Master de Mathématiques et Applications

Spécialité Mathématiques de la Modélisation

Méthodes Numériques Probabilistes

Tony Lelièvre et Julien Reygner

Version du 22 novembre 2022
Part I

Random number simulation and the Monte Carlo method
Chapter 1

Random variables and their numerical simulation

Contents

1.1 General definitions and results ......................................................... 3
1.2 Random number simulation ............................................................ 9
1.3 Random vector simulation ............................................................... 17

This chapter aims at presenting, or reviewing, most basic material concerning the notion of random variable. A particular emphasis is put on their numerical simulation.

Throughout the chapter, we assume that the reader is familiar with basic notions of measure theory, such as the definition of a measure space, the notion of measurable functions, and the construction and properties of the Lebesgue integral. We refer to [4, 5] for details.

1.1 General definitions and results

Throughout the notes, 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 $\mathbb{N}$ of integers.

1.1.1 Probability space

Definition 1.1.1 (Probability space). A probability space is a triple $(\Omega, A, \mathbb{P})$ such that:

- $\Omega$ is a set;
- $A$ is a $\sigma$-field$^1$ on $\Omega$;
- $\mathbb{P}$ is a probability measure$^2$ on $(\Omega, A)$.

Measurable sets $A \in A$ 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 $A$,

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

$^1$That is to say, a collection $A$ of subsets of $\Omega$ such that: $\Omega \in A$; for any $A \in A$, $A^c \in A$; and for any countable family $(A_n)_{n \geq 1}$ of elements of $A$, $\bigcup_{n \geq 1} A_n \in A$.

$^2$That is to say, a mapping $\mathbb{P} : A \to [0, 1]$ such that $\mathbb{P}(\Omega) = 1$ and, for any countable family $(A_n)_{n \geq 1}$ of pairwise distinct elements of $A$, $\mathbb{P}(\bigcup_{n \geq 1} A_n) = \sum_{n \geq 1} \mathbb{P}(A_n)$. The latter property is called $\sigma$-additivity.
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 \( \sigma \)-additivity property of the measure \( \mathbb{P} \), its proof is omitted.

**Proposition 1.1.2 (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 \( B_{n+1} \subseteq B_n \) for any \( n \), then
  \[ \lim_{n \to +\infty} \mathbb{P}(B_n) = \mathbb{P}(\cap_{n \geq 1} B_n). \]
- If \((B_n)_{n \geq 1}\) is nondecreasing, that is to say \( B_n \subseteq B_{n+1} \) for any \( n \), then
  \[ \lim_{n \to +\infty} \mathbb{P}(B_n) = \mathbb{P}(\cup_{n \geq 1} B_n). \]

Proposition 1.1.2 has the following practical corollary.

**Corollary 1.1.3 (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 \( n \),
\[
\mathbb{P}(B_{n+1}^c) = \mathbb{P}((B_n \cap A_{n+1})^c)
= \mathbb{P}(B_n^c \cup A_{n+1}^c)
\leq \mathbb{P}(B_n^c) + \mathbb{P}(A_{n+1}^c)
= \mathbb{P}(B_n^c),
\]
which by an immediate induction shows that \( \mathbb{P}(B_n^c) \leq \mathbb{P}(B_1^c) = \mathbb{P}(A_1^c) = 0 \), and therefore \( \mathbb{P}(B_n) = 1 \) for any \( n \). By Proposition 1.1.2, we conclude that
\[
\mathbb{P}(\cap_{n \geq 1} A_n) = \mathbb{P}(\cap_{n \geq 1} B_n) = \lim_{n \to +\infty} \mathbb{P}(B_n) = 1,
\]
which shows that the event \( \cap_{n \geq 1} A_n \) is almost sure. \qed

### 1.1.2 Conditional probability

**Definition 1.1.4 (Conditional probability).** Let \( B \in \mathcal{A} \) be such that \( \mathbb{P}(B) > 0 \). The conditional probability given \( B \) is the probability measure \( \mathbb{P}(\cdot \mid B) \) defined on \((\Omega, \mathcal{A})\) by
\[
\mathbb{P}(\cdot \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.
\]

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

**Lemma 1.1.5 (Total probability formula).** Let us be given a partition of \( \Omega \) into events \((B_n)_{n \geq 1}\). For any event \( A \),
\[
\mathbb{P}(A) = \sum_{n \geq 1} \mathbb{P}(A \cap B_n) = \sum_{n \geq 1} \mathbb{P}(A \mid B_n) \mathbb{P}(B_n),
\]
with the obvious convention that \( \mathbb{P}(A \mid B_n) \mathbb{P}(B_n) = 0 \) if \( \mathbb{P}(B_n) = 0 \).
1.1 General definitions and results

1.1.3 Random variables

In this subsection, we consider measurable functions defined on \((\Omega, A)\) and taking their values in some measurable space \((E, \mathcal{E})\).

**Definition 1.1.6 (Random variable).** Let \((E, \mathcal{E})\) be a measurable space. A random variable in \(E\) is a measurable function \(X : \Omega \rightarrow E\), that is to say a function such that

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

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

\[
\forall C \in \mathcal{E}, \quad P_X(C) = \mathbb{P}(X^{-1}(C)).
\]

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 \(P_X = P\). Notice that the triple \((E, \mathcal{E}, P_X)\) is a probability space itself.

1.1.4 Density

**Definition 1.1.7 (Absolute continuity).** Let \(\mu\) be a nonnegative \(\sigma\)-finite measure\(^3\) on the measurable space \((E, \mathcal{E})\). A probability measure \(P\) on \((E, \mathcal{E})\) is called absolutely continuous with respect to \(\mu\) if

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

In this case, we write \(P \ll \mu\).

**Theorem 1.1.8 (Radon–Nikodym Theorem).** If \(P \ll \mu\), then there exists a measurable function \(p : E \rightarrow [0, +\infty)\) such that

\[
\forall C \in \mathcal{E}, \quad P(C) = \int_{x \in E} 1_{\{x \in C\}} p(x) d\mu(x).
\]

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

\[
p(x) = \frac{dP}{d\mu}(x),
\]

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

\[
\int_{x \in E} p(x) d\mu(x) = 1.
\]

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

- \(E = \mathbb{R}^d\), \(\mathcal{E}\) is the Borel \(\sigma\)-field \(\mathcal{B}(\mathbb{R}^d)\) and \(\mu\) is the Lebesgue measure;
- \(E\) is countable, \(\mathcal{E}\) is the power set of \(E\) (called the discrete \(\sigma\)-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.

\(^3\)That is to say a \(\sigma\)-additive mapping \(\mu : \mathcal{E} \rightarrow [0, +\infty]\) such that there exists a countable family \((C_n)_{n \geq 1}\) of elements of \(E\) for which \(\mu(C_n) < +\infty\) and \(\bigcup_{n \geq 1} C_n = E\).
1.1.5 Expectation and the Transfer Theorem

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

Definition 1.1.9 (Expectation). Let $X \in L^1(\mathbb{P})$. The expectation of $X$ is the Lebesgue integral

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

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 $L^p(P_X) = L^p(E, \mathcal{E}, P_X)$ are defined similarly to $L^p(\mathbb{P}) = L^p(\Omega, A, \mathbb{P})$.

Theorem 1.1.10 (Transfer Theorem). Let $X$ be a random variable in $E$ and $f : E \to \mathbb{R}$ be a measurable function. Then $f(X) \in L^1(\mathbb{P})$ if and only if $f \in L^1(P_X)$, and

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

In addition, if $X$ has density $p$ with respect to some $\sigma$-finite measure $\mu$ on $E$, then

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

Remark 1.1.11. When $X$ is nonnegative but not necessarily in $L^1(\mathbb{P})$, the integral in Definition 1.1.9 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 $X \in L^1(\mathbb{P})$ 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\(^4\) — but, of course, the converse statement does not hold in general.

Lemma 1.1.12 (Jensen inequality). Let $X \in L^1(\mathbb{P})$ 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 1.1.13. Prove Lemma 1.1.12.

1.1.6 Variance and moments

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

Definition 1.1.15 (Variance). The variance of a random variable $X \in L^2(\mathbb{P})$ is defined by

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

Notice that, by Exercise 1.1.14, the assumption that $X \in L^2(\mathbb{P})$ ensures that $\mathbb{E}[X]$ is well-defined.

Exercise 1.1.16. Show that, for any $X \in L^2(\mathbb{P})$, for any $a, b \in \mathbb{R}$, $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

\(^4\)See Exercise 1.1.21 for an illustrative example.
### 1.1 General definitions and results

#### 1.1.7 Independence

**Definition 1.1.17 (Independence).** Let $X_1, \ldots, X_k$ be random variables taking their values in respective measurable spaces $(E_1, \mathcal{E}_1), \ldots, (E_k, \mathcal{E}_k)$. These variables are called independent if, for any $C_1 \in \mathcal{E}_1, \ldots, C_k \in \mathcal{E}_k$,

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

It is clear that, equivalently, the random variables $X_1, \ldots, X_k$ are independent if the law of $(X_1, \ldots, X_k) \in E_1 \times \cdots \times E_k$ is the product measure $P_{X_1} \otimes \cdots \otimes P_{X_k}$. When $dP_{X_i} = p_i(x_i) \, du_i(x_i)$ 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 $X_1, \ldots, X_k$ are independent, then for any functions $f_1 \in L^1(P_{X_1}), \ldots, f_k \in L^1(P_{X_k})$, the random variable $f_1(X_1) \cdots f_k(X_k)$ is integrable and satisfies

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

**Exercise 1.1.18.** For any independent variables $X, Y \in L^2(\mathbb{P})$, show that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. *What is the value of $\text{Var}(X - Y)$?*

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 $\{i_1, \ldots, i_k\}$, the variables $X_{i_1}, \ldots, X_{i_k}$ 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 1.1.19 (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 $A_1, \ldots, A_k$ to be independent.

Notice that the first question in Exercise 1.1.19 shows in particular that two events $A$ and $B$, with $\mathbb{P}(B) > 0$, are independent if and only if $\mathbb{P}(A | B) = \mathbb{P}(A)$: the knowledge that the event $B$ is realised does not affect the probability of $A$.

#### 1.1.8 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 $\phi$)? 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 1.1.10, for any measurable and bounded function $f : F \to \mathbb{R}$, we have

$$\mathbb{E}[f(Y)] = \mathbb{E}[f(\phi(X))] = \int_{x \in \mathbb{R}^d} f(\phi(x))p_X(x) \, dx.$$
Assume that by a suitable change of variable \( x \rightarrow y = \phi(x) \), one is able to rewrite the right-hand side under the form
\[
\int_{y \in F} f(y)q(y)\,d\mu(y)
\]
for some \( \sigma \)-finite measure \( \mu \) on \( F \). Then necessarily \( q \) is a probability density with respect to \( \mu \), and it is the density of \( Y \) since we have written that
\[
\mathbb{E}[f(Y)] = \int_{y \in F} f(y)q(y)\,d\mu(y)
\]
for all bounded and measurable functions \( f \).

If \( p_X \) vanishes outside some open subset \( U \) of \( \mathbb{R}^d \) and \( \phi \) 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
\[
\int_{x \in \mathbb{R}^d} f(\phi(x))p_X(x)\,dx = \int_{x \in U} f(\phi(x))p_X(x)\,dx = \int_{y \in V} f(y)p_X(\phi^{-1}(y))|J_{\phi^{-1}}(y)|\,dy,
\]
where \( J_{\phi^{-1}} \) is the Jacobian determinant of \( \phi^{-1} \). In this case, we deduce that \( Y \) has density
\[
1_{\{y \in U\}} p_X(\phi^{-1}(y))|J_{\phi^{-1}}(y)|
\]
with respect to the Lebesgue measure on \( \mathbb{R}^d \).

If \( \phi \) 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. Many examples of such situations will be seen in the sequel of the course.

**Exercise 1.1.20.** Let \( X \) and \( Y \) be two independent random variables in \( \mathbb{R} \) with respective densities \( p \) and \( q \). Show that \( Z = X + Y \) has density
\[
p \ast q(z) = \int_{y \in \mathbb{R}} p(z-y)q(y)\,dy.
\]

1.1.9 * The Borel–Cantelli Lemmas

**Exercise 1.1.21** (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
\[
X = \sum_{n \geq 1} 1_{A_n}
\]

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 \( \omega \) 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 \( \sigma \)-field and \( \mathbb{P} \) the Lebesgue measure. Let \( A_n = [0, \epsilon_n] \) for some sequence \( \epsilon_n \) 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 1.1.22 (Second Borel–Cantelli Lemma). Let \((A_n)_{n \geq 1}\) be a sequence of independent events, which satisfies the condition that
\[
\sum_{n \geq 1} P(A_n) = + \infty. \tag{1.2}
\]

Our aim is to show that in this case, the event \(\limsup_{n \to +\infty} A_n\) introduced in Exercise 1.1.21 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 \(B_N = \bigcap_{n \geq N} A_n^c\). Show that
\[
P(B) \leq \sum_{N \geq 1} P(B_N).
\]

2. For any \(N \geq 1\), show that
\[
P(B_N) \leq \liminf_{k \to +\infty} \prod_{n=N}^{N+k-1} (1 - P(A_n)).
\]

3. Deduce that, for any \(N \geq 1\), \(P(B_N) = 0\).

1.2 Random number simulation

1.2.1 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.’\(^5\). Hence, pseudo-random number generators are algorithms which, starting from a seed \(x_0\), return a sequence \(x_1, x_2, \ldots\) of numbers which looks ‘as random as possible’. By the way, how to determine whether a given sequence ‘looks random’ is already a nontrivial issue, which we will not discuss here. Because of the finiteness of the memory of a computer, a pseudo-random number generator is necessarily periodic, that is to say that there exists \(t \geq 0\) such that for all \(n \geq 0\), \(x_{n+t} = x_n\). Since ‘truly random’ sequences should not be periodic, it is an intuitive statement that a ‘good’ pseudo-random number generator should have a large 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 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\(^6\), it is based on the arithmetic properties of Mersenne numbers and its period is \( 2^{19937} - 1 \).

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 1.1.17;
(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 \( \frac{1}{m} \).

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

\[
\mathbb{P}(U_n \in C) = \frac{1}{m} \sum_{x=0}^{m-1} \mathbb{1}_{\{x/m \in C\}} \approx \int_{u=0}^{1} \mathbb{1}_{\{u \in C\}} \, du.
\]

This motivates the following definition.

**Definition 1.2.1** (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 \( U \sim U[0, 1] \).

**Exercise 1.2.2.** Let \( U \sim U[0, 1] \). Compute \( \mathbb{E}[U] \) and \( \text{Var}(U) \).

**Exercise 1.2.3.** Let \( U \sim U[0, 1] \). Show that the random variable \( 1 - U \) has the same distribution as \( U \).

**Remark 1.2.4** (Difference between variable and law). Exercise 1.2.3 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.

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 1.2.5** (Uniform distribution). The uniform distribution on the interval \([a, b]\), denoted by \( U[a, b] \), is the probability measure with density

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

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

1.2 Random number simulation

1.2.2 Classical discrete distributions

We first introduce several discrete distributions.

Definition 1.2.6 (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 \( T \sim \mathcal{Geo}(p) \).

Exercise 1.2.7 (Properties of Bernoulli, binomial and geometric distributions). Let \( X, S \) and \( T \) be as in Definition 1.2.6.

1. Compute \( \mathbb{E}[X] \) and \( \text{Var}(X) \).
2. Compute \( \mathbb{E}[S] \) and \( \text{Var}(S) \).
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 \( \text{Var}(T) \).

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

Exercise 1.2.8. 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 \( T \sim \mathcal{Geo}(p) \)?

1.2.3 The inverse CDF method

Definition 1.2.9 (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

\[
\forall x \in \mathbb{R}, \quad F_X(x) := \mathbb{P}(X \leq x).
\]

Remark 1.2.10. Since the Borel \( \sigma \)-field on \( \mathbb{R} \) is generated by the intervals\(^7\) of the form \(( -\infty, x ]\), two random variables have the same CDF if and only if they have the same law\(^8\).

Exercise 1.2.11 (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 1.2.9 yields the identity

\[
\forall x \in \mathbb{R}, \quad F_X(x) = \int_{y=-\infty}^{x} p(y)\,dy,
\]

which shows that \( F_X \) is continuous and \( dx \)-almost everywhere differentiable, with \( F_X' = p \).

---

\(^7\)See [5, Exercice p. 10].

\(^8\)It is a practical consequence of the Dynkin System Theorem [5, Théorème 1.4.1, Corollaire 1.4.2] that if a \( \sigma \)-field \( \mathcal{B} \) is generated by a set \( \mathcal{C} \) which is stable by intersection, then any two probability measures which agree on \( \mathcal{C} \) actually agree on \( \mathcal{B} \).
Definition 1.2.12. 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
\[
\forall u \in [0, 1], \quad F_X^{-1}(u) := \inf\{x \in \mathbb{R} : F_X(x) \geq u\},
\]
with the conventions that \( \inf \mathbb{R} = -\infty \) and \( \inf \emptyset = +\infty \).

