16.*

A famous application of Slutsky's Lemma is the *Delta method*, introduced in the next exercise.

🏠 **Exercise 3.1.19** (The Delta method)**.** *Let $(X_n)_{n\geq 1}$, x and Y in $\mathbb{R}^d$ be such that $a_n(X_n - x) \to Y$ in distribution, for some deterministic sequence $(a_n)_{n\geq 1}$ which grows to $+\infty$.*

1. *Show that $X_n \to x$ in probability.*
2. *Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be C^1. Show that*

$$\lim_{n\to+\infty} a_n \left(f(X_n) - f(x)\right) = \nabla f(x) Y, \qquad \textit{in distribution,}$$

where $\nabla f(x) \in \mathbb{R}^{k\times d}$ is the matrix with coordinates $\partial f_i(x)/\partial x_j$.

We complete this subsection by mentioning a sufficient condition for convergence in distribution which is often easy to check.

☕ **Exercise 3.1.20** (Scheffé's Lemma)**.** *Let $(p_n)_{n\geq 1}$ be a sequence of probability densities with respect to some σ-finite measure μ on $\mathbb{R}^d$, such that*

$$\mu\textit{-almost everywhere,} \qquad p_n \to p,$$

for some probability density p with respect to μ on $\mathbb{R}^d$.

1. *Show that for any $n \geq 1$,*

$$\int_{x\in\mathbb{R}^d} |p_n(x) - p(x)| \mathrm{d}\mu(x) = 2 \int_{x\in\mathbb{R}^d} [p_n(x) - p(x)]_- \mathrm{d}\mu(x).$$

2. *Deduce that*

$$\lim_{n\to+\infty} \int_{x\in\mathbb{R}^d} |p_n(x) - p(x)| \mathrm{d}\mu(x) = 0,$$

and then that if X_n has density p_n and X has density p then $X_n \to X$ in distribution.

Remark 3.1.21 (Convergence in distribution in discrete spaces)**.** *Definition 3.1.12 is given for random variables in $\mathbb{R}^d$, but the notion of convergence in distribution also makes sense (and is of interest) for discrete random variables, that is to say variables taking their values in a countable set E endowed with the σ-field of all its subsets. In this case, the natural topology on E is the one making all functions $f : E \to \mathbb{R}$ continuous. Then, with similar arguments as in Exercise 3.1.20 (taking for μ the counting measure $\sum_{x\in E} \delta_x$), it may be shown that the following statements are equivalent:*

(i) *$X_n \to X$ in distribution, that is to say $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for any bounded function f;*
(ii) *for any $x \in E$, $\mathbb{P}(X_n = x) \to \mathbb{P}(X = x)$;*
(iii) *$\sum_{x\in E} |\mathbb{P}(X_n = x) - \mathbb{P}(X = x)| \to 0$.*

3.1.4 Convergence of moments

In this subsection we let $(X_n)_{n\geq 0}$ be random variables in $\mathbb{R}^d$ which converge in distribution to X, and for $f : \mathbb{R}^d \to \mathbb{R}$, we look for conditions under which $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$. Of course, by Definition 3.1.12, it is the case if f is continuous and bounded. We shall study how to relax both conditions. We start with the continuity condition.

Proposition 3.1.22 (Mapping theorem)**.** *Let $X_n \to X$ in distribution and $f : \mathbb{R}^d \to \mathbb{R}$ be bounded. Denote by C_f the set of $x \in \mathbb{R}^d$ such that f is continuous at x. If $\mathbb{P}(X \in C_f) = 1$ then $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$.*

We leave the proof of Proposition 3.1.22 aside but insist on an important corollary.

Corollary 3.1.23 (Convergence of CDFs)**.** *Let X_n be a sequence of random variables in $\mathbb{R}$ which converge in distribution to X. Denote by F_n (resp. F) the Cumulative Distribution Function of X_n (resp. X). For any x such that $\mathbb{P}(X = x) = 0$, or equivalently $F(x^-) = F(x)$,*

$$\lim_{n\to+\infty} F_n(x) = F(x).$$

In particular,

(i) $F_n(x) \to F(x)$, $\mathrm{d}x$-almost everywhere;

(ii) if X has a density with respect to the Lebesgue measure on $\mathbb{R}$, then $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$.

Corollary 3.1.23 is a straighforward consequence of Proposition 3.1.22, and for the point (i), of the observation that the set of discontinuity points of F is at most countable and therefore negligible for the Lebesgue measure.

We now turn our attention to functions f which are continuous but not necessarily bounded. In this case, since $f(X_n)$ converges in distribution to $f(X)$, up to renaming $f(X_n)$ in X_n we may directly study conditions under which $\mathbb{E}[X_n]$ converges to $\mathbb{E}[X]$ for $X_n, X \in \mathbb{R}$.

Definition 3.1.24 (Uniform integrability)**.** *A sequence of random variables $(X_n)_{n\geq 1}$ in $\mathbb{R}$ is called* uniformly integrable *if*

$$\lim_{M\to+\infty} \sup_{n\geq 1} \mathbb{E}\left[|X_n|\mathbb{1}_{\{|X_n|\geq M\}}\right] = 0.$$

🗎 **Exercise 3.1.25.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables in $\mathbb{R}$. This sequence is said to be bounded in $\mathbf{L}^p(\mathbb{P})$ if $\sup_{n\geq 1} \mathbb{E}[|X_n|^p] < +\infty$.*

1. *Show that if $(X_n)_{n\geq 1}$ is uniformly integrable then it is bounded in $\mathbf{L}^1(\mathbb{P})$.*
2. *Construct a sequence which is bounded in $\mathbf{L}^1(\mathbb{P})$ but not uniformly integrable.*
3. *If there exists $p > 1$ such that $(X_n)_{n\geq 1}$ is bounded in $\mathbf{L}^p(\mathbb{P})$, show that $(X_n)_{n\geq 1}$ is uniformly integrable.*

Uniform integrability is the key property to deduce the convergence of moments from the convergence in distribution.

Proposition 3.1.26 (Convergence of expectations)**.** *If $X_n \to X$ in distribution and the sequence $(X_n)_{n\geq 1}$ is uniformly integrable, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.*

🏠 **Exercise 3.1.27.** *Prove Proposition 3.1.26.*

3.2 Limit theorems

Throughout this section, we consider a sequence $(X_n)_{n\geq 1}$ of iid random variables in $\mathbb{R}^d$, and for any $n \geq 1$ we denote by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \in \mathbb{R}^d$$

the *empirical mean* of $X_1, \ldots, X_n$.

🗎 **Exercise 3.2.1.** *If $X_1 \in \mathbf{L}^1(\mathbb{P}; \mathbb{R}^d)$, compute $\mathbb{E}[\overline{X}_n]$ and if $X_1 \in \mathbf{L}^2(\mathbb{P}; \mathbb{R}^d)$, compute $\mathrm{Cov}[\overline{X}_n]$.*

The two main results of this section, the Law of Large Numbers and the Central Limit Theorem, describe the asymptotic behaviour of $\overline{X}_n$.

3.2.1 Laws of Large Numbers

There are two distinct statement of the Law of Large Numbers (LLN): a weak and a strong form.

Proposition 3.2.2 (Weak Law of Large Numbers)**.** *If* $X_1 \in \mathbf{L}^2(\mathbb{P};\mathbb{R}^d)$ *then* $\overline{X}_n \to \mathbb{E}[X_1]$ *in* $\mathbf{L}^2$.

Proof. The proof of the weak LLN is elementary. For the sake of simplicity we assume that $d = 1$. Then, by Exercise 3.2.1,

$$\mathbb{E}\left[\left|\overline{X}_n - \mathbb{E}[X_1]\right|^2\right] = \mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_1)}{n},$$

which converges to 0. □

Theorem 3.2.3 (Strong Law of Large Numbers)**.** *If* $X_1 \in \mathbf{L}^1(\mathbb{P};\mathbb{R}^d)$ *then* $\overline{X}_n \to \mathbb{E}[X_1]$ *almost surely.*

Theorem 3.2.3 is certainly a cornerstone of probability theory, but its proof is far from trivial. It is sketched and discussed in Subsection 3.2.2.

🏠 **Exercise 3.2.4** (Law of Large Numbers in $\mathbf{L}^p$)**.** *Let* $p \geq 1$ *and assume that* $X_1 \in \mathbf{L}^p(\mathbb{P};\mathbb{R}^d)$. *The aim of this exercise is to prove that* $\overline{X}_n \to \mathbb{E}[X_1]$ *in* $\mathbf{L}^p$. *Clearly, there is no loss of generality in assuming that* $\mathbb{E}[X_1] = 0$.

1. *Show that, for any* $n \geq 1$, *for any* $M \geq 1$,

$$\mathbb{E}\left[|\overline{X}_n|^p \mathbb{1}_{\{|\overline{X}_n|^p \geq M\}}\right] \leq \mathbb{E}\left[|X_1|^p \mathbb{1}_{\{|\overline{X}_n|^p \geq M\}}\right].$$

2. *Deduce that the sequence* $(|\overline{X}_n|^p)_{n\geq 1}$ *is uniformly integrable, and complete the proof of the claimed statement.*

3.2.2 On the proof of the strong Law of Large Numbers

The first proof of the strong LLN is due to Kolmogorov in 1933. It is decomposed in the following steps. As a preliminary remark, we assume without loss of generality that $\mathbb{E}[X_1] = 0$, which is possible up to replacing X_i with $X_i - \mathbb{E}[X_i]$.

1. Truncation: one sets $X_i' = X_i \mathbb{1}_{\{|X_i| \leq i\}}$ and proves that Theorem 3.2.3 is equivalent to the statement that

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^n X_i' = 0, \qquad \text{almost surely.} \tag{3.1}$$

The justification of this equivalence relies on the Borel–Cantelli Lemma.

2. Centering: one sets $Z_i = X_i' - \mathbb{E}[X_i]$ and proves that (3.1) is equivalent to

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^n Z_i = 0, \qquad \text{almost surely.} \tag{3.2}$$

Besides, an explicit computation yields

$$\sum_{n\geq 1} \mathrm{Var}\left(\frac{Z_n}{n}\right) < +\infty. \tag{3.3}$$

3. Auxiliary results: the following three statements, which are of independent interest, are used to complete the proof.

Lemma 3.2.5 (Kolmogorov's maximal inequality). *Let $(Y_i)_{i\geq 1}$ be a sequence of independent and centered variables, and $W_n = Y_1 + \cdots + Y_n$. Then for any $x > 0$,*

$$\mathbb{P}\left(\sup_{n\geq 1} |W_n| > x\right) \leq \frac{1}{x^2} \sum_{i=1}^{\infty} \mathrm{Var}(Y_i).$$

Lemma 3.2.5, combined with the Borel–Cantelli Lemma, allows to prove the following statement.

Lemma 3.2.6 (Convergence criterion). *Let $(U_n)_{n\geq 1}$ be a sequence of independent and centered random variables. If*

$$\sum_{n=1}^{\infty} \mathrm{Var}(U_n) < +\infty,$$

then there exists a random variable T such that

$$\lim_{N\to+\infty} \sum_{n=1}^{N} U_n = T, \qquad \textit{almost surely.}$$

The last auxiliary result is Kronecker's Lemma (which is purely deterministic).

Lemma 3.2.7 (Kronecker Lemma). *Let $(a_n)_{n\geq 1}$ be a sequence of positive numbers which decreases to 0. For any sequence $(u_n)_{n\geq 1}$, if the sequence $(\sum_{n=1}^{N} a_n u_n)_{N\geq 1}$ has a finite limit then $a_n \sum_{i=1}^{n} u_i$ converges to 0.*

4. Conclusion of the proof: combining (3.3) with Lemma 3.2.6, we get that $\sum_{n=1}^{N} Z_n/n$ has a finite limit, which by the Kronecker Lemma then yields (3.2).

Shorter and more elementary proofs have been proposed since Kolmogorov's original proof. A particularly famous one is due to Etemadi in 1981[4]. A recent preprint by Fitzsimmons[5] discusses another elementary sketch and its relation with previous similar arguments in the literature.

3.2.3 The Central Limit Theorem

In this subsection we assume that $X_1 \in \mathbf{L}^2(\mathbb{P};\mathbb{R}^d)$ and set $K = \mathrm{Cov}[X_1] \in \mathbb{R}^{d\times d}$. In the next statement it is convenient to write $X_n \to P$, in distribution, when $X_n \to X$ in distribution and $X \sim P$. We recall that Gaussian measures on $\mathbb{R}^d$ are introduced in Chapter 2.

Theorem 3.2.8 (Central Limit Theorem). *We have*

$$\lim_{n\to+\infty} \sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_1]\right) = \mathcal{N}_d(0, K).$$

Proof. For all $i \geq 1$, let $Y_i = X_i - \mathbb{E}[X_1]$, so that $\mathbb{E}[Y_i] = 0$ and $\mathrm{Cov}[Y_i] = K$. We also denote

$$Z_n = \sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_1]\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,$$

so that the characteristic function of Z_n writes, for all $u \in \mathbb{R}^d$,

$$\Psi_{Z_n}(u) = \mathbb{E}\left[\exp\left(\mathrm{i}\left\langle u, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i\right\rangle\right)\right] = \mathbb{E}\left[\exp\left(\mathrm{i}\left\langle \frac{u}{\sqrt{n}}, Y_1\right\rangle\right)\right]^n = \Phi_{Y_1}(u/\sqrt{n})^n,$$

[4] https://link.springer.com/article/10.1007/BF01013465

[5] https://arxiv.org/abs/2111.05766

where we have used the fact that the variables Y_i are iid. By Lemma 2.2.10, writing $Y_1 = (Y_{1,1}, \dots, Y_{1,d})$, we have

$$\frac{\partial \Psi_{Y_1}}{\partial u_i}(0) = \mathrm{i}\mathbb{E}[Y_{1,i}] = 0, \qquad \frac{\partial^2 \Psi_{Y_1}}{\partial u_i \partial u_j}(0) = \mathrm{i}^2\mathbb{E}[Y_{1,i}Y_{1,j}] = -K_{i,j},$$

so that the function Ψ_{Y_1} satisfies Taylor's expansion

$$\Psi_{Y_1}(u/\sqrt{n}) = 1 - \frac{\langle u, Ku\rangle}{2n} + \mathrm{o}\left(\frac{1}{n}\right)$$

when $n \to +\infty$. Using Lemma 3.2.9 below, we deduce that

$$\lim_{n\to+\infty} \Psi_{Z_n}(u) = \exp\left(-\frac{\langle u, Ku\rangle}{2}\right),$$

which by Proposition 2.2.14 is the characteristic function of $Z \sim \mathcal{N}_d(0, K)$. As a consequence, Proposition 2.2.7 ensures that Z_n converges in distribution to Z. □

In the proof of Theorem 3.2.8, we have used the following technical result.

Lemma 3.2.9 (An exponential limit for complex sequences)**.** *Let $\theta \in \mathbb{R}$ and $(\epsilon_n)_{n\geq 1}$ be a sequence of complex numbers which converges to* 0*. Then*

$$\lim_{n\to+\infty} \left(1 + \frac{\theta}{n} + \frac{\epsilon_n}{n}\right)^n = \mathrm{e}^{\theta}.$$

Proof. Using Taylor's expansion for the logarithm, it is standard to show that

$$\lim_{n\to+\infty} \left(1 + \frac{\theta}{n}\right)^n = \mathrm{e}^{\theta}.$$

This argument cannot be applied directly to $(1 + (\theta + \epsilon_n)/n)^n$ because ϵ_n is a complex number. However we may compare both prelimits by writing

$$\begin{aligned}\left(1 + \frac{\theta}{n} + \frac{\epsilon_n}{n}\right)^n - \left(1 + \frac{\theta}{n}\right)^n &= \int_{u=0}^{1} \frac{\mathrm{d}}{\mathrm{d}u}\left(1 + \frac{\theta}{n} + \frac{u\epsilon_n}{n}\right)^n \mathrm{d}u\\ &= \epsilon_n \int_{u=0}^{1} \left(1 + \frac{\theta}{n} + \frac{u\epsilon_n}{n}\right)^{n-1} \mathrm{d}u,\end{aligned}$$

and it follows from the estimate

$$\left|1 + \frac{\theta + u\epsilon_n}{n}\right|^{n-1} \leq \left(1 + \frac{|\theta| + |\epsilon_n|}{n}\right)^{n-1} \leq \exp\left(n \log\left(1 + \frac{|\theta| + |\epsilon_n|}{n}\right)\right)$$

that the sequence $\sup_{u\in[0,1]} |1 + (\theta + u\epsilon_n)/n|^{n-1}$ is bounded, which proves the lemma. □

🏠 **Exercise 3.2.10** (Stronger convergence in the CLT)**.** *With the notation of the proof of Theorem 3.2.8, it is a natural question to wonder whether there exists a random variable Z such that Z_n converges to Z almost surely. Notice that if such a variable exists, then necessarily $Z \sim \mathcal{N}_d(0, K)$.*

1. *Let $Z'_n = \frac{1}{\sqrt{n}} \sum_{i=n+1}^{2n} Y_i$. Show that Z'_n converges in distribution to some random variable Z' and explicit the law of Z'.*
2. *If Z_n converges almost surely to some random variable Z, show that Z'_n converges almost surely and express its limit in terms of Z.*
3. *What do you conclude?*

Chapter 4

The Monte Carlo method and variance reduction

Contents

The Monte Carlo method is designed to approximate integrals of the form

$$\mathcal{I} = \int_{x \in E} f(x) \mathrm{d}P(x),$$

where P is a probability measure on a measurable space $(E, \mathcal{E})$, and $f \in \mathbf{L}^1(P)$. This integral naturally rewrites

$$\mathcal{I} = \mathbb{E}\left[f(X)\right], \qquad X \sim P,$$

therefore by the strong Law of Large Numbers, it is the $n \to +\infty$ almost sure limit of

$$\widehat{\mathcal{I}}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

with $X_1, \ldots, X_n$ independent copies of X. The numerical approximation of $\mathcal{I}$ by $\widehat{\mathcal{I}}_n$ is the essence of the Monte Carlo method.

4.1 Accuracy of the Monte Carlo method

4.1.1 Asymptotic confidence intervals

Throughout the sequel, we assume that $f \in \mathbf{L}^2(P)$ and denote by σ^2 the variance of $f(X)$. We assume that $\sigma^2 > 0$ (otherwise the numerical computation of $\mathcal{I}$ is rather trivial). The Central Limit Theorem asserts that

$$\lim_{n \to +\infty} \frac{\sqrt{n}}{\sigma} \left(\widehat{\mathcal{I}}_n - \mathcal{I} \right) = \mathcal{N}(0,1), \qquad \text{in distribution,}$$

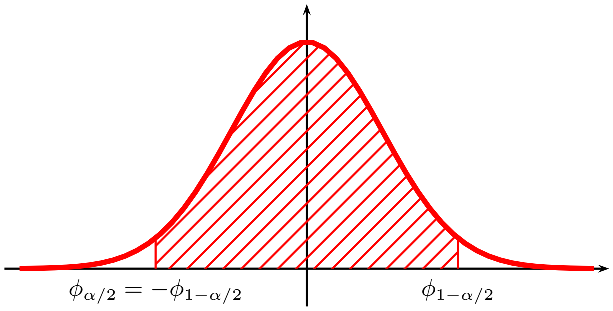
and therefore, by Corollary 3.1.23, for any $\phi > 0$, the interval

$$I_n = \left[\widehat{\mathcal{I}}_n - \phi\frac{\sigma}{\sqrt{n}}, \widehat{\mathcal{I}}_n + \phi\frac{\sigma}{\sqrt{n}}\right]$$

satisfies

$$\lim_{n\to+\infty} \mathbb{P}\left(\mathcal{I} \in I_n\right) = \frac{1}{\sqrt{2\pi}}\int_{u=-\phi}^{\phi} \mathrm{e}^{-u^2/2}\mathrm{d}u.$$

In particular, if one fixes $\alpha \in (0, 1/2)$ and takes for ϕ the *quantile* $\phi_{1-\alpha/2}$ of order $1-\alpha/2$ of the standard Gaussian distribution, then the value of the limit is $1-\alpha$. Standard values of $\phi_{1-\alpha/2}$ are presented on Figure 4.1.



$1-\alpha$	$\phi_{1-\alpha/2}$
90%	1.65
95%	1.96
99%	2.58

Figure 4.1: Quantiles of the standard Gaussian distribution. The hatched area on the figure is equal to $1-\alpha$.

In short, if n is large enough, then there is a 95% probability that the quantity $\mathcal{I}$, which we aim at evaluating, lies between $\widehat{\mathcal{I}}_n - 1.96\sigma/\sqrt{n}$ and $\widehat{\mathcal{I}}_n + 1.96\sigma/\sqrt{n}$. Two problems remain with this statement: the assumption that 'n is large enough' is rather vague, and the variance σ^2 is not necessarily known, as its computation also involves evaluating an integral over $\mathbb{R}^d$. We first address this second point in the next statement.

Proposition 4.1.1 (Asymptotic confidence interval)**.** *Let*

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \widehat{\mathcal{I}}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2 - \widehat{\mathcal{I}}_n^2$$

be the empirical variance *of the sample $f(X_1), \ldots, f(X_n)$. For any $\alpha \in (0, 1/2)$, the interval*

$$I'_n = \left[\widehat{\mathcal{I}}_n - \phi_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\mathcal{I}}_n + \phi_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right]$$

satisfies

$$\lim_{n\to+\infty} \mathbb{P}\left(\mathcal{I} \in I'_n\right) = 1-\alpha.$$

Proposition 4.1.1 shows that the confidence intervals I_n and I'_n share the same asymptotic properties, so that we do not lose anything estimating the variance σ^2 by its empirical version $\widehat{\sigma}_n^2$. Since the latter estimator is easily computed from the sample $f(X_1), \ldots, f(X_n)$, error bars given by the interval I'_n should always be provided together with the result $\widehat{\mathcal{I}}_n$ of a Monte Carlo estimation.

Proof of Proposition 4.1.1. We first write

$$\frac{\sqrt{n}}{\widehat{\sigma}_n}\left(\widehat{\mathcal{I}}_n - \mathcal{I}\right) = \frac{\sigma}{\widehat{\sigma}_n}\frac{\sqrt{n}}{\sigma}\left(\widehat{\mathcal{I}}_n - \mathcal{I}\right).$$

By the strong LLN, the ratio $\sigma/\widehat{\sigma}_n$ converges to 1, almost surely. Therefore, Slutsky's Lemma 3.1.14 (ii) implies that the right-hand side above converges in distribution to a standard Gaussian variable, and thus the conclusion follows from the same application of Corollary 3.1.23 as for I_n. □

4.1.2 Nonasymptotic confidence intervals

The statement of Proposition 4.1.1 is asymptotic, and therefore it is natural to ask how large should n be chosen for the probability that $\mathcal{I} \in I'_n$ to be close to $1-\alpha$. The *Berry–Essen Theorem* provides a first answer.

Theorem 4.1.2 (Berry–Essen Theorem)**.** *There exists a universal constant $C > 0$ such that, for any $n \geq 1$,*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\widehat{\mathcal{I}}_n - \mathcal{I}\right) \leq x\right) - \mathbb{P}(G \leq x)\right| \leq \frac{C}{\sigma^3\sqrt{n}}\mathbb{E}[|f(X) - \mathcal{I}|^3],$$

where $G \sim \mathcal{N}(0,1)$.

In view of constructing confidence intervals, one may also derive nonasymptotic bounds from *concentration inequalities*, and get intervals J_n which are such that

$$\mathbb{P}\left(\mathcal{I} \in J_n\right) \geq 1 - \alpha. \tag{4.1}$$

An elementary such example is provided by Tchebychev's inequality:

$$\forall a > 0, \qquad \mathbb{P}\left(|\widehat{\mathcal{I}}_n - \mathcal{I}| \geq a\right) \leq \frac{\mathrm{Var}(\widehat{\mathcal{I}}_n)}{a^2} = \frac{\sigma^2}{na^2},$$

which easily follows from Markov's inequality, and from which we deduce that

$$J_n = \left[\widehat{\mathcal{I}}_n - \frac{\sigma}{\sqrt{\alpha n}}, \widehat{\mathcal{I}}_n + \frac{\sigma}{\sqrt{\alpha n}}\right]$$

satisfies the estimate (4.1). Since the latter is an inequality, and not an equality, the interval J_n is more conservative than I_n: the probability that $\mathcal{I} \in J_n$ could be much larger than $1-\alpha$, but it is *at least* $1-\alpha$. On the other hand, the bound (4.1) holds for any value of n, and does not rely on the Central Limit Theorem. We plot on Figure 4.2 the ratio $(1/\sqrt{\alpha})/\phi_{1-\alpha/2}$ between the widths of J_n and I_n, as a function of α: when α is not too small, J_n is only a few times larger than I_n.

As for I_n, the bounds of J_n depend on σ, which is not known in general. Similarly to Proposition 4.1.1, it may be estimated by the empirical variance of the sample. On the other hand, if $f(X)$ is bounded, then universal bounds are available.

🏠 **Exercise 4.1.3.** *Let $Y \in \mathbf{L}^2(\mathbb{P})$.*

1. *Show that*

$$\mathrm{Var}(Y) = \min_{y\in\mathbb{R}}\mathbb{E}\left[(Y-y)^2\right].$$

2. *Deduce that if Y takes its values in a bounded interval $[a,b]$, then*

$$\mathrm{Var}(Y) \leq \frac{(b-a)^2}{4}.$$

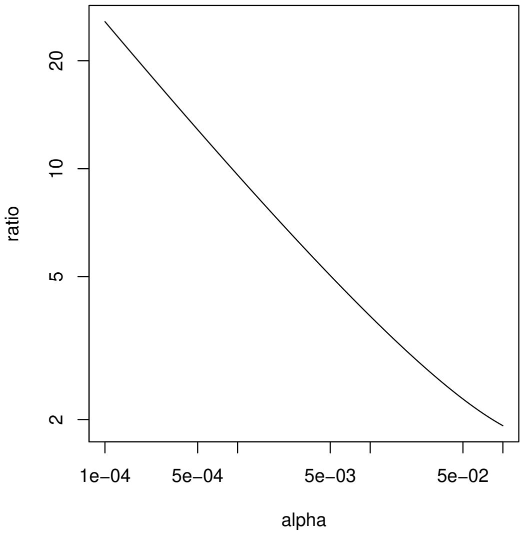


Figure 4.2: Ratio between the width of the confidence intervals J_n and I_n.

3. *Show that this inequality is sharp by exhibiting a random variable Y for which it is an equality.*

In fact, when $f(X)$ is bounded, concentration inequalities are available which are more powerful than Tchebychev's inequality, in the sense that they provide smaller confidence intervals.

☕ **Exercise 4.1.4** (The Hoeffding inequality)**.** *Throughout the exercise, we let $Y_1, \ldots, Y_n$ be iid random variables which take their values in $[0,1]$. We set $Z_i = Y_i - \mathbb{E}[Y_i]$ and, for any $\lambda \geq 0$, define*

$$F(\lambda) = \log \mathbb{E}\left[\exp(\lambda Z_1)\right].$$

1. *Show that $F'(\lambda) = \mathbb{E}_\lambda[Z_1]$ and $F''(\lambda) = \mathrm{Var}_\lambda(Z_1)$ for some probability measure $\mathbb{P}_\lambda$ to be defined.*
2. *Using Exercise 4.1.3, deduce that, for any $\lambda \geq 0$, $\mathbb{E}[\exp(\lambda Z_1)] \leq \exp(\lambda^2/8)$.*
3. *Deduce that, for any $r \geq 0$ and $n \geq 1$,*

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \geq r\sqrt{n}\right) \leq \exp\left(\frac{\lambda^2 n}{8} - \lambda r\sqrt{n}\right).$$

4. *Optimising in $\lambda \geq 0$, conclude that*

$$\mathbb{P}\left(\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \geq r\sqrt{n}\right) \leq \exp(-2r^2).$$

   This inequality is called Hoeffding's inequality.
5. *If $f(X)$ takes its values in some bounded interval $[a,b]$, deduce from Hoeffding's inequality a confidence interval for $\mathcal{I}$.*
6. *Compare the width of this confidence interval with those given by Tchebychev's inequality, or the Central Limit Theorem.*

4.2 Variance reduction

Neglecting the error induced by the approximation of σ^2 by $\widehat{\sigma}_n^2$, the length of the confidence interval I_n obtained in Section 4.1 for $\mathcal{I}$ is

$$\ell_n := 2\phi_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Let $X \sim \mathcal{N}(0,1)$. Assume that the quantity which we are trying to approximate is

$$\mathcal{I} = \mathbb{P}(X \geq 20) = \mathbb{E}[f(X)], \qquad f(x) = \mathbb{1}_{\{x \geq 20\}}.$$

On the one hand, an upper bound on $\mathcal{I}$ can be obtained analytically by writing

$$\mathcal{I} = \frac{1}{\sqrt{2\pi}}\int_{y=20}^{+\infty} \mathrm{e}^{-y^2/2}\mathrm{d}y \leq \frac{1}{\sqrt{2\pi}}\int_{y=20}^{+\infty} \frac{y}{20}\mathrm{e}^{-y^2/2}\mathrm{d}y = \frac{\mathrm{e}^{-20^2/2}}{20\sqrt{2\pi}} \simeq 2.8 \times 10^{-89}.$$

On the other hand, the Monte Carlo method consists in drawing iid realisations $X_1, \ldots, X_n$ of $\mathcal{N}(0,1)$ and approximate $\mathcal{I}$ with

$$\widehat{\mathcal{I}}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Exercise 4.2.1. *What is the law of the random variable* $N = \inf\{n \geq 1 : \widehat{\mathcal{I}}_n \neq 0\}$*? What is its expectation?*

Since $f(X_i) \sim \mathcal{B}(\mathcal{I})$, we have $\sigma^2 = \mathrm{Var}(f(X_1)) = \mathcal{I}(1-\mathcal{I}) \simeq \mathcal{I}$, so that the length of the Monte Carlo confidence interval writes

$$\ell_n \simeq 2\phi_{1-\alpha/2}\sqrt{\frac{\mathcal{I}}{n}}.$$

Assume that we want this length to be smaller than $\epsilon\mathcal{I}$, in order for the estimation of $\mathcal{I}$ to have a relative precision of ϵ. Then we need to take n such that

$$2\phi_{1-\alpha/2}\sqrt{\frac{\mathcal{I}}{n}} \leq \epsilon\mathcal{I},$$

that is to say

$$n \geq \left(\frac{2\phi_{1-\alpha/2}}{\epsilon}\right)^2 \frac{1}{\mathcal{I}}.$$

For $\epsilon = 0.01$ and $\alpha = 0.05$, using the analytic bound on $\mathcal{I}$ we obtain that n should be at least 5.6×10^{93}, which is impossible to realise in practice.

In this section, we present *variance reduction* techniques which allow to construct estimators of $\mathcal{I}$ with a smaller variance σ^2, and therefore yield smaller confidence intervals.

4.2.1 Control variate

In this subsection, we assume that in addition to $X_1, \ldots, X_n$, we are able to sample iid random variables $Y_1, \ldots, Y_n$ whose common expectation $\mathbb{E}[Y]$ is known analytically. Then, for all $\beta \in \mathbb{R}$,

$$\mathcal{I} = \mathbb{E}[f(X)] = \mathbb{E}[f(X) - \beta Y] + \beta\mathbb{E}[Y],$$

which suggests to approximate $\mathcal{I}$ by the estimator

$$\widehat{\mathcal{I}}_n^{\mathsf{CV},\beta} := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \beta Y_i) + \beta\mathbb{E}[Y].$$

The variance of this estimator is $(\sigma^{\mathsf{CV},\beta})^2/n$, where

$$(\sigma^{\mathsf{CV},\beta})^2 = \mathrm{Var}(f(X) - \beta Y) = \sigma^2 - 2\beta\,\mathrm{Cov}(f(X), Y) + \beta^2\,\mathrm{Var}(Y).$$

We may already remark that if $\mathrm{Cov}(f(X), Y) = 0$ then $(\sigma^{\mathsf{CV},\beta})^2$ is always larger than the variance σ^2 associated with the original Monte Carlo estimator: for the control variate method to be efficient, it is thus necessary that $f(X)$ and Y be correlated. The choice of β for which $(\sigma^{\mathsf{CV},\beta})^2$ is minimal is then

$$\beta^* = \frac{\mathrm{Cov}(f(X), Y)}{\mathrm{Var}(Y)},$$

which yields the variance

$$(\sigma^{\mathsf{CV},\beta^*})^2 = \sigma^2\left(1 - \rho^2\right),$$

where

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

is the correlation coefficient between $f(X)$ and Y. As a consequence, the more $f(X)$ and Y are correlated, the better the variance reduction. Typically, one may choose Y of the form $g(X)$, where the function g is close to f in regions where X has a high probability to take its values, while being 'simpler' than f, in the sense that $\mathbb{E}[g(X)]$ is easier to compute than $\mathbb{E}[f(X)]$ – see Exercise 4.2.3 for an illustration.

In practice, the optimal choice of β depends on the quantity $\mathrm{Cov}(f(X), Y)$ which may need to be estimated. Let us introduce

$$\widehat{C}_n = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \widehat{\mathcal{I}}_n)(Y_i - \overline{Y}_n).$$

The strong LLN shows that

$$\widehat{\beta}_n^* := \frac{\widehat{C}_n}{\mathrm{Var}(Y)}$$

converges to β^* almost surely, and Slutsky's Lemma 3.1.14 (ii) then yields the following result.

Proposition 4.2.2 (Control variate method)**.** *Let $(X_i, Y_i)_{1\leq i\leq n}$ be a sequence of iid pairs such that $f(X_i), Y_i \in \mathbf{L}^2(\mathbb{P})$. For all $n \geq 1$, let*

$$\widehat{\mathcal{I}}_n^{\mathsf{CV}} := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \widehat{\beta}_n^* Y_i) + \widehat{\beta}_n^*\mathbb{E}[Y],$$

with $\widehat{\beta}_n^*$ defined above. The interval

$$I_n^{\mathsf{CV}} = \left[\widehat{\mathcal{I}}_n^{\mathsf{CV}} - \phi_{1-\alpha/2}\sqrt{\frac{(\widehat{\sigma}_n^{\mathsf{CV}})^2}{n}}, \widehat{\mathcal{I}}_n^{\mathsf{CV}} + \phi_{1-\alpha/2}\sqrt{\frac{(\widehat{\sigma}_n^{\mathsf{CV}})^2}{n}}\right],$$

where

$$(\widehat{\sigma}_n^{\mathsf{CV}})^2 = \widehat{\sigma}_n^2\left(1 - \frac{\widehat{C}_n^2}{\widehat{\sigma}_n^2\,\mathrm{Var}(Y)}\right) \to \sigma^2(1 - \rho^2),$$

satisfies

$$\lim_{n\to+\infty}\mathbb{P}\left(\mathcal{I} \in I_n^{\mathsf{CV}}\right) = 1 - \alpha.$$

🏠 **Exercise 4.2.3.** *Let $X \sim \mathcal{N}(0,1)$. For all $t > 0$, we define*

$$f_t(x) = \frac{1}{1+tx^2},$$

and set

$$\mathcal{I} = \mathbb{E}\left[f_t(X)\right] = \frac{1}{\sqrt{2\pi}} \int_{x\in\mathbb{R}} \frac{\mathrm{e}^{-x^2/2}}{1+tx^2}\mathrm{d}x.$$

Let $X_1, \dots, X_n$ be independent $\mathcal{N}(0,1)$ variables, and let $Y_i = 1 - tX_i^2$.

1. *Compute $\mathbb{E}[Y_1]$.*
2. *Compare numerically the variances of the Monte Carlo estimator $\widehat{\mathcal{I}}_n$ and of the control variate estimator $\widehat{\mathcal{I}}_n^{\mathsf{CV}}$.*
3. *How does this comparison vary with t? What is your interpretation of this fact?*

</> This exercise is detailed in the notebook `VarianceReduction.ipynb`.

4.2.2 Importance sampling

Importance sampling is based on the remark that, for any probability measure Q on E such that $P \ll Q$,

$$\mathcal{I} = \int_{x\in E} f(x)\mathrm{d}P(x) = \int_{x\in E} f(x)w(x)\mathrm{d}Q(x),$$

where the function w is simply the density

$$w(x) = \frac{\mathrm{d}P}{\mathrm{d}Q}(x).$$

As a consequence, the quantity

$$\widehat{\mathcal{I}}_n^{\mathsf{IS}} := \frac{1}{n}\sum_{i=1}^n f(Y_i)w(Y_i),$$

where $Y_1, \dots, Y_n$ are iid with law Q, converges almost surely to $\mathcal{I}$. In fact, this construction may be applied with a more general class of probability measures Q, namely those for which one has

$$\mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}P(x) \ll \mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}Q(x), \tag{4.2}$$

and for which we still denote by w the associated density.

📄 **Exercise 4.2.4.** *Show that if $P \ll Q$, then Q satisfies (4.2), but that the converse does not hold true in general.*

The whole game of importance sampling then consists in choosing Q in order to make the asymptotic variance

$$(\sigma_Q^{\mathsf{IS}})^2 := \mathrm{Var}(f(Y)w(Y))$$

as small as possible.

Proposition 4.2.5 (Optimal choice of Q)**.** *Let $\overline{\mathcal{I}} = \mathbb{E}[|f(X)|]$, assume that this quantity is positive, and define the probability measure Q^* by*

$$\mathrm{d}Q^*(x) = \frac{|f(x)|}{\overline{\mathcal{I}}}\mathrm{d}P(x).$$

(i) *Q^* satisfies (4.2) and $(\sigma^{\mathsf{IS}}_{Q^*})^2 = \bar{\mathfrak{I}}^2 - \mathfrak{I}^2$.*
(ii) *For any probability measure Q which also satisfies (4.2), $(\sigma^{\mathsf{IS}}_{Q^*})^2 \leq (\sigma^{\mathsf{IS}}_{Q})^2$.*
(iii) *If f has constant sign P-almost everywhere, then $(\sigma^{\mathsf{IS}}_{Q^*})^2 = 0$.*

Proof. As a preliminary remark, we note that for any Q satisfying (4.2),

$$(\sigma^{\mathsf{IS}}_{Q})^2 = \mathbb{E}\left[(f(Y)w(Y))^2\right] - \mathfrak{I}^2, \qquad Y \sim Q. \tag{4.3}$$

First, it is easily checked that $\mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}P(x)$ has density

$$w^*(x) = \mathbb{1}_{\{f(x)\neq 0\}}\frac{\bar{\mathfrak{I}}}{|f(x)|}$$

with respect to $\mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}Q^*(x)$, therefore Q^* satisfies (4.2) and besides, if $Y^* \sim Q^*$, then

$$\begin{aligned}
\mathbb{E}\left[(f(Y^*)w^*(Y^*))^2\right] &= \int_{x\in E} \mathbb{1}_{\{f(x)\neq 0\}}|f(x)|^2\left(\frac{\bar{\mathfrak{I}}}{|f(x)|}\right)^2 \mathrm{d}Q^*(x)\\
&= \bar{\mathfrak{I}}^2\int_{x\in E} \mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}Q^*(x)\\
&= \bar{\mathfrak{I}}^2,
\end{aligned}$$

which, together with (4.3), proves (i). The point (iii) then immediately follows.

Second, let us fix Q which satisfies (4.2) and denote by w the associated density. By definition of $\bar{\mathfrak{I}}$ and w, and the Cauchy–Schwarz inequality,

$$\begin{aligned}
\bar{\mathfrak{I}}^2 &= \left(\int_{x\in E} |f(x)|\mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}P(x)\right)^2\\
&= \left(\int_{x\in E} |f(x)|w(x)\mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}Q(x)\right)^2\\
&\leq \int_{x\in E} |f(x)|^2 w(x)^2\mathbb{1}_{\{f(x)\neq 0\}}\mathrm{d}Q(x)\\
&= \mathbb{E}\left[(f(Y)w(Y))^2\right],
\end{aligned}$$

with $Y \sim Q$. Combined with (4.3), this estimate completes the proof of (ii). □

In practice it is impossible to implement the method with the optimal measure Q^* since the latter depends explicitly on the quantity $\bar{\mathfrak{I}}$, which is likely to be unknown — and, in the case where f is nonnegative P-almost everywhere, is exactly the quantity $\mathfrak{I}$ which we aim to estimate. Still, this lemma suggests that a ‘good’ choice of Q would be one which has a large mass under the measure $|f(x)|\mathrm{d}P(x)$.