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 1.2.13 (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 \( u \leq F_X(x) \). Reciprocally, if \( u \leq F_X(x) \), then by the definition of \( F_X^{-1} \), \( F_X^{-1}(u) \leq x \). \(\square\)

Corollary 1.2.14 (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 1.2.13 and Definition 1.2.1, 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 1.2.10 we conclude that they have the same distribution. \(\square\)

We illustrate this method on the exponential distribution.

Definition 1.2.15 (Exponential distribution). Let \( \lambda > 0 \). A random variable \( X \) in \([0, +\infty)\) is called exponential with parameter \( \lambda \) if it has the density
\[
p(x) = \mathbb{1}_{x > 0} \lambda e^{-\lambda x}.
\]
We denote \( X \sim \mathcal{E}(\lambda) \).

Exercise 1.2.16 (Properties of exponential distributions). Let \( X \sim \mathcal{E}(\lambda) \).
1. Compute \( \mathbb{E}[X] \) and \( \text{Var}(X) \).
2. If \( a > 0 \), what is the law of \( aX \)?

An immediate computation shows that the CDF of \( X \) writes
\[
F_X(x) = \begin{cases} 
0 & \text{if } x \leq 0, \\
1 - e^{-\lambda x} & \text{otherwise}.
\end{cases}
\]
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 1.2.3, it is also equivalent to return \( -\frac{1}{\lambda} \ln(U) \).
Exercise 1.2.17 (Other standard densities). Apply the inverse CDF method to the following standard probability densities.

1. The Pareto distribution with parameter $\alpha > 0$, with density $f(x) = \begin{cases} \frac{\alpha x^{-\alpha-1}}{x_\min^{\alpha+1}} & \text{if } x > x_\min \\ 0 & \text{otherwise} \end{cases}$.

2. The Cauchy distribution with parameter $\nu > 0$, with density $f(x) = \frac{1}{\pi (\nu^2 + x^2)}$.

3. The Weibull distribution with parameter $\lambda > 0$, with density $f(x) = \lambda x^{\lambda-1} \exp(-x/\lambda)$.

4. The Rayleigh distribution with parameter $\sigma > 0$, with density $f(x) = \frac{x}{\sigma^2} \exp(-x^2/2\sigma^2)$.

Exercise 1.2.18 (Poisson distribution). A random variable $N \in \mathbb{N}$ is distributed according to the Poisson distribution with parameter $\lambda > 0$ if, for any $k \in \mathbb{N}$,

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

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

1. Compute $E[N]$ and $\text{Var}(N)$.

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 $k \geq 0$, $P(S_n = k) \to P(N = k)$. 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 $\lambda$. 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 $[0, 1]$.

1.2.4 Gaussian random variables

We recall that the Gauss integral is equal to

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

Definition 1.2.19 (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 1.2.20. If $G$ is a standard Gaussian variable, show that $G \in L^p(\mathbb{P})$ for any $p \in [1, +\infty)$ and compute $E[G]$ and $\text{Var}(G)$.

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

Definition 1.2.21 (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 $\mu$ and variance $\sigma^2$. 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 $\sigma$ through $\sigma^2$ is justified by the following result.

---

9We use the notation $\mathbb{N} = \{0, 1, \ldots\}$.
10Do not hesitate to redo the computation just to be sure that you still know how to!
Exercise 1.2.22. Show that if $X \sim N(\mu, \sigma^2)$ with $\sigma^2 > 0$, then $X$ has density

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

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 $\mu$. In this case, the law of $X$ is the Dirac measure $\delta_\mu$ 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 $\Phi : \mathbb{R} \rightarrow [0,1]$ denote its CDF, given by

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

It is known that $\Phi$ 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 1.2.23 (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

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

are independent and follow the standard Gaussian distribution.

Proof. We use the dummy function method introduced in Subsection 1.1.8 and let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be measurable and bounded. Since $R$ and $\Theta$ are independent, the law of the pair $(R, \Theta)$ is the product of the marginal densities, and therefore

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

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

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

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 $\Theta$ can be sampled using the inverse CDF method, Proposition 1.2.23 provides a method to sample $X$ and $Y$ from two independent uniform random variables on $[0,1]$. 

1.2 Random number simulation

1.2.5 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

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

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 1.1):

(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).

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.

\begin{itemize}
  \item \textbf{Exercise 1.2.24.} Show that $N \sim \text{Geo}(|D|/|R|)$.
\end{itemize}

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

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

Figure 1.1: The domain $D$ framed into a rectangle $R$. Random points are drawn in $R$, only those falling into $D$ are kept.
Since the random variables $X_1, \ldots, X_n$ are independent, each term of the sum rewrites
\[
\mathbb{P}(X_1 \notin D, \ldots, X_{n-1} \notin D, X_n \in C \cap D) = \mathbb{P}(X_1 \notin D) \cdot \cdots \cdot \mathbb{P}(X_{n-1} \notin D) \mathbb{P}(X_n \in C \cap D)
\]
\[
= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \int_{x \in R} \mathbbm{1}_{\{x \in C \cap D\}} \frac{dx}{|R|}
\]
\[
= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \frac{|D|}{|R|} \int_{x \in \mathbb{R}^d} \mathbbm{1}_{\{x \in C\}} p(x) dx,
\]
where $p$ denotes the uniform density on $D$. Summing over $n$, we deduce that
\[
\mathbb{P}(X_N \in C) = \int_{x \in \mathbb{R}^d} \mathbbm{1}_{\{x \in C\}} p(x) dx,
\]
which shows that $X_N$ has density $p$.

\section*{Exercise 1.2.25}

Show that the random variables $X_N$ and $N$ are independent.

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

\section*{Theorem 1.2.26 (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, $\mathbb{R}^d$-almost everywhere, $p(x) \leq k q(x)$. 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
\[
N := \inf\{n \geq 1 : k q(X_n) U_n \leq p(X_n)\}.
\]

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

The proof of Theorem 1.2.26 follows from the same computation as in the example of uniform distributions, for which $q$ is the uniform distribution on the rectangle, and $k = |R|/|D|$. It 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 $\frac{dp}{dq} \leq k$, $Q$-almost everywhere. Then the statement of Theorem 1.2.26 remains in force, with $N$ defined as the first index for which $k U_n \leq \frac{dp}{dq}(X_n)$.

Rejection sampling is useful when one is not able to sample directly from $p$, but can find $q$ such that $p \leq k q$ 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 $p$.

\section*{Exercise 1.2.27 (Gamma distribution)}

The Gamma distribution with (shape) parameter $a > 0$ is the probability measure on $\mathbb{R}$ with density
\[
p(x) = \mathbbm{1}_{\{x > 0\}} \frac{1}{\Gamma(a)} x^{a-1} e^{-x},
\]
where $\Gamma$ is Euler’s function
\[
\Gamma(a) := \int_{x=0}^{+\infty} x^{a-1} e^{-x} dx.
\]

We assume that $a > 1$ and want to implement the rejection sampling method with $q$ the density of the exponential distribution with parameter $\lambda$. Which value of $\lambda$ should we take? What will be the resulting value of $k$?
1.3 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 $L^p(\mathbb{P}; \mathbb{R}^d)$ the set of random vectors whose coordinates are random variables in $L^p(\mathbb{P})$. If $X = (X_1, \ldots, X_n) \in L^1(\mathbb{P}; \mathbb{R}^d)$, we denote by $E[X]$ the vector $(E[X_1], \ldots, E[X_d])$.

1.3.1 Covariance

**Definition 1.3.1** (Covariance between two random variables). Let $X, Y \in L^2(\mathbb{P})$. The covariance between $X$ and $Y$ is defined by

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

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

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

Besides, we have the formula

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

and the Cauchy–Schwarz inequality yields

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

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

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

is always between $-1$ and $1$.

**Exercise 1.3.2.**

1. Show that, if $X$ and $Y$ are independent, then $\text{Cov}(X, Y) = 0$.

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

**Definition 1.3.3** (Covariance matrix). Let $X = (X_1, \ldots, X_d) \in L^2(\mathbb{P}; \mathbb{R}^d)$. The covariance matrix of $X$ is the $d \times d$ matrix $\text{Cov}[X]$ with coefficients $\text{Cov}(X_i, X_j)$.

Clearly, a covariance matrix is symmetric. Exercise 1.3.4 below shows that it is also nonnegative. We shall see in Proposition 1.3.17 that, conversely, any symmetric and nonnegative matrix is the covariance matrix of a (Gaussian) random vector.

**Exercise 1.3.4.** Show that if $X \in L^2(\mathbb{P}; \mathbb{R}^d)$ has covariance matrix $K$, then for any $u \in \mathbb{R}^d$,

$$\text{Var}(\langle u, X \rangle) = \langle u, K u \rangle.$$

**Exercise 1.3.5.** Let $X \in L^2(\mathbb{P}; \mathbb{R}^d)$ with covariance matrix $K$, and let $b \in \mathbb{R}^k$, $A \in \mathbb{R}^{k \times d}$. Show that

$$\text{Cov}[b + AX] = AK A^\top.$$
1.3.2 Characteristic function

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

**Definition 1.3.6** (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

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

By Theorem 1.1.10, we get that

\[
\forall u \in \mathbb{R}^d, \quad \Psi_X(u) = \int_{x \in \mathbb{R}^d} e^{i(u,x)}dP_X(x),
\]

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

\[
\forall u \in \mathbb{R}^d, \quad \Psi_X(u) = \Psi_Y(u).
\]

**Exercise 1.3.8.** 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 1.3.9.** Compute the characteristic function of \( X \) when \( X \sim \mathcal{B}(p), \mathcal{B}(n,p), \mathcal{G}(p), \mathcal{P}(\lambda), \mathcal{U}(a,b), \mathcal{E}(\lambda) \).

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. A more elementary approach is proposed in Exercise 1.3.11 below. It first requires the following technical statement, which will also be used to prove the Central Limit Theorem in Chapter 2.

**Lemma 1.3.10** (Derivatives of the characteristic function). If \( X \in L^p(\mathbb{P}; \mathbb{R}^d) \) for some integer \( p \geq 1 \), then \( \Psi_X \) is of class \( C^p \) on \( \mathbb{R}^d \) and, for any multi-index \( q = (q_1, \ldots, q_d) \) with \( |q| = q_1 + \cdots + q_d \leq p \),

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

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

\[
\frac{\partial^{|q|}}{\partial u_1^{q_1} \cdots \partial u_d^{q_d}} e^{i(u,X)} = i^{|q|} X_1^{q_1} \cdots X_d^{q_d} e^{i(u,X)}.
\]

The right-hand side satisfies

\[
\left|i^{|q|} X_1^{q_1} \cdots X_d^{q_d} e^{i(u,X)}\right| = |X_1|^{q_1} \cdots |X_d|^{q_d}.
\]

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

\[
\mathbb{E}\left[|X_1|^{q_1} \cdots |X_d|^{q_d}\right] \leq \mathbb{E}\left[|X_1|^{\alpha_1}\right]^{1/\alpha_1} \cdots \mathbb{E}\left[|X_d|^{\alpha_d}\right]^{1/\alpha_d} = \mathbb{E}\left[|X_1|^p\right]^{1/\alpha_1} \cdots \mathbb{E}\left[|X_d|^p\right]^{1/\alpha_d} < +\infty.
\]

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. \(\Box\)
Exercise 1.3.11 (Characteristic function of Gaussian random variables). Let \( G \sim \mathcal{N}(0, 1) \).

1. Show that \( \Psi_G \) 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 1.3.12 (Characterisation of independence). Two random vectors \( X \in \mathbb{R}^d \) and \( Y \in \mathbb{R}^k \) are independent if and only if
\[
\forall u \in \mathbb{R}^d, \quad \forall v \in \mathbb{R}^k, \quad \Psi_{(X,Y)}(u,v) = \Psi_X(u)\Psi_Y(v).
\]

1.3.3 Gaussian vectors

Definition 1.3.13 (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 1.2.21.

Let \( X \in L^2(\mathbb{P}; \mathbb{R}^d) \). Set \( m = \mathbb{E}[X] \in \mathbb{R}^d \) and \( K = \text{Cov}[X] \in \mathbb{R}^{d \times d} \). For any \( u \in \mathbb{R}^d \), it is immediate that
\[
\mathbb{E}[\langle u, X \rangle] = \langle u, m \rangle,
\]
and by Exercise 1.3.4,
\[
\text{Var}(\langle u, X \rangle) = \langle u, Ku \rangle.
\]
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 1.3.11,
\[
\Psi_X(u) = \mathbb{E}[e^{i\langle u, X \rangle}] = \exp \left( i\langle u, m \rangle - \frac{1}{2}\langle u, Ku \rangle \right).
\]
We deduce the following statement.

Proposition 1.3.14 (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 \( u \in \mathbb{R}^d \),
\[
\Psi_X(u) = \exp \left( i\langle u, m \rangle - \frac{1}{2}\langle u, Ku \rangle \right).
\]
In this case, we have \( m = \mathbb{E}[X] \) and \( K = \text{Cov}[X] \), and we denote by \( \mathcal{N}_d(m, K) \) the law of \( X \).

Exercise 1.3.15 (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 1.3.16 (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
\[
\forall i \in \{1, \ldots, d\}, \quad \forall j \in \{1, \ldots, k\}, \quad \text{Cov}(X_i, Y_j) = 0.
\]

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 1.3.2. Conversely, assume that \( K_{X,Y} = 0 \), and set \( m_X = \mathbb{E}[X] \), \( m_Y = \mathbb{E}[Y] \). Then by Exercise 1.3.2 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. \(\square\)

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 1.2.23 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 1.3.17 (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 1.3.11, 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,
\[
\text{Var}(\langle u, X \rangle) = \sum_{i=1}^d \lambda_i \langle u, e_i \rangle^2 = \langle u, Ku \rangle,
\]
which shows that \( \mathbb{E}[X] = m \) and \( \text{Cov}[X] = K \). \(\square\)

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 \( X \).

Proposition 1.3.17 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 1.3.18. Show that, if $K$ is invertible, $X \sim N_d(m, K)$ 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?
Chapter 2

Limit theorems and the Monte Carlo method

Contents

2.1 Convergence of random variables ...................................... 23
2.2 Limit theorems .................................................................. 31
2.3 Principle of the Monte Carlo method .................................. 35
2.4 Variance reduction ............................................................ 38

In this chapter we introduce the Monte Carlo method in its most basic form, where one approximates an integral (interpreted as an expectation) by the empirical mean of iid samples. The theoretical justification of this method is provided by the Law of Large Numbers, and the quantification of its efficiency relies on the Central Limit Theorem. We therefore start by introducing these two important results, and thereby also provide background on the different notions of convergence of random variables. We assume knowledge of the main convergence results for Lebesgue integral: Fatou’s Lemma, the Monotone Convergence Theorem, and the Dominated Convergence Theorem.

2.1 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\(^1\).

2.1.1 Convergence of random variables: definitions and basic results

Definition 2.1.1 (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

\[
\forall \omega \in A, \quad \lim_{n \to +\infty} X_n(\omega) = X(\omega).
\]

\(^1\)To the notable exception of results involving characteristic functions, which would require a linear structure with duality properties.
(ii) $X_n$ converges to $X$ in probability if, for any $\epsilon > 0$,
\[
\lim_{n \to +\infty} \mathbb{P}(\|X_n - X\| \geq \epsilon) = 0.
\]

(iii) For any $p \in [1, +\infty)$, $X_n$ converges to $X$ in $L^p$ if\footnote{We should rather write in $L^p(\mathbb{P}; \mathbb{R}^d)$ to be completely consistent with the notation of Chapter 1, however in order not to overweight the exposition we shall adopt this shorthand notation throughout the chapter.}
\[
\lim_{n \to +\infty} \mathbb{E}[\|X_n - X\|^p] = 0.
\]

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 $L^p$ is the standard strong convergence in the linear space $L^p(\mathbb{P}; \mathbb{R}^d)$. Convergence in probability is a bit more unusual from this point of view\footnote{It is referred to as convergence in measure in analysis.}, however it plays a pivotal role in the articulation of the various modes of convergence.

**Proposition 2.1.2** (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 $L^q$, then $X_n \to X$ in $L^p$.

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

![Hierarchy of various modes of convergence (including the convergence in distribution which will be seen in Subsection 2.1.3) and partial converse statements.](image)

The hierarchy between these modes of convergence is summarised on Figure 2.1. In the proof of Proposition 2.1.2, we shall need the following two results.

**Lemma 2.1.3** (Dominated Convergence Theorem for random variables). Assume that $X_n \to X$ almost surely and that there exists $Y \in L^1(\mathbb{P})$ such that $|X_n| \leq Y$ for any $n$. Then $\mathbb{E}[X_n]$ converges to $\mathbb{E}[X]$.

The statement of Lemma 2.1.3 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 2.1.4** (Markov’s inequality). Let $Y \in L^1(\mathbb{P})$ be such that $Y \geq 0$, almost surely. For any $a > 0$, we have
\[
\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}.
\]
2.1 Convergence of random variables

Proof. Observe that, for any \( y \geq 0 \),
\[
\mathbb{I}_{\{y \geq a\}} \leq \frac{y}{a}
\]
and take the expectation of this inequality evaluated in \( y = Y \).

We are now in position to prove Proposition 2.1.2.

Proof of Proposition 2.1.2. 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 \( \omega \). For any \( \epsilon > 0 \), for any \( \omega \in A \), we have \(|X_n(\omega) - X(\omega)| < \epsilon\) for \( n \) large enough, and therefore
\[
\lim_{n \to +\infty} \mathbb{I}_{\{|X_n(\omega) - X(\omega)| \geq \epsilon\}} = 0.
\]
As a consequence, the random variable \( \mathbb{I}_{\{|X_n - X| \geq \epsilon\}} \) converges to 0, almost surely, and therefore by Lemma 2.1.3,
\[
\Pr(|X_n - X| \geq \epsilon) = \mathbb{E}\left[\mathbb{I}_{\{|X_n - X| \geq \epsilon\}}\right] \to 0.
\]
This proves the first point.

The second point is an immediate consequence of Exercise 1.1.14.

To prove the third point, we assume that \( X_n \to X \) in \( L^1 \) and fix \( \epsilon > 0 \). By Lemma 2.1.4, we then have
\[
\Pr(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\epsilon} \to 0,
\]
and the proof is completed.

Proposition 2.1.5 (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 2.1.5, we point out several remarks.

- You should first convince yourself that the three statements of Proposition 2.1.5 become trivial if convergence in probability is replaced with almost sure convergence.
- If convergence in probability is replaced with convergence in \( L^p \) 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 \( X_n + Y_n \to X + Y \), \( X_n Y_n \to XY \) (if \( d = 1 \), and so on.

Proof of Proposition 2.1.5. 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
\[
\Pr(|f(X_n) - f(X)| \geq \epsilon) = \Pr(|f(X_n) - f(X)| \geq \epsilon, |X| > M) + \Pr(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M) \leq \Pr(|X| > M) + \Pr(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M).
\]
The second term in the right-hand side rewrites
\[
\Pr(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M) = \Pr(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| \leq \delta_{M, \epsilon}) + \Pr(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| > \delta_{M, \epsilon}).
\]
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

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

On the other hand, it is immediate that

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

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 $\epsilon > 0$,

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

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 1.1.3 that almost surely, $|X - Y|$ is smaller than any $\epsilon_M$, and therefore $X = Y$. \(\square\)

### 2.1.2 * Convergence of random variables: complements

In general, none of the converse statements to those of Proposition 2.1.2 hold true: counter-examples are provided in Exercise 2.1.6. However, some partial converse statements are gathered in Proposition 2.1.7 and Exercise 2.1.9.

**Exercise 2.1.6** (Counter-examples to converse statements to Proposition 2.1.2). 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 1.1.22.

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

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

(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 $L^p$ but not in $L^q$.

(c) Choose $a$ and $b$ so that $X_n \to 0$ in probability but not in $L^1$.

**Proposition 2.1.7** (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 $X$. 

—
2.1 Convergence of random variables

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

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

where \((\epsilon_M)_{M \geq 1}\) is a deterministic sequence of positive numbers which converges to 0. Since, by Corollary 1.1.3, a countable intersection of almost sure events remains almost sure, we deduce that it suffices to construct \((n_k)_{k \geq 1}\) such that

\[
\forall M \geq 1, \quad \mathbb{P} (\exists K \geq 1 : \forall k \geq K, |X_{n_k} - X| \leq \epsilon_M) = 1.
\]

By the Borel–Cantelli Lemma (see Exercise 1.1.21), 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

\[
\forall k \geq 1, \quad \mathbb{P} (|X_{n_{k,M}} - X| > \epsilon_M) \leq \frac{1}{k^2},
\]

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 \(\epsilon_M\) is assumed to decrease, we have, for any \(M \geq 1\),

\[
\sum_{k=1}^{+\infty} \mathbb{P}(|X_{n_k} - X| > \epsilon_M) \leq M + \sum_{k=M+1}^{+\infty} \mathbb{P}(|X_{n_{k,M}} - X| > \epsilon_M)
\]

\[
\leq M + \sum_{k=M+1}^{+\infty} \mathbb{P}(|X_{n_{k,k}} - X| > \epsilon_k)
\]

\[
\leq M + \sum_{k=M+1}^{+\infty} \frac{1}{k^2},
\]

which completes the proof. \(\square\)

Proposition 2.1.7 allows to prove the following generalisation of Lemma 2.1.3, in which the almost sure convergence requirement is relaxed to convergence in probability, and which will be useful in the sequel.

\(\blacklozenge\) Exercise 2.1.8 (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 \(|X_n| \leq Y\) for any \(n\). The purpose of this exercise is to prove that \(\mathbb{E}[X_n]\) converges to \(\mathbb{E}[X]\).

1. Using Proposition 2.1.7, 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 \(\mathbb{E}[Y1_{\{|Y|>M\}}] \leq \epsilon\).
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 \(\mathbf{L}^p\).

\(\blacklozenge\) Exercise 2.1.9 (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 \(2^p(|X_n|^p + |X|^p) - |X_n - X|^p \geq 0\), 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.
2.1.3 Convergence in distribution

Convergence in distribution is a bit different from the modes of convergence introduced in Definition 2.1.1, because it concerns the law of $X_n$ rather than the variable $X_n$ itself.

**Definition 2.1.10** (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 $f : \mathbb{R}^d \to \mathbb{R}$,

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

**Definition 2.1.11** (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 $f : \mathbb{R}^d \to \mathbb{R}$,

$$\lim_{n \to +\infty} \int_{x \in \mathbb{R}^d} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

In order to compare convergence in distribution with the modes of convergence introduced in Definition 2.1.1, one should compare the following three statements with the contents of Proposition 2.1.5.

**Proposition 2.1.12** (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 have the same law.

The first and third points of Proposition 2.1.12 are straightforward consequences of Definition 2.1.11. 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 2.1.13** (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 2.1.13 is postponed to Exercise 2.1.16. We first detail important properties and characterisations of the convergence in distribution.

**Proposition 2.1.14** (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.
2.1 Convergence of random variables

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 \( \epsilon > 0 \),

\[
|E[f(X_n)] - E[f(X)]| \leq E[|f(X_n) - f(X)|] \\
= E[|f(X_n) - f(X)|I_{|f(X_n) - f(X)| < \epsilon}] \\
+ E[|f(X_n) - f(X)|I_{|f(X_n) - f(X)| \geq \epsilon}] \\
\leq \epsilon + 2\|f\|_{\infty} P(|f(X_n) - f(X)| \geq \epsilon).
\]

By Proposition 2.1.5 (i), \( P(|f(X_n) - f(X)| \geq \epsilon) \) goes to 0 when \( n \to \infty \). Therefore

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

for any \( \epsilon \), 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 \( \psi_\epsilon(r) \geq I_{\{r \geq \epsilon\}} \) 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

\[
\forall x' \in \mathbb{R}^d, \quad I_{\{|x' - x| \geq \epsilon\}} \leq f(x'), \quad \text{and} \quad f(x) = 0.
\]

Applying Definition 2.1.11 with this function, we get

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

which shows that \( X_n \to x \) in probability. \( \square \)

Proposition 2.1.15 (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 (slightly modified) Fourier transform \( \mathcal{F} \), so that the claim to prove reduces to showing some continuity property of the inverse transform \( \mathcal{F}^{-1} \). \( \square \)

Exercise 2.1.16 (Proof of Lemma 2.1.13). Prove the two statements of Lemma 2.1.13 using Proposition 2.1.15.

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

Exercise 2.1.17 (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, for any \( \epsilon > 0 \) and any \( M \geq 0 \),

\[
\limsup_{n \to +\infty} P(|X_n - x| \geq \epsilon) \leq P(|Y| \geq M),
\]

and deduce that \( X_n \to x \) in probability.

2. Let \( f : \mathbb{R}^d \to \mathbb{R}^k \) be \( C^1 \). Show that

\[
\lim_{n \to +\infty} a_n (f(X_n) - f(x)) = \nabla f(x)Y,
\]

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.

\textbf{Exercise 2.1.18} (Scheffé’s Lemma). Let \((p_n)_{n \geq 1}\) be a sequence of probability densities with respect to some \(\sigma\)-finite measure \(\mu\) on \(\mathbb{R}^d\), such that

\[
\mu\text{-almost everywhere, } \quad p_n \rightarrow p,
\]

for some probability density \(p\) with respect to \(\mu\) on \(\mathbb{R}^d\).

\begin{enumerate}[(i)]
    \item Show that for any \(n \geq 1\),
    \[
    \int_{x \in \mathbb{R}^d} |p_n(x) - p(x)| \, d\mu(x) = 2 \int_{x \in \mathbb{R}^d} [p_n(x) - p(x)]_+ \, d\mu(x).
    \]
    \item Deduce that
    \[
    \lim_{n \rightarrow +\infty} \int_{x \in \mathbb{R}^d} |p_n(x) - p(x)| \, d\mu(x) = 0,
    \]
    and then that if \(X_n\) has density \(p_n\) and \(X\) has density \(p\) then \(X_n \rightarrow X\) in distribution.
\end{enumerate}

\textbf{Remark 2.1.19} (Convergence in distribution in discrete spaces). Definition 2.1.11 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 \(\sigma\)-field of all its subsets. In this case, the natural topology on \(E\) is the one making all functions \(f : E \rightarrow \mathbb{R}\) continuous. Then, with similar arguments as in Exercise 2.1.18 (taking for \(\mu\) the counting measure \(\sum_{x \in E} \delta_x\)), it may be shown that the following statements are equivalent:

\(i\) \(X_n \rightarrow X\) in distribution, that is to say \(\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]\) for any bounded function \(f\);  

\(ii\) for any \(x \in E\), \(\mathbb{P}(X_n = x) \rightarrow \mathbb{P}(X = x)\);  

\(iii\) \(\sum_{x \in E} |\mathbb{P}(X_n = x) - \mathbb{P}(X = x)| \rightarrow 0\).

**2.1.4 * 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 \(f : \mathbb{R}^d \rightarrow \mathbb{R}\), we look for conditions under which \(\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]\). Of course, by Definition 2.1.11, it is the case if \(f\) is continuous and bounded. We shall study how to relax both conditions. We start with the continuity condition.

\textbf{Proposition 2.1.20} (Mapping theorem). Let \(X_n \rightarrow X\) in distribution and \(f : \mathbb{R}^d \rightarrow \mathbb{R}\) 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 \(\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]\).

We leave the proof of Proposition 2.1.20 aside but insist on an important corollary.

\textbf{Corollary 2.1.21} (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 \(F(x^-) = F(x)\),

\[
\lim_{n \rightarrow +\infty} F_n(x) = F(x).
\]

In particular,

\(i\) \(F_n(x) \rightarrow F(x), \; \text{d}x\text{-almost everywhere};\)
(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 2.1.21 is a straighforward consequence of Proposition 2.1.20, 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 $E[X_n]$ converges to $E[X]$ for $X_n, X \in \mathbb{R}$.

**Definition 2.1.22** (Uniform integrability). A sequence of random variables $(X_n)_{n \geq 1}$ in $\mathbb{R}$ is called uniformly integrable if

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

**Exercise 2.1.23.** Let $(X_n)_{n \geq 1}$ be a sequence of random variables in $\mathbb{R}$. This sequence is said to be bounded in $L^p(\mathbb{P})$ if $\sup_{n \geq 1} E[|X_n|^p] < +\infty$.

1. Show that if $(X_n)_{n \geq 1}$ is uniformly integrable then it is bounded in $L^1(\mathbb{P})$.
2. Construct a sequence which is bounded in $L^p(\mathbb{P})$ but not uniformly integrable.
3. If there exists $p > 1$ such that $(X_n)_{n \geq 1}$ is bounded in $L^p(\mathbb{P})$, 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 2.1.24** (Convergence of expectations). If $X_n \to X$ in distribution and the sequence $(X_n)_{n \geq 1}$ is uniformly integrable, then $E[X_n] \to E[X]$.

**Exercise 2.1.25.** Prove Proposition 2.1.24.

### 2.2 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 $n$ 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 2.2.1.** If $X_1 \in L^1(\mathbb{P}; \mathbb{R}^d)$, compute $E[\overline{X}_n]$ and if $X_1 \in L^2(\mathbb{P}; \mathbb{R}^d)$, compute $\text{Cov}[\overline{X}_n]$.

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

#### 2.2.1 Laws of Large Numbers

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

**Proposition 2.2.2** (Weak Law of Large Numbers). If $X_1 \in L^2(\mathbb{P}; \mathbb{R}^d)$ then $\overline{X}_n \to E[X_1]$ in $L^2$. 

**Proof.** The proof of the weak LLN is elementary. For the sake of simplicity we assume that \( d = 1 \). Then, by Exercise 2.2.1,
\[
\mathbb{E} \left[ |X_n - \mathbb{E}[X_1]|^2 \right] = \text{Var}(X_n) = \frac{\text{Var}(X_1)}{n},
\]
which converges to 0.

**Theorem 2.2.3** (Strong Law of Large Numbers). If \( X_1 \in L^1(\mathbb{P}; \mathbb{R}^d) \) then \( \overline{X}_n \to \mathbb{E}[X_1] \) almost surely.

Theorem 2.2.3 is certainly a cornerstone of probability theory, but its proof is far from trivial. It is sketched and discussed in Subsection 2.2.2.

**Exercise 2.2.4** (Law of Large Numbers in \( L^p \)). Let \( p \geq 1 \) and assume that \( X_1 \in L^p(\mathbb{P}; \mathbb{R}^d) \). The aim of this exercise is to prove that \( \overline{X}_n \to \mathbb{E}[X_1] \) in \( L^p \). 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 \( M \geq 1 \),
\[
\mathbb{E} \left[ |X_n|^p \mathbb{1}_{\{|X_n|^p \geq M\}} \right] \leq \mathbb{E} \left[ |X_1|^p \mathbb{1}_{\{|X_n|^p \geq M\}} \right].
\]

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

**2.2.2 * 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 2.2.3 is equivalent to the statement that

\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} X'_i = 0, \quad \text{almost surely.} \tag{2.1}
\]

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 (2.1) is equivalent to

\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = 0, \quad \text{almost surely.} \tag{2.2}
\]

Besides, an explicit computation yields

\[
\sum_{n \geq 1} \text{Var} \left( \frac{Z_n}{n} \right) < +\infty. \tag{2.3}
\]

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

**Lemma 2.2.5** (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 \( x > 0 \),

\[
\mathbb{P} \left( \sup_{n \geq 1} |W_n| > x \right) \leq \frac{1}{x^2} \sum_{i=1}^{\infty} \text{Var}(Y_i).
\]
Lemma 2.2.5, combined with the Borel–Cantelli Lemma, allows to prove the following statement.

**Lemma 2.2.6** (Convergence criterion). Let \((U_n)_{n \geq 1}\) be a sequence of independent and centered random variables. If
\[
\sum_{n=1}^{\infty} \text{Var}(U_n) < +\infty,
\]
then there exists a random variable \(T\) such that
\[
\lim_{N \to +\infty} \sum_{n=1}^{N} U_n = T, \quad \text{almost surely.}
\]

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

**Lemma 2.2.7** (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
\[
a_n \sum_{i=1}^{n} u_i \text{ converges to } 0.
\]

4. Conclusion of the proof: combining (2.3) with Lemma 2.2.6, we get that \(\sum_{n=1}^{N} Z_n/n\) has a finite limit, which by the Kronecker Lemma then yields (2.2).

Shorter and more elementary proofs have been proposed since Kolmogorov’s original proof. A particularly famous one is due to Etemadi in 1981⁴. A recent preprint by Fitzsimmons⁵ discusses another elementary sketch and its relation with previous similar arguments in the literature.

### 2.2.3 The Central Limit Theorem

In this subsection we assume that \(X_1 \in L^2(\mathbb{P}; \mathbb{R}^d)\) and set \(K = \text{Cov}[X_1] \in \mathbb{R}^{d \times d}\). 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 1.