Exercise 4.2.6. *For the example of the estimation of $\mathbb{P}(X \geq 20)$, for $X \sim \mathcal{N}(0,1)$, given in the introduction of this section, the optimal density is proportional to $\mathbb{1}_{\{y\geq 20\}}\mathrm{e}^{-y^2/2}$. We take $q(y)$ the density of the law $\mathcal{N}(20,1)$.*

1. *Compute the associated variance $(\sigma^{\mathsf{IS}})^2$.*
2. *What is the minimal number of samples to draw with this method in order to construct a confidence interval of level 0.95 which has a relative precision $\epsilon = 0.01$?*

We complete this subsection with a few remarks on the role of the condition (4.2) for the optimality property stated in Proposition 4.2.5. On the one hand, it is necessary to work with measures Q which satisfy the condition (4.2) rather than $P \ll Q$. Indeed, the optimal density Q^* does not necessarily satisfy the latter condition. On the other hand, when $\overline{\mathcal{I}} > |\mathcal{I}|$ (which implies that $f(X)$ changes sign), it is possible to find importance sampling estimators of the form

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_i)w(Y_i), \qquad Y_i \text{ iid according to } Q,$$

such that for $Y \sim Q$,

$$\mathbb{E}[f(Y)w(Y)] = \mathcal{I} \qquad \text{and} \qquad \mathrm{Var}\,(f(Y)w(Y)) < \overline{\mathcal{I}}^2 - \mathcal{I}^2.$$

A trivial example would be to assume that there exists $x_0 \in E$ such that $f(x_0) = \mathcal{I}$, and set $Q = \delta_{x_0}$, $w(x) = 1$ for any x. Of course, in this case, the measure Q does not satisfy (4.2).

Part II

Markov chains and MCMC methods

Chapter 5

The Markov property and ergodic theorems in discrete spaces

Contents

Let X be a random variable with values in some measurable space $(E, \mathcal{E})$ and let $f \in \mathbf{L}^1(P_X)$. The Monte Carlo method consists in using the Law of Large Numbers in order to approximate the integral

$$\mathcal{I} = \int_{x \in E} f(x) P_X(\mathrm{d}x) = \mathbb{E}[f(X)]$$

by the empirical mean of iid samples $f(X_1), \ldots, f(X_n)$.

The next few chapters are dedicated to the case where it is not possible, or at least too complicated, to sample iid realisations $X_1, X_2, \ldots$ of X. The theory of *Markov chains* provides an appropriate extension of the Law of Large Numbers (and the Central Limit Theorem) to sequences $X_1, X_2, \ldots$ that are neither independent nor identically distributed. This allows to implement the (Markov Chain) Monte Carlo method to evaluate the integral $\mathcal{I}$ in some cases where iid samples are not available.

Throughout the next three chapters, we consider random sequences defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking their values in a **countable** set E, endowed with the *discrete* σ-field

$\mathcal{E}$. For any measure μ on E and $x \in E$, we shall write $\mu(x)$ for $\mu(\{x\})$, so that for any $C \in \mathcal{E}$, $\mu(C) = \sum_{x \in C} \mu(x)$. The set of probability measures on $(E, \mathcal{E})$ is denoted by $\mathcal{P}(E)$.

5.1 The Markov property

5.1.1 Preliminary: conditional expectation in discrete spaces

We recall that for an event B such that $\mathbb{P}(B) > 0$, the conditional probability given B is defined, for any event A, by

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

We first extend this definition to the notion of conditional *expectation* given an event, by defining, for any random variable $X \in \mathbf{L}^1(\mathbb{P})$,

$$\mathbb{E}[X|B] = \frac{\mathbb{E}\left[X \mathbb{1}_B\right]}{\mathbb{P}(B)}.$$

The consistency of this definition with conditional probabilities lies in the remark that, for any event A,

$$\mathbb{P}(A|B) = \mathbb{E}\left[\mathbb{1}_A | B\right].$$

Now, given a random variable Z taking its values in some discrete space $(F, \mathcal{F})$, let us first set

$$F_Z := \{z \in F : \mathbb{P}(Z = z) > 0\}.$$

Then, for any $z \in F_Z$ one may define the quantity

$$\varphi(z) := \mathbb{E}[X|Z = z],$$

which is a deterministic function of z (and measurable, since we work with the discrete σ-field $\mathcal{F}$). The *conditional expectation of X given Z* is then the random variable

$$\mathbb{E}[X|Z] := \varphi(Z).$$

It is almost surely well-defined, since $\mathbb{P}(Z \in F_Z) = 1$.

The main properties of conditional expectations on which we shall rely are gathered in the next statement.

Lemma 5.1.1 (Properties of conditional expectations)**.** *Let $X \in \mathbf{L}^1(\mathbb{P})$ and $Z \in F$.*

(i) $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Z]]$.

(ii) For any measurable and bounded function $\psi : F \to \mathbb{R}$, $\mathbb{E}[\psi(Z)X|Z] = \psi(Z)\mathbb{E}[X|Z]$, almost surely.

Proof. By definition,

$$\mathbb{E}[X] = \sum_{z \in F_Z} \mathbb{E}[X \mathbb{1}_{\{Z=z\}}] = \sum_{z \in F_Z} \mathbb{E}[X|Z = z]\mathbb{P}(Z = z) = \mathbb{E}[\varphi(Z)],$$

which yields (i). As far as (ii) is concerned, we have, for any $z \in F_Z$,

$$\mathbb{E}[\psi(Z)X|Z = z] = \frac{\mathbb{E}\left[\psi(Z)X\mathbb{1}_{\{Z=z\}}\right]}{\mathbb{P}(Z = z)} = \psi(z)\frac{\mathbb{E}\left[X\mathbb{1}_{\{Z=z\}}\right]}{\mathbb{P}(Z = z)} = \psi(z)\varphi(z),$$

which yields the claimed statement. □

For any event A, we may next naturally define $\mathbb{P}(A|Z) := \mathbb{E}[\mathbb{1}_A|Z]$.

5.1.2 Stochastic matrices

Loosely speaking, a Markov chain is a sequence of random variables $(X_n)_{n\geq 0}$ with values in the set E, such that at each step $n \geq 0$, if $X_n = x$ then the next state for X_{n+1} is chosen randomly, according to some probability measure $P_{n+1}(x, \cdot)$ on E. This is translated in more formal terms in Definition 5.1.4 below. We first introduce the notion of *stochastic matrix*[1].

Definition 5.1.2 (Stochastic matrix)**.** *A* stochastic matrix *on E is a $E \times E$ matrix P with coefficients $(P(x, y))_{x,y\in E}$ which satisfy:*
(i) for all $x, y \in E$, $P(x, y) \geq 0$;
(ii) for all $x \in E$, $\sum_{y\in E} P(x, y) = 1$.

In other words, each row of the matrix P represents a probability measure $P(x, \cdot)$ on E. For this reason, we shall take the convention to identify measures on E with row vectors of $\mathbb{R}^E$, and dually, functions from E to $\mathbb{R}$ will be identified with column vectors of $\mathbb{R}^E$. These conventions then allow us to employ usual matrix/vector product notation: for example, if μ is a probability measure on E, X a random variable in E with distribution μ and $f : E \to \mathbb{R}$ is in $\mathbf{L}^1(\mu)$, then $\mathbb{E}[f(X)] = \sum_{x\in E} \mu(x)f(x)$ simply rewrites μf. We also denote by $\mathbf{1} \in \mathbb{R}^E$ the column vector of which all coordinates are equal to 1.

Exercise 5.1.3 (Properties of stochastic matrices)**.** *Let P be a stochastic matrix.*
1. *Show that $P\mathbf{1} = \mathbf{1}$.*
2. *Show that, for any $\mu \in \mathcal{P}(E)$, $\mu P \in \mathcal{P}(E)$.*
3. *Show that, for any stochastic matrix Q on E, PQ remains a stochastic matrix.*

5.1.3 Markov chains

We may now introduce the notion of *Markov chain*.

Definition 5.1.4 (Markov chain)**.** *Let $(P_n)_{n\geq 1}$ be a sequence of stochastic matrices. A sequence of random variables $(X_n)_{n\geq 0}$ in E is called a* Markov chain with sequence of transition matrices *$(P_n)_{n\geq 1}$ if, for all $n \geq 0$, for any $x_0, \dots, x_n \in E$ such that $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) > 0$, for all $x_{n+1} \in E$,*

$$\begin{aligned}\mathbb{P}(X_{n+1} = x_{n+1}|X_0 = x_0, \dots, X_n = x_n) &= \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n)\\ &= P_{n+1}(x_n, x_{n+1}).\end{aligned} \tag{5.1}$$

If all stochastic matrices P_n are equal to some stochastic matrix P, the chain is said to be homogeneous.

Equation (5.1) is called the *Markov property*. It expresses the fact that the law of the *future* value X_{n+1} only depends on the *past* trajectory $X_0, \dots, X_n$ through the *current* state X_n.

Remark 5.1.5. *With the notation introduced in Subsection 5.1.1, the Markov property rewrites*

$$\mathbb{P}(X_{n+1} = x_{n+1}|X_0, \dots, X_n) = \mathbb{P}(X_{n+1} = x_{n+1}|X_n) = P_{n+1}(X_n, x_{n+1}), \qquad \textit{almost surely.}$$

To prove that a random sequence is a Markov chain, it is often useful to write it under the form of a *random dynamical system*.

[1] In the case where E is countably infinite, we slightly abuse terminology and still call *matrix* an infinite array indexed by $E \times E$.

Proposition 5.1.6 (Random dynamical system)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables in E which satisfy the condition*

$$\forall n \geq 0, \qquad X_{n+1} = f_{n+1}(X_n, U_{n+1})$$

for some sequence of measurable functions $f_n : E \times \mathcal{U} \to E$, $n \geq 1$, and a sequence $(U_n)_{n\geq 1}$ which is iid in some measurable space $\mathcal{U}$, which is independent from X_0. Then $(X_n)_{n\geq 0}$ is a Markov chain with sequence of transition matrices defined by $P_n(x,y) = \mathbb{P}(f_n(x,U_1) = y)$. If f_n does not depend on n, then the chain is homogeneous.

Example 5.1.7 (Random walk on the discrete torus)**.** *Let $N \geq 1$ and $\mathbb{T}_N := \mathbb{Z}/N\mathbb{Z}$ be the associated* discrete torus *with size N. Given a parameter $p \in [0,1]$ and a sequence of iid random variables $(U_i)_{i\geq 1}$ such that $\mathbb{P}(U_1 = 1) = p$, $\mathbb{P}(U_1 = -1) = 1-p$, the random sequence defined by*

$$X_{n+1} = X_n + U_{n+1} \mod N$$

is called the random walk *in $\mathbb{T}_N$. If $p = 1/2$, this walk is* symmetric. *It is a homogeneous Markov chain, with transition matrix given by*

$$P(x,y) = \begin{cases} p & \text{if } y = x+1, \\ 1-p & \text{if } y = x-1, \\ 0 & \text{otherwise.} \end{cases}$$

Any homogeneous Markov chain can be represented by a directed graph, with set of vertices E and set of edges $\{(x,y) \in E \times E : P(x,y) > 0\}$. Each edge is endowed with the weight $P(x,y)$. The graph associated with the random walk on $\mathbb{T}_N$ is represented on Figure 5.1.

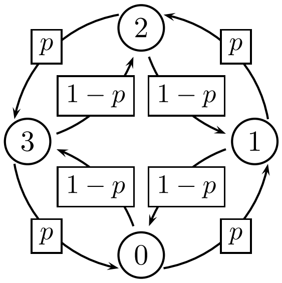


Figure 5.1: Graph associated with the random walk on the discrete torus $\mathbb{Z}/4\mathbb{Z}$.

Example 5.1.8 (The Ehrenfest urn)**.** *Consider a box divided into two compartments, called* A *and* B*, and which contains N particles, see Figure 5.2. At each step, one particle is chosen uniformly at random and moved to the other compartment. There are at least two ways to describe this dynamics.*

The microscopic description *consists in recording the compartment in which each particle is located, so that a configuration is a vector $x = (x^1, \dots, x^N) \in E_{\text{micro}} := \{\mathrm{A}, \mathrm{B}\}^N$. The transition matrix of the dynamics is given by*

$$P(x,y) = \begin{cases} \frac{1}{N} & \text{if } x \text{ and } y \text{ differ from exactly one coordinate,} \\ 0 & \text{otherwise.} \end{cases}$$

The macroscopic description *consists in recording merely the number of particles contained in the compartment* A*, so that the configuration space is $E_{\text{macro}} = \{0, \dots, N\}$, and the transition matrix is given by*

$$P(k,k+1) = \frac{N-k}{N}, \qquad P(k,k-1) = \frac{k}{N},$$

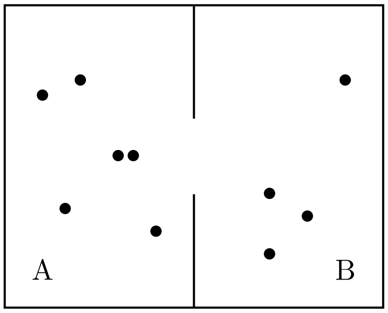


Figure 5.2: The Ehrenfest urn with $N = 10$ particles.

and the other coefficients are 0.

Exercise 5.1.9. *Let π be a probability measure on E.*

1. *Let $(X_n)_{n\geq 0}$ be a sequence of iid random variables with law π. Show that $(X_n)_{n\geq 0}$ is a homogeneous Markov chain and describe its transition matrix.*
2. *Let ξ be a random variable with law π, and let $(Y_n)_{n\geq 0}$ be the random sequence defined by $Y_n = \xi$ for all $n \geq 0$. Show that $(Y_n)_{n\geq 0}$ is a homogeneous Markov chain and describe its transition matrix.*
3. *What can you say about the law of X_n and Y_n, for any $n \geq 0$? And what about the law of the vectors $(X_0, \dots, X_n)$ and $(Y_0, \dots, Y_n)$?*

From Definition 5.1.4 we deduce the following properties related with the law of the sequence $(X_n)_{n\geq 0}$.

Proposition 5.1.10 (Marginal distributions of a Markov chain)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with sequence of transition matrices $(P_n)_{n\geq 1}$. For all $n \geq 0$, let $\mu_n \in \mathcal{P}(E)$ denote the law of the random variable X_n.*

(i) For all $n \geq 0$, for all $x_0, \dots, x_n \in E$,

$$\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \mu_0(x_0)P_1(x_0, x_1)\cdots P_n(x_{n-1}, x_n). \tag{5.2}$$

(ii) For all $n \geq 0$, $\mu_{n+1} = \mu_n P_{n+1}$.

Before detailing the proof of Proposition 5.1.10, we emphasise a few of its consequences.

Remark 5.1.11. *(i) The first assertion shows that the law of any vector $(X_0, \dots, X_n)$ is entirely characterised by two objects: the initial distribution μ_0 and the sequence of transition matrices $(P_n)_{n\geq 1}$.*

(ii) The second assertion immediately yields the identity $\mu_n = \mu_0 P_1 \cdots P_n$.

(iii) In particular, if the chain is homogeneous with transition matrix P, then for any $f \in \mathbf{L}^1(\mu_n)$, $\mathbb{E}[f(X_n)] = \mu_0 P^n f$.

Proof of Proposition 5.1.10. We prove the first assertion by induction on $n \geq 0$. For $n = 0$ this is immediate. Let $n \geq 0$ be such that (5.2) holds, and let $x_0, \dots, x_n, x_{n+1} \in E$. If $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = 0$, then on the one hand the fact that $\{X_0 = x_0, \dots, X_{n+1} = x_{n+1}\} \subset \{X_0 = x_0, \dots, X_n = x_n\}$ ensures that the former event has also probability 0, while on the other hand the identity (5.2) implies that $\mu_0(x_0)P_1(x_0, x_1)\cdots P_n(x_{n-1}, x_n) = 0$ and therefore this quantity remains 0 when multiplied by $P_{n+1}(x_n, x_{n+1})$. If $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) > 0$, we have

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, \ldots, X_{n+1} = x_{n+1})\\&= \mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n)\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n)\\&= \mu_0(x_0)P_1(x_0, x_1) \cdots P_{n+1}(x_n, x_{n+1}),\end{aligned}$$

where we have used (5.1) and (5.2) at the last line.

The second assertion follows from the computation

$$\mathbb{P}(X_{n+1} = y) = \sum_{x \in E} \mathbb{P}(X_{n+1} = y | X_n = x)\mathbb{P}(X_n = x) = \sum_{x \in E} P_{n+1}(x, y)\mu_n(x) = \mu_n P_{n+1}(y),$$

in which we have used (5.1). □

In practice we shall often compare homogeneous Markov chains with the same transition matrix P but different initial distributions. It will then be helpful to use the notation $\mathbb{P}_\mu, \mathbb{E}_\mu, \ldots$ to emphasise the fact that the initial distribution of the chain is μ. When this initial distribution is a Dirac distribution at some $x \in E$ (that is to say that $X_0 = x$ almost surely), we shall write $\mathbb{P}_x, \mathbb{E}_x, \ldots$ rather than $\mathbb{P}_{\delta_x}, \mathbb{E}_{\delta_x}, \ldots$. As an example, we may observe from Proposition 5.1.10 and Remark 5.1.11 that when $X_0 = x$, the law of X_n is related with the n-th power of P by the identity

$$\forall x, y \in E, \qquad \mathbb{P}_x(X_n = y) = P^n(x, y).$$

Proposition 5.1.12 (Homogeneous Markov property)**.** *Let $(X_n)_{n\geq 0}$ be a homogeneous Markov chain with transition matrix P. With the notation of Subsection 5.1.1, for any $\mu \in \mathcal{P}(E)$, for any $n, m \geq 0$ and for any bounded $G : E^{m+1} \to \mathbb{R}$,*

$$\mathbb{E}_\mu\left[G(X_n, X_{n+1}, \ldots, X_{n+m}) | X_0, \ldots, X_n\right] = \mathbb{E}_{X_n}\left[G(X_0, \ldots, X_m)\right], \qquad \mathbb{P}_\mu\text{-almost surely.}$$

The statement of Proposition 5.1.12 must be understood as the fact that, conditionally on the trajectory $(X_0, \ldots, X_n)$, the sequence $(X_n, X_{n+1}, \ldots)$ is a Markov chain with starting point X_n and transition matrix P.

Proof of Proposition 5.1.12. Let $x_0, \ldots, x_n \in E$ such that $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0$. We have

$$\begin{aligned}&\mathbb{E}_\mu\left[G(X_n, X_{n+1}, \ldots, X_{n+m}) | X_0 = x_0, \ldots, X_n = x_n\right]\\&= \frac{\mathbb{E}_\mu\left[G(X_n, X_{n+1}, \ldots, X_{n+m})\mathbb{1}_{\{X_0 = x_0, \ldots, X_n = x_n\}}\right]}{\mathbb{P}_\mu(X_0 = x_0, \ldots, X_n = x_n)}.\end{aligned}$$

Let us focus on the numerator, which rewrites

$$\begin{aligned}&\mathbb{E}_\mu\left[G(X_n, X_{n+1}, \ldots, X_{n+m})\mathbb{1}_{\{X_0 = x_0, \ldots, X_n = x_n\}}\right]\\&= \sum_{x_{n+1}, \ldots, x_{n+m} \in E} G(x_n, x_{n+1}, \ldots, x_{n+m})\mathbb{P}_\mu(X_0 = x_0, \ldots, X_{n+m} = x_{n+m})\\&= \sum_{x_{n+1}, \ldots, x_{n+m} \in E} G(x_n, x_{n+1}, \ldots, x_{n+m})\mu(x_0)P(x_0, x_1) \cdots P(x_{n+m-1}, x_{n+m}).\end{aligned}$$

Now, the product in the right-hand side satisfies

$$\begin{aligned}&\mu(x_0)P(x_0, x_1) \cdots P(x_{n+m-1}, x_{n+m})\\&= \mathbb{P}_\mu(X_0 = x_0, \ldots, X_n = x_n)P(x_n, x_{n+1}) \cdots P(x_{n+m-1}, x_{n+m})\\&= \mathbb{P}_\mu(X_0 = x_0, \ldots, X_n = x_n)\mathbb{P}_{x_n}(X_0 = x_n, \ldots, X_m = x_{n+m}),\end{aligned}$$

so that

$$\begin{aligned}
&\frac{\mathbb{E}_\mu\left[G(X_n, X_{n+1}, \ldots, X_{n+m})\mathbb{1}_{\{X_0=x_0,\ldots,X_n=x_n\}}\right]}{\mathbb{P}_\mu(X_0 = x_0, \ldots, X_n = x_n)}\\
&= \sum_{x_{n+1},\ldots,x_{n+m}\in E} G(x_n, x_{n+1}, \ldots, x_{n+m})\mathbb{P}_{x_n}(X_0 = x_n, \ldots, X_m = x_{n+m})\\
&= \mathbb{E}_{x_n}\left[G(X_0, \ldots, X_m)\right].
\end{aligned}$$

This completes the proof. □

5.2 Stationary distribution

From now on, we only consider homogeneous Markov chains, and omit the precision when referring to ‘Markov chains’. The first step to establish a connection between Markov chains and the Monte Carlo method is the notion of *stationary distribution*.

5.2.1 Definition

Definition 5.2.1 (Stationary distribution)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain in E with transition matrix P. A probability measure π on E is called a* stationary distribution *for $(X_n)_{n\geq 0}$ if it satisfies*

$$\pi P = \pi.$$

The denomination ‘stationary’ comes from the following result.

Proposition 5.2.2 (Stationary distribution)**.** *Let π be a stationary distribution for $(X_n)_{n\geq 0}$. For any $n \geq 0$,*

$$\forall x \in E, \qquad \mathbb{P}_\pi(X_n = x) = \pi(x);$$

in other words, if $X_0 \sim \pi$ then $X_n \sim \pi$ for all $n \geq 0$.

Proof. It is a straightforward consequence of the second assertion of Remark 5.1.11. □

🗎 **Exercise 5.2.3.** *Show that if π is a stationary distribution for $(X_n)_{n\geq 0}$, then the whole sequence is actually stationary in the sense that for any $k \geq 0$ and $n \geq 0$, the vectors $(X_0, \ldots, X_n)$ and $(X_k, \ldots, X_{k+n})$ have the same distribution under $\mathbb{P}_\pi$.*

🗎 **Exercise 5.2.4** (Random walk on the torus)**.** *Show that, whatever the value of p, the uniform measure on $\mathbb{T}_N$ is stationary for the random walk on $\mathbb{T}_N$ introduced in Example 5.1.7.*

🏠 **Exercise 5.2.5.** *Consider the Ehrenfest urn from Example 5.1.8.*

1. *Show that the uniform distribution on E_{micro} is stationary for the microscopic description.*
2. *If $X = (X^1, \ldots, X^N)$ is a random vector uniformly distributed in E_{micro}, what is the law of the corresponding macroscopic configuration $K = \sum_{i=1}^N \mathbb{1}_{\{X^i=\mathrm{A}\}}$?*
3. *Show that the law of K is stationary for the macroscopic description.*

In the sequel of this section, we study the existence and uniqueness of stationary distributions. We restrict ourselves to the case of **finite** state spaces and defer the case of countably infinite state spaces to Section 5.4.

5.2.2 Existence

Proposition 5.2.6 (Existence of stationary distribution)**.** *If E is finite, then every Markov chain in E admits at least one stationary distribution.*

We provide two different proofs of Proposition 5.2.6. Both rely on the observation that the set $\mathcal{P}(E)$ can be identified with the finite-dimensional *simplex* $\{\mu \in [0,1]^E : \sum_{x \in E} \mu(x) = 1\}$, and is therefore convex and compact.

Proof by Brouwer's Fixed Point Theorem. By Exercise 5.1.3, the mapping $\mu \mapsto \mu P$ is continuous from $\mathcal{P}(E)$ to $\mathcal{P}(E)$. As a consequence, Brouwer's Fixed Point Theorem ensures that it admits a fixed point in $\mathcal{P}(E)$. □

Elementary proof. Let $\mu \in \mathcal{P}(E)$. For all $n \geq 1$, set

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=0}^{n-1} \mu P^i,$$

so that for all $f : E \to \mathbb{R}$,

$$\widehat{\mu}_n f = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}_\mu[f(X_i)].$$

By Exercise 5.1.3 and convexity, $\widehat{\mu}_n \in \mathcal{P}(E)$ and by compactness, there exists an increasing sequence $(n_\ell)_{\ell \geq 1}$ such that $\widehat{\mu}_{n_\ell}$ converges to some $\pi \in \mathcal{P}(E)$ when $\ell \to +\infty$. Since, for all $\ell \geq 1$,

$$\widehat{\mu}_{n_\ell} P = \frac{1}{n_\ell} \sum_{i=0}^{n_\ell - 1} \mu P^{i+1} = \widehat{\mu}_{n_\ell} + \frac{1}{n_\ell} \left(\mu P^{n_\ell} - \mu \right),$$

we deduce using the boundedness of $\mu P^{n_\ell} - \mu$ that $\pi P = \pi$, which is the expected result. □

5.2.3 Irreducibility and uniqueness

What may prevent a stationary distribution from being unique? Let E_1, E_2 be two disjoint subsets of the finite space E, P_1 and P_2 be stochastic matrices respectively defined on E_1 and E_2, and let π_1, π_2 be some associated stationary distributions. On the space $E' = E_1 \cup E_2$, define the stochastic matrix P' by the block decomposition

$$P' = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix};$$

similarly, define the probability measures

$$\pi_1' = (\pi_1 \quad 0), \qquad \pi_2' = (0 \quad \pi_2).$$

Then it is clear that both π_1' and π_2' (and as a consequence, all their convex combinations) are stationary distributions for P'. Observe that in this situation, if the chain starts at some point $x \in E_1$, then X_n will remain in E_1 for all $n \geq 1$, see also Figure 5.3.

This remark motivates the following definition.

Definition 5.2.7 (Irreducibility)**.** *A Markov chain with transition matrix P on E is called* irreducible *if, for all $x, y \in E$, there exist $n \geq 1$ and $x = x_0, x_1, \ldots, x_n = y$ such that*

$$P(x_0, x_1) \cdots P(x_{n-1}, x_n) > 0.$$

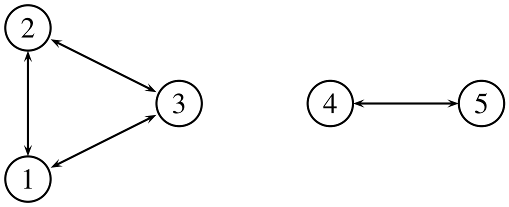


Figure 5.3: On the space $E = \{1, 2, 3, 4, 5\}$, arrows represent the possible moves of the Markov chain. Clearly, a chain started in $E_1 = \{1, 2, 3\}$ can never go to $E_2 = \{4, 5\}$.

We shall also say that the matrix P is irreducible.

Exercise 5.2.8. *Check that the condition that there exist $x = x_0, x_1, \dots, x_n = y$ such that $P(x_0, x_1) \cdots P(x_{n-1}, x_n) > 0$ is equivalent to $\mathbb{P}_x(X_n = y) > 0$.*

The main result of this section is the following statement.

Proposition 5.2.9 (Uniqueness of a stationary distribution)**.** *If the space E is finite and the stochastic matrix P is irreducible, then it possesses a unique stationary distribution.*

Existence was shown in Propsoition 5.2.6 so we focus on uniqueness. We start the proof with the following exercise.

Exercise 5.2.10. *Let P be an irreducible stochastic matrix, and let π be an associated stationary distribution. Show that for all $x \in E$, $\pi(x) > 0$.*

We now introduce a useful object.

Definition 5.2.11 (Dirichlet form)**.** *Let P be a stochastic matrix, and let π be an associated stationary distribution. The* Dirichlet form *of (P, π) is the quadratic form $\mathcal{E}_\pi$ defined on $\mathbb{R}^E$ by*

$$\mathcal{E}_\pi(f) = \frac{1}{2}\mathbb{E}_\pi\left[(f(X_1) - f(X_0))^2\right] = \frac{1}{2}\sum_{x,y \in E}(f(y) - f(x))^2\pi(x)P(x, y).$$

If E is countably infinite, then $\mathcal{E}_\pi(f)$ is well-defined in $[0, +\infty]$. For the sequel of the proof to make sense, it is however more convenient to restrict ourselves to the case where E is finite, so as not to discuss the convergence of sums over E.

Lemma 5.2.12 (Another expression for $\mathcal{E}_\pi$)**.** *If E is finite, then for all $f \in \mathbb{R}^E$,*

$$\mathcal{E}_\pi(f) = -\sum_{x \in E} f(x)(P - I)f(x)\pi(x).$$

Proof. From Definition 5.2.11, we write

$$\begin{aligned}\mathcal{E}_\pi(f) &= \frac{1}{2}\sum_{x,y \in E}(f(y)^2 - 2f(y)f(x) + f(x)^2)\pi(x)P(x, y) \\ &= \frac{1}{2}\sum_{y \in E} f(y)^2\pi P(y) - \sum_{x \in E} f(x)\pi(x)Pf(x) + \frac{1}{2}\sum_{x \in E} f(x)^2\pi(x),\end{aligned}$$

where we have used the fact that $\sum_{y \in E} P(x, y) = 1$ at the last line. Since π is stationary, we may furthermore write $\pi P = \pi$ in the first term, so that

$$\mathcal{E}_\pi(f) = \sum_{x \in E} f(x)^2\pi(x) - \sum_{x \in E} f(x)\pi(x)Pf(x) = -\sum_{x \in E} f(x)(P - I)f(x)\pi(x),$$

which is the claimed expression. □

Lemma 5.2.13 (Kernel of $P - I$ for irreducible matrices)**.** *Assume that E is finite and let P be an irreducible stochastic matrix. For all $f \in \mathbb{R}^E$, if $Pf = f$ then there exists $c \in \mathbb{R}$ such that $f = c\mathbf{1}$.*

Proof. Let π be a stationary distribution for P. If $Pf = f$, then Lemma 5.2.12 immediately shows that $\mathcal{E}_\pi(f) = 0$, therefore by Definition 5.2.11 and Exercise 5.2.10, $f(x) = f(y)$ for all pairs (x, y) such that $P(x, y) > 0$. We now take arbitrary $x, y \in E$ and let $x = x_0, x_1, \ldots, x_n = y$ be given by Definition 5.2.7. From this definition, $P(x_i, x_{i+1}) > 0$ for all $i = 0, \ldots, n-1$, so that by the argument above, $f(x_0) = \cdots = f(x_n)$ and thus f is a constant function on E. □

We are now ready to complete the proof of Proposition 5.2.9.

Proof of Proposition 5.2.9. By Lemma 5.2.13 and the Rank-Nullity Theorem, 1 is a simple eigenvalue for both left- and right-multiplication, and any stationary distribution for P is in the kernel (for the left multiplication) of $P - I$. So any two stationary distributions are necessarily collinear, and since both are probability measures, they must coincide. □

🏠 **Exercise 5.2.14** (The coupon collector)**.** *A brand of chocolate eggs hides surprise gifts in each egg. There are N different models of gifts, each of which is equally likely to be hidden in a given egg. We denote by $X_n \in \{0, \ldots, N\}$ the number of different gifts that you have collected after eating n eggs, and $\tau_N = \inf\{n \geq 0 : X_n = N\}$ the time at which you have found all eggs.*

1. *Show that $(X_n)_{n \geq 0}$ is a Markov chain and write its transition matrix.*
2. *Is this chain irreducible?*
3. *Describe the set of its stationary distributions.*
4. *Compute $\mathbb{E}_0[\tau_N]$ and give an equivalent of this quantity when $N \to +\infty$.* Hint: define $\eta_0 = 0$ and, for $i \in \{1, \ldots, N\}$, $\eta_i = \inf\{n \geq 1 : X_{\eta_{i-1}+n} = i\}$. How to express τ_N in terms of $\eta_1, \ldots, \eta_N$? What is the law of each η_i?
5. *Show that, for any $c > 0$, $\mathbb{P}(\tau_N > \lceil N \ln N + cN \rceil) \leq \mathrm{e}^{-c}$.* Hint: for $i \in \{1, \ldots, N\}$ and $k \geq 1$, introduce the event $A_i^k = \{$no gift of the i-th type has been found in the first k eggs$\}$.

5.3 Ergodic theorems in finite state spaces

Let $(X_n)_{n \geq 0}$ be a homogeneous Markov chain, with transition matrix P. In this section, we assume that X_n takes its values in some **finite** state space E.

5.3.1 Return time

For any $x \in E$, let us define the random variable

$$\tau_x = \inf\{n \geq 1 : X_n = x\}, \tag{5.3}$$

which may take the value $+\infty$ and is called the *return time* to x.

Lemma 5.3.1 (Integrability of return times in finite state spaces)**.** *If the chain $(X_n)_{n \geq 0}$ is irreducible and the state space E is finite, then for any $\mu \in \mathcal{P}(E)$ and $x \in E$, $\mathbb{E}_\mu[\tau_x] < +\infty$.*

Proof. Let $x \in E$. For all $x' \in E$, Definition 5.2.7 implies that there exists $n_{x'} \geq 1$ such that $P^{n_{x'}}(x', x) > 0$. Let

$$\kappa := \min_{x' \in E} P^{n_{x'}}(x', x) > 0, \qquad m := \max_{x' \in E} n_{x'} < +\infty,$$

so that whatever the initial state x', the probability for the chain to return to x before the time m is at least κ. Indeed,

$$\kappa \le P^{n_{x'}}(x',x) = \mathbb{P}_{x'}(X_{n_{x'}} = x) \le \mathbb{P}_{x'}(\tau_x \le n_{x'}) \le \mathbb{P}_{x'}(\tau_x \le m).$$

Hence, for any $\ell \ge 1$,

$$\begin{aligned}\mathbb{P}_\mu(\tau_x > \ell m) &= \mathbb{P}_\mu\left(X_1 \neq x, \dots, X_{\ell m} \neq x\right)\\ &= \mathbb{E}_\mu\left[\mathbb{1}_{\{X_1\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\mathbb{1}_{\{X_{(\ell-1)m+1}\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\right].\end{aligned}$$

Using Lemma 5.1.1 and then Proposition 5.1.12, we get

$$\begin{aligned}&\mathbb{E}_\mu\left[\mathbb{1}_{\{X_1\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\mathbb{1}_{\{X_{(\ell-1)m+1}\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\right]\\ &= \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathbb{1}_{\{X_1\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\mathbb{1}_{\{X_{(\ell-1)m+1}\neq x,\cdots,X_{(\ell-1)m}\neq x\}}|X_0,\dots,X_{(\ell-1)m}\right]\right]\\ &= \mathbb{E}_\mu\left[\mathbb{1}_{\{X_1\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\mathbb{E}_\mu\left[\mathbb{1}_{\{X_{(\ell-1)m+1}\neq x,\cdots,X_{(\ell-1)m}\neq x\}}|X_0,\dots,X_{(\ell-1)m}\right]\right]\\ &= \mathbb{E}_\mu\left[\mathbb{1}_{\{X_1\neq x,\cdots,X_{(\ell-1)m}\neq x\}}\mathbb{P}_{X_{(\ell-1)m}}\left(X_1 \neq x,\cdots,X_m\neq x\right)\right].\end{aligned}$$

But since, almost surely,

$$\mathbb{P}_{X_{(\ell-1)m}}\left(X_1 \neq x,\cdots,X_m\neq x\right) = \mathbb{P}_{X_{(\ell-1)m}}\left(\tau_x > m\right) \le 1-\kappa,$$

we deduce that

$$\mathbb{P}_\mu(\tau_x > \ell m) \le (1-\kappa)\mathbb{P}_\mu(\tau > (\ell-1)m),$$

and thus

$$\mathbb{P}_\mu(\tau_x > \ell m) \le (1-\kappa)^\ell.$$

We complete the proof by remarking that, by the Fubini–Tonelli Theorem,

$$\begin{aligned}\mathbb{E}_\mu[\tau_x] &= \mathbb{E}_\mu\left[\sum_{n=0}^{+\infty}\mathbb{1}_{\{n<\tau_x\}}\right]\\ &= \sum_{n=0}^{+\infty}\mathbb{P}_\mu\left(\tau_x > n\right)\\ &= \sum_{\ell=0}^{+\infty}\sum_{k=0}^{m-1}\mathbb{P}_\mu\left(\tau_x > \ell m + k\right)\\ &\le \sum_{\ell=0}^{+\infty} m\mathbb{P}_\mu\left(\tau_x > \ell m\right)\\ &\le \sum_{\ell=0}^{+\infty} m(1-\kappa)^\ell\\ &= \frac{m}{\kappa}.\end{aligned}$$

□

🏠 **Exercise 5.3.2** (Exponential moments). *Under the assumptions of Lemma 5.3.1, show that there exists $\epsilon > 0$ such that $\mathbb{E}_\mu[\exp(\epsilon\tau_x)] < +\infty$. Deduce that for all $p \ge 1$, $\mathbb{E}_\mu[(\tau_x)^p] < +\infty$.*

5.3.2 Law of Large Numbers

From now on we assume that the chain $(X_n)_{n\geq 0}$ is irreducible. By Proposition 5.2.9, it admits a unique stationary distribution π.

Theorem 5.3.3 (Law of Large Numbers for Markov chains in finite state spaces)**.** *If the chain $(X_n)_{n\geq 0}$ is irreducible and the space E is finite, then for any $f \in \mathbb{R}^E$,*

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(X_i) = \pi f, \qquad \textit{almost surely.}$$

Furthermore, the stationary distribution π of $(X_n)_{n\geq 0}$ satisfies the identity

$$\forall x \in E, \qquad \pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]},$$

where we recall the definition of the return time τ_x in (5.3).

Theorem 5.3.3 obviously generalises the usual strong LLN to Markov chains, and emphasises the key role played by the stationary distribution in this perspective. We insist on the technical point that in the statement of this theorem, we do not make explicit the initial distribution of the chain: it is to be understood that for any initial measure $\mu_0 \in \mathcal{P}(E)$, the convergence holds almost surely.

To prove Theorem 5.3.3, we shall show the next two facts:

$$\forall x \in E, \qquad \lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=x\}} = \frac{1}{\mathbb{E}_x[\tau_x]}, \tag{5.4}$$

and

$$\text{the measure } \widetilde{\pi}(x) := \frac{1}{\mathbb{E}_x[\tau_x]} \text{ is stationary.} \tag{5.5}$$

To proceed we first define

$$\tau_x^0 := \inf\{n \geq 0 : X_n = x\}, \qquad \tau_x^{\ell+1} := \inf\{n \geq \tau_x^\ell + 1 : X_n = x\}.$$

Lemma 5.3.4 below shows that $\tau_x^\ell < +\infty$, almost surely, for any $\ell \geq 0$, so the sequence $(\tau_x^\ell)_{\ell\geq 0}$ is well-defined, almost surely.

Lemma 5.3.4 (On the sequence $(\tau_x^\ell)_{\ell\geq 0}$)**.** *Under the assumptions of Theorem 5.3.3, we have $\tau_x^0 < +\infty$, almost surely, and the sequence $(\tau_x^{\ell+1} - \tau_x^\ell)_{\ell\geq 0}$ is iid, with law the distribution of τ_x under $\mathbb{P}_x$.*

Proof. We first notice that

$$\tau_x^0 = \begin{cases} 0 & \text{if } X_0 = x, \\ \tau_x & \text{otherwise,} \end{cases}$$

so $\tau_x^0 \leq \tau_x$ and by Lemma 5.3.1, $\tau_x^0 < +\infty$, almost surely. Now, for any $n, m \geq 1$, by Proposition 5.1.12,

$$\begin{aligned}
\mathbb{P}(\tau_x^1 - \tau_x^0 = n, \tau_x^2 - \tau_x^1 = m) &= \sum_{p=0}^{+\infty} \mathbb{P}(\tau_x^0 = p, \tau_x^1 - \tau_x^0 = n, \tau_x^2 - \tau_x^1 = m) \\
&= \sum_{p=0}^{+\infty} \mathbb{P}\left(\begin{array}{l} X_0 \neq x, \ldots, X_{p-1} \neq x, X_p = x, \\ X_{p+1} \neq x, \ldots, X_{p+n-1} \neq x, X_{p+n} = x, \\ X_{p+n+1} \neq x, \ldots, X_{p+n+m-1} \neq x, X_{p+n+m} = x \end{array} \right) \\
&= \sum_{p=0}^{+\infty} \mathbb{P}(\tau_x^0 = p)\mathbb{P}_x(\tau_x = n)\mathbb{P}_x(\tau_x = m) \\
&= \mathbb{P}_x(\tau_x = n)\mathbb{P}_x(\tau_x = m),
\end{aligned}$$

where we have used the fact that $\mathbb{P}(\tau_x^0 < +\infty) = 1$ twice. This proves that $\tau_x^1 - \tau_x^0$ and $\tau_x^2 - \tau_x^1$ are independent with law $\mathbb{P}_x(\tau_x = \cdot)$, and the computation easily generalises to an arbitrary number of variables $\tau_x^{\ell+1} - \tau_x^{\ell}$. □

We are now ready to complete the proof of Theorem 5.3.3.

Proof of Theorem 5.3.3. We first prove (5.4). For $n \geq \tau_x^0$, let $L_n \geq 0$ be such that

$$\tau_x^{L_n} \leq n < \tau_x^{L_n+1}.$$

Since the sequence of integers $(\tau_x^{\ell})_{\ell \geq 0}$ is increasing, L_n is well-defined and $L_n \to +\infty$ when $n \to +\infty$. Furthermore,

$$\sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} = L_n + 1.$$

We deduce that for n large enough,

$$\frac{L_n + 1}{\tau_x^{L_n+1}} < \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} \leq \frac{L_n + 1}{\tau_x^{L_n}}.$$

By the strong Law of Large Numbers and Lemma 5.3.4, we have

$$\frac{\tau_x^L}{L} = \frac{\tau_x^0}{L} + \frac{1}{L} \sum_{\ell=0}^{L-1} (\tau_x^{\ell+1} - \tau_x^{\ell}) \to \mathbb{E}_x[\tau_x], \qquad \text{almost surely,}$$

when $L \to +\infty$. As a consequence, almost surely, both bounds in the inequality above converge to the same limit $\widetilde{\pi}(x)$. This proves (5.4).