**Theorem 2.2.8** (Central Limit Theorem). We have
\[
\lim_{n \to +\infty} \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \mathcal{N}_d(0, K).
\]

**Proof.** For all \(i \geq 1\), let \(Y_i = X_i - \mathbb{E}[X_1]\), so that \(\mathbb{E}[Y_i] = 0\) and \(\text{Cov}[Y_i] = K\). We also denote
\[
Z_n = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,
\]
so that the characteristic function of \(Z_n\) writes, for all \(u \in \mathbb{R}^d\),
\[
\Psi_{Z_n}(u) = \mathbb{E} \left[ \exp \left( i \left( u, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \right) \right) \right] = \mathbb{E} \left[ \exp \left( i \left( u, \sqrt{n} Y_1 \right) \right) \right] = \Phi_{Y_1}(u/\sqrt{n})^n,
\]
where we have used the fact that the variables \(Y_i\) are iid. By Lemma 1.3.10, writing \(Y_1 = (Y_{1,1}, \ldots, Y_{1,d})\), we have
\[
\frac{\partial \Psi_{Y_1}(0)}{\partial u_i} = i \mathbb{E}[Y_{1,i}] = 0, \quad \frac{\partial^2 \Psi_{Y_1}(0)}{\partial u_i \partial u_j} = i^2 \mathbb{E}[Y_{1,i}Y_{1,j}] = -K_{i,j},
\]

⁴https://link.springer.com/article/10.1007/BF01013465
⁵https://arxiv.org/abs/2111.05766
so that the function $\Psi_{Y_1}$ satisfies Taylor’s expansion

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

when $n \to +\infty$. Using Lemma 2.2.9 below, we deduce that

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

which by Proposition 1.3.14 is the characteristic function of $Z \sim \mathcal{N}_d(0, K)$. As a consequence, Proposition 1.3.7 ensures that $Z_n$ converges in distribution to $Z$.

In the proof of Theorem 2.2.8, we have used the following technical result.

**Lemma 2.2.9** (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

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

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

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

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

$$\left(1 + \frac{\theta}{n} + \frac{\epsilon_n}{n}\right)^n - \left(1 + \frac{\theta}{n}\right)^n = \int_{u=0}^{1} \frac{d}{du} \left(1 + \frac{\theta}{n} + \frac{u\epsilon_n}{n}\right)^n du = \epsilon_n \int_{u=0}^{1} \left(1 + \frac{\theta}{n} + \frac{u\epsilon_n}{n}\right)^{n-1} du,$$

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. $\square$

**Exercise 2.2.10** (Stronger convergence in the CLT). With the notation of the proof of Theorem 2.2.8, 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?
2.3 Principle of the Monte Carlo method

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

$$I = \int_{x \in E} f(x) dP(x),$$

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

$$I = \mathbb{E}[f(X)], \quad X \sim P,$$

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

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

with $X_1, \ldots, X_n$ independent copies of $X$. The numerical approximation of $I$ by $\hat{I}_n$ is the essence of the Monte Carlo method. In this section, we discuss its accuracy.

2.3.1 Asymptotic confidence intervals

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

$$\lim_{n \to +\infty} \frac{\sqrt{n}}{\sigma^2} (\hat{I}_n - I) = N(0, 1),$$

in distribution, and therefore, by Corollary 2.1.21, for any $\phi > 0$, the interval

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

satisfies

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

In particular, if one fixes $\alpha \in (0, 1/2)$ and takes for $\phi$ 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 2.2.

In short, if $n$ is large enough, then there is a 95% probability that the quantity $I$, which we aim at evaluating, lies between $\hat{I}_n - 1.96\sigma/\sqrt{n}$ and $\hat{I}_n + 1.96\sigma/\sqrt{n}$. Two problems remain with this statement: the assumption that ‘$n$ is large enough’ is rather vague, and the variance $\sigma^2$ 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 2.3.1** (Asymptotic confidence interval). Let

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

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

$$I'_n = \left[ \hat{I}_n - \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{I}_n + \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$$
satisfies
\[ \lim_{n \to +\infty} \mathbb{P} \left( J \in I'_n \right) = 1 - \alpha. \]

Proposition 2.3.1 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 \( \sigma^2 \) by its empirical version \( \hat{\sigma}_n^2 \). 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 \( \hat{I}_n \) of a Monte Carlo estimation.

**Proof of Proposition 2.3.1.** We first write
\[ \frac{\sqrt{n}}{\hat{\sigma}_n} \left( \hat{\sigma}_n - \hat{\sigma}_n \right) = \frac{\sigma}{\hat{\sigma}_n} \frac{\sqrt{n}}{\sigma} \left( \hat{\sigma}_n - \hat{\sigma}_n \right). \]

By the strong LLN, the ratio \( \sigma/\hat{\sigma}_n \) converges to 1, almost surely. Therefore, Slutsky’s Lemma 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 2.1.21 as for \( I_n \). \( \square \)

### 2.3.2 Nonasymptotic confidence intervals

The statement of Proposition 2.3.1 is asymptotic, and therefore it is natural to ask how large should \( n \) be chosen for the probability that \( J \in I'_n \) to be close to \( 1 - \alpha \). In general, there is no really satisfactory answer to this question. However, nonasymptotic bounds may be derived from concentration inequalities, and yield intervals \( J_n \) which are such that
\[ \mathbb{P} \left( J \in J_n \right) \geq 1 - \alpha. \] (2.4)

An elementary such example is provided by Tchebychev’s inequality:
\[ \forall a > 0, \quad \mathbb{P} \left( |\hat{J}_n - \hat{J}_n| \geq a \right) \leq \frac{\text{Var}(\hat{J}_n)}{a^2} = \frac{\sigma^2}{na^2}, \]
which easily follows from Markov’s inequality, and from which we deduce that
\[ J_n = \left[ \hat{J}_n - \frac{\sigma}{\sqrt{\alpha n}}, \hat{J}_n + \frac{\sigma}{\sqrt{\alpha n}} \right] \]
satisfies the estimate (2.4). Since the latter is an inequality, and not an equality, the interval \( J_n \) is more conservative than \( I_n \): the probability that \( J \in J_n \) could be much larger than \( 1 - \alpha \), but it is
2.3 Principle of the Monte Carlo method

Figure 2.3: Ratio between the width of the confidence intervals $J_n$ and $I_n$.

at least $1 - \alpha$. On the other hand, the bound (2.4) holds for any value of $n$, and does not rely on the Central Limit Theorem. We plot on Figure 2.3 the ratio $(1/\sqrt{\alpha})/\phi_{1-\alpha/2}$ between the widths of $J_n$ and $I_n$, as a function of $\alpha$: when $\alpha$ 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 $\sigma$, which is not known in general. Similarly to Proposition 2.3.1, 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 2.3.2. Let $Y \in L^2(\mathbb{P})$.

1. Show that

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

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

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

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 2.3.3 (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

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

1. Show that $F'(\lambda) = \mathbb{E}_\lambda[Z_1]$ and $F''(\lambda) = \text{Var}_\lambda(Z_1)$ for some probability measure $\mathbb{P}_\lambda$ to be defined.

2. Using Exercise 2.3.2, deduce that, for any $\lambda \geq 0$, $\mathbb{E}[^{\text{exp}}(\lambda Z_1)] \leq \exp(\lambda^2/8)$. 
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.

### 2.4 Variance reduction

The length of the confidence interval \( I_n \) obtained in Section 2.3 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
\[
\mathcal{I} = \mathbb{P}(X \geq 20) = \mathbb{E}[f(X)], \quad f(x) = 1_{\{x \geq 20\}}.
\]

On the one hand, an upper bound on \( \mathcal{I} \) can be obtained analytically by writing
\[
\mathcal{I} = \frac{1}{\sqrt{2\pi}} \int_{y=20}^{+\infty} e^{-y^2/2} \, dy \leq \frac{1}{\sqrt{2\pi}} \int_{y=20}^{+\infty} \frac{y}{20} e^{-y^2/2} \, dy = \frac{e^{-20^2/2}}{20\sqrt{2\pi}} \simeq 2.8 \times 10^{-89}.
\]

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
\[
\hat{\mathcal{I}}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).
\]

Exercise 2.4.1. What is the law of the random variable \( N = \inf\{n \geq 1 : \hat{\mathcal{I}}_n \neq 0\} \)? What is its expectation?

Since \( f(X_i) \sim \mathcal{B}(\mathcal{I}) \), we have \( \sigma^2 = \text{Var}(f(X_1)) = \mathcal{I}(1-\mathcal{I}) \simeq \mathcal{I} \), 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 \( \epsilon \). Then we need to take \( n \) such that
\[
2\phi_{1-\alpha/2} \sqrt{\frac{\mathcal{I}}{n}} \leq \epsilon \mathcal{I},
\]
2.4 Variance reduction

that is to say
\[ n \geq \left( \frac{2\phi_1 - \alpha/2}{\epsilon} \right)^2 \frac{1}{J}. \]

For \( \epsilon = 0.01 \) and \( \alpha = 0.05 \), using the analytic bound on \( J \) we obtain that \( n \) should be at least \( 5.6 \times 10^3 \), which is impossible to realise in practice.

In this section, we present variance reduction techniques which allow to construct estimators of \( J \) with a smaller variance \( \sigma^2 \), and therefore yield smaller confidence intervals.

2.4.1 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 \( \beta \in \mathbb{R} \),
\[ J = \mathbb{E}[f(X)] = \mathbb{E}[f(X) - \beta Y] + \beta \mathbb{E}[Y], \]
which suggests to approximate \( J \) by the estimator
\[ \hat{J}_{n}^{\text{CV}, \beta} := \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \beta Y_i) + \beta \mathbb{E}[Y]. \]

The variance of this estimator is \( (\sigma_{\text{CV}, \beta}^2)/n \), where
\[ (\sigma_{\text{CV}, \beta}^2) = \text{Var}(f(X) - \beta Y) = \sigma^2 - 2\beta \text{Cov}(f(X), Y) + \beta^2 \text{Var}(Y). \]

We may already remark that if \( \text{Cov}(f(X), Y) = 0 \) then \( (\sigma_{\text{CV}, \beta}^2) \) is always larger than the variance \( \sigma^2 \) 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 \( \beta \) for which \( (\sigma_{\text{CV}, \beta}^2) \) is minimal is then
\[ \beta^* = \frac{\text{Cov}(f(X), Y)}{\text{Var}(Y)}, \]
which yields the variance
\[ (\sigma_{\text{CV}, \beta^*}^2) = \sigma^2 \left( 1 - \rho^2 \right), \]
where
\[ \rho = \frac{\text{Cov}(f(X), Y)}{\sqrt{\text{Var}(f(X))} \sqrt{\text{Var}(Y)}} \in [-1, 1] \]
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 2.4.3 for an illustration.

In practice, the optimal choice of \( \beta \) depends on the quantity \( \text{Cov}(f(X), Y) \) which may need to be estimated. Let us introduce
\[ \hat{C}_n = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \hat{J}_n)(Y_i - \bar{Y}_n). \]

The strong LLN shows that
\[ \hat{\beta}_n^* := \frac{\hat{C}_n}{\text{Var}(Y)} \]
converges to \( \beta^* \) almost surely, and Slutsky’s Lemma then yields the following result.
Proposition 2.4.2 (Control variate method). Let \((X_i, Y_i)_{1 \leq i \leq n}\) be a sequence of iid pairs such that \(f(X_i), Y_i \in L^2(P)\). For all \(n \geq 1\), let

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

with \(\hat{\beta}^*_{n}\) defined above. The interval

\[
I^{\text{CV}}_n = \left[ \hat{I}^{\text{CV}}_n - \phi_{1-\alpha/2} \sqrt{\frac{(\hat{\sigma}^2_{n})^2}{n}}, \hat{I}^{\text{CV}}_n + \phi_{1-\alpha/2} \sqrt{\frac{(\hat{\sigma}^2_{n})^2}{n}} \right],
\]

where

\[
(\hat{\sigma}^2_{n})^2 = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 \text{Var}(Y)}
\]

satisfies

\[
\lim_{n \to +\infty} \mathbb{P}(I \in I^{\text{CV}}_n) = 1 - \alpha.
\]

Exercise 2.4.3. Let \(X \sim \mathcal{N}(0, 1)\). For all \(\lambda > 0\), we define

\[
f_{\lambda}(x) = \frac{1}{1 + \lambda x^2},
\]

and set

\[
\mathcal{J}_{\lambda} = \mathbb{E}[f_{\lambda}(X)] = \frac{1}{\sqrt{2\pi}} \int_{x \in \mathbb{R}} \frac{e^{-x^2/2}}{1 + \lambda x^2} \, dx.
\]

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

1. Compute \(\mathbb{E}[Y_1]\).
2. Compare numerically the variances of the Monte Carlo estimator \(\hat{I}_n\) and of the control variate estimator \(\hat{I}^{\text{CV}}_n\).
3. How does this comparison vary with \(\lambda\)? What is your interpretation of this fact?

2.4.2 Importance sampling

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

\[
\mathcal{J} = \int_{x \in E} f(x) \, dP(x) = \int_{x \in E} f(x) w(x) \, dQ(x),
\]

where the function \(w\) is simply the density

\[
w(x) = \frac{dP}{dQ}(x).
\]

As a consequence, the quantity

\[
\hat{\mathcal{J}}^{\text{IS}}_n := \frac{1}{n} \sum_{i=1}^{n} f(Y_i) w(Y_i),
\]

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

\[
1_{\{f(x) \neq 0\}} \, dP(x) \ll 1_{\{f(x) \neq 0\}} \, dQ(x),
\]

and for which we still denote by \(w\) the associated density.
Exercise 2.4.4. Show that if \( P \ll Q \), then \( Q \) satisfies (2.5), 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
\[
(\sigma_Q^{IS})^2 := \text{Var}(f(Y)w(Y))
\]
as small as possible.

**Proposition 2.4.5** (Optimal choice of \( Q \)). Let \( \mathcal{J} = \mathbb{E}[|f(X)|] \), assume that this quantity is positive, and define the probability measure \( Q^* \) by
\[
dQ^*(x) = \frac{|f(x)|}{\mathcal{J}} dP(x).
\]

(i) \( Q^* \) satisfies (2.5) and \( (\sigma_Q^{IS})^2 = \mathcal{J}^2 - \mathcal{J}^2 \).

(ii) For any probability measure \( Q \) which also satisfies (2.5), \( (\sigma_Q^{IS})^2 \leq (\sigma_Q^{IS})^2 \).

(iii) If \( f \) has constant sign \( P \)-almost everywhere, then \( (\sigma_Q^{IS})^2 = 0 \).

**Proof.** As a preliminary remark, we note that for any \( Q \) satisfying (2.5),
\[
(\sigma_Q^{IS})^2 = \mathbb{E}[(f(Y)w(Y))^2] - \mathcal{J}^2, \quad Y \sim Q. \quad (2.6)
\]

First, it is easily checked that \( \mathbb{I}_{\{f(x) \neq 0\}} dP(x) \) has density
\[
w^*(x) = \mathbb{I}_{\{f(x) \neq 0\}} \frac{\mathcal{J}}{|f(x)|}
\]
with respect to \( \mathbb{I}_{\{f(x) \neq 0\}} dQ^*(x) \), therefore \( Q^* \) satisfies (2.5) and besides, if \( Y^* \sim Q^* \), then
\[
\mathbb{E}[(f(Y^*)w^*(Y^*))^2] = \int_{x \in E} \mathbb{I}_{\{f(x) \neq 0\}} |f(x)|^2 \left( \frac{\mathcal{J}}{|f(x)|} \right)^2 dQ^*(x)
\]
\[
= \mathcal{J}^2 \int_{x \in E} \mathbb{I}_{\{f(x) \neq 0\}} dQ^*(x)
\]
\[
= \mathcal{J}^2,
\]
which, together with (2.6), proves (i). The point (iii) then immediately follows.

Second, let us fix \( Q \) which satisfies (2.5) and denote by \( w \) the associated density. By definition of \( \mathcal{J} \) and \( w \), and the Cauchy–Schwarz inequality,
\[
\mathcal{J}^2 = \left( \int_{x \in E} |f(x)| \mathbb{I}_{\{f(x) \neq 0\}} dP(x) \right)^2
\]
\[
= \left( \int_{x \in E} |f(x)| w(x) \mathbb{I}_{\{f(x) \neq 0\}} dQ(x) \right)^2
\]
\[
\leq \int_{x \in E} |f(x)|^2 w(x)^2 \mathbb{I}_{\{f(x) \neq 0\}} dQ(x)
\]
\[
= \mathbb{E}[(f(Y)w(Y))^2],
\]
with \( Y \sim Q \). Combined with (2.6), this estimate completes the proof of (ii). \( \square \)
In practice it is impossible to implement the method with the optimal measure $Q^*$ since the latter depends explicitly on the quantity $I$, which is likely to be unknown — and, in the case where $f$ is nonnegative $P$-almost everywhere, is exactly the quantity $I$ 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 $|f(x)|dP(x)$.

**Exercise 2.4.6.** For the example of the estimation of $\Pr(X \geq 20)$, for $X \sim \mathcal{N}(0, 1)$, given in the introduction of this section, the optimal density is proportional to $1_{\{y \geq 20\}}e^{-y^2/2}$. We take $q(y)$ the density of the law $\mathcal{N}(20, 1)$.

1. Compute the associated variance $(\sigma_{IS})^2$.
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 (2.5) for the optimality property stated in Proposition 2.4.5. On the one hand, it is necessary to work with measures $Q$ which satisfy the condition (2.5) rather than $P \ll Q$. Indeed, the optimal density $Q^*$ does not necessarily satisfy the latter condition. On the other hand, when $\overline{I} > |I|$ (which implies that $f(X)$ changes sign), it is possible to find importance sampling estimators of the form

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

such that for $Y \sim Q$,

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

A trivial example would be to assume that there exists $x_0 \in E$ such that $f(x_0) = I$, and set $Q = \delta_{x_0}$, $w(x) = 1$ for any $x$. Of course, in this case, the measure $Q$ does not satisfy (2.5).
Part II

Markov chains and MCMC methods
Chapter 3

Markov chains and ergodic theorems in discrete spaces

Contents

3.1 Definitions and the Markov property ................. 45
3.2 Stationary distribution ............................. 48
3.3 Ergodic theorems .................................. 55
3.4 Convergence to equilibrium .......................... 59

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

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

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

Chapter 3 to 4 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 $I$ in some cases where iid samples are not available.

Throughout the next two 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 $\sigma$-field $\mathcal{E}$. For any measure $\mu$ on $E$ and $x \in E$, we shall write $\mu(x)$ for $\mu(\{x\})$, so that for any $C \in \mathcal{E}$, $\mu(C) = \sum_{x \in C} \mu(x)$.

3.1 Definitions and the Markov property

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 3.1.3 below. We first introduce the notion of stochastic matrix$^1$.

$^1$In the case where $E$ is countably infinite, we slightly abuse terminology and still call matrix an infinite array indexed by $E \times E$. 
**Definition 3.1.1** (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 \( \mu \) is a probability measure on \( E \), \( X \) a random variable in \( E \) with distribution \( \mu \) and \( f : E \to \mathbb{R} \) is in \( L^1(\mu) \), then \( \mathbb{E}[f(X)] = \sum_{x \in E} \mu(x)f(x) \) simply rewrites \( \mu f \). We also denote by \( 1 \in \mathbb{R}^E \) the column vector of which all coordinates are equal to 1, and recall that the set of probability measures on \( E \) is denoted by \( \mathcal{P}(E) \).

**Exercise 3.1.2** (Properties of stochastic matrices). Let \( P \) be a stochastic matrix.

1. Show that \( P 1 = 1 \).
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.

We may now introduce the notion of Markov chain.

**Definition 3.1.3** (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 \( x_0, \ldots, x_n \in E \) such that \( \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0 \), for all \( x_{n+1} \in E \),

\[
\mathbb{P}(X_{n+1} = x_{n+1}|X_0 = x_0, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n) = P_{n+1}(x_n, x_{n+1}).
\]

(3.1)

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

Equation (3.1) 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 \( X_0, \ldots, X_n \) through the current state \( X_n \).

**Example 3.1.4** (The Ehrenfest urn). Consider a box divided into two compartments, called A and B, and which contains \( N \) particles, see Figure 3.1. 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.

![Figure 3.1: The Ehrenfest urn with \( N = 10 \) particles.](image)
The microscopic description consists in recording the compartment in which each particle is located, so that a configuration is a vector \( x = (x^1, \ldots, x^N) \in E_{\text{micro}} := \{A, B\}^N \). 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 \( E_{\text{macro}} = \{0, \ldots, N\} \), and the transition matrix is given by

\[
P(k, k + 1) = \frac{N - k}{N}, \quad P(k, k - 1) = \frac{k}{N},
\]

and the other coefficients are 0.

**Exercise 3.1.5.** Let \( \pi \) be a probability measure on \( E \).
1. Let \((X_n)_{n \geq 0}\) be a sequence of iid random variables with law \( \pi \). Show that \((X_n)_{n \geq 0}\) is a homogeneous Markov chain and describe its transition matrix.
2. Let \( \xi \) be a random variable with law \( \pi \), 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 \((X_0, \ldots, X_n)\) and \((Y_0, \ldots, Y_n)\)?

From Definition 3.1.3 we deduce the following properties related with the law of the sequence \((X_n)_{n \geq 0}\).

**Proposition 3.1.6** (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 \( x_0, \ldots, x_n \in E \),

\[
\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \mu_0(x_0)P_1(x_0, x_1) \cdots P_n(x_{n-1}, x_n). \tag{3.2}
\]

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

Before detailing the proof of Proposition 3.1.6, we emphasise a few of its consequences.

**Remark 3.1.7.**
(i) The first assertion shows that the law of any vector \((X_0, \ldots, X_n)\) is entirely characterised by two objects: the initial distribution \( \mu_0 \) 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 \( f \in L^1(\mu_0) \), \( \mathbb{E}[f(X_n)] = \mu_0 P^n f \).

**Proof of Proposition 3.1.6.** We prove the first assertion by induction on \( n \geq 0 \). For \( n = 0 \) this is immediate. Let \( n \geq 0 \) be such that (3.2) holds, and let \( x_0, \ldots, x_n, x_{n+1} \in E \). If \( \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = 0 \), then on the one hand the fact that \( \{X_0 = x_0, \ldots, X_{n+1} = x_{n+1}\} \subset \{X_0 = x_0, \ldots, X_n = x_n\} \) ensures that the former event has also probability 0, while on the other hand the identity (3.2) 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 $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0$, we have

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

where we have used (3.1) and (3.2) 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 (3.1).

\hfill \Box

Exercise 3.1.8 (Random mapping representation). Let $(F, \mathcal{F})$ be a measurable space (not necessarily finite, nor even countably infinite) and $h : E \times F \to E$ be a measurable function. Let $X_0$ be a random variable in $E$ with law $\mu_0$, and $(U_n)_{n \geq 1}$ be a sequence of iid random variables in $F$, independent from $X_0$. Consider the sequence $(X_n)_{n \geq 0}$ defined by the recursive relation

$$
\forall n \geq 1, \quad X_n = h(X_{n-1}, U_n).
$$

1. Check that $(X_n)_{n \geq 0}$ is a homogeneous Markov chain and write its transition matrix.

2. Conversely, show that any homogeneous Markov chain in $E$ admits such a representation\(^2\).

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 $\mu$. 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 3.1.6 and Remark 3.1.7 that when $X_0 = x$, the law of $X_n$ is related with the $n$-th power of $P$ by the identity

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

Exercise 3.1.9 (A reformulation of the Markov property for homogeneous chains). Let $(X_n)_{n \geq 0}$ be a homogeneous Markov chain with transition matrix $P$. Let $n, m \geq 0$, $F : E^{n+1} \to \mathbb{R}$ and $G : E^{m+1} \to \mathbb{R}$. For all $\mu \in \mathcal{P}(E)$ and $x \in E$, show that

$$
\mathbb{E}_\mu \left[ F(X_0, \ldots, X_n) \mathbbm{1}_{\{X_n = x\}} G(X_n, \ldots, X_{n+m}) \right] = \mathbb{E}_\mu \left[ F(X_0, \ldots, X_n) \mathbbm{1}_{\{X_n = x\}} \right] \mathbb{E}_x \left[ G(X_0, \ldots, X_m) \right].
$$

Exercise 3.1.9 shows that given the information that $X_n = x$, the sequence $(Y_m)_{m \geq 0}$ defined by $Y_m = X_{n+m}$ is a Markov chain with initial state $x$ and transition matrix $P$, independent from the sequence $(X_k)_{0 \leq k \leq n}$.

### 3.2 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.

\(^2\)More precisely, that for any $\mu_0 \in \mathcal{P}(E)$ and stochastic matrix $P$, one may construct a measurable space $(F, \mathcal{F})$ and an iid sequence $(U_n)_{n \geq 1}$ in $F$ such that the sequence $(X_n)_{n \geq 0}$ defined as above is a homogeneous Markov chain with transition matrix $P$ and initial distribution $\mu_0$. 


3.2 Stationary distribution

3.2.1 Definition

Definition 3.2.1 (Stationary distribution). Let \((X_n)_{n \geq 0}\) be a Markov chain in \(E\) with transition matrix \(P\). A probability measure \(\pi\) 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 3.2.2 (Stationary distribution). Let \(\pi\) be a stationary distribution for \((X_n)_{n \geq 0}\). For any \(n \geq 0\),

\[ \forall x \in E, \quad \mathbb{P}_\pi(X_n = x) = \pi(x); \]

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 3.1.7.

Exercise 3.2.3. Show that if \(\pi\) 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 3.2.4. Consider the Ehrenfest urn from Example 3.1.4.

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

Natural questions about stationary distributions are uniqueness and existence. They are respectively addressed in the next two subsections.

3.2.2 Irreducibility and uniqueness of stationary distributions

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 \(\pi_1\), \(\pi_2\) 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

\[ \pi'_1 = (\pi_1 \ 0), \quad \pi'_2 = (0 \ \pi_2). \]

Then it is clear that both \(\pi'_1\) and \(\pi'_2\) (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 3.2.

This remark motivates the following definition.

Definition 3.2.5 (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 3.2: 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 3.2.6.** Check that the condition that there exist $x = x_0, x_1, \ldots, x_n = y$ such that $P(x_0, x_1) \cdots P(x_{n-1}, x_n) > 0$ is equivalent to $\mathbb{P}_x(X_n = y) > 0$.

**Proposition 3.2.7** (Uniqueness of a stationary distribution). If the stochastic matrix $P$ is irreducible, then it possesses a unique stationary distribution.

We start the proof of Proposition 3.2.7 with the following exercise.

**Exercise 3.2.8.** Let $P$ be an irreducible stochastic matrix, and let $\pi$ be an associated stationary distribution. Show that for all $x \in E$, $\pi(x) > 0$.

We now introduce a useful object.

**Definition 3.2.9** (Dirichlet form). Let $P$ be a stochastic matrix, and let $\pi$ be an associated stationary distribution. The Dirichlet form of $(P, \pi)$ is the quadratic form $\mathcal{E}_\pi$ defined on $\mathbb{R}^E$ by

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

**Lemma 3.2.10** (Another expression for $\mathcal{E}_\pi$). 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 3.2.9, we write

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

$$= \frac{1}{2} \sum_{y \in E} f(y)^2 \pi P(y) - \sum_{x \in E} f(x) \pi(x) P f(x) + \frac{1}{2} \sum_{x \in E} f(x)^2 \pi(x),$$

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

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

which is the claimed expression. \qed

**Lemma 3.2.11** (Dimension of $\ker(P - I)$ for irreducible matrices). Let $P$ be an irreducible stochastic matrix. For all $f \in \mathbb{R}^E$, if $P f = f$ then there exists $c \in \mathbb{R}$ such that $f = c 1$. 
3.2 Stationary distribution

**Proof.** Let \( \pi \) be a stationary distribution for \( P \). If \( Pf = f \), then Lemma 3.2.10 immediately shows that \( \mathcal{E}_\pi(f) = 0 \), therefore by Definition 3.2.9 and Exercise 3.2.8, \( f(x) = f(y) \) 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 3.2.5. 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 \( E \). \( \square \)

We are now ready to complete the proof of Proposition 3.2.7. We first note that if \( E \) is finite, then the proof is a straightforward consequence of Lemma 3.2.11 because by 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 \( P - I \). In the general case, resorting to the Rank-Nullity Theorem may be delicate, however an explicit duality argument can still be employed.

**Proof of Proposition 3.2.7.** Let \( \pi \) be a stationary distribution for the irreducible stochastic matrix \( P \). By Exercise 3.2.8, \( \pi(x) > 0 \) for any \( x \in E \) and therefore the matrix \( Q \) with nonnegative coefficients

\[
Q(x, y) = \frac{\pi(y)P(y, x)}{\pi(x)}
\]

is well-defined. It is easily checked that this matrix is stochastic and irreducible. Now let \( \mu \) be a second stationary distribution for \( P \), and set \( f(x) = \mu(x)/\pi(x) \) for any \( x \in E \). We have

\[
Qf(x) = \sum_{y \in E} Q(x, y)f(y) = \sum_{y \in E} P(y, x)\frac{\mu(y)}{\pi(x)} = \frac{\mu P(x)}{\pi(x)} = f(x).
\]

Hence, applying Lemma 3.2.11 to \( Q \), we deduce that \( \mu \) and \( \pi \) are proportional to each other, and therefore equal. \( \square \)

**Exercise 3.2.12** (The coupon collector). A brand of chocolate eggs hides surprise gifts in each 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 \( \tau_N \) in terms of \( \eta_1, \ldots, \eta_N \)? What is the law of each \( \eta_i \)?
5. Show that, for any \( c > 0 \), \( \mathbb{P}(\tau_N > [N \ln N + cN]) \leq e^{-c} \). Hint: for \( i \in \{1, \ldots, N\} \) and \( k \geq 1 \), introduce the event \( A^k_i = \{ \text{no gift of the } i^\text{th} \text{ type has been found in the first } k \text{ eggs} \} \).

### 3.2.3 Recurrence, transience and existence of a stationary distribution

In this subsection, we address the question of the existence of stationary distributions.

**The case of finite state space**

If the set \( E \) is finite then the situation is rather simple.

**Proposition 3.2.13** (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 3.2.13. 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 3.1.2, 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 $\hat{\mu}_n = \frac{1}{n} \sum_{k=0}^{n-1} \mu P^k$, so that for all $f : E \to \mathbb{R}$,

\[ \hat{\mu}_n f = \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_\mu[f(X_k)]. \]

By Exercise 3.1.2 and convexity, $\hat{\mu}_n \in \mathcal{P}(E)$ and by compactness, there exists an increasing sequence $(n_\ell)_{\ell \geq 1}$ such that $\hat{\mu}_{n_\ell}$ converges to some $\pi \in \mathcal{P}(E)$ when $\ell \to +\infty$. Since, for all $\ell \geq 1$,

\[ \hat{\mu}_{n_\ell} P = \frac{1}{n_\ell} \sum_{k=0}^{n_\ell-1} \mu P^{k+1} = \hat{\mu}_{n_\ell} + \frac{1}{n_\ell} (\mu P^{n_\ell} - \mu), \]

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

**Recurrence and transience**

When $E$ is infinite, the compactness arguments used in both proofs of Proposition 3.2.13 no longer holds, and in fact, there are natural examples of Markov chains which do not admit a stationary distribution.

**Example 3.2.14 (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

\[ P(x, y) = \begin{cases} \frac{1}{d} & \text{if } |x - y| = 1, \\ 0 & \text{otherwise}, \end{cases} \]

where $| \cdot |$ denotes the Euclidean norm on $\mathbb{R}^d$.

**Exercise 3.2.15.** Show that a stationary distribution $\pi$ 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

\[ \tau_x = \inf\{n \geq 1 : X_n = x\}, \quad x \in E. \tag{3.3} \]

**Definition 3.2.16 (Recurrent and transient states).** For a given stochastic matrix $P$ on $E$, a state $x \in E$ is called transient if $\mathbb{P}_x(\tau_x = +\infty) > 0$ and recurrent if $\mathbb{P}_x(\tau_x = +\infty) = 0$. Furthermore, recurrent states are called null if $\mathbb{E}_x[\tau_x] = +\infty$ and positive if $\mathbb{E}_x[\tau_x] < +\infty$. 


The main results regarding recurrence, transience and stationary distributions are gathered in the next proposition. In this statement, we refer to the fact of being transient, null recurrent or positive recurrent as the nature of a state.

**Proposition 3.2.17 (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

\[
\forall x \in E, \quad \pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]}
\]

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.

Most of the proof of Proposition 3.2.17 is postponed to the next section. Here we shall only prove that if \(P\) is irreducible, then either all states are recurrent or they are transient. Our argument relies on the following characterisation of recurrence and transience.

**Exercise 3.2.18 (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 \(\sum_{n=0}^{+\infty} P^n(x, x) = +\infty\)).

**Partial proof of Proposition 3.2.17 (i).** 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 \(n \geq 0\),

\[
\begin{align*}
P^{n+p+q}(x, x) &\geq P^p(x, y)P^n(y, y)P^q(y, x), \\
P^{n+p+q}(y, y) &\geq P^q(y, x)P^n(x, x)P^p(x, y),
\end{align*}
\]

the series \(\sum P^n(x, x)\) and \(\sum P^n(y, y)\) have the same nature, and therefore by Exercise 3.2.18, either all states are transient, or all states are recurrent. \(\square\)

Taking the statement of Proposition 3.2.17 for granted, Proposition 3.2.13 implies that in a finite state space, all irreducible chains are necessarily positive recurrent. This can also be observed directly.