To prove (5.5), we deduce from (5.4) and the Dominated Convergence Theorem applied to the bounded random variable that

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}}\right] = \frac{1}{n} \sum_{i=0}^{n-1} \mu_0 P^i(x) \to \widetilde{\pi}(x)$$

when $n \to +\infty$. But by the proof of Proposition 5.2.6, any limit of $\frac{1}{n}\sum_{i=0}^{n-1} \mu_0 P^i$ must be a stationary distribution. Therefore, by the uniqueness result of Proposition 5.2.9, $\widetilde{\pi} = \pi$.

To complete the proof of Theorem 5.3.3 it remains to observe that thanks to (5.4) and (5.5), for any $f \in \mathbb{R}^E$,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) = \sum_{x\in E} f(x)\left(\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=x\}}\right) \to \sum_{x\in E} f(x)\pi(x), \qquad \text{almost surely,}$$

which is elementary since E is assumed to be finite. □

Remark 5.3.5 (Representation formula for π)**.** *For $f \in \mathbb{R}^E$, Theorem 5.3.3 implies that, with the notation above,*

$$\lim_{L\to+\infty} \frac{1}{\tau_x^L}\sum_{i=0}^{\tau_x^L-1} f(X_i) = \pi f, \qquad \textit{almost surely.}$$

But with the same arguments as in the proof of Theorem 5.3.3, one may write

$$\frac{1}{L}\sum_{i=0}^{\tau_x^L-1} f(X_i) = \frac{1}{L}\sum_{i=0}^{\tau_x^0-1} f(X_i) + \frac{1}{L}\sum_{\ell=0}^{L-1}\sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1} f(X_i),$$

and

$$\frac{\tau_x^L}{L} = \frac{\tau_x^0}{L} + \frac{1}{L}\sum_{\ell=0}^{L-1}(\tau_x^{\ell+1} - \tau_x^\ell).$$

The sequence of random variables $(\sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1} f(X_i))_{\ell\geq 0}$ is iid, with law the distribution of $\sum_{i=0}^{\tau_x-1} f(X_i)$ under $\mathbb{P}_x$. Therefore, applying the strong Law of Large Numbers twice, we get the identity

$$\pi f = \frac{\mathbb{E}_x\left[\sum_{i=0}^{\tau_x-1} f(X_i)\right]}{\mathbb{E}_x[\tau_x]},$$

which rewrites

$$\forall x \in E, \qquad \mathbb{E}_x\left[\sum_{i=0}^{\tau_x-1} f(X_i)\right] = \frac{\pi f}{\pi(x)}.$$

In particular, if $f(\cdot) = \mathbb{1}_{\{\cdot=y\}}$ for some $y \in E$, we get

$$\mathbb{E}_x\left[\sum_{i=0}^{\tau_x-1} \mathbb{1}_{\{X_i=y\}}\right] = \frac{\pi(y)}{\pi(x)},$$

so that the ratio $\pi(y)/\pi(x)$ may be seen as the average number of returns in y by the chain between two consecutive returns in x.

5.3.3 Central Limit Theorem

The proof of the Markov Chain LLN relies on the application of the usual LLN to the decomposition of the empirical mean $\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)$ into iid *excursions* outside x. Therefore it is natural to expect the application of the Central Limit Theorem to these excursions to lead to a Central Limit Theorem for the the empirical mean of the Markov chain.

For simplicity, we assume that $X_0 = x$ and we consider the empirical mean at times $n = \tau_x^L$, so that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) - \pi f\right) = \frac{1}{\sqrt{\tau_x^L}}\sum_{i=0}^{\tau_x^L-1}(f(X_i)-\pi f)$$
$$= \sqrt{\frac{L}{\tau_x^L}}\frac{1}{\sqrt{L}}\sum_{\ell=0}^{L-1} Z_x^\ell,$$

where the random variables $Z_x^\ell = \sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1}(f(X_i)-\pi f)$ are iid, with law the distribution of $Z_x := \sum_{i=0}^{\tau_x-1}(f(X_i)-\pi f)$ under $\mathbb{P}_x$. In particular, by Remark 5.3.5, the variable Z_x^ℓ are centered. We denote by $v(x) := \mathbb{E}_x[Z_x^2]$ their variance. Since $\tau_x^L/L \to 1/\pi(x)$, almost surely, we deduce from Slutsky's Lemma that

$$\lim_{L\to+\infty}\frac{1}{\sqrt{\tau_x^L}}\sum_{i=0}^{\tau_x^L-1}(f(X_i)-\pi f) = \mathcal{N}(0,\pi(x)v(x)), \qquad \text{in distribution.}$$

It turns out that the quantity $\sigma^2(f) := \pi(x)v(x) \geq 0$ does not depend on x, and that the convergence actually holds without the restriction to the (random) subsequence $n = \tau_x^L$.

Theorem 5.3.6 (Markov chain Central Limit Theorem)**.** *Let the assumptions of Theorem 5.3.3 hold. For any $f \in \mathbb{R}^E$,*

$$\lim_{n\to+\infty}\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) - \pi f\right) = \mathcal{N}(0,\sigma^2(f)), \qquad \textit{in distribution.}$$

We refer to Theorems 17.2.2, 17.4.4 and 17.5.3 in the book *Markov Chains and Stochastic Stability* by Meyn and Tweedie for a proof. An alternative expression for $\sigma^2(f)$ will be given in Chapter 6.

5.4 Ergodic theorems in countably infinite state spaces

In this section, we discuss the extension of the results of Sections 5.2 and 5.3 to the case where E is *countably infinite*. The first difficulty lies in the existence of stationary distributions, which may fail.

5.4.1 Recurrence and transience

When E is infinite, the compactness arguments used in both proofs of Proposition 5.2.6 no longer holds, and in fact, there are natural examples of Markov chains which do not admit a stationary distribution.

Example 5.4.1 (Simple random walk on $\mathbb{Z}^d$)**.** *The* simple random walk on $\mathbb{Z}^d$ *is the random sequence $(X_n)_{n\geq 0}$ which at each step picks up its next state uniformly among its neighbours. More precisely, it is the Markov chain in $\mathbb{Z}^d$ with transition matrix*

$$P(x,y) = \begin{cases} \frac{1}{2d} & \textit{if } |x-y| = 1, \\ 0 & \textit{otherwise,} \end{cases}$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$.

Exercise 5.4.2. *Show that a stationary distribution π for the simple random walk on $\mathbb{Z}$ necessarily satisfies $\pi(x+1) = \pi(x)$ for all $x \in \mathbb{Z}$ and therefore cannot exist. Generalise the argument to $\mathbb{Z}^d$ for any $d \geq 1$.*

In this context, the existence of stationary distributions is related with the tail of the return times τ_x defined in (5.3).

Definition 5.4.3 (Recurrent and transient states)**.** *For a given stochastic matrix P on E, a state $x \in E$ is called:*

- transient *if $\mathbb{P}_x(\tau_x = +\infty) > 0$;*
- recurrent *if $\mathbb{P}_x(\tau_x = +\infty) = 0$.*

Furthermore, recurrent states are called:

- null *if $\mathbb{E}_x[\tau_x] = +\infty$;*
- positive *if $\mathbb{E}_x[\tau_x] < +\infty$.*

In the sequel, we refer to the fact of being transient, null recurrent or positive recurrent as the *nature* of a state.

Exercise 5.4.4 (Characterisation of recurrence and transience)**.** *For any $x \in E$, set*

$$N_x = \sum_{n=0}^{+\infty} \mathbb{1}_{\{X_n = x\}}.$$

1. *If x is transient, show that N_x has a geometric distribution, with parameter $\mathbb{P}_x(\tau_x = +\infty)$.*
2. *If x is recurrent, show that $N_x = +\infty$, $\mathbb{P}_x$-almost surely.*
3. *Deduce that x is transient (resp. recurrent) if and only if $\sum_{n=0}^{+\infty} P^n(x,x) < +\infty$ (resp $\sum_{n=0}^{+\infty} P^n(x,x) = +\infty$).*

5.4.2 Stationary distribution and LLN

The first result of this subsection is the following generalisation of (5.4), which does not require irreducibility to hold.

Lemma 5.4.5 (Limit of the empirical measure)**.** *For any $x \in E$, we have*

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} = \frac{1}{\mathbb{E}_x[\tau_x]}, \qquad \textit{almost surely.}$$

If x is transient, then by Exercise 5.4.4, $N_x < +\infty$ almost surely, so that for n large enough,

$$\frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} = \frac{N_x}{n} \to 0 = \frac{1}{\mathbb{E}_x[\tau_x]}.$$

On the other hand, if x is recurrent, then by Exercise 5.4.4 again, one may define the sequence of successive return times $\tau_x^0, \tau_x^1, \ldots$ of X_n as in Section 5.3, and the statement of Lemma 5.4.5 follows from the very same proof[2] as for (5.4). If x is positive recurrent, then the representation formula

$$\forall x, y \in E, \qquad \mathbb{E}_x\left[\sum_{i=0}^{\tau_x - 1} \mathbb{1}_{\{X_i = y\}}\right] = \frac{\mathbb{E}_x[\tau_x]}{\mathbb{E}_y[\tau_y]} \tag{5.6}$$

[2] If x is null recurrent, we use the following variant of the LLN: if $(X_n)_{n \geq 1}$ is a sequence of nonnegative random variables such that $\mathbb{E}[X_1] = +\infty$, then $\frac{1}{n} \sum_{i=0}^{n} X_n \to +\infty$, almost surely. This statement can easily be deduced from the usual strong Law of Large Numbers, applied to the sequence $\min(X_n, M)$, and the Monotone Convergence Theorem to show that $\lim_{M \to +\infty} \mathbb{E}[\min(X_1, M)] = +\infty$.

remains true.

With Lemma 5.4.5 at hand, we may now state and prove the main two results regarding recurrence, transience, stationary distribution and LLN.

Proposition 5.4.6 (Recurrence, transience and stationary distributions)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix P. We assume that P is irreducible.*

(i) All states have the same nature, and therefore being transient, null recurrent or positive recurrent is a property of the chain.

(ii) If the chain is positive recurrent, then setting

$$\forall x \in E, \qquad \pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]}$$

defines a probability measure on E, which is the unique stationary distribution of the chain.

(iii) If the chain is transient or null recurrent, then it does not admit a stationary distribution.

Remark 5.4.7. *In the finite state space case, Lemma 5.3.1 shows that all irreducible chains are positive recurrent.*

Theorem 5.4.8 (Ergodic theorem)**.** *If the chain $(X_n)_{n\geq 0}$ is irreducible and positive recurrent, then for any $f \in \mathbf{L}^1(\pi)$,*

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(X_i) = \pi f, \qquad \textit{almost surely.}$$

Proof of Proposition 5.4.6. Let us assume that P is irreducible and fix $x, y \in E$. Then there exist $p, q \geq 1$ such that $P^p(x,y) > 0$ and $P^q(y,x) > 0$. Since, for any $n \geq 0$,

$$P^{n+p+q}(x,x) \geq P^p(x,y)P^n(y,y)P^q(y,x),$$
$$P^{n+p+q}(y,y) \geq P^q(y,x)P^n(x,x)P^p(x,y),$$

the series $\sum P^n(x,x)$ and $\sum P^n(y,y)$ have the same nature, and therefore by Exercise 5.4.4, either all states are transient, or all states are recurrent. We now assume that there is a positive recurrent state x. Then $\pi(x) > 0$ and, for any $y \in E$, the representation formula (5.6) yields

$$\pi(y) = \pi(x)\mathbb{E}_x\left[\sum_{n=0}^{\tau_x-1} \mathbb{1}_{\{X_n=y\}}\right].$$

Assume that $\mathbb{E}_x[\sum_{n=0}^{\tau_x-1} \mathbb{1}_{\{X_n=y\}}] = 0$. Then necessarily $\sum_{n=0}^{\tau_x-1} \mathbb{1}_{\{X_n=y\}} = 0$, $\mathbb{P}_x$-almost surely, which implies that starting from x, the state y cannot be reached. This is in contradiction with the assumption that the chain is irreducible and thereby implies that $\pi(y) > 0$, so that all states are positive recurrent.

To prove both (ii) and (iii), we note that the boundedness of the variable $\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}_{\{X_n=x\}}$ allows to apply the Dominated Convergence Theorem to the statement of Lemma 5.4.5, and thus deduce that, for any $x \in E$,

$$\begin{aligned}\pi(x) &= \mathbb{E}\left[\lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=x\}}\right]\\ &= \lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{P}(X_i = x)\\ &= \lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mu_0 P^i(x)\\ &= \lim_{n\to+\infty} \widehat{\mu}_n(x),\end{aligned}$$

with $\widehat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n-1} \mu_0 P^i$, and where we recall that μ_0 is the law of X_0. If the chain is positive recurrent[3], then by Scheffé's Lemma (see Remark 3.1.21), for any $y \in E$,

$$|\widehat{\mu}_n P(y) - \pi P(y)| = \left|\sum_{x\in E} (\widehat{\mu}_n(x) - \pi(x))P(x,y)\right| \leq \sum_{x\in E} |\widehat{\mu}_n(x) - \pi(x)| \to 0,$$

and therefore

$$\lim_{n\to+\infty} \widehat{\mu}_n P(y) = \pi P(y).$$

But on the other hand,

$$\widehat{\mu}_n P(y) = \frac{1}{n}\sum_{i=0}^{n-1} \mu_0 P^i(y) = \widehat{\mu}_n(y) + \frac{1}{n}\left(\mu_0 P^n(y) - \mu_0(y)\right) \to \pi(y),$$

which finally yields $\pi(y) = \pi P(y)$ and completes the proof of (ii). In the case where $\pi = 0$, assuming that μ_0 is a stationary distribution yields $\widehat{\mu}_n = \mu_0$ for any n and therefore immediately contradicts the fact that $\widehat{\mu}_n(x) \to \pi(x)$, which proves (iii). □

Proof of Theorem 5.4.8. If f is bounded then the conclusion follows from Scheffé's Lemma, still applied under the form of Remark 3.1.21. In the general case where $f \in \mathbf{L}^1(\pi)$, the theorem is obtained by using the same decomposition of $\sum_{i=0}^{n-1} f(X_i)$ into excursions as in the proof of Lemma 5.4.5, and using the strong Law of Large Numbers again. □

☕ **Exercise 5.4.9** (Recurrence and transience of the simple random walk)**.** *The purpose of this exercise is to show that for $d \in \{1,2\}$, the random walk on $\mathbb{Z}^d$ is null recurrent, while for $d \geq 3$ it is transient.* In all cases, we shall use the characterisation of recurrence and transience provided by Exercise 5.4.4.

1. *For the simple random walk in dimension $d = 1$, compute $\mathbb{P}_0(X_n = 0)$ and conclude.*
2. *We let $d = 2$ and denote by (X_n^1, X_n^2) the coordinates of the simple random walk.*
   (a) *Let $U_n = (X_{n+1}^1 - X_n^1) + (X_{n+1}^2 - X_n^2)$ and $V_n = (X_{n+1}^1 - X_n^1) - (X_{n+1}^2 - X_n^2)$. Write the law of the pair (U_n, V_n) and show that the sequence $(U_n, V_n)_{n\geq 0}$ is iid.*
   (b) *For n odd, what is the value of $\mathbb{P}_0(X_n = 0)$?*
   (c) *For n even, express the event $\{X_n = 0\}$ in terms of $\sum_{k=0}^{n-1} U_k$ and $\sum_{k=0}^{n-1} V_k$ and deduce the value of $\mathbb{P}_0(X_n = 0)$.*
   (d) *Conclude.*

[3] The argument here is the same as in the second proof of Proposition 5.2.6, extended to the infinite state space case thanks to Scheffé's Lemma.

3. *We now assume that* $d \geq 3$ *and denote by* φ *the characteristic function of* X_1 *under* $\mathbb{P}_0$*, defined by*

$$\forall u \in \mathbb{R}^d, \qquad \varphi(u) := \mathbb{E}_0\left[\mathrm{e}^{\mathrm{i}\langle u, X_1\rangle}\right].$$

   (a) *Show that, for all* $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$*,*

$$\varphi(u) = \frac{1}{d}\left(\cos u_1 + \cdots + \cos u_d\right).$$

   (b) *Show that*

$$\sum_{k=0}^{+\infty} \mathbb{P}_0(X_{2k} = 0) = \frac{1}{(2\pi)^d} \int_{u \in (-\pi,\pi)^d} \frac{\mathrm{d}u}{1 - \varphi^2(u)}.$$

   (c) *Conclude.*

Chapter 6

Convergence to equilibrium of Markov chains

Contents

In this Chapter, we study the long-time behaviour of homogeneous Markov chains.

Exercise 6.0.1. *Show that if $(X_n)_{n\geq 0}$ is an irreducible Markov chain such that $X_n \to \pi$ in distribution, then π is necessarily a stationary distribution for $(X_n)_{n\geq 0}$.*

From a numerical point of view, our motivation is the following: if the law of X_n converges to π, then for n large enough, X_n can be used as an approximate random number generator under π. Quantifying the distance between the law of X_n and π would then allow us to control the approximation error made in this procedure. The main result of the Chapter in this perspective is Theorem 6.2.4, which provides a geometrically decreasing error estimate between the law of X_n and its stationary distribution π. Somewhat unexpectedly, this result also allows us to establish rigorously a formula for the asymptotic variance $\sigma^2(f)$ in the Markov chain CLT, which is also an important point from the Markov Chain Monte Carlo point of view since it is directly related to the construction of confidence intervals for the estimation of πf by trajectorial averages of Markov chains.

6.1 Periodicity

Consider the sequence $(X_n)_{n\geq 0}$ defined in the two-point space $E = \{-1, 1\}$ by $X_n = (-1)^n X_0$. It is an irreducible Markov chain with transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and unique stationary distribution π such that $\pi(-1) = \pi(1) = 1/2$.

Exercise 6.1.1. *Check that the chain $(X_n)_{n\geq 0}$ satisfies the Law of Large Numbers of Theorem 5.3.3. What do you think of the Central Limit Theorem?*

Yet, for any initial distribution μ_0 on E which is not π, the law μ_n of X_n under $\mathbb{P}_{\mu_0}$ does not converge to π. This is related to the phenomenon of periodicity.

We recall that the *greatest common divisor* of a set of nonnegative integers $\mathcal{N}$ is defined by

$$\gcd \mathcal{N} = \max\{k \geq 1 : \forall n \in \mathcal{N}, k|n\},$$

where the notation $k|n$ means that there exists $\ell \in \mathbb{N}$ such that $n = k\ell$.

Definition 6.1.2 (Period)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix P. For all $x \in E$, set $\mathcal{N}(x) = \{n \geq 1 : P^n(x, x) > 0\}$. The* period *of x is $\gcd \mathcal{N}(x)$.*

Exercise 6.1.3. *Compute the period of the states -1 and 1 in the example of the two-point space described above. More generally, what is the period of the states in the random walk on $\mathbb{T}_N$ introduced in Example 5.1.7?*

Lemma 6.1.4 (Period of an irreducible chain)**.** *If the chain $(X_n)_{n\geq 0}$ is irreducible, then all states have the same period, which is thus called the* period *of the chain. The chain is called* aperiodic *if its period is 1 and* periodic *otherwise.*

The proof of Lemma 6.1.4 relies on the inequality

$$\forall x, y, z \in E, \quad \forall n, m \geq 1, \qquad P^{n+m}(x, y) \geq P^n(x, z)P^m(z, y),$$

which comes from the fact that, by Proposition 5.1.12,

$$\begin{aligned} P^{n+m}(x, y) &= \mathbb{P}_x(X_{n+m} = y) \\ &\geq \mathbb{P}_x(X_n = z, X_{n+m} = y) \\ &= \mathbb{P}_x(X_n = z)\mathbb{P}_z(X_m = y) \\ &= P^n(x, z)P^m(z, y). \end{aligned}$$

Proof of Lemma 6.1.4. Let $x, y \in E$. By Definition 5.2.7, there exist $r \geq 1$ and $\ell \geq 1$ such that $P^r(x, y) > 0$ and $P^\ell(y, x) > 0$. Let $m = r + \ell$. Then

$$P^m(x, x) \geq P^r(x, y)P^\ell(y, x) > 0, \qquad P^m(y, y) \geq P^\ell(y, x)P^r(x, y) > 0,$$

so $m \in \mathcal{N}(x) \cap \mathcal{N}(y)$. Moreover, if $n \in \mathcal{N}(x)$ then

$$P^{n+m}(y, y) \geq P^\ell(y, x)P^n(x, x)P^r(x, y) > 0,$$

so $n + m \in \mathcal{N}(y)$. Now let $k = \gcd \mathcal{N}(y)$. Since $m \in \mathcal{N}(y)$ we have $k|m$. Besides, for any $n \in \mathcal{N}(x)$, since $n + m \in \mathcal{N}(y)$ we have $k|n + m$. Therefore $k|n$ and thus $k = \gcd \mathcal{N}(y) \leq \gcd \mathcal{N}(x)$. By the same arguments, $\gcd \mathcal{N}(x) \leq \gcd \mathcal{N}(y)$ and the proof is completed. □

6.2 Geometric convergence of finite state space Markov chains

In this section, we assume that E is finite, and denote by m its cardinality. Recall that we denote by $\mathbb{R}^E$ the space of functions $E \to \mathbb{R}$, which are seen as m-dimensional column vectors. We also denote by $\mathcal{M}(E)$ the space of signed measures on E, which are seen as m-dimensional row vectors.

There are two possible approaches to quantify the convergence in distribution of X_n to π, either by bounding $\mu_0 P^n - \pi$ in $\mathcal{M}(E)$, where $\mu_n = \mu_0 P^n$ is the *law* of X_n, or by bounding $P^n f - \pi f$ in $\mathbb{R}^E$, where $P^n f(x) = \mathbb{E}_x[f(X_n)]$ is called an *observable*[1]. In both cases, the quantity of interest involves the matrix P^n, whose $n \to +\infty$ limit is closely related with the spectral properties of P.

6.2.1 Spectral decomposition of P

Exercise 6.2.1. *Let P be a stochastic matrix. Show that for any (complex) eigenvalue λ of P, $|\lambda| \leq 1$.*

From now on we assume that P is irreducible. Then the eigenvalue 1 is simple and has left and right eigenvectors π and $\mathbf{1}$, respectively. As a consequence, the spaces

$$\begin{aligned} \mathcal{M}_0(E) &:= \{\rho \in \mathcal{M}(E) : \rho\mathbf{1} = 0\} = \mathbf{1}^\perp, \\ \mathbb{R}^E_0 &:= \{g \in \mathbb{R}^E : \pi g = 0\} = \pi^\perp, \end{aligned}$$

are stable by the mappings $\rho \mapsto \rho P$ and $g \mapsto Pg$, respectively. We denote by P_0 the restriction of both mappings to the subsets $\mathcal{M}_0(E)$ and $\mathbb{R}^E_0$, respectively.

Then one the one hand, for any $\mu_0 \in \mathcal{P}(E)$ and for any $f \in \mathbb{R}^E$,

$$\begin{aligned} \mu_0 P^n - \pi &= (\mu_0 - \pi)P^n = (\mu_0 - \pi)P_0^n, \\ P^n f - \pi f &= P^n(f - \pi f) = P_0^n(f - \pi f), \end{aligned} \tag{6.1}$$

so that the convergence of the law $\mu_0 P^n$ to π, or of the observable $P^n f$ to πf, only depends on P_0^n. On the other hand, by construction of the sets $\mathcal{M}_0(E)$ and $\mathbb{R}^E_0$, the set of eigenvalues of P_0 is exactly the set of eigenvalues of P which are not equal to 1. In particular, any eigenvalue λ of P_0 satisfies

$$|\lambda| \leq \lambda_\star := \max\{|\lambda|, \lambda \neq 1 \text{ is an eigenvalue of } P\}.$$

We deduce the following statement.

Lemma 6.2.2 (Geometric decay of P_0^n)**.** *For any $\alpha > \lambda_\star$, $\alpha^{-n} P_0^n \to 0$ when $n \to +\infty$.*

Proof. We detail the proof for P_0 seen as the restriction to $\mathcal{M}_0(E)$ of the mapping $\rho \mapsto \rho P$. By the Dunford Theorem, there is a basis of the $(\mathsf{m} - 1)$-dimensional space $\mathcal{M}_0(E)$ in which P_0 is represented by a matrix of the form $D + N$, where D is diagonal with entries $\lambda \in \mathbb{C}$ such that $|\lambda| \leq \lambda_\star < 1$ and N is a nilpotent matrix, and such that $DN = ND$. Thus, in this basis, for all $n \geq 1$,

$$P_0^n = \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k,$$

[1] The constant function $x \mapsto \pi f$ should be denoted by $\pi f \mathbf{1}$ to emphasise the fact that it is an element of $\mathbb{R}^E$. In order to lighten the notation we will simply write πf.

and since $N^{\mathsf{m}-1} = 0$ we get that as soon as $n \geq \mathsf{m} - 1$,

$$P_0^n = \sum_{k=0}^{\mathsf{m}-1} \binom{n}{k} D^{n-k} N^k.$$

Since the binomial coefficient $\binom{n}{k}$ is equivalent to $n^k/k!$ when $n \to +\infty$ and the diagonal coefficients λ^{n-k} of D^{n-k} satisfy $|\lambda|^{n-k} \leq \lambda_\star^{n-k}$, we deduce that as soon as $\alpha > \lambda_\star$, $\alpha^{-n} P_0^n \to 0$. □

6.2.2 Perron–Frobenius Theorem and geometric convergence to equilibrium

To deduce from Lemma 6.2.2 that $P_0^n \to 0$ at a geometric rate, we now have to check that $\lambda_\star < 1$, that is to say that apart from 1, there is no other eigenvalue of P with modulus 1. This is where aperiodicity comes back in the game.

Proposition 6.2.3 (Perron–Frobenius Theorem)**.** *Let P be the transition matrix of an irreducible Markov chain $(X_n)_{n \geq 0}$, with period $k \geq 1$. The eigenvalues λ of P such that $|\lambda| = 1$ are the k-th roots of unity, and they are all simple.*

We refer to [5, Theorem 3.11] for the proof.

As a corollary, if P is irreducible and aperiodic, then $\lambda_\star < 1$ and we obtain the following statement.

Theorem 6.2.4 (Geometric convergence in finite state spaces)**.** *Let P be an irreducible and aperiodic stochastic matrix with stationary distribution π.*

(i) For any $\alpha \in (\lambda_\star, 1]$ and for any norm $\|\cdot\|$ on $\mathcal{M}(E)$, there exists a constant C_α such that, for any $\mu_0 \in \mathcal{P}(E)$,

$$\forall n \geq 0, \qquad \|\mu_0 P^n - \pi\| \leq C_\alpha \alpha^n \|\mu_0 - \pi\|.$$

(ii) For any $\alpha \in (\lambda_\star, 1]$ and for any norm $\|\cdot\|$ on $\mathbb{R}^E$, there exists a constant C_α such that, for any $f \in \mathbb{R}^E$,

$$\forall n \geq 0, \qquad \|P^n f - \pi f\| \leq C_\alpha \alpha^n \|f - \pi f\|.$$

Proof. In any of the settings (i) and (ii), let $|||\cdot|||$ be the operator norm associated with the norm $\|\cdot\|$. For any $\alpha \in (\lambda_\star, 1]$, Lemma 6.2.2 implies that

$$C_\alpha := \sup_{n \geq 0} \alpha^{-n} |||P_0^n||| < +\infty,$$

which by (6.1) completes the proof. □

Remark 6.2.5. *Owing to the fact that the binomial terms in the Dunford decomposition of P_0 grow polynomially in n, one cannot take $\alpha = \lambda_\star$ in the proof above, except if $N = 0$, that is to say if P_0 is diagonalisable. We shall present an important class of Markov chains for which P_0 is diagonalisable in Section 6.4.*

Remark 6.2.6. *In the statement (i) of Theorem 6.2.4, the quantity $\|\mu_0 - \pi\|$ is bounded uniformly in μ_0 (and π) since $\mathcal{P}(E)$ is compact. Therefore, up to increasing the value of the constant C_α, we deduce that the geometric convergence to π of μ_n holds uniformly in the initial condition μ_0.*

6.2.3 Another formula for the asymptotic variance in the Markov chain CLT

In this subsection, we consider an irreducible Markov chain $(X_n)_{n\geq 0}$, with transition matrix P. Then by Theorem 5.3.6, for any $f \in \mathbb{R}^E$, there exists $\sigma^2(f) \geq 0$ such that

$$\frac{1}{\sqrt{n}}\left(\sum_{i=0}^{n-1} f(X_i) - \pi f\right) \to \mathcal{N}(0, \sigma^2(f)), \qquad \text{in distribution.}$$

Our main result is the following statement.

Proposition 6.2.7 (Asymptotic variance in the Markov chain CLT)**.** *Assume that P is aperiodic. Then we have*

$$\sum_{n=1}^{+\infty} |\operatorname{Cov}_\pi(f(X_0), f(X_n))| < +\infty, \tag{6.2}$$

and

$$\sigma^2(f) = \operatorname{Var}_\pi(f(X_0)) + 2\sum_{n=1}^{+\infty} \operatorname{Cov}_\pi(f(X_0), f(X_n)). \tag{6.3}$$

Proof. Throughout the proof we assume without loss of generality that $\pi f = 0$. Then by Theorem 6.2.4 (ii) applied with $\|f\| = \max_{x\in E}|f(x)|$, and since P is aperiodic, there exists $\alpha < 1$ and $C_\alpha \geq 0$ such that for any $f \in \mathbb{R}^E$,

$$\forall x \in E, \quad \forall n \geq 0, \qquad |P^n f(x)| \leq C_\alpha \alpha^n \|f\|. \tag{6.4}$$

Step 1. We first check (6.2). By (6.4),

$$\begin{aligned}
|\operatorname{Cov}_\pi(f(X_0), f(X_n))| &= |\mathbb{E}_\pi\left[f(X_0)f(X_n)\right]| \\
&= \left|\sum_{x\in E} f(x)P^n f(x)\pi(x)\right| \\
&\leq \sum_{x\in E} |f(x)|\,|P^n f(x)|\pi(x) \\
&\leq C_\alpha \alpha^n \|f\|^2,
\end{aligned}$$

which proves (6.2).

Step 2. We now show that

$$\lim_{n\to+\infty} \operatorname{Var}\left(\frac{1}{\sqrt{n}}\left(\sum_{i=0}^{n-1} f(X_i)\right)\right) = \operatorname{Var}_\pi(f(X_0)) + 2\sum_{n=1}^{+\infty} \operatorname{Cov}_\pi(f(X_0), f(X_n)),$$

and postpone the conclusion of the proof to Step 3. First, we have

$$\operatorname{Var}\left(\frac{1}{\sqrt{n}}\left(\sum_{i=0}^{n-1} f(X_i)\right)\right) = \frac{1}{n}\sum_{i=0}^{n-1} \operatorname{Var}\left(f(X_i)\right) + \frac{2}{n}\sum_{0\leq i<j\leq n-1} \operatorname{Cov}\left(f(X_i), f(X_j)\right).$$

On the one hand,

$$\operatorname{Var}\left(f(X_i)\right) = \mathbb{E}\left[f(X_i)^2\right] - \mathbb{E}\left[f(X_i)\right]^2,$$

and by Theorem 6.2.4, the right-hand side converges to $\pi(f^2) - (\pi f)^2 = \operatorname{Var}_\pi(f(X_0))$ when $i \to +\infty$, so by the Césaro Lemma,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} \operatorname{Var}\left(f(X_i)\right) = \operatorname{Var}_\pi(f(X_0)).$$

On the other hand,

$$\mathrm{Cov}\left(f(X_i), f(X_j)\right) = \mathbb{E}\left[f(X_i)f(X_j)\right] - \mathbb{E}\left[f(X_i)\right]\mathbb{E}\left[f(X_j)\right].$$

First, by (6.4),

$$\left|\frac{2}{n}\sum_{0\leq i<j\leq n-1}\mathbb{E}\left[f(X_i)\right]\mathbb{E}\left[f(X_j)\right]\right| \leq \frac{1}{n}\left(\sum_{i=0}^{n-1} C_\alpha \alpha^i \|f\|\right)^2 \to 0.$$

Second, for $i<j$,

$$\mathbb{E}\left[f(X_i)f(X_j)\right] = \mathbb{E}\left[\mathbb{E}\left[f(X_i)f(X_j)|X_i\right]\right] = \mathbb{E}\left[f(X_i)P^{j-i}f(X_i)\right],$$

so that

$$\begin{aligned}\sum_{j=i+1}^{n-1}\mathbb{E}\left[f(X_i)f(X_j)\right] &= \sum_{k=1}^{n-1-i}\mathbb{E}\left[f(X_i)P^k f(X_i)\right] \\ &= \sum_{k=1}^{+\infty}\mathbb{E}[f(X_i)P^k f(X_i)] - \sum_{k=n-i}^{+\infty}\mathbb{E}[f(X_i)P^k f(X_i)],\end{aligned}$$

Thus, using (6.4) and the Césaro Lemma again,

$$\begin{aligned}\lim_{n\to+\infty}\frac{2}{n}\sum_{0\leq i<j\leq n-1}\mathbb{E}\left[f(X_i)f(X_j)\right] &= \lim_{n\to+\infty}\frac{2}{n}\sum_{i=0}^{n-1}\sum_{k=1}^{+\infty}\mathbb{E}[f(X_i)P^k f(X_i)] \\ &= \sum_{k=1}^{+\infty}\mathbb{E}_\pi\left[f(X_0)P^k f(X_0)\right] \\ &= \sum_{k=1}^{+\infty}\mathrm{Cov}_\pi(f(X_0), f(X_k)),\end{aligned}$$

which completes the proof of the claimed identity.

Step 3. A tedious but elementary computation shows that

$$\sup_{n\geq 1}\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} f(X_i)\right)^4\right] < +\infty.$$

By Proposition 3.1.26 and Theorem 5.3.6, we deduce that the moments of order $p<4$ of $\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} f(X_i)$ converge to the moments of order p of $\mathcal{N}(0,\sigma^2(f))$, which in particular imply that $\sigma^2(f)$ coincides with the limit of $\mathrm{Var}(\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} f(X_i))$ computed in Step 2. □

6.2.4 Poisson equation

In this Subsection we give yet another formula (in fact, two) for $\sigma^2(f)$, which depends on the solution to the so-called *Poisson equation*.

Lemma 6.2.8 (Poisson equation)**.** *Assume that P is irreducible and aperiodic, and let $f \in \mathbb{R}^E$. Set $\widetilde{f} = f - \pi f \in \mathbb{R}^E$. Then there exists a unique $g \in \mathbb{R}^E_0$ such that*

$$-(P-I)g = \widetilde{f} \tag{6.5}$$

and it writes

$$\forall x \in E, \qquad g(x) = \sum_{n=0}^{+\infty} \mathbb{E}_x[\widetilde{f}(X_n)]. \tag{6.6}$$

Proof. Since we look for a solution to the Poisson equation (6.5) in $\mathbb{R}^E_0$, the latter rewrites

$$(I_0 - P_0)g = \widetilde{f},$$

with I_0 the identity of $\mathbb{R}^E_0$. Under the assumptions of the Lemma, all eigenvalues λ of P_0 satisfy $|\lambda| \leq \lambda_\star < 1$, which implies that $I_0 - P_0$ is invertible and has inverse

$$(I_0 - P_0)^{-1} = \sum_{n=0}^{+\infty} P_0^n,$$

which directly yields (6.6). □

In order to present the formula for $\sigma^2(f)$ in terms of g, we introduce the notation $\langle \cdot, \cdot \rangle_\pi$ and $\| \cdot \|_\pi$ on $\mathbb{R}^E$ defined by

$$\langle f, g \rangle_\pi := \sum_{x \in E} f(x)g(x)\pi(x), \qquad \|f\|_\pi := \sqrt{\langle f, f \rangle_\pi}.$$

Proposition 6.2.9 (Asymptotic variance and Poisson equation)**.** *Under the assumptions of Proposition 6.2.7 and with the notation of Lemma 6.2.8,*

$$\sigma^2(f) = 2\langle \widetilde{f}, g \rangle_\pi - \|\widetilde{f}\|^2_\pi.$$

Proof. We start from the result of Proposition 6.2.7 and write

$$\sigma^2(f) = \mathrm{Var}_\pi(f(X_0)) + 2\sum_{n=1}^{+\infty} \mathrm{Cov}_\pi(f(X_0), f(X_n)) = 2\sum_{n=0}^{+\infty} \mathrm{Cov}_\pi(f(X_0), f(X_n)) - \mathrm{Var}_\pi(f(X_0)).$$

Now on the one hand, $\mathrm{Var}_\pi(f(X_0)) = \|\widetilde{f}\|^2_\pi$, while on the other hand,

$$\sum_{n=0}^{+\infty} \mathrm{Cov}_\pi(f(X_0), f(X_n)) = \sum_{n=0}^{+\infty} \mathbb{E}_\pi\left[\widetilde{f}(X_0)P^n\widetilde{f}(X_0)\right] = \sum_{n=0}^{+\infty} \langle \widetilde{f}, P_0^n \widetilde{f} \rangle_\pi = \langle \widetilde{f}, g \rangle_\pi,$$

which completes the proof. □

This Proposition will be used in particular in Subsection 7.2.2. It is also used in the next exercise.

🏠 **Exercise 6.2.10.** *Show that* $\sigma^2(f) = \pi(g^2) - \pi((Pg)^2)$.

6.3 Convergence to equilibrium in the countably infinite case

If E is now assumed to be infinite, then the spectral analysis of P is less easy to manipulate. However, the statement of Theorem 6.2.4 remains essentially correct, although no rate of convergence is, in general, available.

Theorem 6.3.1 (Convergence to equilibrium)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible and positive recurrent Markov chain, with unique stationary distribution π. If the chain is aperiodic, then*

$$\lim_{n\to+\infty} X_n = \pi, \qquad \textit{in distribution.}$$

The proof of Theorem 6.3.1 relies on the following lemma.

Lemma 6.3.2 (Product chain)**.** *Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be two independent Markov chains in the respective discrete spaces E and F, with respective transition matrices P and Q.*

(i) The sequence $(X_n, Y_n)_{n\geq 0}$ is a Markov chain in $E\times F$, with transition matrix

$$P\otimes Q((x,y),(x',y')) = \mathbb{P}(X_{n+1}=x', Y_{n+1}=y'|X_n=x, Y_n=y) = P(x,x')Q(y,y').$$

(ii) If π is a stationary probability for $(X_n)_{n\geq 0}$ and ψ is a stationary probability for $(Y_n)_{n\geq 0}$, then

$$\pi\otimes\psi(x,y) = \pi(x)\psi(y)$$

is a stationary distribution for $(X_n, Y_n)_{n\geq 0}$.

(iii) If the chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ are irreducible and at least one of them is aperiodic, then the chain $(X_n, Y_n)_{n\geq 0}$ is irreducible.

Proof. The points (i) and (ii) are immediate to check. To prove the point (iii), we assume that $(X_n)_{n\geq 0}$ is aperiodic and fix $(x,y),(x',y')\in E\times F$. By irreducibility, there exist $p,q,r\geq 1$ such that $P^p(x,x')>0$, $Q^q(y,y')>0$ and $Q^r(y',y')>0$. Note that for any $k\geq 1$, $Q^{kr}(y',y')>0$. Besides, since $(X_n)_{n\geq 0}$ is aperiodic, by Lemma 6.3.3 below, for k large enough, $P^{q+kr-p}(x,x)>0$ and therefore, with $n=q+kr$, we have

$$\begin{aligned}\mathbb{P}_{(x,y)}(X_n=x', Y_n=y') &= P^n(x,x')Q^n(y,y')\\ &\geq P^{q+kr-p}(x,x)P^p(x,x')Q^q(y,y')Q^{kr}(y',y')>0,\end{aligned}$$

which proves irreducibility. □

Lemma 6.3.3 (Schur's Theorem)**.** *Let $\mathcal{N}\subset\mathbb{N}$ be closed under addition and such that* $\gcd\mathcal{N}=1$. *Then the set $\mathbb{N}\setminus\mathcal{N}$ is finite.*

Lemma 6.3.3 is purely number theoretic. We refer to [10, Proposition 1.7] for details. We are now ready to complete the proof of Theorem 6.3.1.

Proof of Theorem 6.3.1. Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent and aperiodic Markov chain with initial distribution μ_0, transition matrix P and unique stationary distribution π. Let $(Y_n)_{n\geq 0}$ be a Markov chain with transition matrix P and initial distribution π, independent from $(X_n)_{n\geq 0}$. Since $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ have the same transition matrix, the chain $(Y_n)_{n\geq 0}$ is irreducible, aperiodic, and its stationary distribution is π. In fact, for any n, $Y_n\sim\pi$, and therefore for any $y\in E$,

$$\mathbb{P}(X_n=y)-\pi(y) = \mathbb{P}(X_n=y)-\mathbb{P}(Y_n=y).$$

By Lemma 6.3.2, the chain $(X_n,Y_n)_{n\geq 0}$ is irreducible and positive recurrent. Therefore, for any x, the random time

$$\tau_x = \inf\{n\geq 1 : (X_n,Y_n)=(x,x)\}$$

is finite, almost surely. Besides, for any $y\in E$,

$$\mathbb{P}(X_n=y) = \mathbb{P}(X_n=y,\tau_x\leq n)+\mathbb{P}(X_n=y,\tau_x>n)\leq\mathbb{P}(X_n=y,\tau_x\leq n)+\mathbb{P}(\tau_x>n).$$

By Proposition 5.1.12, we then have

$$\begin{aligned}
\mathbb{P}(X_n = y, \tau_x \leq n) &= \sum_{k=0}^{n} \mathbb{P}(X_n = y, \tau_x = k) \\
&= \sum_{k=0}^{n} \mathbb{P}(\tau_x = k)\mathbb{P}_x(X_{n-k} = y) \\
&= \sum_{k=0}^{n} \mathbb{P}(\tau_x = k)\mathbb{P}_x(Y_{n-k} = y) \\
&= \mathbb{P}(Y_n = y, \tau_x \leq n),
\end{aligned}$$

where we have used the fact that the chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ have the same transition matrices to get $\mathbb{P}_x(X_{n-k} = y) = \mathbb{P}_x(Y_{n-k} = y)$. We deduce that

$$\mathbb{P}(X_n = y) \leq \mathbb{P}(Y_n = y, \tau_x \leq n) + \mathbb{P}(\tau_x > n) \leq \mathbb{P}(Y_n = y) + \mathbb{P}(\tau_x > n),$$

and then by symmetry

$$|\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| \leq \mathbb{P}(\tau_x > n).$$

Since τ_x is finite, the right-hand side converges to 0 when $n \to +\infty$, which implies that

$$\forall y \in E, \qquad \lim_{n\to+\infty} \mathbb{P}(X_n = y) = \pi(y),$$

and the conclusion follows from Scheffé’s Lemma (see Remark 3.1.21). □

6.4 Reversibility

In this section, we introduce and study the particular class of *reversible* Markov chains, which enjoy several useful properties. In particular, their transition matrix is symmetric for a certain scalar product, which enables to use the Spectral Theorem to study their long time behaviour. In the finite state space case, this slightly improves the statement of Theorem 6.2.4.

6.4.1 Definition and general remarks

For the moment, E can be either finite or countably infinite.

Definition 6.4.1 (Reversibility)**.** *A Markov chain $(X_n)_{n\geq 0}$ with transition matrix P is said to be* reversible *with respect to $\pi \in \mathcal{P}(E)$ if, for any $x, y \in E$,*

$$\pi(x)P(x,y) = \pi(y)P(y,x). \tag{6.7}$$

Equation (6.7) is called the *detailed balance equation*. The denomination ‘reversibility’ is explained by the following result.

Proposition 6.4.2 (Reversibility)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix P, reversible with respect to π. For any $n \geq 0$, the vectors $(X_0, \ldots, X_n)$ and $(X_n, \ldots, X_0)$ have the same distribution under $\mathbb{P}_\pi$.*

Proof. For any $x_0, \dots, x_n \in E$, we deduce from Proposition 5.1.10 that

$$\mathbb{P}_\pi(X_0 = x_0, \dots, X_n = x_n) = \pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n).$$

Applying Definition 6.4.1 once shows that $\pi(x_0)P(x_0, x_1) = P(x_1, x_0)\pi(x_1)$, and iterating this procedure leads to the identity

$$\pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1}) \cdots P(x_1, x_0),$$

the right-hand side of which is $\mathbb{P}_\pi(X_0 = x_n, \dots, X_n = x_0)$ by Proposition 5.1.10 again. □

Looking only at the marginal distribution of the first coordinate of the vectors $(X_0, \dots, X_n)$ and $(X_n, \dots, X_0)$, we deduce the following link between the notions of reversibility and stationary distribution.

Corollary 6.4.3 (Reversibility and stationary distribution)**.** *If $(X_n)_{n\geq 0}$ is reversible with respect to π, then π is a stationary distribution for $(X_n)_{n\geq 0}$.*

Notice that this result can also be obtained by the direct computation

$$\pi P(y) = \sum_{x\in E} \pi(x)P(x, y) = \sum_{x\in E} P(y, x)\pi(y) = \pi(y),$$

which uses the fact that $\sum_{x\in E} P(y, x) = 1$.

🏠 **Exercise 6.4.4.** *Show that for the Ehrenfest urn, both the microscopic and the macroscopic descriptions are reversible with respect to the stationary distributions from Exercise 5.2.5.*

🗎 **Exercise 6.4.5.** *Under which condition is the random walk on $\mathbb{T}_N$ reversible?*

6.4.2 Spectral characterisation of reversibility

Fix $\pi \in \mathcal{P}(E)$. Recall that we denote by $\mathbf{L}^p(\pi)$ the set $\{f \in \mathbb{R}^E : \sum_{x\in E} |f(x)|^p \pi(x) < +\infty\}$.

Lemma 6.4.6 ($\mathbf{L}^p$-contractivity of stochastic matrices)**.** *Let π be a stationary distribution for a stochastic matrix P. For any $p \geq 1$ and $f \in \mathbf{L}^p(\pi)$,*

$$\sum_{x\in E} |Pf(x)|^p \pi(x) \leq \sum_{x\in E} |f(x)|^p \pi(x).$$

Proof. By Jensen's inequality, for any $x \in E$,

$$|Pf(x)|^p = |\mathbb{E}_x[f(X_1)]|^p \leq \mathbb{E}_x[|f(X_1)|^p] = P(|f|^p)(x),$$

so by stationarity

$$\sum_{x\in E} |Pf(x)|^p \pi(x) \leq \sum_{x\in E} P(|f|^p)(x)\pi(x) = \sum_{x,y\in E} \pi(x)P(x, y)|f(y)|^p = \sum_{y\in E} \pi(y)|f(y)|^p. \quad \square$$

From now on, we endow the space $\mathbf{L}^2(\pi)$ with the symmetric and bilinear form

$$\langle f, g\rangle_\pi := \sum_{x\in E} f(x)g(x)\pi(x).$$

An operator A is called *symmetric in* $\mathbf{L}^2(\pi)$ if $\langle Af, g\rangle_\pi = \langle f, Ag\rangle_\pi$, for all $f, g \in \mathbf{L}^2(\pi)$.

Remark 6.4.7. *If π is the stationary distribution of an irreducible stochastic matrix P, then $\pi(x) > 0$ for all $x \in E$ and thus $\langle \cdot, \cdot \rangle_\pi$ is a scalar product. In this case, the space $\mathbf{L}^2(\pi)$ is a Hilbert space, and the associated Euclidean norm is denoted by $\| \cdot \|_\pi$.*

Proposition 6.4.8 (Spectral characterisation of reversibility)**.** *Let $\pi \in \mathcal{P}(E)$. A Markov chain with transition matrix P is reversible with respect to π if and only if P is symmetric in $\mathbf{L}^2(\pi)$.*

Proof. Assume that a Markov chain with transition matrix P is reversible with respect to π. Then for all $f, g \in \mathbf{L}^2(\pi)$,

$$\begin{aligned}
\langle Pf, g \rangle_\pi &= \sum_{x \in E} \left(\sum_{y \in E} P(x,y) f(y) \right) g(x) \pi(x) \\
&= \sum_{x,y \in E} f(y) g(x) \pi(x) P(x,y) \\
&= \sum_{x,y \in E} f(y) g(x) \pi(y) P(y,x) \\
&= \sum_{y \in E} f(y) \left(\sum_{x \in E} P(y,x) g(x) \right) \pi(y) \\
&= \langle f, Pg \rangle_\pi .
\end{aligned}$$

Conversely, assume that P is symmetric in $\mathbf{L}^2(\pi)$, fix $x, y \in E$ and take $f(z) = \mathbb{1}_{\{z=x\}}$, $g(z) = \mathbb{1}_{\{z=y\}}$. Then

$$\langle Pf, g \rangle_\pi = P(x,y)\pi(y), \qquad \langle f, Pg \rangle_\pi = P(y,x)\pi(x),$$

so that Equation (6.7) is satisfied. □

6.4.3 Geometric convergence for finite reversible chains

In this subsection we assume that E is finite, with cardinality m. Then the Spectral Theorem yields the following statement.

Proposition 6.4.9 (Eigenvalues of reversible chains)**.** *Let P be the transition matrix of an irreducible chain which is reversible with respect to its invariant measure π. The eigenvalues of P are real and can be labelled $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{\mathsf{m}} \geq -1$, and there exists an orthonormal basis $(f_1, \ldots, f_{\mathsf{m}})$ of $\mathbf{L}^2(\pi)$ such that $Pf_i = \lambda_i f_i$ for all i.*

Remark 6.4.10. *If the chain is aperiodic, then by Proposition 6.2.3, $\lambda_{\mathsf{m}} > -1$. As a consequence, $\lambda_\star = \max\{|\lambda_2|, |\lambda_{\mathsf{m}}|\} < 1$.*

From Proposition 6.4.9, we deduce another statement for the geometric convergence of $P^n f$ to πf, to be compared with Theorem 6.2.4 (ii).

Proposition 6.4.11 (Rate of convergence for reversible chains)**.** *Under the assumptions of Proposition 6.4.9, for all $f \in \mathbf{L}^2(\pi)$, for all $n \geq 0$,*

$$\|P^n f - \pi f\|_\pi \leq \lambda_\star^n \|f - \pi f\|_\pi .$$

Notice that this result only provides the geometric convergence of $P^n f$ to πf if $\lambda_\star < 1$, that is to say if the chain is aperiodic (see Remark 6.4.10), which is of course in accordance with Theorem 6.2.4.

Proof. Notice that in Proposition 6.4.9 we may take $f_1 = \mathbf{1}$, in which case $\langle f, f_1\rangle_\pi = \pi f$, so that writing the orthogonal decomposition

$$P^n f = \sum_{i=1}^{\mathsf{m}} \langle P^n f, f_i\rangle_\pi f_i = \sum_{i=1}^{\mathsf{m}} \langle f, P^n f_i\rangle_\pi f_i = \sum_{i=1}^{\mathsf{m}} \lambda_i^n \langle f, f_i\rangle_\pi f_i$$

yields

$$P^n f - \pi f = \sum_{i=2}^{\mathsf{m}} \lambda_i^n \langle f, f_i\rangle_\pi f_i.$$

As a consequence,

$$\|P^n f - \pi f\|_\pi^2 = \sum_{i=2}^{\mathsf{m}} \left(\lambda_i^n \langle f, f_i\rangle_\pi\right)^2 \leq \lambda_\star^{2n} \sum_{i=2}^{\mathsf{m}} \langle f, f_i\rangle_\pi^2 = \lambda_\star^{2n} \|f - \pi f\|_\pi^2. \qquad \square$$

🏠 **Exercise 6.4.12.** *The* chi-square distance *on $\mathcal{P}(E)$ is defined by*

$$\chi_2(\mu|\pi) = \begin{cases} \displaystyle\sum_{x\in E} \left(\frac{\mu(x)}{\pi(x)} - 1\right)^2 \pi(x) & \text{if } \mu \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that it is not a distance, because it is not symmetric in μ and π. Show that under the assumptions of Proposition 6.4.9, for any initial distribution $\mu \in \mathcal{P}(E)$,

$$\chi_2(\mu P^n|\pi) \leq \lambda_\star^{2n} \chi_2(\mu|\pi).$$

Remark 6.4.13 (Asymptotic variance in the Markov chain CLT... again!)**.** *With the notation of Proposition 6.4.9, for any $f \in \mathbb{R}^E$ we have*

$$\mathrm{Var}_\pi(f(X_0)) = \|f - \pi f\|_\pi^2 = \sum_{i=2}^{\mathsf{m}} \langle f, f_i\rangle_\pi^2.$$

On the other hand, for any $n \geq 1$,

$$\begin{aligned} \mathrm{Cov}_\pi\left(f(X_0), f(X_n)\right) &= \mathbb{E}_\pi\left[f(X_0) P^n f(X_0)\right] - (\pi f)^2 \\ &= \langle f, P^n f\rangle_\pi - \langle f, f_1\rangle_\pi^2 \\ &= \sum_{i=2}^{\mathsf{m}} \lambda_i^n \langle f, f_i\rangle_\pi^2. \end{aligned}$$

Therefore, under the assumptions of Proposition 6.2.7, the asymptotic variance in the Markov chain CLT rewrites

$$\sigma^2(f) = \sum_{i=2}^{\mathsf{m}} \langle f, f_i\rangle_\pi^2 + 2\sum_{n=1}^{+\infty} \sum_{i=2}^{\mathsf{m}} \lambda_i^n \langle f, f_i\rangle_\pi^2 = \sum_{i=2}^{\mathsf{m}} \frac{1+\lambda_i}{1-\lambda_i} \langle f, e_i\rangle_\pi^2.$$

In particular, if all eigenvalues of P are nonnegative, then for any $f \in \mathbb{R}^E$, $\sigma^2(f) \geq \mathrm{Var}_\pi(f(X_0))$, so the convergence of the empirical mean of observables of the Markov chain is slower than the convergence of the empirical mean of iid samples of $f(X)$ with $X \sim \pi$. On the other hand, if there is a negative eigenvalue λ_i, then for the observable $f = f_i$, one gets $\sigma^2(f) < \mathrm{Var}_\pi(f(X_0))$, so the convergence in the Markov chain LLN is faster *than the convergence in the independent case.*

Chapter 7

The Markov chain Monte Carlo method

Contents

Let π be a probability measure on the finite space E. Assume that we want to either compute an expectation of the form

$$\mathcal{I} = \sum_{x \in E} f(x)\pi(x),$$

for some function $f : E \to \mathbb{R}$, or generate iid random variables $X_1, X_2, \ldots$ distributed according to π. Both tasks are virtually elementary because the finiteness of E allows them to be handled by a simple enumeration procedure. However when E is large, this procedure may have a computational cost which makes it impractical.

An alternative approach, called the *Markov chain Monte Carlo* (MCMC) method, consists in constructing a Markov chain $(X_n)_{n\geq 0}$ of which π is a stationary distribution, and using either the Law of Large Numbers and Central Limit Theorem from Chapter 5 to compute an estimator and a confidence interval for $\mathcal{I}$, or the convergence theorems from Chapter 6 to sample independent random variables $\widetilde{X}_1, \widetilde{X}_2, \ldots$ which are approximately distributed according to π by running independent realisations of the chain $(X_n)_{n\geq 0}$ on long enough times.

Throughout the chapter, we assume that the state space E is discrete, but not necessarily finite.

7.1 Gibbs measures

7.1.1 Definition and notation

A *Gibbs measure* is a probability measure π_β on E which writes under the form

$$\pi_\beta(x) = \frac{1}{Z_\beta}\mathrm{e}^{-\beta V(x)},$$

where $\beta > 0$ is the *inverse temperature* parameter, $V : E \to \mathbb{R} \cup \{+\infty\}$ is called the *potential* and

$$Z_\beta = \sum_{x \in E} \mathrm{e}^{-\beta V(x)}$$

is called the *partition function*. This terminology comes from statistical physics, where $\beta = (kT)^{-1}$ with T the temperature and k the Boltzmann constant.

Obviously, any probability measure π on E writes under this form, since it suffices to set $\beta = 1$ and $V(x) = -\ln \pi(x)$. In the sequel, we shall work under the following two assumptions which are related with the computational complexity of the underlying model:

(i) the function V is easy to evaluate;

(ii) the constant Z_β is not easy to compute.

These assumptions make the enumeration procedure discussed in the introduction impossible to implement.

7.1.2 Example: the Ising model

The Ising model[1] is a seemingly very simple model to describe ferromagnetism. In this model, the material is represented by an undirected graph $(\mathcal{V}, \mathcal{E})$ in which each vertex $v \in \mathcal{V}$ has a *spin* $x_v \in \{-1, 1\}$. Locally, the spins tend to align with their neighbours: the *potential* of a *configuration* $x = (x_v)_{v \in \mathcal{V}} \in E = \{-1, 1\}^{\mathcal{V}}$ is defined by

$$V(x) := -\sum_{\{v,w\} \in \mathcal{E}} x_v x_w,$$

so that configurations with lowest potential energy are those in which all spins have the same value. The Ising model is then the probability measure π_β defined on E by

$$\pi_\beta(x) = \frac{\mathrm{e}^{-\beta V(x)}}{Z_\beta}, \qquad Z_\beta = \sum_{x \in E} \mathrm{e}^{-\beta V(x)}.$$

Notice that the cardinality of E is $2^{|\mathcal{V}|}$, where $|\mathcal{V}|$ is the cardinality of $\mathcal{V}$. Thus, this quantity grows very fast as a function of $|\mathcal{V}|$, which explains why it is impossible, in practice, to compute Z_β.

The *magnetisation* of a configuration x is defined by

$$m(x) = \frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} x_v \in [-1, 1].$$

Exercise 7.1.1 (High- and low-temperature limits). *1. Show that π_0 is the uniform measure on E, under which $(x_v)_{v \in \mathcal{V}}$ is iid with $\pi_0(x_v = 1) = \pi_0(x_v = -1) = 1/2$.*

[1] See the introduction of https://cel.archives-ouvertes.fr/cel-00392289/ for a detailed presentation.

2. *Show that when $\beta \to +\infty$, π_β converges to*
$$\pi_\infty = \frac{1}{2}\delta_{x^+} + \frac{1}{2}\delta_{x^-},$$
where the configurations x^+ and x^- are defined by $x_v^\pm = \pm 1$ for any $v \in \mathcal{V}$.
3. *Describe the limit of m under π_0 when $|\mathcal{V}| \to +\infty$.*
4. *Describe the law of m under π_∞.*

The system's macroscopic behaviour, measured by the magnetisation, therefore seems radically different depending on temperature: at *high* temperature $\beta = (kT)^{-1} = 0$, spins behave independently from each other; at *low* temperature $\beta = \infty$, spins are strongly aligned with each other and macroscopic *droplets* appear. These two phases are illustrated in Figure 7.1 below. In the $|\mathcal{V}| \to +\infty$ limit, called the *thermodynamic limit*, this phase transition occurs abruptly (with respect to β). To describe this phenomenon more formally, let us start by specifying the graph $(\mathcal{V}, \mathcal{E})$ with which we work. Fix a dimension $d \geq 1$ and, for any $N \geq 1$, set

$$\mathcal{V}_{d,N} := (\mathbb{Z}/N\mathbb{Z})^d$$

the d-dimensional discrete torus, so that $|\mathcal{V}_{d,N}| = N^d$. Vertices are neighbours if their Euclidean distance is 1. We next set

$$m_N(\beta) := \mathbb{E}\left[|m(X)|\right], \qquad X \sim \pi_\beta,$$

and

$$m_*(\beta) := \lim_{N \to +\infty} m_N(\beta).$$

We do not justify the existence of this limit. Note that, by Exercise 7.1.1, $m_*(0) = 0$ and $m_*(+\infty) = 1$. The system is said to exhibit a *phase transition* if there is $\beta_\mathrm{c} \in (0, +\infty)$, called the *critical inverse temperature*, such that $m_*(\beta) = 0$ for $\beta < \beta_\mathrm{c}$ and $m_*(\beta) > 0$ for $\beta > \beta_\mathrm{c}$, as on the example of Figure 7.2.

In dimension $d = 1$, there is no phase transition since $m_*(\beta) = 0$ for $\beta \in [0, +\infty)$. This result was proved by Ising in 1925. In 1936, Peierls showed the existence of a phase transition for any dimension $d \geq 2$, and Onsager proved in 1944 that for $d = 2$,

$$m_*(\beta) = \max\left\{0, 1 - \left(\frac{2(1-p_\beta)}{p_\beta(2-p_\beta)}\right)^4\right\}^{1/8}, \qquad p_\beta := 1 - \mathrm{e}^{-2\beta}.$$

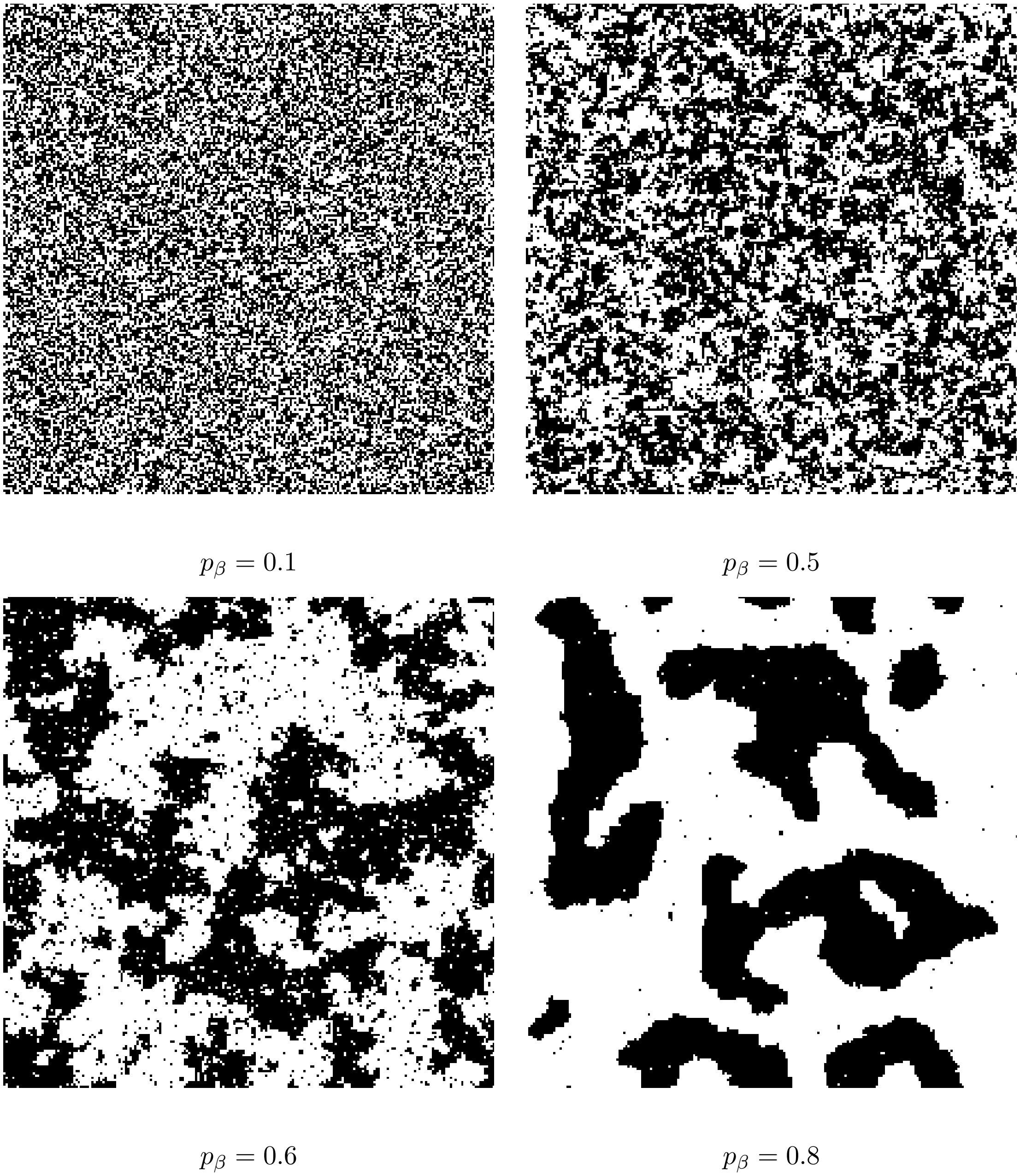
The phase transition occurs at the critical inverse temperature

$$\beta_c = \frac{1}{2}\ln(1+\sqrt{2}) \simeq 0.440687,$$

which corresponds to the value $\sqrt{2}/(1+\sqrt{2}) \simeq 0.585786$ for the parameter p_β. For $d \geq 3$, the value theoretical of the critical inverse temperature is not known and it has to be estimated by the Monte Carlo method.

7.2 The Metropolis algorithm

The Metropolis algorithm provides a method to construct a Markov chain which is reversible with respect to a given probability measure π on a finite space E. In the sequel, we shall still assume that $\pi(x) > 0$ for all $x \in E$. This is not a restrictive assumption since if $\pi(x) = 0$ then one may simply remove x from E.



$p_\beta = 0.1$

$p_\beta = 0.5$

$p_\beta = 0.6$

$p_\beta = 0.8$

Figure 7.1: Typical configurations of the Ising model in dimension $d = 2$ with $N = 250$, for different values of the parameter $p_\beta \in [0, 1]$. Droplets of aligned spins appear for $p_\beta > p_{\beta_c} \simeq 0.59$.

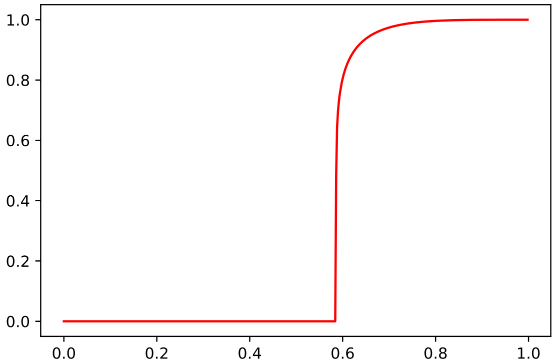


Figure 7.2: Value of the magnetisation $m_*(\beta)$ as a function of p_β, for the Ising model in dimension 2.

7.2.1 Definition and properties

The basic ingredients of the construction of the *Metropolis chain* $(X_n)_{n\geq 0}$ are:

- an irreducible stochastic matrix Q on E such that $Q(x,y) > 0$ if and only if $Q(y,x) > 0$, called the *proposal matrix*;
- an *acceptance function* $F : (0,+\infty) \to (0,1]$ such that

$$\forall \rho > 0, \qquad F(\rho) = \rho F(1/\rho). \tag{7.1}$$

Common acceptance functions are $F(\rho) = \min(\rho, 1)$ (the *Metropolis–Hastings rule*) and $F(\rho) = \rho/(1+\rho)$ (the *Barker rule*).

When the chain is in the state $x \in E$, the next state is computed as follows:

(i) draw a state y with probability $Q(x,y)$,

(ii) move the chain to y with probability

$$a(x,y) := F(r(x,y)), \qquad r(x,y) := \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)},$$

otherwise remain at x.

The condition on Q ensures that, almost surely, the ratio $r(x,y)$ takes its values in $(0,+\infty)$.

Proposition 7.2.1 (Reversibility of the Metropolis chain)**.** *The Metropolis chain $(X_n)_{n\geq 0}$ is irreducible and reversible with respect to π.*

As a consequence, π is the unique stationary distribution of $(X_n)_{n\geq 0}$ and all convergence results from Chapter 5 can be applied to this chain.

Proof. Let P denote the transition matrix of the Metropolis chain. It follows from the description of this chain that for all $x, y \in E$,

$$P(x,y) = \begin{cases} Q(x,y)a(x,y) & \text{if } x \neq y, \\ 1 - \sum_{z\neq x} Q(x,z)a(x,z) & \text{if } x = y. \end{cases}$$

We first check irreducibility. Let $x, y \in E$. Since Q is irreducible, there exist $n \geq 1$ and $x = x_0, \ldots, x_n = y \in E$ such that $Q(x_i, x_{i+1}) > 0$ for all $i \in \{0, \ldots, n-1\}$. Clearly, there is no loss of generality in assuming that $x_i \neq x_{i+1}$. Then, as a consequence of the assumption on

Q, we also have $Q(x_{i+1}, x_i) > 0$. Therefore all ratios $r(x_i, x_{i+1})$ are positive, and so are their images by F, so that

$$P^n(x, y) \geq \mathbb{P}_x(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=0}^{n-1} Q(x_i, x_{i+1}) a(x_i, x_{i+1}) > 0.$$

We now check reversibility. For all $x, y \in E$ such that $x \neq y$, the property $F(\rho) = \rho F(1/\rho)$ yields

$$\begin{aligned} \pi(x)P(x,y) &= \pi(x)Q(x,y)F\left(\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right) \\ &= \pi(x)Q(x,y)\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}F\left(\frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}\right) \\ &= \pi(y)Q(y,x)a(y,x) \\ &= \pi(y)P(y,x), \end{aligned}$$

which ensures that the detailed balance equation holds. □

🏠 **Exercise 7.2.2.** *Show that if Q is aperiodic, then P is also aperiodic.*

Remark 7.2.3. *If Q is already reversible with respect to π, the chain constructed with the Metropolis–Hastings rule has transition matrix $P = Q$, while the chain constructed with the Barker rule has transition matrix $P = (Q + I)/2$. In the latter case, P is aperiodic even if Q is not, which shows that the converse statement to Exercise 7.2.2 does not hold.*

If π has the form a Gibbs measure π_β as is described in Section 7.1, then simulating the Metropolis chain $(X_n)_{n\geq 0}$ requires to compute the ratio

$$r(x, y) = \mathrm{e}^{-\beta(V(y)-V(x))}\frac{Q(y,x)}{Q(x,y)},$$

which does not depend on the partition function Z_β but only on the potential V. Therefore, the chain $(X_n)_{n\geq 0}$ can be simulated in practice.

7.2.2 Optimality of the Metropolis–Hastings rule

Let $F : (0, +\infty) \to (0, 1]$ be an arbitrary function which satisfies the condition (7.1). Since $F(\rho) \leq 1$ by construction and $F(\rho) = \rho F(1/\rho) \leq \rho$ we have

$$\forall \rho > 0, \qquad F(\rho) \leq \min(\rho, 1) =: F_{\mathrm{MH}}(\rho),$$

with F_{MH} the *Metropolis–Hastings* (MH) acceptance rule. So, for a given proposal matrix Q, the MH acceptance rule accepts more jumps than any other acceptance function. As a consequence, the chain tends to explore the space more rapidly under the MH acceptance rule, so it may be expected to converge faster. This is indeed the case, and we shall illustrate this fact on both the rate of convergence to equilibrium and the asymptotic variance in the Markov chain CLT.

In the sequel, we fix a target measure π on the finite space E, and assume that the proposal matrix Q is irreducible and aperiodic. For an arbitrary acceptance function F which satisfies (7.1), we denote by P the transition matrix of the associated Metropolis chain $(X_n)_{n\geq 0}$, and let λ_2 be the second eigenvalue of P, and $\sigma^2(f)$ the asymptotic variance in the Markov chain CLT given by Proposition 6.2.7. When $F = F_{\mathrm{MH}}$, we use the notation $\lambda_{2,\mathrm{MH}}$ and $\sigma^2_{\mathrm{MH}}(f)$.

Theorem 7.2.4 (Optimality of the Metropolis–Hastings rule)**.** *With the notation introduced above, we have:*

(i) $\lambda_{2,\mathrm{MH}} \leq \lambda_2$;

(ii) *for any* $f \in \mathbb{R}^E$, $\sigma^2_{\mathrm{MH}}(f) \leq \sigma^2(f)$.

When $\lambda_\star = \lambda_2$ and $\lambda_{\star,\mathrm{MH}} = \lambda_{2,\mathrm{MH}}$[2], the assertion (i) means that the rate of convergence to equilibrium is faster for the MH rule. The assertion (ii) means that the fluctuations of $\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)$ around πf are smaller under the MH rule.

The proof of both assertions relies on the following remark. We recall that the Dirichlet form is introduced in Definition 5.2.11.

Lemma 7.2.5 (Comparison of Dirichlet forms)**.** *In the setting above, let $\mathcal{E}_\pi$ (resp. $\mathcal{E}_{\pi,\mathrm{MH}}$) denote the Dirichlet form associated with the Metropolis chain with arbitrary acceptance function (resp. with the MH acceptance rule). For any $f \in \mathbb{R}^E$,*

$$\mathcal{E}_\pi(f) \leq \mathcal{E}_{\pi,\mathrm{MH}}(f).$$

Proof. It directly follows from Definition 5.2.11, the proof of Proposition 7.2.1 and the remark made above that for any $f \in \mathbb{R}^E$,

$$\begin{aligned}\mathcal{E}_\pi(f) &= \frac{1}{2}\sum_{x\neq y}(f(y)-f(x))^2\pi(x)Q(x,y)F(r(x,y))\\ &\leq \frac{1}{2}\sum_{x\neq y}(f(y)-f(x))^2\pi(x)Q(x,y)F_{\mathrm{MH}}(r(x,y)) = \mathcal{E}_{\pi,\mathrm{MH}}(f).\end{aligned}$$ □

We first prove Theorem 7.2.4 (i).

Proof of Theorem 7.2.4 (i). For any $f \in \mathbb{R}^E$, by Lemma 5.2.12 and with the notation of Proposition 6.4.9,

$$\mathcal{E}_\pi(f) = \langle f, (I-P)f\rangle_\pi = \sum_{i=2}^{\mathsf{m}}(1-\lambda_i)\langle f, f_i\rangle_\pi$$

where we have used the fact that $\lambda_1 = 1$. Since $\lambda_i \leq \lambda_2$ for any $i \geq 2$, we deduce the bound

$$\mathcal{E}_\pi(f) \geq (1-\lambda_2)\|f - \pi f\|^2_\pi,$$

which is reached for $f = f_2$. Therefore, the quantity $1-\lambda_2$ admits the variational formulation

$$1-\lambda_2 = \inf_{f:\mathrm{Var}_\pi(f(X_0))>0}\frac{\mathcal{E}_\pi(f)}{\mathrm{Var}_\pi(f(X_0))}.$$

Combining this formulation with Lemma 7.2.5 leads to the claimed inequality. □

The proof of Theorem 7.2.4 (ii) is due to Peskun[3]. It is based on the following variational formulation of $\sigma^2(f)$.

Proposition 7.2.6 (Variational formulation of $\sigma^2(f)$)**.** *Under the assumptions of Proposition 6.2.7 and if, in addition, P is reversible with respect to π, then*

$$\sigma^2(f) = \sup_{g\in\mathbb{R}^E_0}\{4\langle f, g\rangle_\pi - 2\mathcal{E}_\pi(g)\} - \mathrm{Var}_\pi(f(X_0)),$$

where we recall that $\mathbb{R}^E_0$ is the set of functions $g \in \mathbb{R}^E$ such that $\pi g = 0$.

[2] This happens in particular if P and P_{MH} have nonnegative eigenvalues.

[3] Peskun, P.H., Optimal Monte-Carlo sampling using Markov chains. *Biometrika*, 1973

Proof. By Proposition 6.2.9,

$$\sigma^2(f) = 2\langle \widetilde{f}, (I_0 - P_0)^{-1}\widetilde{f}\rangle_\pi - \mathrm{Var}_\pi(f(X_0)). \tag{7.2}$$

We first establish the variational formula

$$\langle \widetilde{f}, (I_0 - P_0)^{-1}\widetilde{f}\rangle_\pi = \sup_{g\in\mathbb{R}_0^E} 2\langle g, \widetilde{f}\rangle_\pi - \langle g, (I_0 - P_0)g\rangle_\pi. \tag{7.3}$$

Let $\mathcal{J}_{\widetilde{f}} : \mathbb{R}_0^E \to \mathbb{R}$ be defined by $\mathcal{J}_{\widetilde{f}}(g) = 2\langle g, \widetilde{f}\rangle_\pi - \langle g, (I_0 - P_0)g\rangle_\pi$. Since P is reversible with respect to π, Proposition 6.4.9 applies and shows that $\mathbb{R}_0^E = \mathrm{Span}(f_i, 2 \leq i \leq \mathsf{m})$, so that in particular, $I_0 - P_0$ is positive definite on $(\mathbb{R}_0^E, \langle\cdot,\cdot\rangle_\pi)$. Moreover,

$$\nabla_\pi \mathcal{J}_{\widetilde{f}}(g) = 2\widetilde{f} - 2(I_0 - P_0)g,$$

where ∇_π denotes the gradient on the space $(\mathbb{R}_0^E, \langle\cdot,\cdot\rangle_\pi)$. We deduce that $\nabla_\pi \mathcal{J}_{\widetilde{f}}(g) = 0$ if and only if $g = (I_0 - P_0)^{-1}\widetilde{f}$, which then yields (7.3).

Combining (7.2) and (7.3), we deduce that

$$\sigma^2(f) = \sup_{g\in\mathbb{R}_0^E} \left\{4\langle g, \widetilde{f}\rangle_\pi - 2\langle g, (I_0 - P_0)g\rangle_\pi\right\} - \mathrm{Var}_\pi(f(X_0)).$$

For any $g \in \mathbb{R}_0^E$,

$$\langle g, \widetilde{f}\rangle_\pi = \langle g, f - \pi f\rangle_\pi = \langle g, f\rangle_\pi - \pi g\, \pi f = \langle g, f\rangle_\pi,$$

while, for any $g \in \mathbb{R}_0^E$,

$$\langle g, (I_0 - P_0)g\rangle_\pi = \langle g, (I - P)g\rangle_\pi = \mathcal{E}_\pi(g),$$

thanks to Lemma 5.2.12. The conclusion follows. □

The proof of Theorem 7.2.4 (ii) immediately follows from the combination of Proposition 7.2.6 with Lemma 7.2.5.

7.2.3 Application to the Ising model

We take as a proposal matrix Q the transition matrix of the Markov chain which at each step picks a vertex u uniformly in $\mathcal{V}$ and flips its spin. Defining, for any $x \in E$ and $u \in \mathcal{V}$, the configuration x^u by

$$\forall v \in \mathcal{V}, \qquad x^u_v = \begin{cases} x_v & \text{if } v \neq u, \\ -x_v & \text{if } v = u, \end{cases}$$

the matrix Q writes

$$Q(x,y) = \begin{cases} \dfrac{1}{|\mathcal{V}|} & \text{if there exists } u \in \mathcal{V} \text{ such that } y = x^u, \\ 0 & \text{otherwise.} \end{cases}$$

It satisfies the assumptions of Subsection 7.2.1. Besides, it is easily seen that $Q(x,y) = Q(y,x)$ for any $x, y \in E$. The acceptance ratio therefore rewrites, for $y = x^u$,

$$a(x, x^u) = \frac{\pi_\beta(x^u)}{\pi_\beta(x)} = \exp\left(-\beta(V(x^u) - V(x))\right).$$

From the explicit definition of the Ising potential V and the definition of x^u, we get

$$\begin{aligned} V(x^u) - V(x) &= -\sum_{\{v,w\}\in\mathcal{E}} x^u_v x^u_w + \sum_{\{v,w\}\in\mathcal{E}} x_v x_w \\ &= -\sum_{v\in\mathcal{V}:\{u,v\}\in\mathcal{E}} x^u_u x^u_v + \sum_{v\in\mathcal{V}:\{u,v\}\in\mathcal{E}} x_u x_v \\ &= 2x_u \Sigma(x,u), \end{aligned}$$

with

$$\Sigma(x,u) = \sum_{v\in\mathcal{V}:\{u,v\}\in\mathcal{E}} x_v.$$

Thus, under the Metropolis–Hastings rule, the spin x_u is changed to $-x_u$ with probability

$$\min\left\{1, \exp(2\beta x_u \Sigma(x,u))\right\};$$

under the Barker rule, the spin x_u is changed to $-x_u$ with probability

$$\frac{1}{1+\exp(-2\beta x_u \Sigma(x,u))}.$$

7.2.4 Simulated annealing for optimisation

Consider the Gibbs measure

$$\pi_\beta(x) = \frac{1}{Z_\beta}\mathrm{e}^{-\beta V(x)},$$

for some function $V : E \to \mathbb{R}$. When $\beta \to +\infty$, π_β converges to the uniform distribution on the set

$$\operatorname{argmin} V := \{x \in E : \forall y \in E, V(y) \geq V(x)\}.$$

If one is interested in finding the global minima of V, then a first approach may consist in taking a 'large' value of β, constructing a Metropolis chain $(X_n)_{n\geq 0}$ reversible with respect to the Gibbs measure π and running it on a long enough time for X_n to be essentially concentrated on the global minima of V.

Observe that if the algorithm uses the Metropolis–Hastings rule with a symmetric proposal matrix Q, then the probability to accept a move from x to y rewrites $\exp(-\beta[V(y)-V(x)]_+)$ and the following two phenomena occur.

- Moves that make the value of $V(X_n)$ decrease are always accepted. This brings the chain toward 'local minima' of V on a short time scale, in accordance with the idea of gradient descent algorithms. Here, the notion of a 'local' minimum has to be understood with respect to the graph structure induced on E by the pairs (x,y) such that $Q(x,y) > 0$.
- Moves that make the value of $V(X_n)$ increase are accepted with an exponentially small (but nonetheless positive) probability. This allows the chain to 'escape' local minima on long time scales and go exploring other local minima. This behaviour is an essential feature of stochastic algorithms.

The idea of *simulated annealing* is a refinement of the Metropolis algorithm, in which the parameter β increases with time. Given a proposal matrix Q on E, an acceptance function F, and a deterministic sequence $(\beta_n)_{n\geq 1}$ growing to $+\infty$ which we call a *cooling scheme*, it can be described as follows: for all $n \geq 0$, given the current state $X_n = x \in E$,

(i) select a state y with probability $Q(x,y)$;

(ii) set $X_{n+1} = y$ with probability

$$a_n(x, y) = F\left(\frac{\pi_{\beta_{n+1}}(y)Q(y, x)}{\pi_{\beta_{n+1}}(x)Q(x, y)}\right),$$

otherwise set $X_{n+1} = x$.