**Lemma 3.2.19 (Positive recurrence 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 3.2.5 implies that there exists \(n_{x'} \geq 1\) such that \(P^{n_{x'}}(x', x) > 0\). Let

\[
\kappa := \min_{x' \in E} P^{n_{x'}}(x', x) > 0, \quad m := \max_{x' \in E} n_{x'} < +\infty,
\]

so that whatever the initial state \(x'\), the probability for the chain to return to \(x\) before the time \(m\) is at least \(\kappa\). Indeed,

\[
\kappa \leq P^{n_{x'}}(x', x) = \mathbb{P}_{x'}(X_{n_{x'}} = x) \leq \mathbb{P}_{x'}(\tau_x \leq n_{x'}) \leq \mathbb{P}_{x'}(\tau_x \leq m).
\]
Hence, for any $\ell \geq 1$,
\[
\mathbb{P}_\mu(\tau_x > \ell m) = \mathbb{P}_\mu(X_1 \neq x, \ldots, X_{\ell m} \neq x) = \sum_{x' \neq x} \mathbb{E}_\mu \left[ F(X_1, \ldots, X_{(\ell-1)m}) \mathbb{1}_{(X_{(\ell-1)m}=x')} G(X_{(\ell-1)m+1}, \ldots, X_{\ell m}) \right],
\]
where
\[
F(x_1, \ldots, x_{(\ell-1)m}) = \mathbb{1}_{\{x_1 \neq \ldots, x_{(\ell-1)m} \neq x\}}, \quad G(x_1, \ldots, x_m) = \mathbb{1}_{\{x_1 \neq \ldots, x_m \neq x\}}.
\]
Using Exercise 3.1.9, we get
\[
\mathbb{E}_\mu \left[ F(X_1, \ldots, X_{(\ell-1)m}) \mathbb{1}_{(X_{(\ell-1)m}=x')} G(X_{(\ell-1)m+1}, \ldots, X_{\ell m}) \right] = \mathbb{P}_\mu \left[ (X_1 \neq x, \ldots, X_{(\ell-1)m-1} \neq x, X_{(\ell-1)m} = x') \mathbb{P}_{x'}(X_{(\ell-1)m} = x') \mathbb{P}_x'(\tau > m) \right]
\]
\[
\leq (1 - \kappa) \mathbb{P}_\mu(\tau > (\ell - 1)m) \mathbb{P}_\mu(\tau_x > (\ell - 1)m),
\]
so that
\[
\mathbb{P}_\mu(\tau_x > \ell m) \leq (1 - \kappa) \mathbb{P}_\mu(\tau > (\ell - 1)m),
\]
and thus
\[
\mathbb{P}_\mu(\tau_x > \ell m) \leq (1 - \kappa)^\ell.
\]
We complete the proof by remarking that, by the Fubini–Tonelli Theorem,
\[
\mathbb{E}_\mu[\tau_x] = \mathbb{E}_\mu \left[ \sum_{n=0}^{+\infty} \mathbb{1}_{\{n < \tau_x\}} \right]
\]
\[
= \sum_{n=0}^{+\infty} \mathbb{P}_\mu(\tau_x > n) = \sum_{\ell=0}^{+\infty} \sum_{k=0}^{m-1} \mathbb{P}_\mu(\tau_x > \ell m + k)
\]
\[
< \sum_{\ell=0}^{+\infty} m \mathbb{P}_\mu(\tau_x > \ell m) = \sum_{\ell=0}^{+\infty} m (1 - \kappa)^\ell = \frac{m}{\kappa}.
\]

\begin{enumerate}
\item[\textclubsuit] Exercise 3.2.20 (Exponential moments). Under the assumptions of Lemma 3.2.19, show that there exists $\epsilon > 0$ such that $\mathbb{E}_\mu[\exp(\epsilon \tau_x)] < +\infty$. Deduce that for all $p \geq 1$, $\mathbb{E}_\mu[(\tau_x)^p] < +\infty$.
\item[\textdegrees] Exercise 3.2.21 (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 3.2.18.
\end{enumerate}
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 \((X^1_n, X^2_n)\) the coordinates of the simple random walk.
   
   (a) Let \( U_n = (X^1_{n+1} - X^1_n) + (X^2_{n+1} - X^2_n) \) and \( V_n = (X^1_{n+1} - X^1_n) - (X^2_{n+1} - X^2_n) \).
   
   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. We now assume that \( d \geq 3 \) and denote by \( \varphi \) the characteristic function of \( X_1 \) under \( \mathbb{P}_0 \), defined by

\[
\forall u \in \mathbb{R}^d, \quad \varphi(u) := \mathbb{E}_0[e^{i\langle u, X_1 \rangle}].
\]

(a) Show that, for all \( u = (u_1, \ldots, u_d) \in \mathbb{R}^d \),

\[
\varphi(u) = \frac{1}{d} (\cos u_1 + \cdots + \cos u_d).
\]

(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.

### 3.3 Ergodic theorems

Throughout this section, we let \((X_n)_{n \geq 0}\) be a Markov chain with transition matrix \( P \), and we set

\[
\forall x \in E, \quad \pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]} \in [0, 1], \quad (3.4)
\]

where we recall the definition of the return time \( \tau_x \) in (3.3). We point out the fact that \( \pi(x) > 0 \) if and only if \( x \) is positive recurrent.

#### 3.3.1 The Law of Large Numbers

**Statement and comments**

The main result of this subsection is the following statement.

**Theorem 3.3.1 (Ergodic theorem).** If the chain \((X_n)_{n \geq 0}\) is irreducible and positive recurrent, then for any \( f \in L^1(\pi) \),

\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \pi f, \quad \text{almost surely},
\]

where we recall from Proposition 3.2.17 that \( \pi \) is the unique stationary distribution.

Theorem 3.3.1 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. The same remark applies to Lemma 3.3.2 below.
Proofs

The main ingredient for the proof of Theorem 3.3.1, which will also be used to complete the proof of Proposition 3.2.17, is the following lemma, which holds without irreducibility assumption, and whatever the nature of the state.

**Lemma 3.3.2** (Limit of the empirical measure). For any \( x \in E \), we have

\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} = \pi(x), \quad \text{almost surely.}
\]

If \( x \) is transient, then by Exercise 3.2.18, \( N_x < +\infty \) almost surely, so that for \( n \) large enough,

\[
\frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} = \frac{N_x}{n} \to 0 = \pi(x).
\]

On the other hand, if \( x \) is recurrent, then by Exercise 3.2.18 again, one may define the sequence of successive return times \( \tau_x^0, \tau_x^1, \ldots \) of \( X_n \) to \( x \) by

\[
\tau_x^0 = \inf\{n \geq 0 : X_n = x\}, \quad \tau_x^{\ell+1} = \inf\{n > \tau_x^\ell : X_n = x\}.
\]

The trajectory of \( X_n \) between times \( \tau_x^\ell \) and \( \tau_x^{\ell+1} - 1 \) is called the \( \ell \)-th excursion of the chain outside \( x \). For any \( C \subset E \), for any \( \ell \geq 0 \), we then introduce the random variable

\[
Z_x^\ell(C) = \sum_{n=\tau_x^\ell}^{\tau_x^{\ell+1}-1} \mathbb{1}_{\{X_n \in C\}},
\]

which counts how many steps of the \( \ell \)-th excursion are in \( C \). We first prove the following result.

**Lemma 3.3.3** (Excursions). For any \( C \subset E \), the integer-valued random variables \( (Z_x^\ell(C))_{\ell \geq 0} \) are iid, with law given by

\[
\forall \ell \geq 0, \quad \forall z \geq 0, \quad \mathbb{P}
\left( Z_x^\ell(C) = z \right) = \mathbb{P}
\left( \sum_{n=0}^{\tau_x^{\ell}-1} \mathbb{1}_{\{X_n \in C\}} = z \right).
\]

Besides, \( Z_x^\ell(x) = 1 \) and \( Z_x^\ell(E) = \tau_x^{\ell+1} - \tau_x^\ell \).

**Proof.** Let \( C \subset E \). For \( k \geq 1 \) and \( z_0, \ldots, z_{k-1} \geq 0 \), we show that

\[
\mathbb{P}
\left( Z_x^0(C) = z_0, \ldots, Z_x^{k-1}(C) = z_{k-1} \right) = \prod_{\ell=0}^{k-1} \mathbb{P}
\left( \sum_{n=0}^{\tau_x^{\ell}-1} \mathbb{1}_{\{X_n \in C\}} = z_\ell, \tau_x = t_{\ell+1} - t_\ell \right),
\]

which proves the first part of the lemma. To proceed, we use Exercise 3.1.9 repeatedly to get

\[
\mathbb{P}
\left( Z_x^0(C) = z_0, \ldots, Z_x^{k-1}(C) = z_{k-1} \right) = \sum_{0 \leq t_0 < \cdots < t_{k-1}} \mathbb{P}
\left( Z_x^0(C) = z_0, \ldots, Z_x^{k-1}(C) = z_{k-1}; \tau_x^0 = t_0, \ldots, \tau_x^{k-1} = t_{k-1} \right)
\]

\[
= \sum_{0 \leq t_0 < \cdots < t_{k-1}} \mathbb{P}(\tau_x^0 = t_0) \prod_{\ell=0}^{k-1} \mathbb{P}_x
\left( \sum_{n=0}^{\tau_x^{\ell}-1} \mathbb{1}_{\{X_n \in C\}} = z_\ell, \tau_x = t_{\ell+1} - t_\ell \right)
\]

\[
= \sum_{t_0 \geq 0} \mathbb{P}(\tau_x^0 = t_0) \prod_{\ell=0}^{k-1} \sum_{s_\ell \geq 1} \mathbb{P}_x
\left( \sum_{n=0}^{\tau_x^{\ell}-1} \mathbb{1}_{\{X_n \in C\}} = z_\ell, \tau_x = s_\ell \right)
\]

\[
= \prod_{\ell=0}^{k-1} \mathbb{P}_x
\left( \sum_{n=0}^{\tau_x^{\ell}-1} \mathbb{1}_{\{X_n \in C\}} = z_\ell \right),
\]
where we have set \( s_\ell = t_{\ell+1} - t_\ell \) in the penultimate line. This leads to the claimed result. The fact that \( Z_x^t(x) = 1 \) and \( Z_x^t(E) = \tau_x^{t+1} - \tau_x^t \) is immediate. \( \square \)

We may now complete the proof of Lemma 3.3.2. Indeed, for \( n \geq \tau_x^0 \), we denote by \( L_n \) the unique integer such that
\[
\tau_x^{L_n} \leq n < \tau_x^{L_n+1}.
\]
Then, for any \( C \), we have
\[
\sum_{i=0}^{\tau_x^0-1} \mathbb{1}_{\{X_i \in C\}} + \sum_{\ell=0}^{L_n-1} Z_{x}^\ell(C) \leq \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i \in C\}} \leq \sum_{i=0}^{\tau_x^0-1} \mathbb{1}_{\{X_i \in C\}} + \sum_{\ell=0}^{L_n} Z_{x}^\ell(C),
\]
which rewrites
\[
\sum_{\ell=0}^{L_n-1} Z_{x}^\ell(C) \leq \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i \in C\}} \leq \tau_x^0 + \sum_{\ell=0}^{L_n} Z_{x}^\ell(C),
\]
using the elementary bound \( 0 \leq \sum_{i=0}^{\tau_x^0-1} \mathbb{1}_{\{X_i \in C\}} \leq \tau_x^0 \).

As a consequence, for \( n \) large enough,
\[
\sum_{\ell=0}^{L_n-1} Z_{x}^\ell(x) \leq \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i = x\}} \leq \tau_x^0 + \sum_{\ell=0}^{L_n} Z_{x}^\ell(E),
\]
so that we have the following representation formula for \( \pi \):
\[
\forall C \in E, \quad \forall x \in E, \quad \pi(C) = \frac{\mathbb{E}_x \left[ \sum_{i=0}^{\tau_x^0} \mathbb{1}_{\{X_i \in C\}} \right]}{\mathbb{E}_x [\tau_x]}, \quad (3.5)
\]
The latter identity proves in particular that \( \pi \) defined by (3.4) is a probability measure on \( E \), which was far from obvious.

We are now in position to complete the proofs of Proposition 3.2.17 and Theorem 3.3.1.

---

3If \( x \) is null recurrent, for the denominator 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 \( \lim_{M \to +\infty} \mathbb{E}[\min(X_1, M)] = +\infty \).
Proof of Proposition 3.2.17. We first complete the proof of (i). We assume that there is a positive recurrent state \( x \). Then \( \pi(x) > 0 \) and, for any \( y \in E \), the representation formula (3.5) yields

\[
\pi(y) = \pi(x) \mathbb{E}_x \left[ \sum_{n=0}^{\tau_y-1} \mathbb{1}_{\{X_n=y\}} \right].
\]

Assume that \( \mathbb{E}_x [\sum_{n=0}^{\tau_y-1} \mathbb{1}_{\{X_n=y\}}] = 0 \). Then necessarily \( \sum_{n=0}^{\tau_y-1} \mathbb{1}_{\{X_n=y\}} = 0 \), \( \mathbb{P}_x\)-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 \( \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=x\}} \) allows to apply the Dominated Convergence Theorem to the statement of Lemma 3.3.2, and thus deduce that, for any \( x \in E \),

\[
\pi(x) = \mathbb{E} \left[ \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=x\}} \right] = \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{P}(X_i = x) = \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mu_0 P^i(x) = \lim_{n \to +\infty} \hat{\mu}_n(x),
\]

with \( \hat{\mu}_n = \frac{1}{n} \sum_{i=0}^{n-1} \mu_0 P^i \), and where we recall that \( \mu_0 \) is the law of \( X_0 \). If the chain is positive recurrent\(^4\), then by Scheffé’s Lemma (see Remark 2.1.19), for any \( y \in E \),

\[
|\hat{\mu}_n P(y) - \pi P(y)| = \left| \sum_{x \in E} (\hat{\mu}_n(x) - \pi(x)) P(x,y) \right| \leq \sum_{x \in E} |\hat{\mu}_n(x) - \pi(x)| \to 0,
\]

and therefore

\[
\lim_{n \to +\infty} \hat{\mu}_n P(y) = \pi P(y).
\]

But on the other hand,

\[
\hat{\mu}_n P(y) = \frac{1}{n} \sum_{i=0}^{n-1} \mu_0 P^i(y) = \hat{\mu}_n(y) + \frac{1}{n} (\mu_0 P^n(y) - \mu_0(y)) \to \pi(y),
\]

which finally yields \( \pi(y) = \pi P(y) \) and completes the proof of (ii). In the case where \( \pi = 0 \), assuming that \( \mu_0 \) is a stationary distribution yields \( \hat{\mu}_n = \mu_0 \) for any \( n \) and therefore immediately contradicts the fact that \( \hat{\mu}_n(x) \to \pi(x) \), which proves (iii).

Proof of Theorem 3.3.1. If \( E \) is finite then Theorem 3.3.1 is an immediate consequence of Lemma 3.3.2. In the infinite case, if \( f \) is bounded then the conclusion follows from Scheffé’s Lemma, still applied under the form of Remark 2.1.19. In the general case where \( f \in L^1(\pi) \), 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 3.3.2, and using the strong Law of Large Numbers again. The details are left to the reader. \( \square \)

\(^4\)The argument here is the same as in the second proof of Proposition 3.2.13, extended to the possibly infinite state space case thanks to Scheffé’s Lemma.
3.3.2 Central Limit Theorem

From now on we assume that the chain \((X_n)_{n \geq 0}\) is irreducible and positive recurrent, and we denote by \(\pi\) its stationary distribution. For \(f \in L^1(\pi)\), our aim is now to study the convergence in distribution of

\[
\sqrt{n} \left( \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \pi f \right).
\]

Letting \(\tilde{f}(x) = f(x) - \pi f\), assuming that \(X_0 = x\) for simplicity, and using the notation of the proof of Lemma 3.3.2, we observe that when \(n = \tau_L x\), this quantity rewrites

\[
\frac{1}{\sqrt{\tau_L x}} \sum_{i=0}^{\tau_L - 1} \tilde{f}(X_i) = \frac{1}{\sqrt{\tau_L x}} \sum_{\ell=0}^{L-1} Z^\ell_x(\tilde{f}), \quad Z^\ell_x(\tilde{f}) = \sum_{i=\tau^\ell_x}^{\tau^{\ell+1}_x} \tilde{f}(X_i) = \sum_{y \in E} Z^\ell_x(y) \tilde{f}(y).
\]

Since the variables \(Z^\ell_x(\tilde{f})\) are iid, as soon as they have a finite second order moment, namely

\[
v(x, f) = \mathbb{E}_x \left[ \left( \sum_{n=0}^{\tau_x-1} (f(X_n) - \pi f) \right)^2 \right] < +\infty, \tag{3.6}
\]

the CLT and Slutsky’s Lemma yield

\[
\frac{1}{\sqrt{\tau_L x}} \sum_{\ell=0}^{L-1} Z^\ell_x(\tilde{f}) = \frac{1}{\sqrt{\tau_L x}} \frac{1}{\sqrt{L}} \sum_{\ell=0}^{L-1} Z^\ell_x(\tilde{f}) \rightarrow \frac{1}{\sqrt{\mathbb{E}[\tau_x]}} \xi, \quad \xi \sim \mathcal{N}(0, v(x, f)).
\]

This argument leads to the following statement, the proof of which relies on technical details which are omitted.