The resulting sequence $(X_n)_{n\geq 0}$ is an inhomogeneous Markov chain. It can be shown that under some assumptions on V, there exist cooling schemes for which $V(X_n)$ converges to $V_{\min} := \min_{x\in E} V(x)$. More precisely, we have the following statement.

Theorem 7.2.7 (). *For any function $V : E \to \mathbb{R}$ and proposal matrix Q, there is a constant C such that the sequence $(X_n)_{n\geq 0}$ constructed with cooling scheme $\beta_n = C\ln(n)$ satisfies*

$$\lim_{n\to+\infty} \mathbb{P}\left(V(X_n) = V_{\min}\right) = 1.$$

We refer to [3, Chapitre 2], [2, Chapitre 5.3] for details.

7.3 The Gibbs sampler

The *Gibbs sampler algorithm* is a MCMC method which provides an alternative to the Metropolis algorithm. It is designed for probability measures π on state spaces E which have the specific form $E = S^{\mathcal{V}}$, where S and $\mathcal{V}$ are finite spaces.

7.3.1 Definition and properties

By analogy with the Ising model, we shall keep calling elements u of $\mathcal{V}$ *vertices*, and denoting configurations $x \in E$ by $x = (x_u)_{u\in\mathcal{V}}$, where $x_u \in S$ is the *spin* of the vertex u. Given a configuration $x \in E$, a vertex $u \in \mathcal{V}$ and a possible value s for the spin, we denote by $x^{u,s}$ the configuration defined by

$$\forall v \in \mathcal{V}, \qquad x_v^{u,s} = \begin{cases} s & \text{if } v = u,\\ x_v & \text{otherwise,}\end{cases}$$

and let

$$E_{x,u} = \{x^{u,s} : s \in S\}$$

be the set of configurations which can be obtained by changing the value of the spin x_u.

Definition 7.3.1 (Gibbs sampler). *Let π be a probability measure on $E = S^{\mathcal{V}}$, such that $\pi(x) > 0$ for all $x \in E$. The* Gibbs sampler *of π is the Markov chain in E defined by, at each step:*

(i) picking a vertex $u \in \mathcal{V}$ uniformly;

(ii) selecting the new value of the spin x_u according to the conditional probability $\pi(\cdot|E_{x,u})$, where x is the current configuration.

Let us provide more detail on the update of the spin x_u, in the case where $\pi = \pi_\beta$ as in Section 7.1. For all $s \in S$, the spin x_u is updated to the value s with probability

$$\begin{aligned}\pi_\beta(x^{u,s}|E_{x,u}) &= \frac{\pi_\beta(x^{u,s})}{\sum_{s'\in S}\pi_\beta(x^{u,s'})}\\ &= \frac{\mathrm{e}^{-\beta V(x^{u,s})}/Z_\beta}{\sum_{s'\in S}\mathrm{e}^{-\beta V(x^{u,s'})}/Z_\beta}\\ &= \frac{1}{1+\sum_{s'\in S\setminus\{s\}}\mathrm{e}^{-\beta(V(x^{u,s'})-V(x^{u,s}))}}.\end{aligned}$$

This identity shows that it is not necessary to know the value of the partition function Z_β to compute the conditional probability $\pi_\beta(\cdot|E_{x,u})$; instead, only $|S|$ evaluations of the potential V are used.

Remark 7.3.2 (Graph structure for $\mathcal{V}$)**.** *This evaluation is particularly fast when the potential V depends on a geometrical structure of the set $\mathcal{V}$. Assume indeed that, as in the Ising model, the latter is the set of vertices of an undirected graph with set of edges $\mathcal{E}$, and that the potential writes under the form*

$$V(x) = \sum_{\{u,v\} \in \mathcal{E}} w(x_u, x_v),$$

for some symmetric function $w : S \times S \to \mathbb{R}$. Then for all $x \in E$, $u \in \mathcal{V}$, and $s, s' \in S$,

$$V(x^{u,s}) - V(x^{u,s'}) = 2 \sum_{v : \{u,v\} \in \mathcal{E}} w(s, x_v) - w(s', x_v),$$

which makes the computation of the conditional probability $\pi_\beta(\cdot|E_{x,u})$ local *in the sense that it only depends on the spins of the neighbouring vertices v of u in the configuration x.*

The interest of the Gibbs sampler is given by the following result.

Proposition 7.3.3 (Reversibility)**.** *Under the assumptions of Definition 7.3.1, the Gibbs sampler of π is irreducible, aperiodic and reversible with respect to π.*

🗎 **Exercise 7.3.4.** *Prove the irreducibility and aperiodicity properties.*

Proof of reversibility. From Definition 7.3.1, we deduce that the transition matrix P of the Gibbs sampler writes, for all $x, y \in E$ with $x \neq y$,

$$P(x,y) = \begin{cases} \dfrac{1}{|\mathcal{V}|}\pi(y|E_{x,u}) & \text{if there exists } u \in \mathcal{V} \text{ such that } y \in E_{x,u}, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, for all $x, y \in E$ and $u \in \mathcal{V}$, we have $y \in E_{x,u}$ if and only if $x \in E_{y,u}$, and in this case $E_{x,u} = E_{y,u}$. As a consequence, in such a case,

$$\pi(x)P(x,y) = \pi(x)\frac{1}{|\mathcal{V}|}\pi(y|E_{x,u}) = \frac{\pi(x)\pi(y)}{|\mathcal{V}|\pi(E_{x,u})} = \frac{\pi(x)\pi(y)}{|\mathcal{V}|\pi(E_{y,u})} = \pi(y)P(y,x),$$

which is the detailed balance condition needed to prove reversibility. If there is no $u \in \mathcal{V}$ such that $y \in E_{x,u}$, or equivalently $x \in E_{y,u}$, then $P(x,y) = 0 = P(y,x)$ which makes the detailed balance also hold in this case, and thus completes the proof. □

☕ **Exercise 7.3.5** (Updating all components at each step)**.** *For the sake of clarity we assume that $\mathcal{V} = \{1, \ldots, V\}$ for some integer $V \geq 1$. Let M_V be a probability measure*[4] *on the set of permutations of $\mathcal{V}$. We consider the following algorithm: given $X_n = (x_1, \ldots, x_V) \in E$,*

- *draw a random permutation σ of $\mathcal{V}$ under M_V;*
- *for $u = 1, \ldots, V$, update the spin $x'_{\sigma(u)}$ according to $\pi(\cdot|x'_{\sigma(1)}, \ldots, x'_{\sigma(u-1)}, x_{\sigma(u+1)}, \ldots, x_{\sigma(V)})$*[5]*;*

[4] There are two natural choices for M_V: either the uniform measure, so at each step of the algorithm the order in which the coordinates are updated is chosen uniformly, or the Dirac measure at some permutation, which without loss of generality may be taken to be the identity, so the coordinates are always updated in the same order.

[5] That is to say, according to $\pi(\cdot|E')$ with $E' := \{y \in E : y_{\sigma(1)} = x'_{\sigma(1)}, \ldots, y_{\sigma(u-1)} = x'_{\sigma(u-1)}, y_{\sigma(u+1)} = x_{\sigma(u+1)}, \ldots, y_{\sigma(V)} = x_{\sigma(V)}\}$.

- *set* $X_{n+1} = x'$.

Under the assumption that $\pi(x) > 0$ *for any* $x \in E$*, the Markov chain* $(X_n)_{n\geq 0}$ *remains irreducible and aperiodic.*

1. *For any* $v \in \mathcal{V}$*, let us set*

$$\widetilde{P}_v(x,y) = \begin{cases} \pi(y|E_{x,u}) & \text{if } y \in E_{x,u}, \\ 0 & \text{otherwise.} \end{cases}$$

   Write the transition matrix P *of* $(X_n)_{n\geq 0}$ *in terms of the stochastic matrices* $\{\widetilde{P}_v : v \in \mathcal{V}\}$*.*
2. *Show that* π *is stationary for* $(X_n)_{n\geq 0}$*.*
3. *Give a sufficient condition over* M_V *which ensures that* P *is reversible with respect to* π*.*
4. *Why would one prefer to employ this algorithm rather than the one described in Definition 7.3.1?*

7.3.2 Application to the Ising model

When the chain is in a configuration x, a vertex u is picked uniformly in $\mathcal{V}$ and the spin is updated according to the probabilities

$$p(+|x,u) := \frac{\mathrm{e}^{-\beta V(x^{u,+})}}{\mathrm{e}^{-\beta V(x^{u,+})} + \mathrm{e}^{-\beta V(x^{u,-})}}, \qquad p(-|x,u) := \frac{\mathrm{e}^{-\beta V(x^{u,-})}}{\mathrm{e}^{-\beta V(x^{u,+})} + \mathrm{e}^{-\beta V(x^{u,-})}}.$$

With the notation $\Sigma(x,u)$ introduced in Subsection 7.2.3, these probabilities rewrite

$$p(+|x,u) = \frac{\mathrm{e}^{\beta\Sigma(x,u)}}{\mathrm{e}^{\beta\Sigma(x,u)} + \mathrm{e}^{-\beta\Sigma(x,u)}}, \qquad p(-|x,u) = \frac{\mathrm{e}^{-\beta\Sigma(x,u)}}{\mathrm{e}^{\beta\Sigma(x,u)} + \mathrm{e}^{-\beta\Sigma(x,u)}}.$$

Exercise 7.3.6. *Check that, on the example of the Ising model, this algorithm coincides with the Metropolis algorithm applied with the Barker rule.*

Chapter 8

Markov chains in continuous spaces

Contents

8.1 Markov kernels and the Markov property

8.2 Harris recurrence

8.3 Convergence to equilibrium

lypounov, approche Hairer–Mattingly

parler aussi de réversibilité?

Part III

Diffusion processes and partial differential equations

Chapter 9

Stochastic processes and Brownian motion

Contents

In this last part of the course we introduce *diffusion processes*, a class of continuous-time processes which also allow to compute integrals of the form

$$\mathcal{I} = \int_{x\in\mathbb{R}^d} f(x)p(x)\mathrm{d}x$$

by means of continuous-time ergodic averages. In fact diffusion processes have a very rich theory and can for instance be used to provide a probabilistic representation of solutions to certain partial differential equations, thereby establishing a connection between the Monte Carlo method and the numerical analysis of such partial differential equations.

9.1 Generalities on stochastic processes

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, I be an arbitrary set of indices, and $(E, \mathcal{E})$ be a measurable space.

9.1.1 Stochastic processes

Definition 9.1.1 (Stochastic process)**.** *A* stochastic process *indexed by I with values in E is a family $X = (X_t)_{t\in I}$ such that for all $t \in I$,*

$$X_t : \begin{cases} \Omega & \to & E \\ \omega & \mapsto & X_t(\omega) \end{cases}$$

is a real-valued random variable.

A stochastic process can also be seen as a random variable with values in the space E^I of functions $I \to E$, as soon as the latter space is endowed with a suitable σ-field.

Definition 9.1.2 (Product σ-field)**.** *For any $k \geq 1$, $t_1, \dots, t_k \in I$, $B_1, \dots, B_k \in \mathcal{E}$, the set*

$$C(t_1, \dots, t_k; B_1, \dots, B_k) = \left\{(x_t)_{t\in I} \in E^I : x_{t_1} \in B_1, \dots, x_{t_k} \in B_k\right\}$$

is called a cylinder. *The* product σ-field *on $\mathbb{R}^I$, denoted by $\mathcal{E}^{\otimes I}$, is defined as the smallest σ-field on E^I containing all cylinders.*

Lemma 9.1.3 (Equivalence of definitions)**.** *Let $(X_t)_{t\in I}$ a family of functions $X_t : \Omega \to E$. The function*

$$X : \begin{cases} \Omega & \to & E^I \\ \omega & \mapsto & (X_t(\omega))_{t\in I} \end{cases}$$

is measurable for the product σ-field if and only if, for any $t \in I$, the function

$$X_t : \begin{cases} \Omega & \to & E \\ \omega & \mapsto & X_t(\omega) \end{cases}$$

is measurable.

🏠 **Exercise 9.1.4.** *Prove Lemma 9.1.3.*

It is therefore equivalent to speak about stochastic processes in the sense of Definition 9.1.1, or to speak about E^I-valued random variables. As such, a stochastic process possesses a *law* (we shall equivalently say a *distribution*), which is a probability measure on the measurable space $(E^I, \mathcal{E}^{\otimes I})$. This law is characterised by the following result, which is an immediate consequence of the Dynkin System Theorem (see Lemma 1.1.4).

Proposition 9.1.5 (Characterisation of the law of a process)**.** *Let $X = (X_t)_{t\in I}$ and $Y = (Y_t)_{t\in I}$ be two stochastic processes such that, for any $k \geq 1$ and $t_1, \dots, t_k \in I$, the random vectors $(X_{t_1}, \dots, X_{t_k})$ and $(Y_{t_1}, \dots, Y_{t_k})$ have the same law in E^k. Then the processes X and Y have the same law in E^I.*

The family of laws of vectors of the form $(X_{t_1}, \dots, X_{t_k})$, $t_1, \dots, t_k \in I$ is called the system of *finite-dimensional distributions* of the process $(X_t)_{t\in I}$. The next result follows from similar arguments.

Proposition 9.1.6 (Characterisation of independence)**.**

(i) A stochastic process $X = (X_t)_{t\in I}$ is independent from a random variable Y if and only if for any $k \geq 1$ and $t_1, \dots, t_k \in I$, the random vector $(X_{t_1}, \dots, X_{t_k})$ is independent from Y.

(ii) Two stochastic processes $X = (X_t)_{t\in I}$ and $Y = (Y_s)_{s\in J}$ (with J an interval of $\mathbb{R}$ which may differ from I, and which may take their values in different measurable spaces) are independent if and only if for any $k, l \geq 1$, $t_1, \dots, t_k \in I$, $s_1, \dots, s_l \in J$, the random vectors $(X_{t_1}, \dots, X_{t_k})$ and $(Y_{s_1}, \dots, Y_{s_l})$ are independent.

The second statement readily generalises to an arbitrary family of processes.

9.1.2 Gaussian processes

We recall that a random vector $X = (X_1, \ldots, X_k) \in \mathbb{R}^k$ is *Gaussian* if any linear combination of the variables $X_1, \ldots, X_k$ is Gaussian. The law of a Gaussian vector X is then characterised by two quantities: its expectation $\mathbb{E}[X] \in \mathbb{R}^k$ and its covariance matrix $\mathrm{Cov}[X] \in \mathbb{R}^{k \times k}$.

Definition 9.1.7 (Gaussian process)**.** *A real-valued stochastic process $(X_t)_{t \in I}$ is called* Gaussian *if, for any $t_1, \ldots, t_k \in I$, the vector $(X_{t_1}, \ldots, X_{t_k})$ is Gaussian.*

Following Proposition 9.1.5, the law of a Gaussian process is then characterised by two functions: its expectation $\mathsf{m}(t) = \mathbb{E}[X_t]$ and its covariance $\mathsf{c}(s,t) = \mathrm{Cov}(X_s, X_t)$.

9.2 The Brownian motion

9.2.1 Definition

Random walk

Let $\Delta t > 0$, $\Delta x > 0$, and $(\xi_n)_{n \geq 1}$ be a sequence of iid random variables such that $\mathbb{P}(\xi_1 = 1) = \mathbb{P}(\xi_1 = -1) = 1/2$. Let us consider the stochastic process $B^{\Delta t, \Delta x} = (B^{\Delta t, \Delta x}_t)_{t \geq 0}$ constructed as follows:

- $B^{\Delta t, \Delta x}_0 = 0$;
- for any $n \geq 1$, $B^{\Delta t, \Delta x}_{n\Delta t} = B^{\Delta t, \Delta x}_{(n-1)\Delta t} + \xi_n \Delta x$, and the function $t \mapsto B^{\Delta t, \Delta x}_t$ is linear on $[(n-1)\Delta t, n\Delta t]$.

This is a scaled version, with time step Δt and space step Δx, of the one-dimensional *random walk* introduced in Chapter 5. Several realisations of its trajectory are drawn on Figure 9.1.

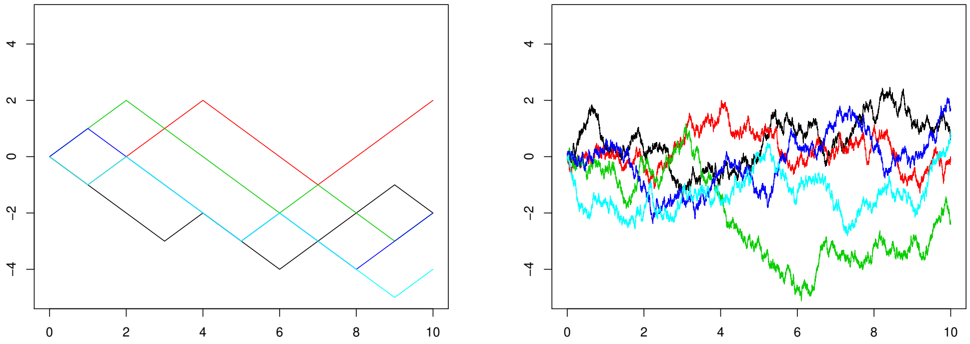


Figure 9.1: Left: five trajectories of the random walk with time step $\Delta t = 1$ and space step $\Delta x = 1$. Right: five trajectories of the Brownian motion on the interval $[0, 10]$.

Exercise 9.2.1. *For any $n \geq 0$, compute $\mathbb{E}[B^{\Delta t, \Delta x}_{n\Delta t}]$ and $\mathrm{Var}(B^{\Delta t, \Delta x}_{n\Delta t})$.*

Definition of the Brownian motion

Exercise 9.2.1 shows that, for any $t \geq 0$, the random variable $B^{\Delta t, \Delta x}_t$ is of the order of magnitude $\sqrt{t \Delta x^2 / \Delta t}$. For this random variable to possess a limit when Δt and Δx go to 0, it is therefore necessary that the ratio $\Delta x^2/\Delta t$ remain non degenerate. We take the convention to keep it equal to 1.

Exercise 9.2.2. *Show that, when $\Delta t \to 0$ and $\Delta x = \sqrt{\Delta t}$:*

1. *for any $t \geq 0$, the random variable $B_t^{\Delta t, \Delta x}$ converges in distribution toward a random variable $B_t \sim \mathcal{N}(0,t)$;*
2. *for all $0 \leq s \leq t$, the pair $(B_s^{\Delta t, \Delta x}, B_t^{\Delta t, \Delta x})$ converges in distribution to a Gaussian vector (B_s, B_t) with covariance $\mathbb{E}[B_s B_t] = s$.*

Exercise 9.2.2 motivates the following definition.

Definition 9.2.3 (Brownian motion)**.** *A real-valued Brownian motion is a Gaussian process $(B_t)_{t\geq 0}$ with expectation*

$$\mathsf{m}(t) = \mathbb{E}[B_t] = 0$$

and covariance function

$$\mathsf{c}(s,t) = \mathbb{E}[B_s B_t] = s \wedge t.$$

In particular, if $(B_t)_{t\geq 0}$ is a Brownian motion, then $B_0 = 0$, almost surely, and for any $t \geq 0$, $B_t \sim \mathcal{N}(0,t)$.

By the results of Section 9.1, Definition 9.2.3 exactly characterises the law (in $\mathbb{R}^{[0,+\infty)}$) of a Brownian motion.

We shall sometimes consider Brownian motions on intervals of the form $[0,T]$ rather than on $[0,+\infty)$, this does not change their definition.

Exercise 9.2.4. *Let G be a random variable with law $\mathcal{N}(0,1)$. For any $t \geq 0$, we set $X_t = \sqrt{t}G$. We also let $(B_t)_{t\geq 0}$ be a Brownian motion.*

1. *Show that, for any $t \geq 0$, the variables X_t and B_t have the same law.*
2. *Show that the process $(X_t)_{t\geq 0}$ is Gaussian and compute its covariance function.*
3. *Do the processes $(X_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ have the same law?*

Exercise 9.2.5 (Transformations of the Brownian motion)**.** *Let $(B_t)_{t\geq 0}$ be a Brownian motion.*

1. *Show that $(-B_t)_{t\geq 0}$ is a Brownian motion.*
2. *For any $c > 0$, show that $(c^{-1/2}B_{ct})_{t\geq 0}$ is a Brownian motion.*
3. *For any $T > 0$, show that $(B_T - B_{T-t})_{t\in[0,T]}$ is a Brownian motion (on $[0,T]$).*
4. *Show that the process $(\widetilde{B}_t)_{t\geq 0}$ defined by*

$$\widetilde{B}_t := \begin{cases} 0 & \text{if } t = 0, \\ tB_{1/t} & \text{if } t > 0, \end{cases}$$

is a Brownian motion.

9.2.2 Increments of the Brownian motion and the Markov property

Exercise 9.2.6. *Let $(B_t)_{t\geq 0}$ be a Brownian motion. Show that for any $0 \leq s \leq t$, the random variables B_s and $B_t - B_s$ are independent, with $B_t - B_s \sim \mathcal{N}(0, t-s)$.*

The increments of the Brownian motion are therefore said to be:

- *stationary*, because for any $0 \leq s \leq t$, the random variables $B_t - B_s$ and B_{t-s} have the same law;
- *independent*, because it is easily deduced by induction that for any $0 = t_0 \leq t_1 \leq \cdots \leq t_k$, the random variables $(B_{t_i} - B_{t_{i-1}})_{1\leq i\leq k}$ are independent.

🗎 **Exercise 9.2.7.** *Show that, conversely, if a random process* $(X_t)_{t\geq 0}$ *is such that* $X_0 = 0$ *almost surely, and for any* $0 = t_0 \leq t_1 \leq \cdots \leq t_k$*, the random variables* $(X_{t_i} - X_{t_{i-1}})_{1\leq i\leq k}$ *are independent with respective law* $\mathcal{N}(0, t_i - t_{i-1})$*, then* $(X_t)_{t\geq 0}$ *is a Brownian motion.*

🏠 **Exercise 9.2.8.** *Show that the* Poisson process $N_t := \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \leq t\}}$*, where* $0 = T_0 < T_1 < \cdots$ *is a random sequence such that the variables* $(T_i - T_{i-1})_{i\geq 1}$ *are iid according to the exponential distribution with parameter* $\lambda > 0$*, has stationary and independent increments.*

Following Proposition 9.1.6, it is not difficult to show that if a random process $(X_t)_{t\geq 0}$ has stationary and independent increments, then for any $t_0, t \geq 0$, the random variable $X'_t := X_{t_0+t} - X_{t_0}$ is independent from the process $(X_r)_{r\in[0,t_0]}$ and has the same law as X_t. For the Brownian motion, this results holds at the level of the *process* $(X'_t)_{t\geq 0}$.

Proposition 9.2.9 (Markov property)**.** *Let* $(B_t)_{t\geq 0}$ *be a Brownian motion. For any* $t_0 \geq 0$*, the process* $(B'_t)_{t\geq 0}$ *defined by* $B'_t := B_{t_0+t} - B_{t_0}$ *is a Brownian motion, independent from* $(B_r)_{r\in[0,t_0]}$*.*

Proof. We start by noting that, for all $t_1, \ldots, t_k \geq 0$, any linear combination of the coordinates of the vector $(B'_{t_1}, \ldots, B'_{t_k})$ is a linear combination of $B_{t_0}, B_{t_0+t_1}, \ldots, B_{t_0+t_k}$, therefore B' is a Gaussian process. Clearly, $\mathbb{E}[B'_t] = \mathbb{E}[B_{t_0+t}] - \mathbb{E}[B_{t_0}] = 0$ and for all $s \leq t$,

$$\begin{aligned}
\mathrm{Cov}(B'_s, B'_t) &= \mathrm{Cov}(B_{t_0+s} - B_{t_0}, B_{t_0+t} - B_{t_0})\\
&= \mathrm{Cov}(B_{t_0+s}, B_{t_0+t}) - \mathrm{Cov}(B_{t_0+s}, B_{t_0}) - \mathrm{Cov}(B_{t_0}, B_{t_0+t}) + \mathrm{Var}(B_{t_0})\\
&= (t_0 + s) \wedge (t_0 + t) - (t_0 + s) \wedge t_0 - t_0 \wedge (t_0 + t) + t_0\\
&= t_0 + s \wedge t - t_0 - t_0 + t_0\\
&= s \wedge t.
\end{aligned}$$

Hence, by Definition 9.2.3, B' is a Brownian motion.

For the same reason, for all $r_1, \ldots, r_l \in [0, t_0]$, the vector $(B_{r_1}, \ldots, B_{r_l}, B'_{t_1}, \ldots, B'_{t_k})$ is Gaussian, and for all $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, l\}$,

$$\begin{aligned}
\mathrm{Cov}(B_{r_j}, B'_{t_i}) &= \mathrm{Cov}(B_{r_j}, B_{t_0+t_j}) - \mathrm{Cov}(B_{r_j}, B_{t_0})\\
&= r_j \wedge (t_0 + t_j) - r_j \wedge t_0\\
&= r_j - r_j\\
&= 0,
\end{aligned}$$

which by Proposition 2.2.17 shows that the vectors $(B_{r_1}, \ldots, B_{r_l})$ and $(B'_{t_1}, \ldots, B'_{t_k})$ are independent. As a consequence, Proposition 9.1.6 implies that the processes $(B'_t)_{t\geq 0}$ and $(B_r)_{r\in[0,t_0]}$ are independent. □

🏠 **Exercise 9.2.10** (Quadratic variation of the Brownian motion)**.** *For all* $n \geq 1$*, let us denote by* $0 = t_0 < t_1 < \cdots < t_n = T$ *the subdivision of the interval* $[0, T]$ *into regular intervals with length* T/n*.*

1. *Show that if* $(B_t)_{t\in[0,T]}$ *is a Brownian motion,*

$$\lim_{n\to+\infty} \sum_{i=0}^{n-1} (B_{t_{i+1}} - B_{t_i})^2 = T, \qquad \textit{in } \mathbf{L}^2.$$

2. *What happens if* $(B_t)_{t\in[0,T]}$ *is replaced by a* C^1 *function* $g : [0, T] \to +\infty$*?*

Exercise 9.2.10 indicates that Brownian trajectories do not behave as C^1 function. The regularity of these trajectories is the subject of the next subsection.

9.2.3 Trajectories of the Brownian motion

So far, we have not said anything about the regularity of the trajectory $t \mapsto B_t(\omega)$ for a given $\omega \in \Omega$. As the right-hand side picture of Figure 9.1 seems to indicate, we shall see that these trajectories are (almost surely) continuous, but not much more regular. These results will be stated without a proof, and their mere formulation already requires some care.

Modification and the Kolmogorov criterion

In this paragraph, we temporarily leave the case of the Brownian motion aside and come back to the general setting of Section 9.1. In particular, I is an arbitrary set of indices.

Definition 9.2.11 (Modification)**.** *Let $(X_t)_{t\in I}$ be a stochastic process. A* modification *of $(X_t)_{t\in I}$ is a stochastic process $(Y_t)_{t\in I}$ such that*

$$\forall t \in I, \qquad \mathbb{P}(X_t = Y_t) = 1.$$

In general, the almost sure event on which $X_t = Y_t$ depends on t. Therefore, the statement that X and Y are modifications of each other is much weaker than the fact that $\mathbb{P}(\forall t \in I, X_t = Y_t) = 1$, in which case X and Y are called *indistinguishable*.

Exercise 9.2.12. *Show that if X and Y are modifications of each other, they have the same law in $\mathbb{R}^I$.*

Exercise 9.2.13. *Let $\Omega = [0,1]$ be equipped with its Borel σ-field and denote by $\mathbb{P}$ the Lebesgue measure on $[0,1]$. Let X be the process on $I = [0,1]$ defined by $X_t(\omega) = \mathbb{1}_{\{\omega=t\}}$. Show that X and 0 are modifications of each other.*

Given a stochastic process, a practical criterion to ensure the existence of a smooth modification is the following. In the next definition, we consider a stochastic process $(X_t)_{t\in I}$ taking its values in a *topological* space E, of which $\mathcal{E}$ is the Borel σ-field, and indexed by an interval $I \subset \mathbb{R}$.

Definition 9.2.14 (Almost surely continuous process)**.** *A stochastic process $(X_t)_{t\in I}$ with values in E is* almost surely continuous *if there exists $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 1$ and, for any $\omega \in A$, the function $t \mapsto X_t(\omega)$ is continuous.*

Theorem 9.2.15 (Kolmogorov criterion)**.** *Let $(X_t)_{t\in I}$ be a real-valued stochastic process. Assume that there exist $\alpha, \beta > 0$ such that, for any bounded interval $J \subset I$,*

$$\sup_{s,t\in J: s\neq t} \frac{\mathbb{E}\left[|X_t - X_s|^\alpha\right]}{|t-s|^{1+\beta}} < +\infty.$$

Then there exists an almost surely continuous modification $(\widetilde{X}_t)_{t\in I}$ of $(X_t)_{t\in I}$.

In fact, the Kolmogorov criterion shows that, almost surely, the mapping $t \mapsto \widetilde{X}_t$ is locally γ-Hölder continuous for any $\gamma < \beta/\alpha$, which means that almost surely, for any bounded interval $J \subset I$,

$$\sup_{s,t\in J; s\neq t} \frac{|\widetilde{X}_t - \widetilde{X}_s|}{|t-s|^\gamma} < +\infty.$$

Continuity of the Brownian trajectories

Let $(B_t)_{t\geq 0}$ be a Brownian motion. For any $\alpha > 0$, for any $s, t \geq 0$,

$$\mathbb{E}\left[|B_t - B_s|^\alpha\right] = |t-s|^{\alpha/2}\mathbb{E}\left[|G|^\alpha\right], \qquad \text{where } G \sim \mathcal{N}(0,1).$$

Therefore, as soon as $\alpha > 2$, the Kolmogorov criterion is satisfied with $\beta = \alpha/2 - 1$. As a consequence, $(B_t)_{t\geq 0}$ admits an almost surely continuous modification. By Exercise 9.2.12, this process remains a Brownian motion. Therefore, up to replacing $(B_t)_{t\geq 0}$ by this modification, **from now on we shall always consider that the trajectories of the Brownian motion are almost surely continuous**.

Besides, optimising the ratio β/α, we deduce from Theorem 9.2.15 that almost surely, $(B_t)_{t\geq 0}$ has locally γ-Hölder continuous trajectories, for any $\gamma < 1/2$. It can be proved that this statement is sharp, in the sense that almost surely, there is no nontrivial interval on which the trajectories of $(B_t)_{t\geq 0}$ are γ-Hölder continuous for $\gamma \geq 1/2$. The particular role of the order $1/2$ can be heuristically understood by noting that for any $t, h \geq 0$,

$$\mathbb{E}[|B_{t+h} - B_t|^2] = h,$$

so that $B_{t+h} - B_t$ is of order $\sqrt{h}$.

Last, and for similar reasons, it can also be proved that **almost surely, the Brownian trajectories are nowhere differentiable**.

The space of continuous sample-paths

Let $I = [0, T]$ for $T > 0$. Up to replacing Ω with the almost sure event A given by Definition 9.2.14, one may see the (continuous modification of the) Brownian motion $(B_t)_{t\in I}$ as a function from Ω to the space $C(I)$ of real-valued continuous functions $I \to \mathbb{R}$. Endowing the latter space with the Borel σ-field induced by the topology of uniform convergence, one may wonder whether $B : \omega \mapsto (B_t(\omega))_{t\in I}$ defines a *random variable* in $C(I)$, and if so, how to characterise its law. The answer is given by the following result.

Proposition 9.2.16 (Law of continuous processes)**.** *Let $X = (X_t)_{t\in I}$ be a real-valued stochastic process.*

(i) The process X has almost surely continuous trajectories if and only if $X : \Omega \to C(I)$ is measurable.

(ii) Two almost surely continuous processes X, Y have the same law in $C(I)$ if and only if they have the same finite-dimensional distributions, that is to say if and only if they have the same law in $\mathbb{R}^I$.

The law of the Brownian motion in $C(I)$ is called the *Wiener measure* on I. It can be defined similarly when $I = [0, +\infty)$, provided that $C([0, +\infty))$ is endowed with the Borel σ-field induced by the topology of the locally uniform convergence[1].

Remark 9.2.17 (Continuity and product σ-field)**.** *While Proposition 9.2.16 shows that two continuous processes which have the same law with respect to the product σ-field on $\mathbb{R}^I$ also have*

[1] It is for instance induced by the distance

$$\mathrm{d}((x_t)_{t\geq 0}, (y_t)_{t\geq 0}) = \sum_{k=1}^{\infty} 2^{-k} 1 \wedge \sup_{t\in[0,k]} (|x_t - y_t| \wedge 1).$$

the same law in $C(I)$, whether or not a process has continuous trajectories is not *an information provided by its law in $\mathbb{R}^I$. Indeed, the example of Exercise 9.2.13 shows two processes which are modifications of each other, and therefore have the same law in $\mathbb{R}^I$, but have continuous trajectories with respective probability 0 and 1:*

$$1 = \mathbb{P}(0 \in C(I)) \neq \mathbb{P}(X \in C(I)) = 0.$$

We therefore deduce from this fact that the subset $C(I)$ of $\mathbb{R}^I$ does not belong *to the product σ-field $\mathcal{B}(\mathbb{R})^{\otimes I}$.*

Since $C(I)$ is a topological space, it is the natural space in which the *convergence* of (almost surely continuous) stochastic processes can be studied. In particular, the convergence of the rescaled random walk to the Brownian motion is stated in this space.

Theorem 9.2.18 (Donsker's invariance principle)**.** *For any $T > 0$, the random walk $(B_t^{\Delta t, \Delta x})_{t \in [0,T]}$ converges in distribution, in the space $C([0,T])$, to the Brownian motion $(B_t)_{t\in[0,T]}$ when $\Delta t \to 0$, $\Delta x \to 0$, with $\Delta x^2/\Delta t = 1$.*

9.2.4 Multidimensional Brownian motion

Definition 9.2.19 (Multidimensional Brownian motion)**.** *For all $d \geq 1$, a* Brownian motion in $\mathbb{R}^d$ *is a process $B = (B^1, \ldots, B^d)$ where $B^1, \ldots, B^d$ are independent Brownian motions.*

Exercise 9.2.20 (Isotropy)**.** *Let B be a Brownian motion in $\mathbb{R}^d$ and let $O \in \mathbb{R}^{d\times d}$ be an orthogonal matrix, that is to say such that $OO^\top = I_d$. Show that the process OB is a Brownian motion in $\mathbb{R}^d$.*

9.3 Filtration, stopping time and $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion

This last section is aimed at presenting the notion of $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion, which will be used in the next lectures. The main new object here is the notion of *filtration*.

9.3.1 More on σ-fields

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A *sub-σ-field $\mathcal{F}$* of $\mathcal{A}$ is a σ-field on Ω which is included in $\mathcal{A}$.

Definition 9.3.1 (Random variables and sub-σ-fields)**.** *Let X be a random variable defined on $(\Omega, \mathcal{A})$ which takes its values in a measurable space $(E, \mathcal{E})$.*

(i) The random variable X is measurable with respect to *a sub-σ-field $\mathcal{F}$ of $\mathcal{A}$ if $\{X \in C\} \in \mathcal{F}$ for all $C \in \mathcal{E}$.*

(ii) The sub-σ-field generated by *X is the sub-σ-field*

$$\sigma(X) := \{\{X \in C\} : C \in \mathcal{E}\}.$$

In the first case, we shall also say that X is *$\mathcal{F}$-measurable*.

Exercise 9.3.2. *Check that $\sigma(X)$ is a sub-σ-field of $\mathcal{A}$.*

Clearly, X is $\mathcal{F}$-measurable if and only if $\sigma(X) \subset \mathcal{F}$, so that $\sigma(X)$ is the smallest sub-σ-field with respect to which X is measurable.

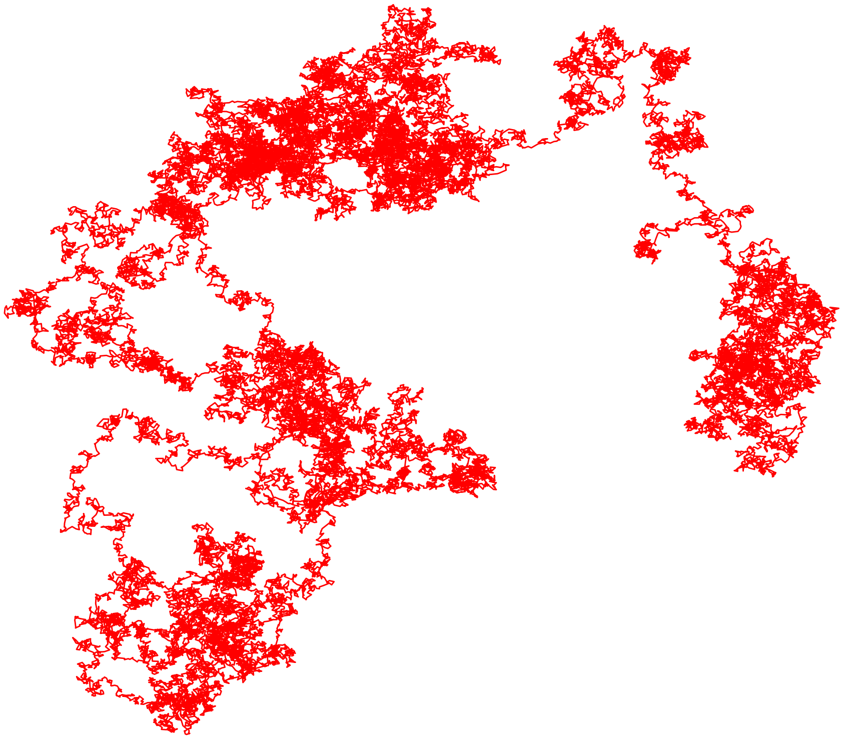


Figure 9.2: A realisation of the Brownian motion in dimension $d = 2$.

Exercise 9.3.3. *Let $c \in \mathbb{R}$ and X be the random variable defined by $X(\omega) = c$ for all $\omega \in \Omega$. What is $\sigma(X)$?*

The following extension of the notion of independence will be useful.

Definition 9.3.4 (Independence between sub-σ-fields)**.** *Two sub-σ-fields $\mathcal{F}$ and $\mathcal{G}$ are* independent *if, for any $A \in \mathcal{F}$, $B \in \mathcal{G}$,*

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

Notice that if $\mathcal{F}$ and $\mathcal{G}$ are the sub-σ-fields respectively generated by some random variables X and Y, then $\mathcal{F}$ and $\mathcal{G}$ are independent if and only if X and Y are independent. This definition can of course be generalised for an arbitrary family of sub-σ-fields. In the sequel of this course we shall sometimes say that a random variable X and a sub-σ-field $\mathcal{F}$ are independent when the sub-σ-fields $\sigma(X)$ and $\mathcal{F}$ are independent.

We shall sometimes require a sub-σ-field to be $\mathbb{P}$-*complete* in the following sense.

Definition 9.3.5 ($\mathbb{P}$-completeness)**.** *A sub-σ-field is called* $\mathbb{P}$-complete *if it contains all negligible events, that is to say all sets $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 0$.*

9.3.2 Filtrations

Let I be a subset of $\mathbb{R}$.

Definition 9.3.6 (Filtration)**.** *A* filtration *is a family of sub-σ-fields $(\mathcal{F}_t)_{t\in I}$ of $\mathcal{A}$ such that, for any $s, t \in I$, if $s \leq t$ then $\mathcal{F}_s \subset \mathcal{F}_t$.*

Definition 9.3.7 (Adapted process)**.** *A stochastic process $X = (X_t)_{t \in I}$ is called* adapted *to a filtration $(\mathcal{F}_t)_{t \in I}$ if, for any $t \in I$, the random variable X_t is $\mathcal{F}_t$-measurable.*

Definition 9.3.8 (Filtration generated by a stochastic process)**.** *Let $X = (X_t)_{t \in I}$ be a stochastic process. The* filtration generated by X *is the filtration $(\mathcal{F}^X_t)_{t \in I}$ defined by*

$$\mathcal{F}^X_t = \sigma\left((X_s)_{s \in I \cap (-\infty, t]}\right),$$

for all $t \in I$.

🗎 **Exercise 9.3.9.** *Check that a process $X = (X_t)_{t \in I}$ is adapted to the filtration $(\mathcal{F}_t)_{t \in I}$ if and only if, for all $t \in I$, $\mathcal{F}^X_t \subset \mathcal{F}_t$.*

9.3.3 $(\mathcal{F}_t)_{t \geq 0}$-Brownian motion

Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration.

Definition 9.3.10 ($(\mathcal{F}_t)_{t \geq 0}$-Brownian motion)**.** *A $(\mathcal{F}_t)_{t \geq 0}$-Brownian motion is a real-valued Brownian motion $B = (B_t)_{t \geq 0}$ such that:*

(i) the process B is adapted to the filtration $(\mathcal{F}_t)_{t \geq 0}$;
(ii) for all $0 \leq s \leq t$, $B_t - B_s$ is independent from $\mathcal{F}_s$.

In particular, every $(\mathcal{F}_t)_{t \geq 0}$-Brownian motion is a Brownian motion in the sense of Definition 9.2.3, and conversely every Brownian motion B in the sense of Definition 9.2.3 is a $(\mathcal{F}^B_t)_{t \geq 0}$-Brownian motion. The notion of $(\mathcal{F}_t)_{t \geq 0}$-Brownian motion is interesting when the filtration $(\mathcal{F}_t)_{t \geq 0}$ is strictly larger (that is to say, contains more information) than the filtration generated by B. This is for instance the case when $(\mathcal{F}_t)_{t \geq 0}$ is the filtration generated by a d-dimensional Brownian motion, which will be a typical situation for the study of stochastic differential equations in $\mathbb{R}^d$. In this case, each coordinate B^i of the multidimensional Brownian motion is a $(\mathcal{F}_t)_{t \geq 0}$-Brownian motion.

9.3.4 Stopping times

Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration.

Definition 9.3.11 (Stopping time)**.** *A function $\tau : \Omega \to [0, +\infty]$ is a $(\mathcal{F}_t)_{t \geq 0}$-*stopping time *if, for any $t \geq 0$, the event $\{\tau \leq t\}$ belongs to $\mathcal{F}_t$.*

🗎 **Exercise 9.3.12.** *1. Show that any deterministic time t_0 is a $(\mathcal{F}_t)_{t \geq 0}$-stopping time.*

2. If τ is a $(\mathcal{F}_t)_{t \geq 0}$-stopping time, show that for all $t \geq 0$, the events $\{\tau > t\}$, $\{\tau < t\}$, $\{\tau = t\}$ and $\{\tau \geq t\}$ belong to $\mathcal{F}_t$.

3. If τ and τ' are two $(\mathcal{F}_t)_{t \geq 0}$-stopping times, check that $\tau \wedge \tau'$ and $\tau \vee \tau'$ are $(\mathcal{F}_t)_{t \geq 0}$-stopping times.

🏠 **Exercise 9.3.13** (An important example)**.** *Let $X = (X_t)_{t \geq 0}$ be a $(\mathcal{F}_t)_{t \geq 0}$-adapted and almost surely continuous process, taking its values in $\mathbb{R}^n$, and F be a closed subset of $\mathbb{R}^n$. Let us moreover assume that $\mathcal{F}_0$ is $\mathbb{P}$-complete in the sense of Definition 9.3.5.*

The purpose of this exercise is to show that

$$\tau = \inf\{t \geq 0 : X_t \in F\}$$

is a $(\mathcal{F}_t)_{t \geq 0}$-stopping time.

1. *Let $t \geq 0$. Write the event $\{\tau \leq t\}$ in terms of the set of events $\{X_s \in F\}$, $s \in [0,t]$, and check that each such event is in $\mathcal{F}_t$.*
2. *For all $x \in \mathbb{R}^n$, let us denote*

$$\mathrm{dist}(x, F) = \inf_{y \in F} |x - y|.$$

Show that there exists $s \in [0,t]$ such that $X_s \in F$ if and only if either $X_t \in F$ or for any $k \geq 1$, there exists $s_k \in [0,t] \cap \mathbb{Q}$ such that $\mathrm{dist}(X_{s_k}, F) \leq 1/k$.

3. *Conclude.*

Remark 9.3.14. *Under the assumptions of Exercise 9.3.13, if G is an open subset of $\mathbb{R}^n$, then the random time $\tau = \inf\{t \geq 0 : X_t \in G\}$ need not be a stopping time. It can however proved to be so if, in addition, the filtration $(\mathcal{F}_t)_{t\geq 0}$ is* right-continuous, *that is to say that for any $t \geq 0$, $\mathcal{F}_t = \cap_{\epsilon>0}\mathcal{F}_{t+\epsilon}$. A filtration which is right-continuous and such that $\mathcal{F}_0$ is $\mathbb{P}$-complete is said to satisfy* the usual conditions.

9.3.5 Strong Markov property and reflection principle

The strong Markov property

Definition 9.3.15 (The σ-field $\mathcal{F}_\tau$)**.** *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration and τ be a $(\mathcal{F}_t)_{t\geq 0}$-stopping time. The σ-field $\mathcal{F}_\tau$ is defined by*

$$\mathcal{F}_\tau = \{A \in \mathcal{A} : \forall t \geq 0, \{\tau \leq t\} \cap A \in \mathcal{F}_t\}.$$

It must be understood as the set of events about which one may decide whether they are realised or not given the information available at time τ. For instance, if $(X_t)_{t\geq 0}$ is an adapted and almost surely continuous process and $\tau = \inf\{t \geq 0 : X_t \in F\}$ for some closed set F, the random variables τ, X_τ, $\sup_{t\in[0,\tau]} |X_t|$, etc. are $\mathcal{F}_\tau$-measurable.

Theorem 9.3.16 (Strong Markov property)**.** *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration on $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathcal{F}_0$ is complete, $(B_t)_{t\geq 0}$ be a $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion, and τ be a $(\mathcal{F}_t)_{t\geq 0}$-stopping time, such that $\tau < +\infty$ almost surely. The process $(B'_t)_{t\geq 0}$ defined by $B'_t = B_{\tau+t} - B_\tau$ is a Brownian motion, independent from $\mathcal{F}_\tau$.*

Proof. Let $t_1, \ldots, t_k \geq 0$, $F : \mathbb{R}^k \to \mathbb{R}$ a bounded and continuous function, and $A \in \mathcal{F}_\tau$. We are going to show that

$$\mathbb{E}\left[\mathbb{1}_A F\left(B'_{t_1}, \ldots, B'_{t_k}\right)\right] = \mathbb{P}(A)\mathbb{E}\left[F\left(B_{t_1}, \ldots, B_{t_k}\right)\right]. \tag{9.1}$$

Applying this identity with $A = \Omega$, we deduce that the vectors $(B'_{t_1}, \ldots, B'_{t_k})$ and $(B_{t_1}, \ldots, B_{t_k})$ have the same law, which by Proposition 9.1.5 ensures that B' is a Brownian motion. We therefore get

$$\mathbb{E}\left[\mathbb{1}_A F\left(B'_{t_1}, \ldots, B'_{t_k}\right)\right] = \mathbb{P}(A)\mathbb{E}\left[F\left(B'_{t_1}, \ldots, B'_{t_k}\right)\right],$$

which shows that the vector $(B'_{t_1}, \ldots, B'_{t_k})$ is independent from the σ-field $\mathcal{F}_\tau$. Following Proposition 9.1.6, this ensures that the process B' is independent from $\mathcal{F}_\tau$.

Let us first write, thanks to the almost sure continuity of the trajectories of B and the fact that $\tau < +\infty$, almost surely,

$$F\left(B'_{t_1}, \ldots, B'_{t_k}\right) = \lim_{n\to+\infty} \sum_{j=1}^{+\infty} \mathbb{1}_{\{(j-1)/n \leq \tau < j/n\}} F\left(B_{j/n+t_1} - B_{j/n}, \ldots, B_{j/n+t_k} - B_{j/n}\right),$$

almost surely. By the Dominated Convergence Theorem, we deduce

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_A F\left(B'_{t_1},\dots,B'_{t_k}\right)\right]\\&=\lim_{n\to+\infty}\sum_{j=1}^{+\infty}\mathbb{E}\left[\mathbb{1}_A\mathbb{1}_{\{(j-1)/n\le\tau<j/n\}}F\left(B_{j/n+t_1}-B_{j/n},\dots,B_{j/n+t_k}-B_{j/n}\right)\right].\end{aligned}$$

For all $n \geq 1$, $j \geq 1$, the event $A \cap \{(j-1)/n \leq \tau < j/n\}$ belongs to $\mathcal{F}_{j/n}$, while by Proposition 9.2.9, the vector $(B_{j/n+t_1} - B_{j/n}, \dots, B_{j/n+t_k} - B_{j/n})$ is independent from $\mathcal{F}_{j/n}$, and has the same law as $(B_{t_1}, \dots, B_{t_k})$. Hence,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_A\mathbb{1}_{\{(j-1)/n\le\tau<j/n\}}F\left(B_{j/n+t_1}-B_{j/n},\dots,B_{j/n+t_k}-B_{j/n}\right)\right]\\&=\mathbb{E}\left[F\left(B_{t_1},\dots,B_{t_k}\right)\right]\mathbb{P}\left(A\cap\{(j-1)/n\le\tau<j/n\}\right),\end{aligned}$$

and we obtain (9.1) taking the sum over j. □

The assumption that τ be a stopping time is crucial for the application of Theorem 9.3.16. Let us indeed consider the random time

$$\vartheta = \sup\{t < 1 : B_t = 0\}.$$

It is clear that this variable is not a stopping time, because in order to decide whether $\vartheta \leq t$ it is necessary to know the trajectory of the Brownian motion on the interval $[t, 1]$. Besides, the process $(B_{t+\vartheta} - B_\vartheta)_{t\geq 0}$ is not a Brownian motion, because it does not touches 0 on the interval $(0, 1-\vartheta)$, see Figure 9.3, while we shall see in Exercise 9.3.21 below that the Brownian motion touches 0 infinitely often in the neighbourhood of $t = 0$.

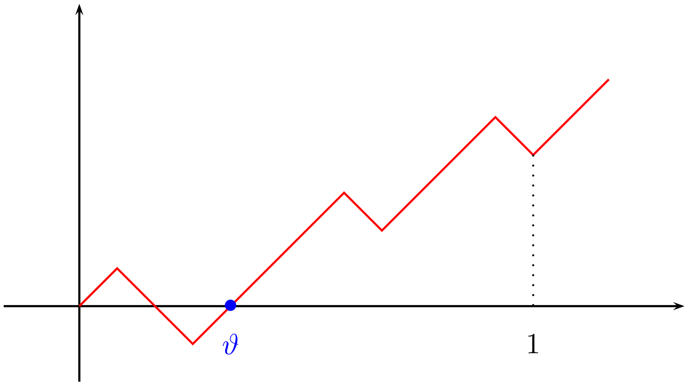


Figure 9.3: The random variable ϑ.

The reflection principle

🏠 **Exercise 9.3.17** (The reflection principle)**.** *Let $(B_t)_{t\geq 0}$ be a Brownian motion, and for any $a > 0$,*

$$\tau_a = \inf\{t \geq 0 : B_t \geq a\}.$$

Let $(B'_t)_{t\geq 0}$ be a Brownian motion, independent from $\mathcal{F}^B_{\tau_a}$, and let $(X_t)_{t\geq 0}$ be defined by

$$X_t = \begin{cases} B_t & \textit{if } t \leq \tau_a, \\ a + B'_{t-\tau_a} & \textit{if } t > \tau_a. \end{cases}$$

1. *Show that X is a Brownian motion.*
2. *Deduce that the process $(Y_t)_{t\geq 0}$ defined by*

$$Y_t = \begin{cases} B_t & \text{if } t \leq \tau_a, \\ 2a - B_t & \text{if } t > \tau_a, \end{cases}$$

   is a Brownian motion. See Figure 9.4.
3. *Let us denote $S_t = \sup_{s\in[0,t]} B_s$. Show that for all $b \in [0,a]$ and $t \geq 0$, $\{S_t > a, B_t < b\} = \{Y_t > 2a - b\}$.*
4. *Deduce the identity $\mathbb{P}(S_t > a, B_t < b) = \mathbb{P}(B_t > 2a - b)$.*

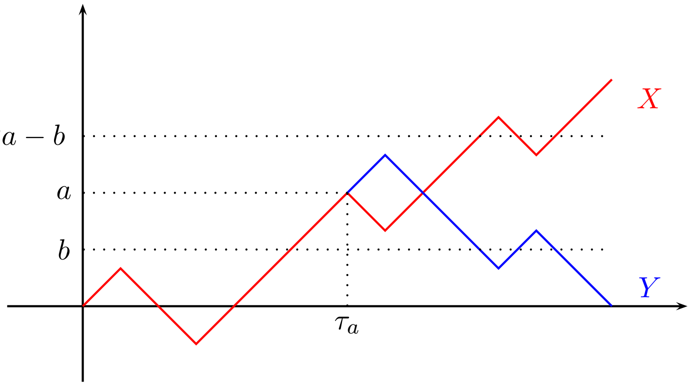


Figure 9.4: The reflection principle: both the red and the blue curve are Brownian motions.

The result of Exercise 9.3.17 is called the *reflection principle*. It has a very useful consequence.

Corollary 9.3.18 (Law of the supremum of the Brownian motion)**.** *For any $t \geq 0$, the random variables $S_t = \sup_{s\in[0,t]} B_s$ and $|B_t|$ have the same law, so that*

$$\forall a > 0, \qquad \mathbb{P}(S_t > a) = \frac{2}{\sqrt{2\pi t}} \int_{x=a}^{+\infty} \mathrm{e}^{-x^2/2t} \mathrm{d}x.$$

Proof. Let $t \geq 0$ et $a > 0$. Let us write

$$\mathbb{P}(S_t > a) = \mathbb{P}(S_t > a, B_t > a) + \mathbb{P}(S_t > a, B_t = a) + \mathbb{P}(S_t > a, B_t < a),$$

and study the three terms of the right-hand side separately: since $S_t \geq B_t$ by construction, $\mathbb{P}(S_t > a, B_t > a) = \mathbb{P}(B_t > a)$; since $\{B_t = a\}$ is a negligible event, $\mathbb{P}(S_t > a, B_t = a) = 0$; finally taking $a = b$ in the conclusion of Exercise 9.3.17, we get $\mathbb{P}(S_t > a, B_t < a) = \mathbb{P}(B_t > a)$. Thus, $\mathbb{P}(S_t > a) = 2\mathbb{P}(B_t > a) = \mathbb{P}(|B_t| > a)$. □

Remark 9.3.19. *If, for all $t \geq 0$, the* variables *S_t and $|B_t|$ have the same law, it is not true that the* processes *$(S_t)_{t\geq 0}$ and $(|B_t|)_{t\geq 0}$ have the same law! Figure 9.5 shows a trajectory of the processes $(B_t)_{t\ \geq 0}$, $(|B_t|)_{t\geq 0}$ and $(S_t)_{t\geq 0}$.*

🏠 **Exercise 9.3.20** (Law of τ_a)**.** *Use Corollary 9.3.18 to compute the density of the random variable τ_a. What can you say about its expectation?*

🏠 **Exercise 9.3.21** (Brownian motion in the neighbourhood of 0)**.** *The purpose of this exercise is to show that almost surely,* $\inf\{t > 0 : B_t = 0\} = 0$.

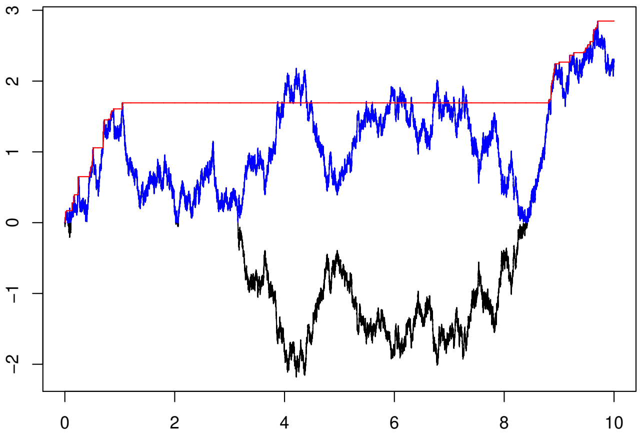


Figure 9.5: A trajectory of the processes $(B_t)_{t\ \geq 0}$ (black), $(|B_t|)_{t\geq 0}$ (blue) and $(S_t)_{t\geq 0}$ (red).

1. *For all* $t \geq 0$*, show that the random variable* $I_t = \inf_{s\in[0,t]} B_s$ *has the same law as* $-S_t$*.*
2. *Assume that there exists* $n \geq 1$ *such that* $B_t \neq 0$ *for all* $t \in (0, 1/n)$*. Deduce that either* $S_{1/n} = 0$ *or* $I_{1/n} = 0$*.*
3. *Conclude.*

Chapter 10

Stochastic calculus

Contents

Let $g : [0,T] \to \mathbb{R}$ be a C^1 function. For any integrable function $h : [0,T] \to \mathbb{R}$, let us introduce the notation

$$\int_0^T h(t)\mathrm{d}g(t) := \int_0^T h(t)g'(t)\mathrm{d}t,$$

which simply corresponds to the standard change-of-variable formula. It satisfies

$$\lim_{n\to+\infty} \sum_{i=0}^{n-1} \xi_i(g(t_{i+1}) - g(t_i)) = \int_0^T h(t)\mathrm{d}g(t)$$

as soon as h^n is a stepwise constant function which writes

$$h^n(t) = \sum_{i=0}^{n-1} \xi_i \mathbb{1}_{\{[t_i,t_{i+1})\}}(t), \qquad 0 = t_0 < t_1 < \cdots < t_n = T,$$

and is such that

$$\lim_{n\to+\infty} \int_0^T |h^n(t) - h(t)|\mathrm{d}t = 0.$$

Notice that the latter construction of the integral of h as the limit of integrals of stepwise constant functions h^n has the advantage of not involving $g'(t)$, and indeed this integral[1] may be constructed for a larger class of functions g, namely the class of functions with *bounded variation*, which must be thought of as cumulative distribution functions of bounded signed measures on $[0,T]$.

The purpose of Itô's calculus is to be able to define a similar integral for stochastic processes $(H_t)_{t\in[0,T]}$, when the smooth function g is replaced with the nonsmooth process $B = (B_t)_{t\in[0,T]}$.

[1] which is sometimes called *Stieltjes integral*

In this case, the change of variable formula $\mathrm{d}B_t = B_t'\mathrm{d}t$ no longer makes sense since the trajectory of the Brownian motion is not differentiable. However, we shall see that the approximation by stepwise constant processes may be employed, under the crucial condition that the process $(H_t)_{t\in[0,T]}$ be *progressively measurable* and (in a first step) in $\mathbf{L}^2$ as a function of (t,ω).

In this lecture, we shall thus:

- construct directly a notion of integral 'with respect to $\mathrm{d}B_t$': the *stochastic integral*;
- deduce an appropriate differential calculus: *Itô's formula*.

10.1 The stochastic integral

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$ and a $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion B. We assume that $\mathcal{F}_0$ is $\mathbb{P}$-complete. We also fix a bounded interval $I = [0,T] \subset [0,+\infty)$. The purpose of this section is to construct the *stochastic integral*

$$\int_{t=0}^{T} H_t \mathrm{d}B_t$$

for a certain class of real-valued stochastic processes $(H_t)_{t\in I}$.

10.1.1 Construction for stepwise constant functions

For all $n \geq 1$, let us denote by $0 = t_0 < \cdots < t_n = T$ a subdivision of the interval I into n intervals. To each process $H^n = (H^n_t)_{t\in I}$ of the form

$$H^n_t = \sum_{i=0}^{n-1} \xi_i \mathbb{1}_{[t_i,t_{i+1})}(t),$$

we associate the random variable

$$S^n_I = \sum_{i=0}^{n-1} \xi_i (B_{t_{i+1}} - B_{t_i}).$$

When H^n converges to some limiting process $H = (H_t)_{t\in I}$, it seems reasonable to be willing to define

$$\int_{t=0}^{T} H_t \mathrm{d}B_t = \lim_{n\to+\infty} S^n_I.$$

Exercise 10.1.1. *Let $t_i = iT/n$ and $\underline{H}^n$ and $\overline{H}^n$ be the processes defined by*

$$\underline{H}^n_t = \sum_{i=0}^{n-1} B_{t_i} \mathbb{1}_{[t_i,t_{i+1})}(t), \qquad \overline{H}^n_t = \sum_{i=0}^{n-1} B_{t_{i+1}} \mathbb{1}_{[t_i,t_{i+1})}(t).$$

Show that the associated sequences $\underline{S}^n_I$ and $\overline{S}^n_I$ do not have the same limit (in $\mathbf{L}^2$).

In order to recover a nonamibiguous notion of limit, we adopt the convention (which is called *Itô's convention*) to restrict ourselves to processes $H = (H_t)_{t\in I}$ which satisfy the following condition.

Definition 10.1.2 (Progressively measurable process)**.** *A stochastic process* $(H_t)_{t\in I}$ *is called* progressively measurable *if, for all* $t \in [0,T]$*, the mapping*

$$\begin{cases} [0,t]\times\Omega & \to & \mathbb{R} \\ (s,\omega) & \mapsto & H_s(\omega) \end{cases}$$

is measurable for the product σ*-field* $\mathcal{B}([0,t])\otimes\mathcal{F}_t$*.*

A progressively measurable process is adapted, but the converse statement does not hold in general. However, since $\mathcal{F}_0$ is assumed to be complete, an adapted process with almost surely continuous trajectories is progressively measurable.

We shall now denote by:

(i) $\mathbf{\Lambda}^2(I)$ the set of progressively measurable processes H such that $\int_{t=0}^T \mathbb{E}[H_t^2]\mathrm{d}t < +\infty$;
(ii) $\mathbf{\Lambda}_0^2(I)$ the subset of *piecewise constant* processes $H \in \mathbf{\Lambda}^2(I)$, that is to say such that there exist $0 = t_0 < t_1 < \cdots < t_n = T$ and random variables $\xi_0,\ldots,\xi_{n-1}$ such that $H_t = \sum_{i=0}^{n-1}\xi_i \mathbb{1}_{[t_i,t_{i+1})}(t)$.

Lemma 10.1.3 (Characterisation of $\mathbf{\Lambda}_0^2(I)$)**.** *A process* $H = (H_t)_{t\in I}$ *which writes under the form* $H_t = \sum_{i=0}^{n-1}\xi_i \mathbb{1}_{[t_i,t_{i+1})}(t)$ *belongs to* $\mathbf{\Lambda}^2(I)$ *(and thus to* $\mathbf{\Lambda}_0^2(I)$*) if and only if, for any* $i \in \{0,\ldots,n-1\}$*, the random variable* ξ_i *is in* $\mathbf{L}^2(\mathbb{P})$ *and* $\mathcal{F}_{t_i}$*-measurable.*

Proof. Let H be a process of the form $H_t = \sum_{i=0}^{n-1}\xi_i \mathbb{1}_{[t_i,t_{i+1})}(t)$, where $\xi_0,\ldots,\xi_{n-1}$ are random variables. It is clear that this process is progressively measurable if and only if it is adapted, and that the latter condition holds if and only if for any $t\in I$, H_t is $\mathcal{F}_t$-measurable, which amounts to asserting that for any $i \in \{0,\ldots,n-1\}$, the random variable ξ_i is measurable with respect to the σ-field

$$\bigcap_{t\in[t_i,t_{i+1})} \mathcal{F}_t = \mathcal{F}_{t_i}.$$

Moreover, since it is assumed that $t_i < t_{i+1}$, it is clear that

$$\int_{t=0}^T \mathbb{E}[H_t^2]\mathrm{d}t = \sum_{i=0}^{n-1}\mathbb{E}[\xi_i^2](t_{i+1}-t_i) < +\infty$$

if and only if $\xi_i \in \mathbf{L}^2(\mathbb{P})$ for all i. □

In Exercise 10.1.1, $\underline{H}^n$ is progressively measurable (and in $\mathbf{\Lambda}_0^2(I)$), but $\overline{H}^n$ is not because it is not adapted.

Definition 10.1.4 (Stochastic integral on $\mathbf{\Lambda}_0^2(I)$)**.** *For all* $H^n = (H_t^n)_{t\in I} \in \mathbf{\Lambda}_0^2(I)$*, we define*

$$\int_{t=0}^T H_t^n \mathrm{d}B_t = \sum_{i=0}^{n-1}\xi_i(B_{t_{i+1}} - B_{t_i}).$$

10.1.2 Extension to $\mathbf{\Lambda}^2(I)$

Let us define on $\mathbf{\Lambda}^2(I)$ the norm[2] $\|\cdot\|_{\mathbf{\Lambda}^2(I)}$ by

$$\|H\|_{\mathbf{\Lambda}^2(I)}^2 = \int_{t=0}^T \mathbb{E}\left[H_t^2\right]\mathrm{d}t.$$

[2] Here, we implicitly identify two processes which coincide $\mathrm{d}t\otimes\mathbb{P}$-almost everywhere.

Lemma 10.1.5 (Approximation properties)**.** *With the previous notation:*

(i) the space $\mathbf{\Lambda}^2(I)$ equipped with the norm $\|\cdot\|_{\mathbf{\Lambda}^2(I)}$ is a Banach space;

(ii) the subset $\mathbf{\Lambda}_0^2(I)$ is dense in $\mathbf{\Lambda}^2(I)$;

(iii) the linear mapping

$$\mathcal{J}_I : \left\{ \begin{array}{ccc} \mathbf{\Lambda}_0^2(I) & \to & \mathbf{L}^2(\mathbb{P}) \\ H & \mapsto & \displaystyle\int_{t=0}^T H_t \mathrm{d}B_t \end{array} \right.$$

is an isometry, that is to say that for all $H \in \mathbf{\Lambda}_0^2(I)$,

$$\|H\|_{\mathbf{\Lambda}^2(I)}^2 = \mathbb{E}\left[\mathcal{J}_I(H)^2\right].$$

Proof. The properties (i) and (ii) are admitted (but they follow from standard arguments). To show the property (iii), let us take $H = (H_t)_{t\in I} \in \mathbf{\Lambda}_0^2(I)$, which writes

$$H_t = \sum_{i=0}^{n-1} \xi_i \mathbb{1}_{[t_i,t_{i+1})}(t).$$

Following Lemma 10.1.3, for all $i \in \{0,\dots,n-1\}$, the random variable ξ_i is $\mathcal{F}_{t_i}$-measurable.

Let us now compute

$$\mathbb{E}\left[\mathcal{J}_I(H)^2\right] = \mathbb{E}\left[\left(\sum_{i=0}^{n-1} \xi_i (B_{t_{i+1}} - B_{t_i})\right)^2\right] = \sum_{i,j=0}^{n-1} \mathbb{E}\left[\xi_i (B_{t_{i+1}} - B_{t_i}) \xi_j (B_{t_{j+1}} - B_{t_j})\right].$$

When $i = j$, $\xi_i\xi_j = \xi_i^2$ is $\mathcal{F}_{t_i}$-measurable, while by Definition 9.3.10, the random variable $(B_{t_{i+1}} - B_{t_i})(B_{t_{j+1}} - B_{t_j}) = (B_{t_{i+1}} - B_{t_i})^2$ is independent from $\mathcal{F}_{t_i}$. Thus,

$$\mathbb{E}\left[\xi_i (B_{t_{i+1}} - B_{t_i}) \xi_j (B_{t_{j+1}} - B_{t_j})\right] = \mathbb{E}\left[\xi_i^2\right] \mathbb{E}\left[(B_{t_{i+1}} - B_{t_i})^2\right] = (t_{i+1} - t_i)\mathbb{E}\left[\xi_i^2\right],$$

since $B_{t_{i+1}} - B_{t_i} \sim \mathcal{N}(0, t_{i+1} - t_i)$. When $i < j$, the random variables ξ_i, $B_{t_{i+1}} - B_{t_i}$ and ξ_j are respectively $\mathcal{F}_{t_i}$-, $\mathcal{F}_{t_{i+1}}$- and $\mathcal{F}_{t_j}$-measurable, while the random variable $B_{t_{j+1}} - B_{t_j}$ is independent from $\mathcal{F}_{t_j}$. But since $\mathcal{F}_{t_i} \subset \mathcal{F}_{t_{i+1}} \subset \mathcal{F}_{t_j}$, we deduce that the random variables $\xi_i(B_{t_{i+1}} - B_{t_i})\xi_j$ and $B_{t_{j+1}} - B_{t_j}$ are independent, so that

$$\mathbb{E}\left[\xi_i (B_{t_{i+1}} - B_{t_i}) \xi_j (B_{t_{j+1}} - B_{t_j})\right] = \mathbb{E}\left[\xi_i (B_{t_{i+1}} - B_{t_i}) \xi_j\right] \mathbb{E}\left[B_{t_{j+1}} - B_{t_j}\right] = 0,$$

since $B_{t_{j+1}} - B_{t_j} \sim \mathcal{N}(0, t_{j+1} - t_j)$. The case $i > j$ is similar, and thus

$$\mathbb{E}\left[\mathcal{J}_I(H)^2\right] = \sum_{i=0}^{n-1} (t_{i+1} - t_i)\mathbb{E}\left[\xi_i^2\right] = \int_{t=0}^T \mathbb{E}[H_t^2]\mathrm{d}t = \|H\|_{\mathbf{\Lambda}^2(I)}^2,$$

which is the expected identity. □

The next result is an immediate consequence[3] of Lemma 10.1.5.

Corollary 10.1.6 (Definition of the stochastic integral on $\mathbf{\Lambda}^2(I)$)**.** *The mapping $\mathcal{J}_I$ can be extended to a unique isometry from $\mathbf{\Lambda}^2(I)$ to $\mathbf{L}^2(\mathbb{P})$, which we shall still denote by $\mathcal{J}_I$.*

[3] Let $(M, \|\cdot\|_M)$, $(L, \|\cdot\|_L)$ be two Banach spaces, and M_0 be a dense linear subspace of M. Assume that there exists a linear function $J : M_0 \to L$ such that $\|J(h)\|_L = \|h\|_M$ for all $h \in M_0$. Then there exists a unique linear function $\widetilde{J} : M \to L$ such that $\widetilde{J}(h) = J(h)$ for all $h \in M_0$, and it satisfies $\|\widetilde{J}(h)\|_L = \|h\|_M$ for all $h \in M$.

For all $H = (H_t)_{t\in I} \in \mathbf{\Lambda}^2(I)$, the random variable $\mathfrak{J}_I(H) \in \mathbf{L}^2(\mathbb{P})$ is called the *stochastic integral of H on* $[0,T]$, it is also denoted by

$$\mathfrak{J}_I(H) = \int_{t=a}^{b} H_t \mathrm{d}B_t.$$

This definition is not very constructive, in practice we shall keep in mind that, for all $H \in \mathbf{\Lambda}^2(I)$,

$$\int_{t=0}^{T} H_t \mathrm{d}B_t = \lim_{n\to+\infty} \int_{t=0}^{T} H_t^n \mathrm{d}B_t, \qquad \text{in } \mathbf{L}^2,$$

as soon as H^n is a sequence of processes in $\mathbf{\Lambda}_0^2(I)$ such that $\|H^n - H\|_{\mathbf{\Lambda}^2(I)} \to 0$.

This construction can be adapted without any change to define the stochastic integral of $(H_t)_{t\in[a,b]}$ on any bounded interval $[a,b] \subset [0,+\infty)$.

In any case, the stochastic integral is defined as an element of $\mathbf{L}^2(\mathbb{P})$, in which random variables which coincide almost surely are identified, so it is uniquely defined only up to a negligible event.

Proposition 10.1.7 (Properties of the stochastic integral on $\mathbf{\Lambda}^2(I)$)**.** *The stochastic integral satisfies the following properties.*

(i) Addition: for any $T' \in [0,T]$,

$$\int_{t=0}^{T'} H_t \mathrm{d}B_t + \int_{t=T'}^{T} H_t \mathrm{d}B_t = \int_{t=0}^{T} H_t \mathrm{d}B_t, \qquad \textit{almost surely.}$$

(ii) Linearity: for all $\lambda, \mu \in \mathbb{R}$ and $H, H' \in \mathbf{\Lambda}^2(I)$,

$$\int_{t=0}^{T} (\lambda H_t + \mu H'_t) \mathrm{d}B_t = \lambda \int_{t=0}^{T} H_t \mathrm{d}B_t + \mu \int_{t=0}^{T} H'_t \mathrm{d}B_t, \qquad \textit{almost surely.}$$

(iii) Measurability: the process $(\int_{s=0}^{t} H_s \mathrm{d}B_s)_{t\in[0,T]}$ is progressively measurable.

(iv) Mean and variance:

$$\mathbb{E}\left[\int_{t=0}^{T} H_t \mathrm{d}B_t\right] = 0 \qquad \textit{and} \qquad \mathbb{E}\left[\left(\int_{t=0}^{T} H_t \mathrm{d}B_t\right)^2\right] = \mathbb{E}\left[\int_{t=0}^{T} H_t^2 \mathrm{d}t\right].$$

The second part of (iv) is called the *Itô isometry*.

Proof. The properties (i), (ii) and (iii) can be obtained by an approximation argument. To show the property (iv), let us first note that if $H = (H_t)_{t\in I} \in \mathbf{\Lambda}_0^2(I)$ writes

$$H_t = \sum_{i=0}^{n-1} \xi_i \mathbb{1}_{[t_i,t_{i+1})}(t),$$

then

$$\mathbb{E}\left[\mathfrak{J}_I(H)\right] = \mathbb{E}\left[\sum_{i=0}^{n-1} \xi_i (B_{t_{i+1}} - B_{t_i})\right] = \sum_{i=0}^{n-1} \mathbb{E}\left[\xi_i (B_{t_{i+1}} - B_{t_i})\right].$$

Since Lemma 10.1.3 asserts that each random variable ξ_i is $\mathcal{F}_{t_i}$-measurable while $B_{t_{i+1}} - B_{t_i}$ is independent from $\mathcal{F}_{t_i}$ and centered, this sum is

$$\sum_{i=0}^{n-1} \mathbb{E}\left[\xi_i\right] \mathbb{E}\left[B_{t_{i+1}} - B_{t_i}\right] = 0.$$

Now let $H \in \mathbf{\Lambda}^2(I)$. According to Lemma 10.1.5, there exists a sequence $(H^n)_{n\geq 1}$ in $\mathbf{\Lambda}_0^2(I)$ which converges to H. For all $n \geq 1$, the triangle inequality, the linearity of $\mathfrak{J}_I$ and the Cauchy–Schwarz inequality yield

$$|\mathbb{E}\left[\mathfrak{J}_I(H)\right] - \mathbb{E}\left[\mathfrak{J}_I(H^n)\right]| \leq \mathbb{E}\left[|\mathfrak{J}_I(H-H^n)|\right] \leq \sqrt{\mathbb{E}\left[|\mathfrak{J}_I(H-H^n)|^2\right]} = \|H - H_n\|_{\mathbf{\Lambda}^2(I)}.$$

Since the previous computation shows that $\mathbb{E}[\mathfrak{J}_I(H^n)] = 0$ for all $n \geq 1$, we conclude that $\mathbb{E}[\mathfrak{J}_I(H)] = 0$, which is the first identity in the property (iv). The second identity immediately rewrites $\mathbb{E}[\mathfrak{J}_I(H)^2] = \|H\|^2_{\mathbf{\Lambda}^2(I)}$ and it is thus a consequence of the fact that the extension of $\mathfrak{J}_I$ to $\mathbf{\Lambda}^2(I)$ remains an isometry. □

Higher-order moments of stochastic integrals can be controlled thanks to the Burkholder–Davis–Gundy inequality, which is admitted.

Lemma 10.1.8 (Burkholder–Davis–Gundy inequality)**.** *For any $p \geq 1$, there exist constants c_p, C_p such that for any $H = (H_t)_{t\in I} \in \mathbf{\Lambda}^2(I)$,*

$$c_p\mathbb{E}\left[\left(\int_{s=0}^T H_s^2 \mathrm{d}s\right)^{p/2}\right] \leq \mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_{s=0}^t H_s \mathrm{d}B_s\right|^p\right] \leq C_p\mathbb{E}\left[\left(\int_{s=0}^T H_s^2 \mathrm{d}s\right)^{p/2}\right].$$

Lemma 10.1.8 allows one to apply the Kolmogorov criterion (Theorem 9.2.15) to deduce that the process $(\int_{s=0}^t H_s \mathrm{d}B_s)_{t\geq 0}$ admits an almost surely continuous modification. From now on we shall systematically work with this modification and therefore consider that the process $(\int_{s=0}^t H_s \mathrm{d}B_s)_{t\geq 0}$ is almost surely continuous.

Exercise 10.1.9. *Using the sequence of processes $\underline{H}^n$ defined in Exercise 10.1.1, show that for all $T \geq 0$,*

$$\int_{t=0}^T B_t \mathrm{d}B_t = \frac{1}{2}(B_T^2 - T), \qquad \textit{almost surely.}$$

Exercise 10.1.10. *Show that for any $H, H' \in \mathbf{\Lambda}^2(I)$,*

$$\int_{t=0}^T H_t \mathrm{d}B_t \int_{t=0}^T H'_t \mathrm{d}B_t \in \mathbf{L}^1(\mathbb{P})$$

and

$$\mathbb{E}\left[\int_{t=0}^T H_t \mathrm{d}B_t \int_{t=0}^T H'_t \mathrm{d}B_t\right] = \mathbb{E}\left[\int_{t=0}^T H_t H'_t \mathrm{d}t\right].$$

Exercise 10.1.11 (Exit from a strip for the Brownian motion)**.** *Let $a < 0 < b$ and define*

$$\tau_a = \inf\{t \geq 0 : B_t \leq a\}, \qquad \tau_b = \inf\{t \geq 0 : B_t \geq b\}.$$

The purpose of the exercise is to compute $\mathbb{P}(\tau_a < \tau_b)$.

1. *Show that $\tau_a \wedge \tau_b$ is a stopping time and that $\tau_a \wedge \tau_b < +\infty$ almost surely.*
2. *Show that, for all $T > 0$, the process $(H_t)_{t\in[0,T]}$ defined by $H_t = \mathbb{1}_{\{t<\tau_a\wedge\tau_b\}}$ belongs to $\mathbf{\Lambda}^2([0,T])$.*
3. *Deduce the identity $0 = a\mathbb{P}(\tau_a < \tau_b) + b\mathbb{P}(\tau_a > \tau_b)$ and conclude.*

Exercise 10.1.12 (Wiener integral)**.** *Let $h : [0,+\infty) \to \mathbb{R}$ be a continuous function, which we rather denote by $(h_t)_{t\geq 0}$ in the sequel. The purpose of the exercise is to study the process $(X_t)_{t\geq 0}$ defined by*

$$X_t = \int_{s=0}^t h_s \mathrm{d}B_s.$$

1. *For all $t > 0$, show that the process $(h^n_s)_{s\in[0,t]}$ defined by*

$$h^n_s = \sum_{i=0}^{n-1} h_{s_i} \mathbb{1}_{[s_i,s_{i+1})}(s), \qquad s_i = \frac{it}{n},$$

converges to h in $\mathbf{\Lambda}^2([0,t])$.
2. *Deduce that X_t is a Gaussian variable.*
3. *Show that for all $0 < s < t$, the random variable $X_t - X_s$ is independent from $\mathcal{F}_s$.*
4. *Conclude that the process $(X_t)_{t\geq 0}$ is Gaussian, with*

$$\mathbb{E}[X_t] = 0, \qquad \mathrm{Cov}(X_s, X_t) = \int_{r=0}^{s\wedge t} h_r^2 \mathrm{d}r.$$

10.1.3 Extension to $\mathbf{\Lambda}_{\mathrm{loc}}$

The condition that the random variable $\int_{t=0}^T H_t^2 \mathrm{d}t$ is in $\mathbf{L}^2(\mathbb{P})$ may seem uselessly restrictive. As an example, given a $\mathcal{F}_0$-measurable random variable ξ_0, it would be natural to be willing to define the stochastic integral of the constant process $H_t = \xi_0$ by

$$\int_{t=0}^T \xi_0 \mathrm{d}B_t = \xi_0 B_T,$$

whether $\mathbb{E}[\xi_0^2] < +\infty$ or not.

In order to relax it in the construction of the stochastic integral, we now introduce the space $\mathbf{\Lambda}_{\mathrm{loc}}$ of processes $H = (H_t)_{t\geq 0}$ which are progressively measurable and such that

$$\forall T > 0, \qquad \int_{t=0}^T H_t^2 \mathrm{d}t < +\infty, \qquad \text{almost surely.}$$

Of course, if H is such that $(H_t)_{t\in[0,T]} \in \mathbf{\Lambda}^2([0,T])$ for all $T > 0$, then $H \in \mathbf{\Lambda}_{\mathrm{loc}}$.

Our construction relies on the notion of *stopping time* introduced in Chapter 9. For $H \in \mathbf{\Lambda}_{\mathrm{loc}}$ and $M \geq 1$, let us denote

$$\tau_M = \inf\left\{T > 0 : \int_{t=0}^T H_t^2 \mathrm{d}t \geq M\right\} \in [0, +\infty].$$

Lemma 10.1.13 (Sequence $(\tau_M)_{M\geq 1}$). *For all $M \geq 1$, τ_M is a $(\mathcal{F}_t)_{t\geq 0}$-stopping time. Besides, almost surely, $\tau_M \to +\infty$ when $M \to +\infty$.*

Proof. By the definition of $\mathbf{\Lambda}^2_{\mathrm{loc}}$, the process $X = (X_T)_{T\geq 0}$ defined by

$$X_T = \int_{t=0}^T H_t^2 \mathrm{d}t$$

is adapted and has continuous trajectories. Exercise 9.3.13 then shows that τ_M is a $(\mathcal{F}_t)_{t\geq 0}$-stopping time.