**Theorem 3.3.4 (CLT for Markov chains).** If the Markov chain \((X_n)_{n \geq 0}\) is irreducible and positive recurrent, then for any \(f\) such that (3.6) holds, we have:

(i) the quantity \(\sigma^2(f) = \pi(x)v(x, f)\) does not depend on \(x \in E\);

(ii) it holds

\[
\lim_{n \to +\infty} \sqrt{n} \left( \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \pi f \right) = \mathcal{N}(0, \sigma^2(f)), \quad \text{in distribution}.
\]

An important remark regarding the assumptions of Theorem 3.3.4 is that if \(E\) is finite, then by Exercise 3.2.20, the condition (3.6) is satisfied for any function \(f\) and therefore the CLT always holds.

While rather natural given the sketch of the proof detailed above, the expression for the asymptotic variance \(\sigma^2(f)\) is not very handful. We shall see in the next section another, more explicit, expression.

3.4 Convergence to equilibrium

In this section, we study the long time behaviour of Markov chains with a slightly different point of view from the previous section. Indeed, we describe the convergence in distribution of \(X_n\).

**Exercise 3.4.1.** Show that if \((X_n)_{n \geq 0}\) is an irreducible Markov chain such that \(X_n \to \pi\) in distribution, then \(\pi\) is necessarily a stationary distribution for \((X_n)_{n \geq 0}\).
3.4.1 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 \(\pi\) such that \(\pi(-1) = \pi(1) = 1/2\).

Exercise 3.4.2. Without using the results of Section 3.3, check directly that the chain \((X_n)_{n \geq 0}\) satisfies the Law of Large Numbers of Theorem 3.3.1. What do you think of the Central Limit Theorem?

Yet, for any initial distribution \(\mu_0\) on \(E\) which is not \(\pi\), the law \(\mu_n\) of \(X_n\) under \(P_{\mu_0}\) does not converge to \(\pi\). 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
\[
\gcd \mathcal{N} = \max\{k \geq 1 : \forall n \in \mathcal{N}, \exists \ell \in \mathbb{N} : n = k\ell\}.
\]

Definition 3.4.3 (Period). Let \((X_n)_{n \geq 0}\) be a Markov chain with transition matrix \(P\). For all \(x \in E\), the period of \(x\) is \(\gcd\{n \geq 1 : P^n(x, x) > 0\}\).

Exercise 3.4.4. Compute the period of the states \(-1\) and \(1\) in the example of the two-point space described above.

Lemma 3.4.5 (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.

Proof. For all \(x \in E\), we denote \(\mathcal{N}(x) = \{n \geq 1 : P^n(x, x) > 0\}\). Let \(x, y \in E\). By Definition 3.2.5, 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 \(m \in \mathcal{N}(x) \cap \mathcal{N}(y)\) and \(\mathcal{N}(x) \subset \mathcal{N}(y) - m, \mathcal{N}(x) + m \subset \mathcal{N}(y)\), so that \(\gcd \mathcal{N}(y)\) divides all elements of \(\mathcal{N}(x)\) and thus \(\gcd \mathcal{N}(y) \leq \gcd \mathcal{N}(x)\). By the same arguments, \(\gcd \mathcal{N}(x) \leq \gcd \mathcal{N}(y)\) and the proof is completed.

3.4.2 * Convergence to equilibrium

Once the pathological periodic case has been removed, it becomes true that \(X_n\) converges to its stationary distribution.

Theorem 3.4.6 (Convergence to equilibrium). Let \((X_n)_{n \geq 0}\) be an irreducible and positive recurrent Markov chain, with unique stationary distribution \(\pi\). If the chain is aperiodic, then
\[
\lim_{n \to +\infty} X_n = \pi, \quad \text{in distribution.}
\]

The proof of Theorem 3.4.6 relies on the following lemma.

Lemma 3.4.7 (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 \( \pi \) is a stationary probability for \((X_n)_{n\geq 0}\) and \( \psi \) 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 3.4.8 below, for \(k\) large enough, \(P^{q+kr-p}(x, x) > 0\) and therefore, with \(n = q + kr\), we have
\[
\mathbb{P}_{(x,y)}(X_n = x', Y_n = y') = P^n(x, x')Q^n(y, y')
\geq P^{q+kr-p}(x, x)P^p(x, x')Q^q(y, y')Q^{kr}(y', y') > 0,
\]
which proves irreducibility. \(\square\)

**Lemma 3.4.8** (Schur’s Theorem). Let \(N' \subset \mathbb{N}\) be closed under addition and such that \(\gcd(N') = 1\). Then the set \(\mathbb{N} \setminus N'\) is finite.

Lemma 3.4.8 is purely number theoretic. We refer to [6, Proposition 1.7] for details. We are now ready to complete the proof of Theorem 3.4.6.

**Proof of Theorem 3.4.6.** Let \((X_n)_{n\geq 0}\) be an irreducible, positive recurrent and aperiodic Markov chain with initial distribution \(\mu_0\), transition matrix \(P\) and unique stationary distribution \(\pi\). Let \((Y_n)_{n\geq 0}\) be a Markov chain with transition matrix \(P\) and initial distribution \(\pi\), 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 \(\pi\). 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 3.4.7, 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 Exercise 3.1.9, we then have
\[
\mathbb{P}(X_n = y, \tau_x \leq n) = \sum_{k=0}^{n} \mathbb{P}(X_n = y, \tau_x = k)
= \sum_{k=0}^{n} \mathbb{P}(\tau_x = k)\mathbb{P}_x(X_{n-k} = y)
= \sum_{k=0}^{n} \mathbb{P}(\tau_x = k)\mathbb{P}_x(Y_{n-k} = y)
= \mathbb{P}(Y_n = y, \tau_x \leq n),
\]
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 \(P_x(X_{n-k} = y) = P_x(Y_{n-k} = y)\). We deduce that
\[
P(X_n = y) \leq P(Y_n = y, \tau_x \leq n) + P(\tau_x > n) \leq P(Y_n = y) + P(\tau_x > n),
\]
and then by symmetry
\[
|P(X_n = y) - P(Y_n = y)| \leq P(\tau_x > n).
\]
Since \(\tau_x\) is finite, the right-hand side converges to 0 when \(n \to +\infty\), which implies that
\[
\forall y \in E, \quad \lim_{n \to +\infty} P(X_n = y) = \pi(y),
\]
and the conclusion follows from Scheffé’s Lemma (see Remark 2.1.19).

In the next two subsections, we use spectral arguments to quantify the speed at which \(X_n\) converges to \(\pi\).

### 3.4.3 Spectrum of stochastic matrices and Perron–Frobenius Theorem in the finite state space case

In this subsection, we assume that \(E\) is finite, with cardinality \(m\). If \((X_n)_{n \geq 0}\) is irreducible (and therefore necessarily positive recurrent) and aperiodic, then by Theorem 3.4.6, \(X_n\) converges in distribution to \(\pi\) and our purpose is to show that the rate of convergence is geometric (that is to say, bounded from above by some quantity which decreases geometrically with \(n\)).

In fact, we shall not use Theorem 3.4.6, and therefore our main statement, Theorem 3.4.12 below, provides an independent proof, restricted to the finite state space case, of this convergence result, but strengthened with a quantitative estimate.

To proceed, we denote by \(M(E)\) the space of signed measures on \(E\), which is identified with the space of row vectors in \(\mathbb{R}^m\), and let \(\| \cdot \|\) be a norm on \(M(E)\). Thus, our aim is to derive an upper bound on \(\|\mu_0 P^n - \pi\|\). To study this quantity, it is natural to expect the spectrum of \(P\) to play a role.

**Exercise 3.4.9.** Let \(P\) be a stochastic matrix. Show that for any (complex) eigenvalue \(\lambda\) of \(P\), \(|\lambda| \leq 1\).

From now on we assume that \(P\) is irreducible. Then 1 is a single eigenvalue, with associated eigenvectors 1 (to the right) and \(\pi\) (to the left). Let us denote by \(M_0(E)\) the subset of \(M(E)\) defined by
\[
M_0(E) = \{ \rho \in M(E) : \rho 1 = 0 \}.
\]
Notice that for any \(\mu, \pi \in P(E), \mu - \pi \in M_0(E)\).

**Exercise 3.4.10.**
1. Show that the linear application \(\rho \mapsto \rho P\) maps \(M_0(E)\) to itself. We denote it by \(P_0\).
2. Show that \(\lambda\) is an eigenvalue of \(P_0\) if and only if it is an eigenvalue of \(P\) and \(\lambda \neq 1\).

The spectrum of \(P_0\) is then described by the Perron–Frobenius Theorem.

**Proposition 3.4.11** (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 \(\lambda\) of \(P\) such that \(|\lambda| = 1\) are the \(k\)-th roots of unity, and they are all simple.
We refer to [3, Theorem 3.11] for the proof. Define
\[ \lambda_* := \max\{|\lambda|, \lambda \neq 1 \text{ is an eigenvalue of } P\} = \max\{|\lambda|, \lambda \text{ is an eigenvalue of } P_0\}, \]
so that if \( P \) is aperiodic, then \( \lambda_* < 1 \). We shall call the quantity \( 1 - \lambda_* \) the spectral gap of \( P \). For any \( \mu_0 \in \mathcal{P}(E) \),
\[ \mu_0 P^n - \pi = (\mu_0 - \pi)P^n = (\mu_0 - \pi)P_0^n, \]
since \( \mu - \pi \in \mathcal{M}_0(E) \). This yields the following geometric convergence result.

**Theorem 3.4.12** (Geometric convergence in finite state spaces). Let \( P \) be an irreducible and aperiodic stochastic matrix with stationary distribution \( \pi \) and spectral gap \( 1 - \lambda_* \in (0, 1] \). For any \( \alpha \in (\lambda_*, 1] \), there exists a constant \( C_\alpha \) such that, for all \( n \geq 0 \),
\[
\sup_{\mu_0 \in \mathcal{P}(E)} \|\mu_0 P^n - \pi\| \leq C_\alpha \alpha^n.
\]

**Proof.** By the Dunford Theorem, there is a basis of the \( m - 1 \)-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 \( |\lambda| \leq \lambda_* < 1 \) and \( N \) is a nilpotent matrix, and such that \( DN = ND \). Thus, for all \( n \geq 1 \),
\[
P_0^n = \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k,
\]
and since \( N^{m-1} = 0 \) we get that as soon as \( n \geq m - 1 \),
\[
P_0^n = \sum_{k=0}^{m-1} \binom{n}{k} D^{n-k} N^k.
\]
Since the binomial coefficient \( \binom{n}{k} \) is equivalent to \( n^k/k! \) when \( n \to +\infty \) and the diagonal coefficients \( \lambda^{n-k} \) of \( D^{n-k} \) satisfy \( |\lambda|^{n-k} \leq \lambda_*^{n-k} \), we deduce that as soon as \( \alpha > \lambda_* \),
\[
\lim_{n \to +\infty} \alpha^{-n} P_0^n = 0.
\]
Let \( ||| \cdot ||| \) be the norm on the space of linear mappings from \( \mathcal{M}_0(E) \) to itself defined by
\[
|||Q||| = \sup \left\{ \|\rho Q\|, \sum_{x \in E} |\rho(x)| \leq 1 \right\}.
\]
We let
\[
C_\alpha = \frac{1}{2} \sup_{n \geq 0} |||\alpha^{-n} P_0^n||| < +\infty,
\]
so that for any \( \mu_0 \in \mathcal{P}(E) \),
\[
\|\mu_0 P^n - \pi\| = \|(\mu_0 - \pi)P_0^n\| \leq |||P_0^n||| \sum_{x \in E} |\mu_0(x) - \pi(x)| \leq C_\alpha \alpha^n,
\]
which completes the proof. \( \square \)

Owing to the fact that the binomial terms in the Dunford decomposition of \( P_0 \) grow polynomially in \( n \), one cannot take \( \alpha = \lambda_* \) 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 the next subsection.
3.4.4 Reversibility

In this subsection, 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 3.4.12.

Definition and general remarks

In this paragraph, $E$ can be either finite or countably infinite.

**Definition 3.4.13 (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 $x, y \in E$,

$$\pi(x)P(x, y) = \pi(y)P(y, x).$$ (3.7)

Equation (3.7) is called the detailed balance equation. The denomination ‘reversibility’ is explained by the following result.

**Proposition 3.4.14 (Reversibility).** Let $(X_n)_{n \geq 0}$ be a Markov chain with transition matrix $P$, reversible with respect to $\pi$. For any $n \geq 0$, the vectors $(X_0, \ldots, X_n)$ and $(X_n, \ldots, X_0)$ have the same distribution under $P_\pi$.

**Proof.** For any $x_0, \ldots, x_n \in E$, we deduce from Proposition 3.1.6 that

$$P_\pi(X_0 = x_0, \ldots, X_n = x_n) = \pi(x_0)P(x_0, x_1)\cdots P(x_{n-1}, x_n).$$

Applying Definition 3.4.13 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 $P_\pi(X_0 = x_n, \ldots, X_n = x_0)$ by Proposition 3.1.6 again. $\Box$

Looking only at the marginal distribution of the first coordinate of the vectors $(X_0, \ldots, X_n)$ and $(X_n, \ldots, X_0)$, we deduce the following link between the notions of reversibility and stationary distribution.

**Corollary 3.4.15 (Reversibility and stationary distribution).** If $(X_n)_{n \geq 0}$ is reversible with respect to $\pi$, then $\pi$ 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 3.4.16.** Show that for the Ehrenfest urn, both the microscopic and the macroscopic descriptions are reversible with respect to the stationary distributions from Exercise 3.2.4.

**Exercise 3.4.17.** Find an elementary example of a Markov chain which is not reversible.
Spectral characterisation of reversibility

From now on, we assume that \( E \) is finite, with cardinality \( m \).

Fix \( \pi \in \mathcal{P}(E) \). We denote by \( \ell^2(\pi) \) the set \( \mathbb{R}^E \) endowed 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 \( \ell^2(\pi) \) if \( \langle Af, g \rangle_{\pi} = \langle f, Ag \rangle_{\pi} \), for all \( f, g \in \mathbb{R}^E \).

**Remark 3.4.18.** If \( \pi \) 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. The associated Euclidean norm is denoted by \( \| \cdot \|_{\pi} \).

**Proposition 3.4.19** (Spectral characterisation of reversibility). Let \( \pi \in \mathcal{P}(E) \). A Markov chain with transition matrix \( P \) is reversible with respect to \( \pi \) if and only if \( P \) is symmetric in \( \ell^2(\pi) \).

**Proof.** Assume that a Markov chain with transition matrix \( P \) is reversible with respect to \( \pi \). Then for all \( f, g \in \ell^2(\pi) \),

\[
\langle Pf, g \rangle_{\pi} = \sum_{x \in E} \left( \sum_{y \in E} P(x, y)f(y) \right) g(x)\pi(x) = \sum_{x, y \in E} f(y)g(x)\pi(x)P(x, y) = \sum_{x, y \in E} f(y)\left( \sum_{x \in E} P(y, x)g(x) \right) \pi(y)
\]

Conversely, assume that \( P \) is symmetric in \( \ell^2(\pi) \), fix \( x, y \in E \) and take \( f(z) = \mathbb{1}_{\{z = x\}} \), \( g(z) = \mathbb{1}_{\{z = y\}} \). Then

\[
\langle Pf, g \rangle_{\pi} = P(x, y)\pi(y), \quad \langle f, Pg \rangle_{\pi} = P(y, x)\pi(x),
\]

so that Equation (3.7) is satisfied. \( \square \)

The spectral study of reversible chains is facilitated by the following result, which is an immediate consequence of the Spectral Theorem.

**Proposition 3.4.20** (Eigenvalues of reversible chains). Let \( P \) be the transition matrix of an irreducible chain which is reversible with respect to its invariant measure \( \pi \). The eigenvalues of \( P \) are real and can be labelled \( 1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_m \geq -1 \), and there exists an orthonormal basis \( (f_1, \ldots, f_m) \) of \( \ell^2(\pi) \) such that \( Pf_i = \lambda_i f_i \) for all \( i \).

**Remark 3.4.21.** If the chain is aperiodic, then by Proposition 3.4.11, \( \lambda_m > -1 \). As a consequence, \( \lambda_* = \max\{ |\lambda_2|, |\lambda_m| \} < 1 \).
Long time behaviour of reversible Markov chains

From Proposition 3.4.20, we deduce another statement for the geometric convergence of $P^n f$ to $\pi f$.

**Proposition 3.4.22** (Rate of convergence for reversible chains). *Under the assumptions of Proposition 3.4.20, for all $f \in \ell^2(\pi)$, for all $n \geq 0$,

$$\|P^n f - \pi f\|_\pi^2 \leq \lambda_*^{2n} \|f - \pi f\|_\pi^2.$$*

Notice that this result only provides the geometric convergence of $P^n f$ to $\pi f$ if $\lambda_* < 1$, that is to say if the chain is aperiodic (see Remark 3.4.21), which is of course in accordance with Theorem 3.4.12.

*Proof.* Notice that in Proposition 3.4.20 we may take $f_1 = 1$, in which case $\langle f, f_1 \rangle_\pi = \pi f$, so that writing the orthogonal decomposition

$$P^n f = \sum_{i=1}^m \langle P^n f, f_i \rangle_\pi f_i = \sum_{i=1}^m \langle f, P^n f_i \rangle_\pi f_i = \sum_{i=1}^m \lambda_i^n \langle f, f_i \rangle_\pi f_i$$

yields

$$P^n f - \pi f = \sum_{i=2}^m \lambda_i^n \langle f, f_i \rangle_\pi f_i.$$ As a consequence,

$$\|P^n f - \pi f\|_\pi^2 = \sum_{i=2}^m (\lambda_i^n \langle f, f_i \rangle_\pi)^2 \leq \lambda_*^{2n} \sum_{i=2}^m \langle f, f_i \rangle_\pi^2 = \lambda_*^{2n} \|f - \pi f\|_\pi^2.$$  

[\square]

\textbf{Exercise 3.4.23.} The chi-square distance on $\mathcal{P}(E)$ is defined by

$$\chi_2(\mu|\pi) = \begin{cases} \sum_{x \in E} \left( \frac{\mu(x)}{\pi(x)} - 1 \right)^2 \pi(x) & \text{if } \mu \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that it is not a distance, because it is not symmetric in $\mu$ and $\pi$. Show that under the assumptions of Proposition 3.4.20, for any initial distribution $\mu \in \mathcal{P}(E)$,

$$\chi_2(\mu P^n|\pi) \leq \lambda_*^{2n} \chi_2(\mu|\pi).$$

3.4.5 * Mixing times of finite state space Markov chains

3.4.6 * Back to the variance of the CLT
Chapter 4

The Markov chain Monte Carlo method

Contents

4.1 Gibbs measures ................................. 67
4.2 The Metropolis algorithm .................... 69
4.3 * The Gibbs sampler ......................... 73
4.4 * More stochastic algorithms ............... 74

Let \( \pi \) 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 \( \pi \). 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 \( \pi \) is a stationary distribution, and using either the Law of Large Numbers and Central Limit Theorem from Section 3.3 to compute an estimator and a confidence interval for \( \mathcal{I} \), or the convergence theorem from Section 3.4 to sample independent random variables \( \tilde{X}_1, \tilde{X}_2, \ldots \) which are approximately distributed according to \( \pi \) 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.

4.1 Gibbs measures

4.1.1 Definition and notation

A Gibbs measure is a probability measure \( \pi_\beta \) on \( E \) which writes under the form

\[
\pi_\beta(x) = \frac{1}{Z_\beta} e^{-\beta V(x)},
\]

where \( \beta > 0 \) is the inverse temperature parameter, \( V : E \to \mathbb{R} \cup \{+\infty\} \) is called the potential and

\[
Z_\beta = \sum_{x \in E} e^{-\beta V(x)}
\]
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 $\pi$ 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 $Z_\beta$ is not easy to compute.

These assumptions make the enumeration procedure discussed in the introduction impossible to implement.

4.1.2 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 $\pi_\beta$ defined on $E$ by

$$\pi_\beta(x) = \frac{e^{-\beta V(x)}}{Z_\beta}, \quad Z_\beta = \sum_{x \in E} e^{-\beta V(x)}.$$

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 $Z_\beta$.

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 4.1.1 (High- and low-temperature limits).\(^2\)

1. Show that $\pi_0$ is the uniform measure on $E$, under which $(x_v)_{v \in \mathcal{V}}$ is iid with $\pi_0(x_v = 1) = \pi_0(x_v = -1) = 1/2$.
2. Show that when $\beta \to +\infty$, $\pi_\beta$ 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 $\pi_0$ when $|\mathcal{V}| \to +\infty$.
4. Describe the law of $m$ under $\pi_\infty$.

The system’s macroscopic behaviour, measures 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 4.1 below. In the $|\mathcal{V}| \to +\infty$ limit, called the thermodynamic limit, this phase transition occurs abruptly (with

\(^1\)See the introduction of https://cel.archives-ouvertes.fr/cel-00392289/ for a detailed presentation.

\(^2\)
4.2 The Metropolis algorithm

The Metropolis algorithm provides a method to construct a Markov chain which is reversible with respect to a given probability measure \( \pi \) 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 \( E \).

4.2.1 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 \( F(\rho) = \rho F(1/\rho) \) for all \( \rho > 0 \).

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 \( Q(x, y) \),
Figure 4.1: 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 $p_{\beta} > p_{\beta_c} \simeq 0.59$. 
4.2 The Metropolis algorithm

Figure 4.2: Value of the magnetisation \( m_*(\beta) \) as a function of \( p_\beta \), for the Ising model in dimension 2.

(ii) move the chain to \( y \) with probability

\[
a(x, y) := F(r(x, y)), \quad r(x, y) := \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)},
\]

otherwise remain at \( x \).

The condition on \( Q \) ensures that, almost surely, the ratio \( r(x, y) \) takes its values in \((0, +\infty)\).

**Proposition 4.2.1** (Reversibility of the Metropolis chain). The Metropolis chain \((X_n)_{n \geq 0}\) is irreducible and reversible with respect to \( \pi \).

As a consequence, \( \pi \) is the unique stationary distribution of \((X_n)_{n \geq 0}\) and all convergence results from Chapter 3 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

\[
\pi(x)P(x, y) = \pi(x)Q(x, y)F\left(\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}\right) \\
= \pi(x)Q(x, y)\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}F\left(\frac{\pi(x)Q(x, y)}{\pi(y)Q(y, x)}\right) \\
= \pi(y)Q(y, x)a(y, x) \\
= \pi(y)P(y, x),
\]

The condition on \( Q \) ensures that, almost surely, the ratio \( r(x, y) \) takes its values in \((0, +\infty)\).
which ensures that the detailed balance equation holds.

Exercise 4.2.2. Show that if \( Q \) is aperiodic, then \( P \) is also aperiodic.

Remark 4.2.3. If \( Q \) is already reversible with respect to \( \pi \), 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 4.2.2 does not hold.

If \( \pi \) has the form a Gibbs measure \( \pi_\beta \) as is described in Section 4.1, then simulating the Metropolis chain \( (X_n)_{n \geq 0} \) requires to compute the ratio

\[
r(x, y) = e^{-\beta(V(y) - V(x))} \frac{Q(y, x)}{Q(x, y)},
\]

which does not depend on the partition function \( Z_\beta \) but only on the potential \( V \). Therefore, the chain \( (X_n)_{n \geq 0} \) can be simulated in practice.

4.2.2 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 \( V \) and flips its spin. Defining, for any \( x \in E \) and \( u \in V \), the configuration \( x^u \) by

\[
\forall v \in V, \quad x^u_v = \begin{cases} x_v & \text{if } v \neq u, \\ -x_v & \text{if } v = u, \end{cases}
\]

the matrix \( Q \) writes

\[
Q(x, y) = \begin{cases} \frac{1}{|V|} & \text{if there exists } u \in V \text{ such that } y = x^u, \\ 0 & \text{otherwise}. \end{cases}
\]

It satisfies the assumptions of Subsection 4.2.1. 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

\[
V(x^u) - V(x) = -\sum_{\{v, w\} \in E} x^u_v x^u_w + \sum_{\{v, w\} \in E} x_v x_w \\
= -\sum_{v \in V - \{u, v\}} x^u_v x^u_v + \sum_{v \in V - \{u, v\}} x_v x_v \\
= 2x_u \Sigma(x, u),
\]

with

\[
\Sigma(x, u) = \sum_{v \in V - \{u, v\}} x_v.
\]

Thus, under the Metropolis–Hastings rule, the spin \( x_u \) is changed to \(-x_u\) with probability

\[
\min \{1, \exp(2\beta x_u \Sigma(x, u))\};
\]

under the Barker rule, the spin \( x_u \) is changed to \(-x_u\) with probability

\[
\frac{1}{1 + \exp(-2\beta x_u \Sigma(x, u))}.
\]
4.3 * The Gibbs sampler

The Gibbs sampler algorithm, sometimes called Glauber dynamics in Ising-like models, is a MCMC method which provides an alternative to the Metropolis algorithm. It is designed for probability measures $\pi$ on state spaces $E$ which have the specific form $E = S^V$, where $S$ and $V$ are finite spaces.

4.3.1 Definition and properties

By analogy with the Ising model, we shall keep calling elements $u$ of $V$ vertices, and denoting configurations $x \in E$ by $x = (x_u)_{u \in V}$, where $x_u \in S$ is the spin of the vertex $u$. Given a configuration $x \in E$, a vertex $u \in V$ and a possible value $s$ for the spin, we denote by $x^{u,s}$ the configuration defined by

$$\forall v \in V, \quad x^{u,s}_v = \begin{cases} s & \text{if } v = u, \\ x_v & \text{otherwise}, \end{cases}$$

and let

$$E_{x,u} = \{x^{u,s} : s \in S\}$$

the set of configurations which can be obtained by changing the value of the spin $x_u$.

**Definition 4.3.1** (Gibbs sampler). Let $\pi$ be a probability measure on $E = S^V$, such that $\pi(x) > 0$ for all $x \in E$. The Gibbs sampler of $\pi$ is the Markov chain in $E$ defined by, at each step:

(i) picking a vertex $u \in V$ 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 4.1. For all $s \in S$, the spin $x_u$ is updated to the value $s$ with probability

$$\pi_\beta(x^{u,s} | E_{x,u}) = \frac{\pi_\beta(x^{u,s})}{\sum_{s' \in S} \pi_\beta(x^{u,s'})} = \frac{e^{-\beta V(x^{u,s})}}{Z_\beta}$$

where $Z_\beta = \sum_{s' \in S} e^{-\beta V(x^{u,s'})}$.

This identity shows that it is not necessary to know the value of the partition function $Z_\beta$ to compute the conditional probability $\pi_\beta(\cdot | E_{x,u})$; instead, only $|S|$ evaluations of the potential $V$ are used.

**Remark 4.3.2** (Graph structure for $V$). This evaluation is particularly fast when the potential $V$ depends on a geometrical structure of the set $V$. Assume indeed that, as in the Ising model, the latter is the set of vertices of an undirected graph with set of edges $E$, and that the potential writes under the form

$$V(x) = \sum_{(u,v) \in E} w(x_u, x_v),$$

for some symmetric function $w : S \times S \to \mathbb{R}$. Then for all $x \in E$, $u \in V$, and $s, s' \in S$,

$$V(x^{u,s}) - V(x^{u,s'}) = 2 \sum_{v: \{u,v\} \in E} w(s, x_v) - w(s', x_v),$$
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 4.3.3 (Reversibility).** Under the assumptions of Definition 4.3.1, the Gibbs sampler of \( \pi \) is irreducible, aperiodic and reversible with respect to \( \pi \).

\[ \text{Exercise 4.3.4. Prove the irreducibility and aperiodicity properties.} \]

\[ \text{Proof of reversibility. From Definition 4.3.1, we deduce that the transition matrix } P \text{ of the Gibbs sampler writes, for all } x, y \in E \text{ with } x \neq y, \]

\[ P(x, y) = \begin{cases} \frac{1}{|V|} \pi(y|E_{x,u}) & \text{if there exists } u \in V \text{ such that } y \in E_{x,u}, \\ 0 & \text{otherwise.} \end{cases} \]

Clearly, for all \( x, y \in E \) and \( u \in V \), 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,

\[ \pi(x)P(x, y) = \pi(x)\frac{1}{|V|}\pi(y|E_{x,u}) = \frac{\pi(x)\pi(y)}{|V|\pi(E_{x,u})} = \frac{\pi(x)\pi(y)}{|V|\pi(E_{y,u})} = \pi(y)P(y, x), \]

which is the detailed balance condition needed to prove reversibility. If there is no \( u \in V \) 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. \( \square \)

### 4.3.2 Application to the Ising model

When the chain is in a configuration \( x \), a vertex \( u \) is picked uniformly in \( V \) and the spin is updated according to the probabilities

\[ p(+)|x, u := \frac{e^{-\beta V(x^u, +)}}{e^{-\beta V(x^u, +)} + e^{-\beta V(x^u, -)}}, \quad p(-)|x, u := \frac{e^{-\beta V(x^u, -)}}{e^{-\beta V(x^u, +)} + e^{-\beta V(x^u, -)}}. \]

With the notation \( \Sigma(x, u) \) introduced in Subsection 4.2.2, these probabilities rewrite

\[ p(+)|x, u := \frac{e^{\beta \Sigma(x, u)}}{e^{\beta \Sigma(x, u)} + e^{-\beta \Sigma(x, u)}}, \quad p(-)|x, u := \frac{e^{-\beta \Sigma(x, u)}}{e^{\beta \Sigma(x, u)} + e^{-\beta \Sigma(x, u)}}, \]

\[ \text{Exercise 4.3.5. Check that, on the example of the Ising model, this algorithm coincides with the Metropolis algorithm applied with the Barker rule. Does this statement seem to be a general fact?} \]

### 4.4 * More stochastic algorithms

The idea of the Metropolis algorithm can be adapted to provide numerical methods for problems that are related with MCMC.
4.4 * More stochastic algorithms

4.4.1 Simulated annealing for optimisation

Consider the Gibbs measure
\[ \pi_\beta(x) = \frac{1}{Z_\beta} e^{-\beta V(x)}, \]
for some function \( V : E \to \mathbb{R} \). When \( \beta \to +\infty \), \( \pi_\beta \) converges to the uniform distribution on the set
\[ \text{argmin } V := \{ x \in E : \forall y \in E, V(y) \geq V(x) \}. \]

If one is interested in finding the global minima of \( V \), then a first approach may consist in taking a ‘large’ value of \( \beta \), constructing a Metropolis chain \( (X_n)_{n \geq 0} \) reversible with respect to the Gibbs measure \( \pi \) 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 \( \beta \) 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 \( \min_{x \in E} V(x) \).

We refer to [2, Chapitre 2], [1, Chapitre 5.3] for details.

4.4.2 Subset simulation for rare events

We come back to the issue, already discussed in Section 2.4, of estimating a small probability \( p \) of the form \( \mathbb{P}(f(X) \geq M) \), where \( X \) is a random variable in \( E \) with distribution \( \mu \), \( f : E \to [0, +\infty) \) and \( M > 0 \). For the sake of simplicity, we remain here in the case where \( E \) is finite, but with a large cardinality \( m \) which prevents any enumerative procedure from being implemented.

In contrast with the remainder of this section, here we assume that we are able to generate iid samples \( X_1, X_2, \ldots \) from \( \mu \). However, the Monte Carlo estimator
\[ \hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} 1_{\{f(X_i) \geq M\}} \]
The Markov chain Monte Carlo method requires too many samples to be accurate — for the sequel of the argument, we recall that it takes at least $1/p$ samples to estimate $p$ accurately.

The idea of subset simulation\textsuperscript{2}, also called multilevel splitting\textsuperscript{3}, consists in introducing $L \geq 1$ levels $0 = M_0 < M_1 < \cdots < M_L = M$, writing

$$\mathbb{P}(f(X) \geq M) = \prod_{\ell=1}^{L} \mathbb{P}(f(X) \geq M_\ell | f(X) \geq M_{\ell-1}),$$

and estimating each term of this product separately.

Assume that the levels are chosen in such a way that all terms $\mathbb{P}(f(X) \geq M_\ell | f(X) \geq M_{\ell-1})$ have approximately the same probability $p^{1/L}$. If one was able to sample iid realisations of $X$ under the conditional probability $\mathbb{P}(\cdot | f(X) \geq M_{\ell-1})$, then it would take $1/p^{1/L}$ such samples to estimate accurately each term $\mathbb{P}(f(X) \geq M_\ell | f(X) \geq M_{\ell-1})$, which would reduce the global complexity of the estimation of $p$ from $1/p$ for the Monte Carlo estimator to $L/p^{1/L}$ for the subset simulation estimator.

**Exercise 4.4.1.** For a fixed value of $p$, what is the value $L^*$ which minimises the complexity $L/p^{1/L}$? Deduce that the resulting optimal complexity is of order of magnitude $-\ln p$.

In practice however, it is not always clear how to sample $X$ under the conditional distribution $\mathbb{P}(\cdot | f(X) \geq M_{\ell-1})$. This is where the Metropolis algorithm intervenes. Let

$$E_\ell := \{ x \in E : f(x) \geq M_{\ell-1} \},$$

and define, for all $x \in E_\ell$,

$$\pi_\ell(x) = \mu(x|E_\ell) = \frac{\mu(x)}{\mu(E_\ell)}.$$

The Metropolis algorithm allows to construct a Markov chain in $E_\ell$ with stationary distribution $\pi_\ell$ without having to compute the quantity $\mu(E_\ell)$, which plays the role of the normalising constant $Z_\beta$ here. This chain may then be used to construct an estimator of $\mathbb{P}(f(X) \geq M_\ell | f(X) \geq M_{\ell-1})$.


Part III

Diffusion processes and partial differential equations
Bibliography