Let us now assume that there exists $T \in (0, +\infty)$ such that $\tau_M \leq T$ for all $M \geq 1$. Then for all $M \geq 1$,

$$M = X_{\tau_M} = \int_{t=0}^{\tau_M} H_t^2 \mathrm{d}t \leq \int_{t=0}^T H_t^2 \mathrm{d}t,$$

which shows that

$$\int_{t=0}^T H_t^2 \mathrm{d}t = +\infty$$

and is a contradiction with the fact that $H \in \mathbf{\Lambda}_{\mathrm{loc}}$. Hence, τ_M goes to $+\infty$. □

We now let $I = [0,T]$ with $T > 0$. By Lemma 10.1.13, for all $M \geq 1$, the process $H^M = (H_t^M)_{t \in I}$ defined by

$$H_t^M = H_t \mathbb{1}_{\{t \leq \tau_M\}}$$

is progressively measurable and naturally satisfies $\|H^M\|_{\mathbf{\Lambda}^2(I)}^2 \leq M < +\infty$, so that $H^M \in \mathbf{\Lambda}^2(I)$: this first remark allows to define the stochastic integral $\mathfrak{J}_I(H^M)$. Furthermore, as soon as $\tau_M \geq T$, the sequence $\mathfrak{J}_I(H^M)$ becomes constant, and therefore it has an almost sure limit.

Definition 10.1.14 (Stochastic integral on $\mathbf{\Lambda}_{\mathrm{loc}}$)**.** *For all $H = (H_t)_{t \geq 0} \in \mathbf{\Lambda}_{\mathrm{loc}}$, for all $I = [0,T]$ with $T > 0$, the stochastic integral of H on I is defined by*

$$\int_{t=0}^{T} H_t \mathrm{d}B_t = \lim_{M \to +\infty} \mathfrak{J}_I(H^M), \qquad \textit{almost surely.}$$

The stochastic integral of H on arbitrary intervals $[a,b] \subset [0,+\infty)$ is then defined by the Chasles relation.

Remark 10.1.15. *Let $H = (H_t)_{t \geq 0} \in \mathbf{\Lambda}_{\mathrm{loc}}$ and $I = [0,T]$, $T > 0$, be such that $(H_t)_{t \in I} \in \mathbf{\Lambda}^2(I)$. Let us check that the Definition 10.1.14 of the stochastic integral of H on I actually agrees with the definition given by Corollary 10.1.6. To this aim, let us write*

$$\mathbb{E}\left[\left(\mathfrak{J}_I(H) - \mathfrak{J}_I(H^M)\right)^2\right] = \mathbb{E}\left[\left(\int_{t=0}^{T} H_t \mathbb{1}_{\{t > \tau_M\}} \mathrm{d}B_t\right)^2\right],$$

where we have used the fact that both processes H and H^M belong to $\mathbf{\Lambda}^2(I)$, and the linearity of $\mathfrak{J}_I$ on $\mathbf{\Lambda}^2(I)$. The Itô isometry now shows that

$$\mathbb{E}\left[\left(\int_{t=0}^{T} H_t \mathbb{1}_{\{t > \tau_M\}} \mathrm{d}B_t\right)^2\right] = \mathbb{E}\left[\int_{t=0}^{T} H_t^2 \mathbb{1}_{\{t > \tau_M\}} \mathrm{d}t\right],$$

and the Dominated Convergence Theorem ensures that the right-hand side converges to 0 when $M \to +\infty$. Thus, the sequence $\mathfrak{J}_I(H^M)$ converges: almost surely to the stochastic integral of H on I in the sense of Definition 10.1.14; and in $\mathbf{L}^2$ to $\mathfrak{J}_I(H)$, that is to say the definition of the stochastic integral of H on I given by Corollary 10.1.6. We deduce that the two limits coincide almost surely.

Remark 10.1.16. *It is easy to check that the properties (i), (ii) and (iii) of Proposition 10.1.7 are preserved by the extension of the stochastic integral to $\mathbf{\Lambda}_{\mathrm{loc}}$. In contrast, the Itô isometry no longer holds necessarily; in fact, if H is only assumed to be in $\mathbf{\Lambda}_{\mathrm{loc}}$, the random variable $\int_{t=a}^{b} H_t \mathrm{d}B_t$ need not even be in $\mathbf{L}^1(\mathbb{P})$. The following practical rule summarises the situation: if $H \in \mathbf{\Lambda}_{\mathrm{loc}}$ and $T > 0$ are such that*

$$\mathbb{E}\left[\int_{t=0}^{T} H_t^2 \mathrm{d}t\right] < +\infty,$$

then we have

$$\mathbb{E}\left[\int_{t=0}^{T} H_t \mathrm{d}B_t\right] = 0 \qquad \textit{and} \qquad \mathbb{E}\left[\left(\int_{t=0}^{T} H_t \mathrm{d}B_t\right)^2\right] = \mathbb{E}\left[\int_{t=0}^{T} H_t^2 \mathrm{d}t\right];$$

otherwise, nothing can be said.

10.2 Itô's formula

10.2.1 Itô's formula for $\Phi(B_t)$

The result of Exercise 10.1.9 rewrites in the differential notation

$$B_t \mathrm{d}B_t = \frac{1}{2}(\mathrm{d}(B_t^2) - \mathrm{d}t),$$

so that with $\Phi(x) = x^2$,

$$\mathrm{d}\Phi(B_t) = \Phi'(B_t)\mathrm{d}B_t + \mathrm{d}t.$$

Therefore, the chain rule formula recalled at the beginning of this section does not apply to the Brownian motion.

Theorem 10.2.1 (Itô's formula for the Brownian motion). *Let $(B_t)_{t\geq 0}$ be a $(\mathcal{F})_{t\geq 0}$-Brownian motion and $\Phi : \mathbb{R} \to \mathbb{R}$ be a C^2 function. For all $T \geq 0$,*

$$\Phi(B_T) = \Phi(0) + \int_{t=0}^{T} \Phi'(B_t)\mathrm{d}B_t + \frac{1}{2}\int_{t=0}^{T} \Phi''(B_t)\mathrm{d}t, \qquad \textit{almost surely,}$$

which we shall also write

$$\mathrm{d}\Phi(B_t) = \Phi'(B_t)\mathrm{d}B_t + \frac{1}{2}\Phi''(B_t)\mathrm{d}t.$$

Remark 10.2.2. *Before detailing the proof of Theorem 10.2.1, we formulate a few remarks.*

(i) Since Φ is C^2, the functions $t \mapsto \Phi'(B_t)$ and $t \mapsto \Phi''(B_t)$ are almost surely continuous on the interval $[0, T]$, and thus they are almost surely bounded. Therefore the right-hand side in Itô's formula is well defined; in particular, the stochastic integral is a priori *understood in the sense of its definition on $\mathbf{\Lambda}_{\mathrm{loc}}$.*

(ii) The differential notation reveals a similarity with a second-order expansion of Φ. We shall indeed exploit this idea in the proof.

(iii) Applying Itô's formula with $\Phi(x) = x^2$, we recover the identity obtained at Exercise 10.1.9, but much more rapidly!

Proof of Theorem 10.2.1. We only detail the proof in the case where the function $|\Phi''|$ is bounded on $\mathbb{R}$ by some $M \geq 0$, and we admit that the result holds in general (see [7, Theorem 3.3] for a complete exposition).

Let us fix $T > 0$, $n \geq 1$, and set $t_i = iT/n$ for any $i \in \{0, \dots, n\}$. Following the Taylor–Lagrange formula, for any i there exists $\theta_i \in [t_i, t_{i+1}]$ such that

$$\begin{aligned}\Phi(B_T) - \Phi(B_0) &= \sum_{i=0}^{n-1} \Phi(B_{t_{i+1}}) - \Phi(B_{t_i})\\ &= \sum_{i=0}^{n-1} \Phi'(B_{t_i})\left(B_{t_{i+1}} - B_{t_i}\right) + \frac{1}{2}\Phi''(B_{\theta_i})\left(B_{t_{i+1}} - B_{t_i}\right)^2.\end{aligned}$$

We first check that

$$\lim_{n\to+\infty} \sum_{i=0}^{n-1} \Phi'(B_{t_i})\left(B_{t_{i+1}} - B_{t_i}\right) = \int_{t=0}^{T} \Phi'(B_t)\mathrm{d}B_t, \qquad \text{in } \mathbf{L}^2.$$

Let us note that, since Φ'' is bounded, the process H^n defined by $H_t^n = \sum_{i=0}^{n-1} \Phi'(B_{t_i})\mathbb{1}_{[t_i,t_{i+1})}(t)$ belongs to the set $\mathbf{\Lambda}_0^2(I)$ with $I = [0,T]$. Hence, by the definition of the stochastic integral, it suffices to show that H^n converges to the process H defined by $H_t = \Phi'(B_t)$ in $\mathbf{\Lambda}^2(I)$. But

$$\begin{aligned}
\|H - H^n\|_{\mathbf{\Lambda}^2(I)}^2 &= \sum_{i=0}^{n-1} \int_{t=t_i}^{t_{i+1}} \mathbb{E}\left[(H_t - H_{t_i})^2\right] \mathrm{d}t \\
&= \sum_{i=0}^{n-1} \int_{t=t_i}^{t_{i+1}} \mathbb{E}\left[(\Phi'(B_t) - \Phi'(B_{t_i}))^2\right] \mathrm{d}t \\
&\leq M^2 \sum_{i=0}^{n-1} \int_{t=t_i}^{t_{i+1}} \mathbb{E}\left[(B_t - B_{t_i})^2\right] \mathrm{d}t \\
&= M^2 \sum_{i=0}^{n-1} \int_{t=t_i}^{t_{i+1}} (t - t_i)\mathrm{d}t = \frac{M^2T^2}{2n},
\end{aligned}$$

which leads to the claimed identity.

We now show that

$$\lim_{n\to+\infty} \sum_{i=0}^{n-1} \Phi''(B_{\theta_i}) \left(B_{t_{i+1}} - B_{t_i}\right)^2 = \int_{t=0}^{T} \Phi''(B_t)\mathrm{d}t, \qquad \text{in probability.}$$

In this purpose, we introduce the notation

$$\begin{aligned}
U_n &= \sum_{i=0}^{n-1} \Phi''(B_{\theta_i}) \left(B_{t_{i+1}} - B_{t_i}\right)^2, \\
V_n &= \sum_{i=0}^{n-1} \Phi''(B_{t_i}) \left(B_{t_{i+1}} - B_{t_i}\right)^2, \\
W_n &= \sum_{i=0}^{n-1} \Phi''(B_{t_i}) \left(t_{i+1} - t_i\right),
\end{aligned}$$

and argue in three steps.

Step 1: $U_n - V_n \to 0$ in $\mathbf{L}^1$. Let us write

$$\begin{aligned}
\mathbb{E}\left[|U_n - V_n|\right] &= \mathbb{E}\left[\left|\sum_{i=0}^{n-1} \left(\Phi''(B_{\theta_i}) - \Phi''(B_{t_i})\right) \left(B_{t_{i+1}} - B_{t_i}\right)^2\right|\right] \\
&\leq \mathbb{E}\left[\max_{0\leq i\leq n-1} \left|\Phi''(B_{\theta_i}) - \Phi''(B_{t_i})\right| \sum_{i=0}^{n-1} \left(B_{t_{i+1}} - B_{t_i}\right)^2\right] \\
&\leq \sqrt{\mathbb{E}\left[\max_{0\leq i\leq n-1} |\Phi''(B_{\theta_i}) - \Phi''(B_{t_i})|^2\right] \mathbb{E}\left[\left(\sum_{i=0}^{n-1} \left(B_{t_{i+1}} - B_{t_i}\right)^2\right)^2\right]},
\end{aligned}$$

by the Cauchy–Schwarz inequality. On the one hand, the almost sure uniform continuity of the function $t \mapsto \Phi''(B_t)$ shows that

$$\lim_{n\to+\infty} \max_{0\leq i\leq n-1} \left|\Phi''(B_{\theta_i}) - \Phi''(B_{t_i})\right|^2 = 0, \qquad \text{almost surely,}$$

and since this function is bounded by $2M$, the Dominated Convergence Theorem ensures that

$$\lim_{n\to+\infty} \mathbb{E}\left[\max_{0\leq i\leq n-1} \left|\Phi''(B_{\theta_i}) - \Phi''(B_{t_i})\right|^2\right] = 0.$$

On the other hand, by Exercise 9.2.10,

$$\lim_{n\to+\infty} \mathbb{E}\left[\left(\sum_{i=0}^{n-1} (B_{t_{i+1}} - B_{t_i})^2\right)^2\right] = T^2.$$

Thus,

$$\lim_{n\to+\infty} \mathbb{E}\left[|U_n - V_n|\right] = 0.$$

Step 2: $V_n - W_n \to 0$ in $\mathbf{L}^2$. Let us now write

$$\mathbb{E}\left[(V_n - W_n)^2\right] = \mathbb{E}\left[\left(\sum_{i=0}^{n-1} \Phi''(B_{t_i})\left[(B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1} - t_i)\right]\right)^2\right]$$

and set $Y_i = (B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1} - t_i)$. Notice that Y_i is independent from $\mathcal{F}_{t_i}$ and that

$$\mathbb{E}[Y_i] = 0, \qquad \mathbb{E}[Y_i^2] = \left(\frac{T}{n}\right)^2 \mathbb{E}[(G^2-1)^2], \quad G \sim \mathcal{N}(0,1).$$

Therefore,

$$\mathbb{E}\left[\left(\sum_{i=0}^{n-1} \Phi''(B_{t_i})Y_i\right)^2\right] = \sum_{i,j=0}^{n-1} \mathbb{E}\left[\Phi''(B_{t_i})Y_i\Phi''(B_{t_j})Y_j\right]$$

and when $i < j$, $\mathbb{E}[\Phi''(B_{t_i})Y_i\Phi''(B_{t_j})Y_j] = \mathbb{E}[\Phi''(B_{t_i})Y_i\Phi''(B_{t_j})]\mathbb{E}[Y_j] = 0$. We deduce that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=0}^{n-1} \Phi''(B_{t_i})Y_i\right)^2\right] &= \sum_{i=0}^{n-1} \mathbb{E}\left[\left(\Phi''(B_{t_i})Y_i\right)^2\right] \\ &\leq M^2 \sum_{i=0}^{n-1} \mathbb{E}\left[Y_i^2\right] = \frac{M^2T^2}{n}\mathbb{E}[(G^2-1)^2],\end{aligned}$$

whence

$$\lim_{n\to+\infty} \mathbb{E}\left[(V_n - W_n)^2\right] = 0.$$

Step 3: almost sure limit of W_n. W_n is a Riemann sum for the almost surely continuous function $t \mapsto \Phi''(B_t)$ on $[0,T]$. Therefore, almost surely,

$$\lim_{n\to+\infty} W_n = \int_{t=0}^{T} \Phi''(B_t)\mathrm{d}t.$$

Conclusion. We deduce from Steps 1, 2 and 3 that U_n converges in probability to $\int_{t=0}^{T} \Phi''(B_t)\mathrm{d}t$, which finally shows that

$$\Phi(B_T) - \Phi(B_0) = \int_{t=0}^{T} \Phi'(B_t)\mathrm{d}B_t + \frac{1}{2}\int_{t=0}^{T} \Phi''(B_t)\mathrm{d}t, \qquad \text{almost surely,}$$

and completes the proof. □

Exercise 10.2.3. *Express $\int_{t=0}^{T} B_t^2\mathrm{d}B_t$ as a function of B_T and $\int_{t=0}^{T} B_t\mathrm{d}t$.*

10.2.2 Itô process and quadratic variation

Definition 10.2.4. *A real-valued* Itô process *is a process $X = (X_t)_{t \geq 0}$ which writes, for any $T \geq 0$,*

$$X_T = X_0 + \int_{t=0}^{T} K_t \mathrm{d}t + \int_{t=0}^{T} H_t \mathrm{d}B_t,$$

where:

- *X_0 is $\mathcal{F}_0$-measurable;*
- *the process $K = (K_t)_{t \geq 0}$ is progressively measurable and, for all $T > 0$, $\int_{t=0}^{T} |K_t| \mathrm{d}t < +\infty$ almost surely;*
- *the process $H = (H_t)_{t \geq 0}$ is progressively measurable and, for all $T > 0$, $\int_{t=0}^{T} H_t^2 \mathrm{d}t < +\infty$ almost surely.*

The last condition equivalently writes $H \in \mathbf{\Lambda}_{\mathrm{loc}}$, which ensures that the stochastic integral is well-defined.

We shall often use the differential notation $\mathrm{d}X_t = K_t \mathrm{d}t + H_t \mathrm{d}B_t$.

The process $T \mapsto \int_{t=0}^{T} K_t \mathrm{d}t$ is called the *bounded variation component* of X.

The following result is admitted.

Lemma 10.2.5 (Decomposition of an Itô process)**.** *An Itô process is adapted and it admits an almost surely continuous modification. Its decomposition (X_0, K, H) is almost surely unique.*

Remark 10.2.6. *As a consequence of Lemma 10.2.5, Itô processes are always progressively measurable.*

Definition 10.2.7 (Quadratic variation)**.** *The* quadratic variation *of an Itô process $X = (X_t)_{t \geq 0}$ is the nonnegative and nondecreasing process $\langle X \rangle = (\langle X \rangle_t)_{t \geq 0}$ defined by*

$$\langle X \rangle_T = \int_{t=0}^{T} H_t^2 \mathrm{d}t.$$

We shall often use the notation $\mathrm{d}\langle X \rangle_t = H_t^2 \mathrm{d}t$.

Exercise 10.2.8. *Show that, for all $\lambda \in \mathbb{R}$, $\langle \lambda X \rangle_t = \lambda^2 \langle X \rangle_t$.*

Exercise 10.2.9.

1. *Show that B is an Itô process, give its decomposition and check that $\langle B \rangle_t = t$.*
2. *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a C^2 function. Show that $X = (\Phi(B_t))_{t \geq 0}$ is an Itô process, give its decomposition and compute its quadratic variation.*

The following generalisation of Itô's formula to Itô processes is admitted (the proof is similar to Theorem 10.2.1).

Theorem 10.2.10 (Itô formula for Itô processes)**.** *Let $(X_t)_{t \geq 0}$ be an Itô process and $\Phi : \mathbb{R} \to \mathbb{R}$ be a C^2 function. The process $(\Phi(X_t))_{t \geq 0}$ is an Itô process, and for any $T \geq 0$,*

$$\Phi(X_T) = \Phi(X_0) + \int_{t=0}^{T} \left(\Phi'(X_t) K_t + \frac{1}{2} \Phi''(X_t) H_t^2 \right) \mathrm{d}t + \int_{t=0}^{T} \Phi'(X_t) H_t \mathrm{d}B_t, \qquad \textit{almost surely,}$$

which is also written under the form

$$\mathrm{d}\Phi(X_t) = \Phi'(X_t) \mathrm{d}X_t + \frac{1}{2} \Phi''(X_t) \mathrm{d}\langle X \rangle_t.$$

🏠 **Exercise 10.2.11** (Geometric Brownian motion)**.** *The goal of this exercise is to find an Itô process X satisfying the identity*

$$\mathrm{d}X_t = \frac{1}{2}X_t\mathrm{d}t + X_t\mathrm{d}B_t, \qquad X_0 = 1,$$

which is a first example of a stochastic differential equation.

1. Analysis. *Assume that there exists a solution X such that almost surely, $X_t > 0$ for all $t > 0$. Compute* $\ln X_t$.
2. Synthesis. *Deduce a solution to this equation.*

☕ **Exercise 10.2.12** (Lévy's characterisation of the Brownian motion)**.** *Let $H = (H_t)_{t\geq 0} \in \mathbf{\Lambda}_{\mathrm{loc}}$ be such that $H_t^2 = 1$ for all $t \geq 0$. The purpose of this exercise is to prove that the process $(X_t)_{t\geq 0}$ defined by the stochastic integral*

$$X_t = \int_{s=0}^t H_s\mathrm{d}B_s$$

is a Brownian motion. This result is called Lévy's characterisation of the Brownian motion.

1. *Let $0 = t_0 \leq t_1 \leq \cdots \leq t_k$ and $u_1, \ldots, u_k \in \mathbb{R}$. For all $t \geq 0$, we define*

   $$\widetilde{H}_t = H_t \sum_{j=1}^k u_j \mathbb{1}_{\{t_{j-1}\leq t<t_j\}},$$

   and

   $$Y_t = \exp\left(\mathrm{i}\int_{s=0}^t \widetilde{H}_s\mathrm{d}B_s + \frac{1}{2}\int_{s=0}^t \widetilde{H}_s^2\mathrm{d}s\right).$$

   Show that $(Y_t)_{t\geq 0}$ is an Itô process with no bounded variation component.
2. *Deduce the value of*

   $$\mathbb{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^k u_j(X_{t_j} - X_{t_{j-1}})\right)\right].$$

3. *Conclude.*

10.2.3 Multidimensional version and applications

Let $(B_t)_{t\geq 0} = (B_t^1, \ldots, B_t^d)_{t\geq 0}$ be a Brownian motion[4] in $\mathbb{R}^d$ such that, for each $k \in \{1, \ldots, d\}$, $(B_t^k)_{t\geq 0}$ is a $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion.

Definition 10.2.13 (Itô process driven by $(B_t)_{t\geq 0}$)**.** *An Itô process* driven by the d-dimensional Brownian motion $(B_t)_{t\geq 0}$ *is a process $(X_t)_{t\geq 0}$ which writes*

$$X_T = X_0 + \int_{t=0}^T K_t\mathrm{d}t + \sum_{k=1}^d \int_{t=0}^T H_t^k\mathrm{d}B_t^k,$$

where:

- *X_0 is $\mathcal{F}_0$-measurable;*
- *the process K is progressively measurable and, for all $T > 0$, $\int_{t=0}^T |K_t|\mathrm{d}t < +\infty$ almost surely;*

[4] We recall from Subsection 9.2.4 that this means that the processes $(B_t^1)_{t\geq 0}, \ldots, (B_t^d)_{t\geq 0}$ are independent Brownian motions.

- *the processes $H^1, \dots, H^d$ belong to $\mathbf{\Lambda}_{\mathrm{loc}}$.*

We shall naturally denote $\mathrm{d}X_t = K_t\mathrm{d}t + \sum_{k=1}^d H_t^k \mathrm{d}B_t^k$, and admit that the decomposition $(X_0, K, (H^k)_{1\leq k\leq d})$ remains unique.

Definition 10.2.14 (Quadratic covariation)**.** *Let $(X_t)_{t\geq 0}$ and $(X'_t)_{t\geq 0}$ be two Itô processes driven by $(B_t)_{t\geq 0}$, with respective decompositions $(X_0, K, (H^k)_{1\leq k\leq d})$ and $(X'_0, K', (H'^k)_{1\leq k\leq d})$. The* quadratic covariation *of X and X' is the process $\langle X, X'\rangle = (\langle X, X'\rangle_t)_{t\geq 0}$ defined by*

$$\langle X, X'\rangle_T = \sum_{k=1}^d \int_{t=0}^T H_t^k H_t'^k \mathrm{d}t.$$

We shall use the notation $\mathrm{d}\langle X, X'\rangle_t = \sum_{k=1}^d H_t^k H_t'^k \mathrm{d}t$.

We also denote $\langle X, X\rangle = \langle X\rangle$, which generalises Definition 10.2.7.

Clearly, the quadratic covariation is symmetric and bilinear.

Exercise 10.2.15. *For $k, l \in \{1, \dots, d\}$, compute $\langle B^k, B^l\rangle$.*

We may now state the Itô formula in its more general formulation.

Theorem 10.2.16 (Multidimensional Itô formula)**.** *Let $(X_t)_{t\geq 0} = (X_t^1, \dots, X_t^n)_{t\geq 0}$ be a process with values in $\mathbb{R}^n$, of which each coordinate $(X_t^i)_{t\geq 0}$ is an Itô process driven by the d-dimensional Brownian motion $(B_t)_{t\geq 0}$, and let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a C^2 function. The Itô formula writes*

$$\mathrm{d}\Phi(X_t) = \sum_{i=1}^n \frac{\partial \Phi}{\partial x_i}(X_t)\mathrm{d}X_t^i + \frac{1}{2}\sum_{i,j=1}^n \frac{\partial^2 \Phi}{\partial x_i \partial x_j}(X_t)\mathrm{d}\langle X^i, X^j\rangle_t.$$

Exercise 10.2.17 (Integration by parts formula)**.** *Show that, if $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ are two Itô processes, then*

$$\mathrm{d}(X_tY_t) = X_t\mathrm{d}Y_t + Y_t\mathrm{d}X_t + \mathrm{d}\langle X, Y\rangle_t.$$

The Itô formula also applies when Φ depends on time. A function $\Phi : [0, +\infty) \times \mathbb{R}^n \to \mathbb{R}$ is called $C^{1,2}$ when its partial derivatives $\frac{\partial \Phi}{\partial t}$, $\frac{\partial \Phi}{\partial x_i}$ and $\frac{\partial^2 \Phi}{\partial x_i \partial x_j}$ exist and are continuous on $[0, +\infty) \times \mathbb{R}^n$.

Proposition 10.2.18 (Time dependent Itô formula)**.** *Let $(X_t)_{t\geq 0} = (X_t^1, \dots, X_t^n)_{t\geq 0}$ be a process with values in $\mathbb{R}^n$, of which each coordinate $(X_t^i)_{t\geq 0}$ is an Itô process driven by the d-dimensional Brownian motion $(B_t)_{t\geq 0}$, and let $\Phi : [0, +\infty) \times \mathbb{R}^n \to \mathbb{R}$ be a $C^{1,2}$ function. We have*

$$\mathrm{d}\Phi(t, X_t) = \frac{\partial \Phi}{\partial t}(t, X_t)\mathrm{d}t + \sum_{i=1}^n \frac{\partial \Phi}{\partial x_i}(t, X_t)\mathrm{d}X_t^i + \frac{1}{2}\sum_{i,j=1}^n \frac{\partial^2 \Phi}{\partial x_i \partial x_j}(t, X_t)\mathrm{d}\langle X^i, X^j\rangle_t.$$

Proof. Let us remark that the process $(\widetilde{X}_t)_{t\geq 0} = (t, X_t^1, \dots, X_t^n)_{t\geq 0}$ with values in $\mathbb{R}^{n+1}$ satisfy the assumptions of Theorem 10.2.16, since the process $t \mapsto t$ is an Itô process with decomposition $(0, 1, 0)$. It is therefore clear that for any $k \in \{1, \dots, d\}$, $\langle t, X^k\rangle = 0$. Thus, if Φ is C^2 on $[0, +\infty) \times \mathbb{R}^n$, the claimed formula follows from the application of Theorem 10.2.16 to the process $\widetilde{X}$. We admit that if Φ is only $C^{1,2}$, it is possible to conclude by a regularisation argument. □

Exercise 10.2.19 (Hitting times for Brownian motion)**.** *Let B be a Brownian motion. For all $a > 0$, let us define the stopping time*

$$\tau_a = \inf\{t \geq 0 : B_t \geq a\}.$$

The purpose of this exercise is to recover some results from Subsection 9.3.5 using stochastic calculus instead of the strong Markov property.

1. *Let $\sigma > 0$. Show that the process $(X_t)_{t\geq 0}$ defined by*
$$X_t = \exp\left(\sigma B_t - \frac{\sigma^2}{2}t\right)$$
is an Itô process with no bounded variation component.
2. *Let $T > 0$. Show that the process $(\widetilde{H}_t)_{t\in[0,T]}$ defined by*
$$\widetilde{H}_t = \mathbb{1}_{\{t<\tau_a\}}\sigma X_t$$
is in $\mathbf{\Lambda}^2([0,T])$.
3. *Deduce that*
$$\mathbb{E}\left[\mathbb{1}_{\{\tau_a<+\infty\}}\exp\left(-\frac{\sigma^2}{2}\tau_a\right)\right] = \exp(-\sigma a).$$
4. *Conclude that $\tau_a < +\infty$ almost surely.*
5. *Show that τ_a has the same law as a^2/G^2, where $G \sim \mathcal{N}(0,1)$.*
Hint: you may first admit that two nonnegative random variables X and Y have the same law if and only if $\mathbb{E}[\mathrm{e}^{-\lambda X}] = \mathbb{E}[\mathrm{e}^{-\lambda Y}]$ for all $\lambda > 0$. Then, set $f(a) = \mathbb{E}[\mathrm{e}^{-\lambda a^2/G^2}]$ and find a clever change of variable to link $f'(a)$ with $f(a)$.

Exercise 10.2.20 (Recurrence and transience of the multidimensional Brownian motion)**.** *Let $d \geq 2$ and $(B_t)_{t\geq 0}$ be a d-dimensional Brownian motion. We fix $x \in \mathbb{R}^d$ and let $X_t^x = x + B_t$. For all $\rho \geq 0$, we define*
$$\tau_\rho^x = \inf\{t \geq 0 : |X_t^x| = \rho\},$$
where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^d$. For $0 < r < R$, we finally write
$$C(r,R) = \{x \in \mathbb{R}^d : r < |x| < R\}.$$

1. Preliminary results. *We assume that $x \in C(r,R)$.*
   (a) *Show that $\tau_R^x < +\infty$, almost surely.*
   (b) *Show that $\lim_{R\to+\infty}\tau_R^x = +\infty$, almost surely.*
   (c) *Show that $\lim_{r\to 0}\tau_r^x = \tau_0^x$, almost surely.*
2. Harmonic functions in $C(r,R)$. *We consider the partial differential equation*
$$\begin{cases} \Delta u(x) = 0, & x \in C(r,R) \\ u(x) = 1, & |x| = r, \\ u(x) = 0, & |x| = R. \end{cases}$$
   (a) *Find an explicit solution of the form $u(x) = f(|x|^2)$, where the function f is continuous on $[r,R]$ and C^2 on (r,R).*
   (b) *Show that $u(x) = \mathbb{P}(\tau_r^x < \tau_R^x)$.*
3. Hitting points. *Deduce that almost surely, $\tau_0^x = +\infty$.*
4. Recurrence in dimension $d = 2$. *If $d = 2$, show that for any $r > 0$, $\tau_r^x < +\infty$ almost surely.*
5. Transience in dimension $d \geq 3$. *If $d \geq 3$, show that for any $r > 0$ and $x \in \mathbb{R}^d$ such that $r \leq |x|$, $\mathbb{P}(\tau_r^x < +\infty) = (r/|x|)^{d-2}$.*

In the case $d = 2$, the result of Exercise 10.2.20 shows that starting from any point $x \in \mathbb{R}^d$, the Brownian motion enters any small ball centered in 0 in finite time. Once it has reached this ball, the strong Markov property shows that it 'starts afresh' and reaches again any other small ball in finite time, see the schematic depiction of Figure 10.1. Iterating this argument, we deduce

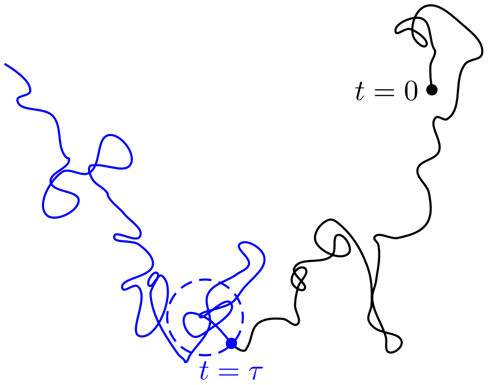


Figure 10.1: Strong Markov property and recurrence for the 2-dimensional Brownian motion: the Brownian motion hits the dashed blue ball at the random time τ. Then the trajectory $(B_{t+\tau} - B_\tau)_{t\geq 0}$ is independent from the trajectory $(B_t)_{t\in[0,\tau]}$.

that the Brownian motion visits all open sets of $\mathbb{R}^d$ infinitely often, which is the reason why it is called *recurrent*.

In dimension $d \geq 3$, on the contrary, one may combine the result of Exercise 10.2.20 with the strong Markov property to show that almost surely, $\lim_{t\to+\infty} |B_t| = +\infty$, so that for any compact set K, there is a finite time after which the Brownian motion no longer comes back to K. It is then called *transient*. Whatever the dimension d, the long time behaviour of the Brownian motion in $\mathbb{R}^d$ is therefore similar to the long time behaviour of the random walk in $\mathbb{Z}^d$ studied in Exercise 5.4.9.

Chapter 11

Stochastic differential equations

Contents

Let $n, d \geq 1$, $I = [0, T]$ with $T > 0$ or $I = [0, +\infty)$, and $b : I \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : I \times \mathbb{R}^n \to \mathbb{R}^{n \times d}$ be measurable functions. Let $(B_t)_{t \in I}$ be a d-dimensional Brownian motion defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ endowed with a filtration $(\mathcal{F}_t)_{t \in I}$ such that $\mathcal{F}_0$ is complete and with respect to which the coordinates of $(B_t)_{t \in I}$ are $(\mathcal{F}_t)_{t \in I}$-Brownian motions.

A *stochastic differential equation* (SDE) is an equation of the form

$$\mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t, \tag{11.1}$$

where the unknown is an n-dimensional Itô process $(X_t)_{t \in I} = (X_t^1, \ldots, X_t^n)_{t \in I}$. It is usually complemented with an initial condition

$$X_0 = \xi, \tag{11.2}$$

where ξ is an $\mathcal{F}_0$-measurable random variable in $\mathbb{R}^n$.

Equivalently, introducing the notation $b = (b_i)_{1 \leq i \leq n}$ and $\sigma = (\sigma_{i,k})_{1 \leq i \leq n, 1 \leq k \leq d}$ for the coordinates of b and σ, the system (11.1)–(11.2) rewrites under the form

$$\forall i \in \{1, \ldots, n\}, \quad \forall t \in I, \qquad X_t^i = \xi_0^i + \int_{s=0}^t b_i(s, X_s)\mathrm{d}s + \sum_{k=1}^d \int_{s=0}^t \sigma_{i,k}(s, X_s)\mathrm{d}B_s^k.$$

We shall call a stochastic process $(X_t)_{t \in I}$ which solves a SDE of the form (11.1) a *diffusion process*. The function b is called the *drift* and σ the *diffusion coefficient* of the SDE.

Throughout this section, we shall always assume that the functions b and σ are bounded on bounded subsets of $I \times \mathbb{R}^n$. The notation $|\cdot|$ shall be used to refer indifferently to the Euclidean norm on $\mathbb{R}^n$ or to the Frobenius norm[1] on $\mathbb{R}^{n \times d}$.

[1] The Frobenius norm of a matrix $s = (s_{i,k})_{1 \leq i \leq n, 1 \leq k \leq d} \in \mathbb{R}^{n \times d}$ is defined by $|s| = (\sum_{i=1}^n \sum_{k=1}^d s_{i,k}^2)^{1/2}$.

11.1 Existence and uniqueness in the Lipschitz case

When $\sigma \equiv 0$, the SDE (11.1) reduces to the classical ODE $\dot{x}_t = b(t, x_t)$, for which well-posedness is ensured by the Cauchy–Lipschitz Theorem. The latter possesses the following stochastic version.

Theorem 11.1.1 (Itô Theorem)**.** *Let $T > 0$, $b : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times d}$ be such that there exists $K \in [0,+\infty)$ for which, for all $t \in [0,T]$ and $x, y \in \mathbb{R}^n$,*

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \leq K|x-y|, \qquad |b(t,x)| + |\sigma(t,x)| \leq K(1+|x|).$$

For any $\mathcal{F}_0$-measurable random variable $\xi \in \mathbb{R}^n$, there is a unique n-dimensional Itô process $(X_t)_{t\in[0,T]} = (X_t^1, \ldots, X_t^n)_{t\in[0,T]}$ such that, for all $t \in [0,T]$,

$$X_t = \xi + \int_{s=0}^{t} b(s, X_s)\mathrm{d}s + \int_{s=0}^{t} \sigma(s, X_s)\mathrm{d}B_s.$$

The proof of Theorem 11.1.1 relies on a fixed-point argument, which will be performed in the space $\mathbf{\Lambda}^2([0,T])$. We therefore need the following a priori estimate.

Lemma 11.1.2 ($\mathbf{\Lambda}^2([0,T])$ a priori estimate)**.** *Under the assumptions of Theorem 11.1.1, if $\mathbb{E}[|\xi|^2] < +\infty$ and $(X_t)_{t\in[0,T]}$ is a solution to (11.1)–(11.2), then for any $i \in \{1,\ldots,n\}$, $(X_t^i)_{t\in[0,T]} \in \mathbf{\Lambda}^2([0,T])$.*

Proof. By Remark 10.2.6, each process $(X_t^i)_{t\in[0,T]}$ is progressively measurable. The proof of square-integrability now relies on the use of a *localisation* procedure: for $M \geq 0$, let $\tau_M := \inf\{t \geq 0 : |X_t| \geq M\}$, with the convention that $\tau_M = +\infty$ if $\sup_{t\in[0,T]} |X_t| < M$. By similar arguments to the proof of Lemma 10.1.13, τ_M is a stopping time and $\tau_M = +\infty$ for M large enough, almost surely. For any $t \geq 0$, using the elementary inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, we have

$$\begin{aligned}
&\mathbb{E}\left[|X_{t\wedge\tau_M}|^2\right] \\
&\leq 3\left(\mathbb{E}\left[|\xi|^2\right] + \mathbb{E}\left[\left|\int_{s=0}^{t\wedge\tau_M} b(s,X_s)\mathrm{d}s\right|^2\right] + \mathbb{E}\left[\left|\int_{s=0}^{t\wedge\tau_M} \sigma(s,X_s)\mathrm{d}B_s\right|^2\right]\right) \\
&= 3\left(\mathbb{E}\left[|\xi|^2\right] + \mathbb{E}\left[\left|\int_{s=0}^{t} \mathbb{1}_{\{s<\tau_M\}} b(s,X_s)\mathrm{d}s\right|^2\right] + \mathbb{E}\left[\left|\int_{s=0}^{t} \mathbb{1}_{\{s<\tau_M\}}\sigma(s,X_s)\mathrm{d}B_s\right|^2\right]\right) \\
&\leq 3\left(\mathbb{E}\left[|\xi|^2\right] + T\int_{s=0}^{t} \mathbb{E}\left[\mathbb{1}_{\{s<\tau_M\}}|b(s,X_s)|^2\right]\mathrm{d}s + \int_{s=0}^{t} \mathbb{E}\left[\mathbb{1}_{\{s<\tau_M\}}|\sigma(s,X_s)|^2\right]\mathrm{d}s\right),
\end{aligned}$$

where we have used the Cauchy–Schwarz inequality and Itô's isometry at the last line. Since $\mathbb{1}_{\{t<\tau_M\}}|X_t|^2 \leq |X_{t\wedge\tau_M}|^2$, we deduce that $u(t) := \mathbb{E}[\mathbb{1}_{\{t<\tau_M\}}|X_t|^2]$ satisfies

$$\forall t \in [0,T], \qquad u(t) \leq C\left(1 + \int_{s=0}^{t} u(s)\mathrm{d}s\right),$$

for some finite constant $C \geq 0$ which depends on $\mathbb{E}[|\xi|^2]$, T and K, but not on M. Moreover, $u(t) \leq M^2$. Therefore, by Gronwall's Lemma,

$$\forall t \in [0,T], \qquad u(t) \leq C\mathrm{e}^{Ct}.$$

By the Monotone Convergence Theorem, taking the $M \to +\infty$ limit yields

$$\forall t \in [0,T], \qquad \mathbb{E}\left[|X_t|^2\right] \leq C\mathrm{e}^{Ct},$$

which finally implies that $\int_{t=0}^{T} \mathbb{E}[|X_t|^2]\mathrm{d}t < +\infty$ and concludes. □

Proof of Theorem 11.1.1. We detail the proof for the simple case $d = n = 1$, but the arguments carry over to any values of d and n immediately. We also assume that $\mathbb{E}[|\xi|^2] < +\infty$ and refer to Remark 11.1.3 below for the extension to any $\mathcal{F}_0$-measurable initial condition ξ. Then by Lemma 11.1.2, it suffices to show that (11.1)–(11.2) has a unique solution in $\mathbf{\Lambda}^2([0,T])$.

Let $X = (X_t)_{t\in[0,T]} \in \mathbf{\Lambda}^2([0,T])$. For any $t \in [0,T]$,

$$|b(t,X_t)| \leq K(1+|X_t|), \qquad |\sigma(t,X_t)|^2 \leq 2K^2(1+|X_t|^2),$$

so that the Itô process $\mathcal{G}X$ defined by

$$\forall t \in [0,T], \qquad (\mathcal{G}X)_t = \xi + \int_{s=0}^{t} b(s,X_s)\mathrm{d}s + \int_{s=0}^{t} \sigma(s,X_s)\mathrm{d}B_s,$$

is well-defined and in $\mathbf{\Lambda}^2([0,T])$. Furthermore, a process $X \in \mathbf{\Lambda}^2([0,T])$ satisfies the SDE (11.1)–(11.2) if and only if $\mathcal{G}X = X$.

For all $X, Y \in \mathbf{\Lambda}^2([0,T])$,

$$\begin{aligned}&\mathbb{E}\left[((\mathcal{G}X)_t - (\mathcal{G}Y)_t)^2\right]\\ &\leq 2\mathbb{E}\left[\left(\int_{s=0}^{t}(b(s,X_s)-b(s,Y_s))\mathrm{d}s\right)^2 + \left(\int_{s=0}^{t}(\sigma(s,X_s)-\sigma(s,Y_s))\mathrm{d}B_s\right)^2\right].\end{aligned}$$

On the one hand, the Cauchy–Schwarz inequality yields

$$\begin{aligned}\left(\int_{s=0}^{t}(b(s,X_s)-b(s,Y_s))\mathrm{d}s\right)^2 &\leq T\int_{s=0}^{t}(b(s,X_s)-b(s,Y_s))^2\mathrm{d}s\\ &\leq TK^2\int_{s=0}^{t}(X_s-Y_s)^2\mathrm{d}s.\end{aligned}$$

On the other hand, by Itô's isometry,

$$\begin{aligned}\mathbb{E}\left[\left(\int_{s=0}^{t}(\sigma(s,X_s)-\sigma(s,Y_s))\mathrm{d}B_s\right)^2\right] &= \mathbb{E}\left[\int_{s=0}^{t}(\sigma(s,X_s)-\sigma(s,Y_s))^2\mathrm{d}s\right]\\ &\leq K^2\mathbb{E}\left[\int_{s=0}^{t}(X_s-Y_s)^2\mathrm{d}s\right].\end{aligned}$$

We deduce that

$$\mathbb{E}\left[((\mathcal{G}X)_t - (\mathcal{G}Y)_t)^2\right] \leq C\int_{s=0}^{t}\mathbb{E}\left[(X_s-Y_s)^2\right]\mathrm{d}s,$$

with $C = 2K^2(T+1)$. Iterating this identity, we deduce that

$$\begin{aligned}\mathbb{E}\left[((\mathcal{G}^2X)_t - (\mathcal{G}^2Y)_t)^2\right] &\leq C\int_{s=0}^{t}\mathbb{E}\left[(\mathcal{G}X_s-\mathcal{G}Y_s)^2\right]\mathrm{d}s\\ &\leq C^2\int_{s_1=0}^{t}\int_{s_2=0}^{s_1}\mathbb{E}\left[(X_{s_2}-Y_{s_2})^2\right]\mathrm{d}s_2\mathrm{d}s_1,\end{aligned}$$

and then by induction, for all $k \geq 1$,

$$\mathbb{E}\left[\left((\mathcal{G}^kX)_t - (\mathcal{G}^kY)_t\right)^2\right] \leq C^k\int_{s_1=0}^{t}\int_{s_2=0}^{s_1}\cdots\int_{s_k=0}^{s_{k-1}}\mathbb{E}\left[(X_{s_k}-Y_{s_k})^2\right]\mathrm{d}s_k\cdots\mathrm{d}s_2\mathrm{d}s_1.$$

The Fubini Theorem allows to rewrite the integral in the right-hand side under the form

$$\begin{aligned}
&\int_{s_k=0}^{t}\int_{s_{k-1}=s_k}^{t}\cdots\int_{s_1=s_2}^{t}\mathbb{E}\left[(X_{s_k}-Y_{s_k})^2\right]\mathrm{d}s_1\cdots\mathrm{d}s_{k-1}\mathrm{d}s_k\\
&=\int_{s_k=0}^{t}\mathbb{E}\left[(X_{s_k}-Y_{s_k})^2\right]\underbrace{\left(\int_{s_{k-1}=s_k}^{t}\cdots\int_{s_1=s_2}^{t}\mathrm{d}s_1\cdots\mathrm{d}s_{k-1}\right)}_{=(t-s_k)^{k-1}/(k-1)!}\mathrm{d}s_k,
\end{aligned}$$

so that

$$\mathbb{E}\left[\left((\mathcal{G}^kX)_t-(\mathcal{G}^kY)_t\right)^2\right]\leq C^k\int_{s_k=0}^{t}\mathbb{E}\left[(X_{s_k}-Y_{s_k})^2\right]\frac{(t-s_k)^{k-1}}{(k-1)!}\mathrm{d}s_k,$$

and then

$$\begin{aligned}
\|\mathcal{G}^kX-\mathcal{G}^kY\|^2_{\mathbf{\Lambda}^2([0,T])}&=\int_{t=0}^{T}\mathbb{E}\left[\left((\mathcal{G}^kX)_t-(\mathcal{G}^kY)_t\right)^2\right]\mathrm{d}t\\
&\leq C^k\int_{s=0}^{T}\mathbb{E}\left[(X_s-Y_s)^2\right]\int_{t=s}^{T}\frac{(t-s)^{k-1}}{(k-1)!}\mathrm{d}t\mathrm{d}s\\
&\leq C^k\frac{T^k}{k!}\|X-Y\|^2_{\mathbf{\Lambda}^2([0,T])}.
\end{aligned}$$

We deduce that as soon as $k\geq 1$ is such that $(CT)^k/k!<1$, the mapping $\mathcal{G}^k$ is a contraction. Since Lemma 10.1.5 asserts that $\mathbf{\Lambda}^2([0,T])$ is a Banach space, Picard's Fixed Point Theorem ensures that $\mathcal{G}$ possesses a unique fixed point in this space. □

Remark 11.1.3. *The proof of Theorem 11.1.1 shows that for any $x\in\mathbb{R}^n$, there is a unique solution to (11.1) with (deterministic) initial condition $X_0=x$. Let us denote by $(X_t^x)_{t\in[0,T]}$ this process. The idea of the extension of the proof of Theorem 11.1.1 to the case where $|\xi|$ is not necessarily in $\mathbf{L}^2(\mathbb{P})$ then consists in defining $X_t(\omega)=X_t^{\xi(\omega)}(\omega)$: by construction, this process solves (11.1)–(11.2); and uniqueness can be checked conditionally on ξ. We leave the technical details apart.*

Remark 11.1.4. *If the coefficients b and σ are defined and satisfy the assumptions of Theorem 11.1.1 on $[0,+\infty)\times\mathbb{R}^n$, then for any $T>0$ the SDE (11.1)–(11.2) has a unique solution $(X_t)_{t\in[0,T]}$ on $[0,T]$, and it is easily checked that if $T'>T$ and $(X'_t)_{t\in[0,T']}$ is the solution on $[0,T']$, then $X_t=X'_t$ for any $t\in[0,T]$. Therefore, in this setting, the SDE (11.1)–(11.2) has a unique solution $(X_t)_{t\geq 0}$ defined on $I=[0,+\infty)$.*

In Exercise 10.2.11, we proved that the Geometric Brownian Motion $X_t=\mathrm{e}^{B_t}$ solves the SDE $\mathrm{d}X_t=\frac{1}{2}X_t\mathrm{d}t+X_t\mathrm{d}B_t$, with $X_0=1$. Theorem 11.1.1 shows that this solution is unique.

Exercise 11.1.5 (The Ornstein–Uhlenbeck process)**.** *Let $\lambda>0$ and $x_0\in\mathbb{R}$. We consider the SDE*

$$\mathrm{d}X_t=-\lambda X_t\mathrm{d}t+\mathrm{d}B_t,\qquad X_0=x_0,\tag{11.3}$$

the unique solution to which is called the Ornstein–Uhlenbeck process.

1. *Determine the set of solutions to the ordinary differential equation $\dot{x}_t=-\lambda x_t$.*
2. *Let $(X_t)_{t\geq 0}$ be the solution to the SDE (11.3) and let $C_t=X_t\mathrm{e}^{\lambda t}$. Compute C_t.*
3. *Deduce an explicit expression for X_t.*

4. *Show that $(X_t)_{t\geq 0}$ is a Gaussian process and compute its expectation and covariance function.*
5. *When $t \to +\infty$, describe the limit (in distribution) of X_t.*

Remark 11.1.6 (Strong and weak solutions)**.** *In the context of Theorem 11.1.1, the Brownian motion $(B_t)_{t\geq 0}$ is given a priori and the solution $(X_t)_{t\in[0,T]}$ to (11.1)–(11.2) given in Theorem 11.1.1 is constructed as a function of $(B_t)_{t\geq 0}$. In particular, it is adapted to the filtration $(\mathcal{F}_t^B)_{t\in[0,T]}$ generated by the Brownian motion $(B_t)_{t\in[0,T]}$: such a solution is called* strong.

There is also a notion of a weak *solution, for which both the process $(X_t)_{t\in\,[0,T]}$ and the Brownian motion $(B_t)_{t\in[0,T]}$ are constructed simultaneously. A strong solution is always a weak solution, but there are cases where the weak solution X_t cannot be expressed as a deterministic function of the Brownian motion $(B_s)_{s\in[0,t]}$, as we show in the following example.*

We consider the stochastic differential equation

$$\mathrm{d}X_t = \mathrm{sgn}(X_t)\mathrm{d}B_t, \qquad X_0 = 0,$$

where $\mathrm{sgn}(x) = \mathbb{1}_{\{x\geq 0\}} - \mathbb{1}_{\{x<0\}}$. We first construct a pair of processes $(X_t, B_t)_{t\geq 0}$ such that $(B_t)_{t\geq 0}$ is a Brownian motion and $(X_t)_{t\geq 0}$ solves the equation. We let $(X_t)_{t\geq 0}$ be a Brownian motion and define

$$B_t = \int_{s=0}^{t} \mathrm{sgn}(X_s)\mathrm{d}X_s.$$

By Lévy's characterisation (see Exercise 10.2.12), $(B_t)_{t\geq 0}$ is a Brownian motion. Besides, since $1/\mathrm{sgn}(x) = \mathrm{sgn}(x)$, we deduce from the relation $\mathrm{d}B_t = \mathrm{sgn}(X_t)\mathrm{d}X_t$ that

$$\mathrm{d}X_t = \frac{1}{\mathrm{sgn}(X_t)}\mathrm{d}B_t = \mathrm{sgn}(X_t)\mathrm{d}B_t.$$

We now check that X_t is not $\mathcal{F}_t^B$-measurable. To this aim, we claim (and leave the proof as an exercise[2]) that, almost surely,

$$\int_{s=0}^{t} \mathrm{sgn}(X_s)\mathrm{d}X_s = |X_t| - \lim_{\epsilon\to 0}\frac{1}{2\epsilon}\int_{s=0}^{t}\mathbb{1}_{\{|X_s|\leq\epsilon\}}\mathrm{d}t,$$

which shows that B_t is $\mathcal{F}_t^{|X|}$-measurable. Therefore if X_t was $\mathcal{F}_t^B$-measurable then we would have $\mathcal{F}_t^X \subset \mathcal{F}_t^{|X|}$ which is not true because the event $\{X_t > 0\}$ is in $\mathcal{F}_t^X$ but not in $\mathcal{F}_t^{|X|}$.

Remark 11.1.7 (Locally Lipschitz continuous coefficients)**.** *Similarly to the Cauchy–Lipschitz Theorem, there is also a version of the Itô Theorem when the coefficients b and σ are only* locally *Lipschitz continuous. Then the fixed point argument from the proof of Theorem 11.1.1 can be adapted to yield a solution $(X_t)_{t\in[0,\tau_*)}$ defined up to some random time τ_* which is an* explosion time, *in the sense that almost surely, if $\tau_* < +\infty$ then $\lim_{t\to\tau_*}|X_t| = +\infty$.*

11.2 The Feynman–Kac formula

11.2.1 Differential operator associated with an SDE

Definition 11.2.1 (Differential operator associated with (11.1))**.** *The* differential operator *associated with the SDE (11.1) is the differential operator L_t on $\mathbb{R}^n$ defined by*

$$L_t\Phi(x) = \sum_{i=1}^{n} b_i(t,x)\frac{\partial\Phi}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^{n} a_{i,j}(t,x)\frac{\partial^2\Phi}{\partial x_i\partial x_j}(x),$$

[2] The interested reader should start by applying the Itô formula to a C^2 approximation $\phi_\epsilon(x)$ of $|x|$.

with $a(t,x) := \sigma(t,x)\sigma^\top(t,x) \in \mathbb{R}^{n\times n}$.

The main reason for the introduction of this notion is the fact that, if $(X_t)_{t\geq 0}$ is a solution to the SDE (11.1), then for any C^2 function $\Phi : \mathbb{R}^n \to \mathbb{R}$, Itô's formula yields the identity

$$\mathrm{d}\Phi(X_t) = L_t\Phi(X_t)\mathrm{d}t + \sigma^\top(t,X_t)\nabla\Phi(X_t)\cdot \mathrm{d}B_t, \tag{11.4}$$

which we shall use several times in the sequel.

Exercise 11.2.2. *What is the differential operator associated with the Brownian motion?*

11.2.2 Probabilistic representation of the solution to backward Cauchy problems

We establish a first connection between SDEs and PDEs through the *Feynman–Kac* formula. Assume that there exists $T > 0$ such that, for all $x \in \mathbb{R}^n$ and $t \in [0,T)$, there exists an Itô process $(X_s^{t,x})_{s\in[t,T]}$ such that, for all $s \in [0,T]$,

$$X_s^{t,x} = x + \int_{r=t}^{s} b(r, X_r^{t,x})\mathrm{d}r + \int_{r=t}^{s} \sigma(r, X_r^{t,x})\mathrm{d}B_r,$$

that is to say, a solution to (11.1) on $[t,T]$ which takes the value x at time t. This is in particular the case if the coefficients b and σ satisfy the assumptions of Theorem 11.1.1.

Theorem 11.2.3 (Feynman–Kac formula for children)**.** *Let* $T > 0$ *and* $f : \mathbb{R}^n \to \mathbb{R}$ *be a continuous function. Assume that there exists a* $C^{1,2}$ *function* $u : [0,T]\times\mathbb{R}^n \to \mathbb{R}$ *such that:*

(i) for any $t \in [0,T)$ *and* $x \in \mathbb{R}^n$, $(\sigma^\top(s, X_s^{t,x})\nabla_x u(s, X_s^{t,x}))_{s\in[t,T]} \in \mathbf{\Lambda}^2([t,T])$*;*

(ii) u solves the parabolic problem

$$\begin{cases} -\dfrac{\partial u}{\partial t}(t,x) = L_t u(t,x), & t \in [0,T], \quad x \in \mathbb{R}^n, \\ u(T,x) = f(x). \end{cases} \tag{11.5}$$

Then, for all $(t,x) \in [0,T]\times\mathbb{R}^n$,

$$u(t,x) = \mathbb{E}\left[f(X_T^{t,x})\right].$$

Proof. For simplicity we write $X_s^{t,x} = X_s = (X_s^1, \ldots, X_s^n)$. Let us fix $t \in [0,T]$ and apply Itô's formula to $u(s,X_s)$ for $s \in [t,T]$. By (11.4), we get

$$\begin{aligned} \mathrm{d}u(s,X_s) &= \left(\frac{\partial u}{\partial t}(s,X_s) + L_s u(s,X_s)\right)\mathrm{d}s + \sigma^\top(s,X_s)\nabla_x u(s,X_s)\cdot\mathrm{d}B_s \\ &= \sigma^\top(s,X_s)\nabla_x u(s,X_s)\cdot\mathrm{d}B_s, \end{aligned}$$

thanks to (ii). As a consequence,

$$u(T,X_T) = u(t,X_t) + \int_{s=t}^{T} \sigma^\top(s,X_s)\nabla_x u(s,X_s)\cdot\mathrm{d}B_s,$$

which rewrites

$$f(X_T) = u(t,x) + \int_{s=t}^{T} \sigma^\top(s,X_s)\nabla_x u(s,X_s)\cdot\mathrm{d}B_s.$$

The assumption (i) now ensures that

$$\mathbb{E}\left[\int_{s=t}^{T} \sigma^{\top}(s, X_s)\nabla_x u(s, X_s) \cdot \mathrm{d}B_s\right] = 0,$$

therefore

$$\mathbb{E}\left[f(X_T)\right] = u(t, x). \qquad \square$$

The Feynman–Kac formula shows that if one is interested in solving the PDE (11.5) in one point (t, x), a possible approach may be to simulate the trajectory of $(X_s^{t,x})_{s\in[t,T]}$ and then to compute the expectation $\mathbb{E}[f(X_T^{t,x})]$ by the Monte Carlo method. This naturally raises the question of the numerical simulation of the solution to SDEs, which is addressed in the next section.

Example 11.2.4 (The Black–Scholes model in mathematical finance)**.** *In mathematical finance, the* Black–Scholes model[3] *assumes that the price of some asset (for instance, an action) is the solution* $(S_t)_{t\geq 0}$ *of the SDE*

$$\mathrm{d}S_t = \sigma S_t \mathrm{d}B_t,$$

whose solution writes $S_t = S_0 \exp(\sigma B_t - \sigma^2 t/2)$*. An* option *with* payoff *function* f *and* maturity T *is a contract between the bank and the client, where at time* T *the bank has to give the client the quantity* $f(S_T)$*. The price that the client has to pay to the bank at time* $t \leq T$ *in order to buy the option is given by* $u(t, s) = \mathbb{E}[f(S_T^{t,s})]$*, where* s *is the value of* S_t*. This quantity can be computed either by the Monte Carlo method, or by solving the parabolic problem*

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, s) + \dfrac{\sigma^2 s^2}{2}\dfrac{\partial^2 u}{\partial s^2}(t, s) = 0, & t \in [0, T), \quad s \geq 0,\\ u(T, s) = f(s). \end{cases}$$

🏠 **Exercise 11.2.5** (Feynman–Kac formula for grown-ups)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *and* $k, g : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ *be continuous functions, with* k *bounded from below. Assume that there exists a* $C^{1,2}$ *function* $u : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ *such that:*

(i) for any $t \in [0, T)$ *and* $x \in \mathbb{R}^n$*,* $(\sigma^{\top}(s, X_s^{t,x})\nabla_x u(s, X_s^{t,x}))_{s\in[t,T]} \in \mathbf{\Lambda}^2([t, T])$*;*

(ii) u solves the parabolic problem

$$\begin{cases} -\dfrac{\partial u}{\partial t}(t, x) = L_t u(t, x) - k(t, x)u(t, x) + g(t, x), & t \in [0, T], \quad x \in \mathbb{R}^n,\\ u(T, x) = f(x). \end{cases} \tag{11.6}$$

Show that for all $(t, x) \in [0, T] \times \mathbb{R}^n$*,*

$$u(t, x) = \mathbb{E}\left[f(X_T^{t,x})\mathrm{e}^{-\int_{u=t}^{T} k(u, X_u^{t,x})\mathrm{d}u} + \int_{s=t}^{T} g(s, X_s^{t,x})\mathrm{e}^{-\int_{u=t}^{s} k(u, X_u^{t,x})\mathrm{d}u}\mathrm{d}s\right].$$

Hint: start by applying Itô's formula to $u(s, X_s^{t,x})\mathrm{e}^{-\int_{u=t}^{s} k(u, X_u^{t,x})\mathrm{d}u}$.

The Feynman–Kac formula (in the form of Exercise 11.2.5) provides a probabilistic representation of a solution to (11.6) which satisfies the integrability condition that

$$\forall (t, x) \in [0, T] \times \mathbb{R}^n, \qquad \mathbb{E}\left[\int_{s=t}^{T} |\sigma^{\top}(s, X_s^{t,x})\nabla_x u(s, X_s^{t,x})|^2 \mathrm{d}s\right] < +\infty. \tag{11.7}$$

[3] F. Black, M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 1973. We also refer to the book [8] for an introduction to mathematical finance.

It is therefore a *uniqueness* result for the PDE (11.6) in the class of solutions which satisfy (11.7). Thus, its application usually requires to find an *existence* result for a smooth solution to this Cauchy problem, in a PDE textbook. For example, the following one can be found in [4, Section 6.5].

Proposition 11.2.6 (Existence of a solution to (11.6))**.** *Fix $T > 0$ and assume that:*

(i) the matrix a is uniformly elliptic: there exists $c > 0$ such that for any $t \in [0,T]$ and $x, \xi \in \mathbb{R}^n$, $\xi \cdot a(t,x)\xi \geq c|\xi|^2$;
(ii) the functions a_{ij} and b_i are bounded on $[0,T] \times \mathbb{R}^n$ and Lipschitz continuous in (t,x) on compact subsets of $[0,T] \times \mathbb{R}^n$;
(iii) the functions a_{ij} are Hölder continuous in x, uniformly in $(t,x) \in [0,T] \times \mathbb{R}^n$;
(iv) k is bounded on $[0,T] \times \mathbb{R}^n$ and Hölder continuous in (t,x) on compact subsets of $[0,T] \times \mathbb{R}^n$;
(v) g is continuous on $[0,T] \times \mathbb{R}^n$ and Hölder continuous in x, uniformly in $(t,x) \in [0,T] \times \mathbb{R}^n$;
(vi) f is continuous on $\mathbb{R}^n$;
(vii) there exists $K, \ell > 0$ such that $|f(x)| + |g(t,x)| \leq K(1+|x|^\ell)$ for any $(t,x) \in [0,T] \times \mathbb{R}^n$.

Then there exists a $C^{1,2}$ function $u : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ which satisfies (11.5). Moreover, there exists $K' \geq 0$ such that this function satisfies

$$\forall (t,x) \in [0,T] \times \mathbb{R}^n, \qquad |u(t,x)| + |\nabla_x u(t,x)| \leq K'(1+|x|^\ell).$$

Once a solution u to (11.5) is given, to show that it admits the probabilistic representation given by Exercise 11.2.5, one needs to check that it satisfies the integrability condition (11.7). This usually requires to combine estimates on the growth of σ and $\nabla_x u$ with moment estimates for $X^{t,x}_s$. For example, assume that in addition to the conditions of Proposition 11.2.6, the diffusion coefficient σ of the SDE (11.1) is bounded on $[0,T] \times \mathbb{R}^n$. Then by the boundedness of σ and the growth condition on $\nabla_x u$ given by Proposition 11.2.6, the integrability condition (11.7) holds if one is able to show that

$$\forall (t,x) \in [0,T] \times \mathbb{R}^n, \qquad \sup_{s \in [t,T]} \mathbb{E}[|X^{t,x}_s|^{2\ell}] < +\infty.$$

To proceed, let us write

$$X^{t,x}_s = x + \int_{r=t}^s b(r, X^{t,x}_r)\mathrm{d}r + \int_{r=t}^s \sigma(r, X^{t,x}_r)\mathrm{d}B_r,$$

and notice that since b is bounded it suffices to show that

$$\sup_{s \in [t,T]} \mathbb{E}\left[\left|\int_{r=t}^s \sigma(r, X^{t,x}_r)\mathrm{d}B_r\right|^{2\ell}\right] < +\infty.$$

By Jensen's inequality there is no loss of generality in assuming that $\ell \geq 1/2$ here, so that the claimed estimate follows from the Burkholder–Davis–Gundy inequality (Lemma 10.1.8). We thus conclude that u admits the probabilistic representation of Exercise 11.2.5. When the coefficients b and σ are not bounded, moment estimates on $X^{t,x}_s$ are usually obtained with arguments similar to the proof of Lemma 11.1.2, which may involve a localisation procedure.

11.2.3 Problems with boundaries

In this subsection, we assume that the coefficients b and σ do not depend on t and we denote by L the associated differential operator. We let D be an open and regular subset of $\mathbb{R}^n$ and, for any solution $(X^x_t)_{t \geq 0}$ of (11.1) with initial condition $x \in D$, we define the stopping time

$$\tau^x := \inf\{t \geq 0 : X^x_t \notin D\}.$$

Proposition 11.2.7 (Probabilistic interpretation of Dirichlet problem)**.** *Let $f : \partial D \to \mathbb{R}$ and $k, g : \overline{D} \to \mathbb{R}^n$ be continuous functions, with $k \geq 0$. Assume that there exists a C^2 function $v : [0,T] \times \overline{D} \to \mathbb{R}$ such that:*

(i) v is bounded;
(ii) for any $t > 0$ and $x \in D$, $(\sigma^\top(X_s^x)\nabla v(X_s^x))_{s\in[0,t]} \in \mathbf{\Lambda}^2([0,t])$;
(iii) v solves the elliptic problem

$$\begin{cases} Lv(x) - k(x)v(x) = -g(x), & x \in D, \\ v(x) = f(x), & x \in \partial D. \end{cases} \tag{11.8}$$

Assume moreover that for any $x \in D$:

(iv) the associated stopping time τ^x is finite, almost surely;
(v) τ^x and g satisfy

$$\int_{t=0}^{+\infty} \mathbb{E}\left[\mathbb{1}_{\{t<\tau^x\}}|g(X_t)|\right] \mathrm{d}t < +\infty.$$

Then for all $x \in D$,

$$v(x) = \mathbb{E}\left[f(X_{\tau^x}^x)\mathrm{e}^{-\int_{u=0}^{\tau} k(u,X_u^x)\mathrm{d}u} + \int_{s=0}^{\tau^x} g(X_s^x)\mathrm{e}^{-\int_{u=0}^{s} k(X_u^x)\mathrm{d}u}\mathrm{d}s\right].$$

Proof. Let us fix $x \in D$ and write $X_t = X_t^x$, $\tau = \tau^x$. Itô's formula applied to $v(X_t)\mathrm{e}^{-\int_{u=0}^{t} k(X_u)\mathrm{d}u}$ yields

$$\begin{aligned} v(X_{t\wedge\tau})\mathrm{e}^{-\int_{u=0}^{t\wedge\tau} k(X_u)\mathrm{d}u} = v(x) &+ \int_{s=0}^{t\wedge\tau} \mathrm{e}^{-\int_{u=0}^{s} k(X_u)\mathrm{d}u}\left(Lv(X_s) - k(X_s)v(X_s)\right)\mathrm{d}s \\ &+ \int_{s=0}^{t\wedge\tau} \mathrm{e}^{-\int_{u=0}^{s} k(X_u)\mathrm{d}u}\sigma^\top(X_s)\nabla v(X_s)\cdot \mathrm{d}B_s, \end{aligned}$$

so that

$$v(x) = \mathbb{E}\left[v(X_{t\wedge\tau})\mathrm{e}^{-\int_{u=0}^{t\wedge\tau} k(X_u)\mathrm{d}u} + \int_{s=0}^{t\wedge\tau} g(X_s)\mathrm{e}^{-\int_{u=0}^{s} k(X_u)\mathrm{d}u}\mathrm{d}s\right].$$

First, since $\tau < +\infty$ almost surely, v is bounded and $k \geq 0$, by the Dominated Convergence Theorem one has

$$\lim_{t\to+\infty} \mathbb{E}\left[v(X_{t\wedge\tau})\mathrm{e}^{-\int_{u=0}^{t\wedge\tau} k(X_u)\mathrm{d}u}\right] = \mathbb{E}\left[v(X_\tau)\mathrm{e}^{-\int_{u=0}^{\tau} k(X_u)\mathrm{d}u}\right] = \mathbb{E}\left[f(X_\tau)\mathrm{e}^{-\int_{u=0}^{\tau} k(X_u)\mathrm{d}u}\right].$$

Second, the integrability condition on g and τ allows to use the Dominated Convergence Theorem again to get

$$\lim_{t\to+\infty} \mathbb{E}\left[\int_{s=0}^{t\wedge\tau} g(X_s)\mathrm{e}^{-\int_{u=0}^{s} k(X_u)\mathrm{d}u}\mathrm{d}s\right] = \mathbb{E}\left[\int_{s=0}^{\tau} g(X_s)\mathrm{e}^{-\int_{u=0}^{s} k(X_u)\mathrm{d}u}\mathrm{d}s\right],$$

which completes the proof. □

Example 11.2.8 (The committor function)**.** *Let $(X_t^x)_{t\geq 0}$ be the solution to the SDE (11.1) with coefficients b and σ which do not depend on t, and with deterministic initial condition $x \in \mathbb{R}^n$. Given two disjoint closed subsets $A, B \subset \mathbb{R}^n$, set*

$$\tau_A^x := \inf\{t \geq 0 : X_t^x \in A\}, \qquad \tau_B^x := \inf\{t \geq 0 : X_t^x \in B\},$$

and define

$$v(x) = \mathbb{P}(\tau_A^x < \tau_B^x).$$

In molecular dynamics, this function is called the committor *function*[4]*: in this context, X_t^x must be thought of as describing the* microscopic *state of a molecular system, and A and B describe particular* macroscopic *configurations. For example, in a protein-ligand system, X_t^x encodes the complete geometry of the protein-ligand, while A and B contain the states which correspond to the system being bound or unbound, respectively. Computing the committor function then allows to determine whether, given an initial state x, it is more likely that the system evolves toward one or the other configuration. Proposition 11.2.7 shows that under regularity assumptions and if $\tau_A^x \wedge \tau_B^x < \infty$, almost surely, then u solves the PDE*

$$\begin{cases} Lv(x) = 0, & x \in \mathbb{R}^n \setminus (A \cup B), \\ v(x) = 1, & x \in A, \\ v(x) = 0, & x \in B. \end{cases}$$

Remark 11.2.9. *The assumption that $k \geq 0$ is crucial in the statement of Proposition 11.2.7. Indeed, consider the case where $n = 1$, $D = (0, 1)$ and $\mathrm{d}X_t = \mathrm{d}B_t$ so that $L = \frac{1}{2}\frac{\partial^2}{\partial x^2}$. It can be directly checked that for any $m \geq 1$, $v_m(x) = \sin(\pi m x)$ satisfies (11.8) with $k = -\frac{1}{2}(\pi m)^2 < 0$ and $f = g = 0$, so that applying the result of Proposition 11.2.7 would yield $v_m = 0$ on D.*

The same remark as for the Feynman–Kac formula applies to Proposition 11.2.7: it is a uniqueness result for v, which requires to find an existence result first, for which it is then necessary to check that the integrability conditions are satisfied. We refer to [4, Section 6.5] again for examples of such results. To apply Proposition 11.2.7 it is moreover necessary to get quantitative estimates on τ^x in order to check the conditions (iv) and (v). One may for instance use the following statement.

Lemma 11.2.10 (Exponential moment for τ^x for bounded domains)**.** *Assume that b and σ are continuous on $\mathbb{R}^n$, that D is bounded and that $a = \sigma\sigma^\top$ is uniformly elliptic: there exists $c > 0$ such that, for any $x, \xi \in \mathbb{R}^n$, $\xi \cdot a(x)\xi \geq c|\xi|^2$. Then there exists $\epsilon > 0$ and $M < +\infty$ such that for any $x \in D$, $\mathbb{E}[\mathrm{e}^{\epsilon\tau^x}] \leq M$.*

11.3 Discretisation of SDEs

The standard explicit Euler scheme for ordinary differential equations possesses a natural generalisation to SDEs, which is sometimes referred to as the (explicit) *Euler–Maruyama* scheme. For an initial condition $X_0 = \xi$, and given a final time $T > 0$, a number of steps $N \geq 1$ and a step size $h = T/N$, the discretisation of (11.1)–(11.2) yields the sequence of random variables $(\theta_k^h)_{0 \leq k \leq N}$ defined by

$$\begin{cases} \theta_0^h = \xi, \\ \theta_{k+1}^h = \theta_k^h + b(kh, \theta_k^h)h + \sigma(kh, \theta_k^h)(B_{(k+1)h} - B_{kh}), \qquad 0 \leq k \leq N-1. \end{cases}$$

This scheme is easy to simulate since the random variables $B_{(k+1)h} - B_{kh}$ are independent and distributed under the law $\mathcal{N}(0, h)$. Its accuracy may for example be measured by the *strong error*

$$\mathcal{E}_T^h := \max_{0 \leq k \leq N} \mathbb{E}\left[|X_{kh} - \theta_k^h|^2\right]^{1/2}.$$

[4] In potential theory, it is the *equilibrium potential*.

For the sake of simplicity we shall assume in the next statement that the coefficients b and σ do not depend on time.

Theorem 11.3.1 (Strong error)**.** *Assume that there exists* $M_b, M_\sigma, L_b, L_\sigma \in [0,+\infty)$ *such that*

$$\forall x \in \mathbb{R}^n, \qquad |b(x)| \leq M_b, \quad |\sigma(x)| \leq M_\sigma, \tag{11.9}$$

and

$$\forall x, y \in \mathbb{R}^n, \qquad |b(x)-b(y)| \leq L_b|x-y|, \quad |\sigma(x)-\sigma(y)| \leq L_\sigma|x-y|. \tag{11.10}$$

For all $T > 0$*, there exists* $C_T \in [0,+\infty)$ *such that*

$$\forall h > 0, \qquad \mathcal{E}^h_T \leq C_T\sqrt{h}.$$

The Euler–Maruyama scheme is therefore said to be of *strong order* $1/2$.

Proof. For the proof it is convenient to work with the adapted interpolation $(\overline{X}^h_t)_{t\in[0,T]}$ of the Euler–Maruyama scheme defined by

$$\overline{X}^h_t = \theta^h_k + b(\theta^h_k)(t-kh) + \sigma(\theta^h_k)(B_t - B_{kh}), \qquad t \in [kh,(k+1)h],$$

which is an Itô process which satisfies

$$\mathrm{d}\overline{X}^h_t = b(\overline{X}^h_{\tau_h(t)})\mathrm{d}t + \sigma(\overline{X}^h_{\tau_h(t)})\mathrm{d}B_t,$$

with $\tau_h(t) = kh$ if $t \in [kh,(k+1)h)$. Therefore we have, for all $t \in [0,T]$,

$$X_t - \overline{X}^h_t = \int_{s=0}^t \left(b(X_s) - b(\overline{X}^h_{\tau_h(s)})\right)\mathrm{d}s + \int_{s=0}^t \left(\sigma(X_s) - \sigma(\overline{X}^h_{\tau_h(s)})\right)\mathrm{d}B_s,$$

so that

$$\left|X_t - \overline{X}^h_t\right|^2 \leq 2\left(\left|\int_{s=0}^t \left(b(X_s) - b(\overline{X}^h_{\tau_h(s)})\right)\mathrm{d}s\right|^2 + \left|\int_{s=0}^t \left(\sigma(X_s) - \sigma(\overline{X}^h_{\tau_h(s)})\right)\mathrm{d}B_s\right|^2\right).$$

Using the Cauchy–Schwarz inequality and (11.9), we first write

$$\begin{aligned}\left|\int_{s=0}^t \left(b(X_s) - b(\overline{X}^h_{\tau_h(s)})\right)\mathrm{d}s\right|^2 &\leq T\int_{s=0}^t \left|b(X_s) - b(\overline{X}^h_{\tau_h(s)})\right|^2\mathrm{d}s\\ &\leq TL_b^2\int_{s=0}^t \left|X_s - \overline{X}^h_{\tau_h(s)}\right|^2\mathrm{d}s;\end{aligned}$$

likewise, since σ is assumed to be bounded in (11.9), Itô's isometry yields

$$\begin{aligned}\mathbb{E}\left[\left|\int_{s=0}^t \left(\sigma(X_s) - \sigma(\overline{X}^h_{\tau_h(s)})\right)\mathrm{d}B_s\right|^2\right] &= \mathbb{E}\left[\int_{s=0}^t \left|\sigma(X_s) - \sigma(\overline{X}^h_{\tau_h(s)})\right|^2\mathrm{d}s\right]\\ &\leq L_\sigma^2\mathbb{E}\left[\int_{s=0}^t \left|X_s - \overline{X}^h_{\tau_h(s)}\right|^2\mathrm{d}s\right],\end{aligned}$$

so that

$$\mathbb{E}\left[\left|X_t - \overline{X}^h_t\right|^2\right] \leq 2(TL_b^2 + L_\sigma^2)\mathbb{E}\left[\int_{s=0}^t \left|X_s - \overline{X}^h_{\tau_h(s)}\right|^2\mathrm{d}s\right].$$

We now write

$$\left|X_s - \overline{X}^h_{\tau_h(s)}\right|^2 \leq 2\left(\left|X_s - \overline{X}^h_s\right|^2 + \left|\overline{X}^h_s - \overline{X}^h_{\tau_h(s)}\right|^2\right),$$

and

$$\overline{X}^h_s - \overline{X}^h_{\tau_h(s)} = b(\overline{X}^h_{\tau_h(s)})(s - \tau_h(s)) + \sigma(\overline{X}^h_{\tau_h(s)})(B_s - B_{\tau_h(s)}).$$

As a consequence, using (11.9), we get

$$\begin{aligned}\mathbb{E}\left[\left|\overline{X}^h_s - \overline{X}^h_{\tau_h(s)}\right|^2\right] &\leq 2\left(\mathbb{E}\left[|b(\overline{X}^h_{\tau_h(s)})(s-\tau_h(s))|^2\right] + \mathbb{E}\left[|\sigma(\overline{X}^h_{\tau_h(s)})(B_s - B_{\tau_h(s)})|^2\right]\right)\\ &\leq 2\left(M_b^2 h^2 + M_\sigma^2 h\right),\end{aligned}$$

and finally

$$\mathbb{E}\left[\left|X_t - \overline{X}^h_t\right|^2\right] \leq 2(TL_b^2 + L_\sigma^2)\int_{s=0}^t \left\{2\mathbb{E}\left[\left|X_s - \overline{X}^h_s\right|^2\right] + 4\left(M_b^2h^2 + M_\sigma^2 h\right)\right\}\mathrm{d}s.$$

We thus deduce from Gronwall's Lemma that, for all $t \in [0,T]$,

$$\mathbb{E}\left[\left|X_t - \overline{X}^h_t\right|^2\right] \leq 8T(TL_b^2 + L_\sigma^2)(M_b^2h^2 + M_\sigma^2h)\mathrm{e}^{4(TL_b^2+L_\sigma^2)T},$$

which completes the proof. □

If one is only interested in the numerical approximation of quantities of the form $\mathbb{E}[f(X_T)]$, as is suggested by the Feynman–Kac formula for the Monte Carlo approximation of the solution to parabolic PDEs, then computing the discretisation error between the *realisations* of the trajectories $(X_t)_{t\in[0,T]}$ and $(\theta^h_k)_{0\leq k\leq N}$ is too demanding, since what really matters here is the discretisation error between their *laws*. Such an error is called *weak*, and is typically measured by quantities of the form

$$\mathbf{e}^h_T := \max_{0\leq k\leq N}\left|\mathbb{E}[f(X_{kh})] - \mathbb{E}[f(\theta^h_k)]\right|,$$

for a certain choice of function f. If f is Lipschitz continuous, then it is immediately observed that the weak order is at least larger than the strong order; in general it is strictly larger.

🏠 **Exercise 11.3.2.** *Recall the Ornstein–Uhlenbeck process $(X_t)_{t\geq 0}$ from Exercise 11.1.5. Let $(\theta^h_k)_{0\leq k\leq N}$ denote the associated Euler–Maruyama scheme on $[0,T]$.*

1. *Recall the law of the random variable X_t. We denote $m_t = \mathbb{E}[X_t]$ and $v_t = \mathrm{Var}(X_t)$.*
2. *Show that for any $k \in \{0,\ldots,N\}$, the random variable θ^h_k is Gaussian, and compute its expectation m^h_k and its variance v^h_k.*
3. *Show that for any Lipschitz continuous function $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz norm 1, for all $k \in \{0,\ldots,N\}$,*

$$\left|\mathbb{E}[f(\theta^h_k)] - \mathbb{E}[f(X_{kh})]\right| \leq \left|m^h_k - m_{kh}\right| + \left|\sqrt{v^h_k} - \sqrt{v_{kh}}\right|.$$

4. *Conclude that the weak error is of order 1.*

Remark 11.3.3. *Assume that, in the definition of the Euler–Maruyama scheme $(\theta^h_k)_{0\leq k\leq N}$, the increments $B_{(k+1)h} - B_{kh}$ are replaced with arbitrary independent random variables ζ_k with expectation 0 and variance h, but not necessarily Gaussian (nor even identically distributed). Then, in the context of Exercise 11.3.2, the random variables θ^h_k are no longer necessarily Gaussian, but the formulas obtained for m^h_k and v^h_k remain true. As a consequence, the weak error remains of order 1 under very mild assumptions on the construction of the Euler–Maruyama scheme.*

Bibliography

[1] C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer. *Sur les inégalités de Sobolev logarithmiques*. Société Mathématique de France, 2000. In French.

[2] D. Chafai and F. Malrieu. *Recueil de modèles aléatoires*, volume 78. Springer, 2016. In French, https://hal.archives-ouvertes.fr/hal-01897577v1.

[3] J.-F. Delmas and B. Jourdain. *Modèles aléatoires. Applications aux sciences de l'ingénieur et du vivant*. Springer, 2006. In French.

[4] Avner Friedman. *Stochastic differential equations and applications. Vol. 1*, volume Vol. 28 of *Probability and Mathematical Statistics*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975.

[5] M. Hairer. Ergodic properties of markov processes. http://www.hairer.org/notes/Markov.pdf, 2006. Lecture given at the University of Warwick.

[6] B. Jourdain. *Probabilités et statistiques*. Ellipses, 2016. In French. http://cermics.enpc.fr/~jourdain/probastat/poly.pdf.

[7] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*. Springer, 1991.

[8] D. Lamberton and B. Lapeyre. *Introduction au calcul stochastique appliqué à la finance*. Ellipses, 1997. In French.

[9] J.-F. Le Gall. Intégration, probabilités et processus aléatoires. https://www.math.u-psud.fr/~jflegall/IPPA2.pdf, 2006. In French. Lecture given at Ecole Normale Supérieure.

[10] D.A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017. Second edition, with contributions by E. L. Wilmer. https://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf.