Master de Mathématiques et Applications

Spécialité Mathématiques de la Modélisation

Méthodes Numériques Probabilistes

Tony Lelièvre et Julien Reygner

Version du 24 octobre 2022
## Contents

I Random number simulation and the Monte Carlo method

1 Random variables and their numerical simulation
   1.1 General definitions and results ........................................ 3
   1.2 Random number simulation .............................................. 9
   1.3 Random vector simulation .............................................. 17

2 Limit theorems and the Monte Carlo method
   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

II Markov chains and MCMC methods ........................................ 43

III Diffusion processes and partial differential equations ............ 45

Bibliography ................................................................. 47
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 [1, 2] 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 $(\Omega, \mathcal{A}, \mathbb{P})$ such that:

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

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

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

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

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

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

\(^2\)That is to say, a mapping $\mathbb{P} : A \rightarrow [0, 1]$ such that $\mathbb{P}(\Omega) = 1$ and, for any countable family $(A_n)_{n \geq 1}$ of pairwise distinct elements of $\mathcal{A}$, $\mathbb{P}(\cup_{n \geq 1} A_n) = \sum_{n \geq 1} \mathbb{P}(A_n)$. The latter property is called $\sigma$-additivity.
We shall sometimes refer to the inequality
\[ P(A \cup B) \leq P(A) + P(B) \]
as the union bound.

The following property relies on the \( \sigma \)-additivity property of the measure \( P \), 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 \( B_{n+1} \subset B_n \) for any \( n \), then
  \[ \lim_{n \to +\infty} P(B_n) = P(\cap_{n \geq 1} B_n). \]
- If \( (B_n)_{n \geq 1} \) is nondecreasing, that is to say \( B_n \subset B_{n+1} \) for any \( n \), then
  \[ \lim_{n \to +\infty} P(B_n) = P(\cup_{n \geq 1} B_n). \]

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

1.1.2 Conditional probability

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

Notice that \((\Omega, A, P(\cdot | B))\) 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 \( A \),
\[ P(A) = \sum_{n \geq 1} P(A \cap B_n) = \sum_{n \geq 1} P(A | B_n) P(B_n), \]
with the obvious convention that \( P(A | B_n) P(B_n) = 0 \) if \( P(B_n) = 0 \).
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 \(X : \Omega \to E\), 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

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

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 \(p : E \to [0, +\infty)\) 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 \(\mu : \mathcal{E} \to [0, +\infty]\) such that there exists a countable family \((C_n)_{n \geq 1}\) of elements of \(\mathcal{E}\) for which \(\mu(C_n) < +\infty\) and \(\cup_{n \geq 1} C_n = E\).
1.1.5 Expectation and the Transfer Theorem

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

**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

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

\(\square\) **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 \( X \).

\(\square\) **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.

\(\square\) **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 $C_1 \in \mathcal{E}_1, \ldots, C_k \in \mathcal{E}_k$,

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

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 $dP_{X_i} = p_i(x_i) d\mu_i(x_i)$ for any $i$, the latter product measure has density $p_1(x_1) \cdots p_k(x_k)$ with respect to the product measure $\mu_1 \otimes \cdots \otimes \mu_k$. Besides, this characterisation shows that if the random variables $X_1, \ldots, X_k$ 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 $P(A_1 \cap A_2) = P(A_1) P(A_2)$.

2. Show that if $k \geq 3$, the identity $P(A_1 \cap \cdots \cap A_k) = P(A_1) \cdots P(A_k)$ 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 $P(B) > 0$, are independent if and only if $P(A|B) = P(A)$: 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 \( x \to y = \phi(x) \), 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
\[
p * q(z) = \int_{y \in \mathbb{R}} p(z - y) q(y) dy.
\]

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

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

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

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

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 \( U_n = x_n/m \in [0, 1) \), 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 \( P(X = 1) = p \) and \( P(X = 0) = 1 - p \) is called a Bernoulli random variable with parameter \( p \). We denote \( X \sim B(p) \).

(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 \( S \sim B(n, p) \).

(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 \( T \sim G eo(p) \).

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

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 \( X \sim B(p) \)?
2. Using a for loop, how to draw a random variable \( S \sim B(n, p) \)?
3. Using a while loop, how to draw a random variable \( T \sim Geo(p) \)?

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 \( F_X : \mathbb{R} \rightarrow [0, 1] \) defined by

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

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 [2, Exercises p. 10].

\(^8\)It is a practical consequence of the Dynkin System Theorem [2, 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 $F_X^{-1}$, $F_X^{-1}(u) \leq x$. \qed

**Corollary 1.2.14 (The inverse CDF method).** Let $F_X$ be the CDF of a random variable $X$, and let $U \sim U[0, 1]$. 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 $x \in \mathbb{R}$,
\[
\mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = \int_{u=0}^{F_X(x)} \frac{dF_X(x)}{dx} = F_X(x),
\]
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. \qed

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
\[
p(x) = 1_{\{x > 0\}} \lambda e^{-\lambda x}.
\]
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 $-\frac{1}{\lambda} \ln(U)$.
1.2 Random number simulation

◮ 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 $f(x) = \frac{1}{x^{\alpha+1}}$.
2. The Cauchy distribution with parameter $a > 0$, with density $f(x) = \frac{a}{\pi (a^2 + x^2)}$.
3. The Weibull distribution with parameter $m > 0$, with density $f(x) = \frac{mx^{m-1}}{\sigma^m} \exp\left(-\frac{x^m}{\sigma^m}\right)$.
4. The Rayleigh distribution with parameter $\sigma^2 > 0$, with density $f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$.

◮ 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 $k \in \mathbb{N}$,
$$
\Pr(N = k) = e^{-\lambda} \frac{\lambda^k}{k!}.
$$
We denote $N \sim \mathcal{P}(\lambda)$.
1. Compute $\mathbb{E}[N]$ 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 $k \geq 0$, $\Pr(S_n = k) \to \Pr(N = k)$. 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 $\mathbb{E}[G]$ 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 $\mathbb{E}[X] = \mu$ 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.

---

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

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

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 $\Phi : \mathbb{R} \to [0, 1]$ denote its CDF, given by

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

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 $f : \mathbb{R}^2 \to \mathbb{R}$ 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

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

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

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

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. \qed

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 $[0, 1]$. 
1.2.5 Rejection sampling

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

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

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

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

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

![Figure 1.1: The domain $D$ framed into a rectangle $R$. Random points are drawn in $R$, only those falling into $D$ are kept.](image)

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.

\[ \text{Exercise 1.2.24. Show that } N \sim \text{Geo}(\frac{|D|}{|R|}). \]

In particular, $\mathbb{E}[N] = \frac{|R|}{|D|}$ 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

$$\mathbb{P}(X_N \in C) = \sum_{n=1}^{+\infty} \mathbb{P}(X_n \in C, N = n)$$

$$= \sum_{n=1}^{+\infty} \mathbb{P}(X_1 \notin D, \ldots, X_{n-1} \notin D, X_n \in C \cap D).$$
Since the random variables $X_1, \ldots, X_n$ are independent, each term of the sum rewrites
\[
\mathbb{P}(X_1 \notin D, \ldots, X_{n-1} \notin D, X_n \in C \cap D) = \mathbb{P}(X_1 \notin D) \cdot \mathbb{P}(X_{n-1} \notin D) \mathbb{P}(X_n \in C \cap D) = 
\left(1 - \frac{|D|}{|R|}\right)^{n-1} \int_{|x| \in R} \mathbb{I}_{\{x \in C \cap D\}} \frac{dx}{|R|}
\]
whence
\[
= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \frac{|D|}{|R|} \int_{x \in \mathbb{R}^d} \mathbb{I}_{\{x \in C\}} p(x) dx,
\]
where $p$ denotes the uniform density on $D$. Summing over $n$, we deduce that
\[
\mathbb{P}(X_N \in C) = \int_{x \in \mathbb{R}^d} \mathbb{I}_{\{x \in C\}} p(x) dx,
\]
which shows that $X_N$ has density $p$.

**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.

**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, $dx$-almost everywhere, $p(x) \leq kq(x)$. Let $(X_n)_{n \geq 1}$ be a sequence of independent random variables in $\mathbb{R}^d$ with density $q$, and $(U_n)_{n \geq 1}$ be a sequence of independent random variables uniformly distributed in $[0, 1]$, independent from $(X_n)_{n \geq 1}$. Let
\[
N := \inf \{n \geq 1 : kq(X_n)U_n \leq p(X_n) \},
\]
We have the following results:

(i) $N \sim \text{Geo}(1/k)$;
(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 $kU_n \leq \frac{dp}{dq}(X_n)$.

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

**Exercise 1.2.27 (Gamma distribution).** The Gamma distribution with (shape) parameter $a > 0$ is the probability measure on $\mathbb{R}$ with density
\[
p(x) = \mathbb{I}_{\{x > 0\}} \frac{1}{\Gamma(a)} x^{a-1} e^{-x},
\]
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 $L^p(P; \mathbb{R}^d)$ the set of random vectors whose coordinates are random variables in $L^p(P)$. If $X = (X_1, \ldots, X_n) \in L^1(P; \mathbb{R}^d)$, 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 $X, Y \in L^2(P)$. The covariance between $X$ and $Y$ is defined by


It is clear that the covariance is symmetric and bilinear on $L^2(P)$, 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 $X = (X_1, \ldots, X_d) \in L^2(P; \mathbb{R}^d)$. 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 $X \in L^2(P; \mathbb{R}^d)$ has covariance matrix $K$, then for any $u \in \mathbb{R}^d$,

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

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

$$\text{Cov}[b + AX] = AKAT.$$
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 

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

By Theorem 1.1.10, we get that 

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

so that up to sign change and dilution, 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 $|q| = q_1 + \cdots + q_d \leq p$,

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

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

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

The right-hand side satisfies

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

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

$$
\mathbb{E}\left[\left|X_1^{q_1} \cdots X_d^{q_d}\right|\right] \leq \mathbb{E}\left[\left|X_1^{q_1}\right|^{\alpha_1}\right] \cdots \mathbb{E}\left[\left|X_d^{q_d}\right|^{\alpha_d}\right]^{1/\alpha_d}.
$$

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.

\[ \square \]
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

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

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

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

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,

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

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 $m = E[X]$ 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.
\textbf{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
\[
\forall i \in \{1, \ldots, d\}, \quad \forall j \in \{1, \ldots, k\}, \quad \text{Cov}(X_i, Y_j) = 0.
\]

\textit{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. \( \Box \)

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 \( i \), \( Ke_i = \lambda_i e_i \).

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

\textit{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 \( \text{Cov}[X] = K \). \( \Box \)

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–Loève 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, $X \sim \mathcal{N}_d(m, K)$ has density

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

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.

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

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\(^3\), 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.

![Figure 2.1: 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_url)

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

\(^3\)It is referred to as *convergence in measure* in analysis.
2.1 Convergence of random variables

Proof. Observe that, for any \( y \geq 0 \),

\[
1(\{y \geq a\}) \leq \frac{y}{a}
\]

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

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

As a consequence, the random variable \( 1(\{|X_n-X| \geq \epsilon\}) \) converges to 0, almost surely, and therefore by Lemma 2.1.3,

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

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

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

and the proof is completed.

\[\square\]

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 \( B(0, M + 1) \), so that there exists \( \delta_{M,\epsilon} \in (0, 1] \) such that for any \( x, x' \in B(0, M + 1) \), if \( |x - x'| \leq \delta_{M,\epsilon} \) then \(|f(x) - f(x')| \leq \epsilon\). We fix \( M \geq 0 \) and first write

\[
P(|f(X_n) - f(X)| \geq \epsilon) = P(|f(X_n) - f(X)| \geq \epsilon, |X| > M) + P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M)
\]

\[
\leq P(|X| > M) + P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M).
\]

The second term in the right-hand side rewrites

\[
P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M) = P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| \leq \delta_{M,\epsilon}) + P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| > \delta_{M,\epsilon}).
\]
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
\[
P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| \leq \delta_{M,\epsilon}) = 0.
\]

On the other hand, it is immediate that
\[
P(|f(X_n) - f(X)| \geq \epsilon, |X| \leq M, |X_n - X| > \delta_{M,\epsilon}) \leq \P(|X_n - X| > \delta_{M,\epsilon}).
\]

Since $X_n \to X$ in probability, the right-hand side above goes to 0 when $n \to +\infty$. Overall, we deduce that
\[
\limsup_{n \to +\infty} \P(|f(X_n) - f(X)| \geq \epsilon) \leq \P(|X| > M).
\]

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 $\epsilon > 0$,
\[
\P(|X_n - X| + |Y_n - Y| \geq \epsilon) \leq \P(|X_n - X| \geq \epsilon/2) + \P(|Y_n - Y| \geq \epsilon/2),
\]
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 $\epsilon > 0$,
\[
\P(|X - Y| \geq \epsilon) \leq \P(|X_n - X| + |X_n - Y| \geq \epsilon)
\leq \P(|X_n - X| \geq \epsilon/2) + \P(|X_n - Y| \geq \epsilon/2),
\]
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 $X = Y$.

### 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

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

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

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

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

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

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

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

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 \(M \geq 1\),

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

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

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

which completes the proof.

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.

\[\textbf{Exercise 2.1.8 (Dominated Convergence Theorem with convergence in probability). Assume that } X_n \to X \text{ in probability, and that there exists } Y \in L^1(\mathbb{P}) \text{ such that } |X_n| \leq Y \text{ for any } n. \text{ The purpose of this exercise is to prove that } E[X_n] \text{ converges to } E[X].
\]

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 \(E[Y 1_{\{|Y| > M\}}] \leq \epsilon\).
3. Deduce that if \(X_n \to X\) in probability and there is a random variable \(Y \in L^p(\mathbb{P})\) such that \(|X_n - X| \leq Y\), then \(X_n \to X\) in \(L^p\).

\[\textbf{Exercise 2.1.9 (Riesz–Scheffé’s Lemma). Assume that } X_n \to X \text{ almost surely, and that } X_n, X \in L^p(\mathbb{P}) \text{ with } E[|X_n|^p] \to E[|X|^p].
\]

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 \(L^p\).
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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

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

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 $X_n \rightarrow X$ in distribution, then for any continuous function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$, $f(X_n)$ converges in distribution to $f(X)$.

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

(iii) If $X_n \rightarrow X$ in distribution and $X_n \rightarrow Y$ 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 $X_n \rightarrow X$ and $Y_n \rightarrow Y$ 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 $X_n \rightarrow X$, $Y_n \rightarrow Y$ 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 $X_n \rightarrow X$ in distribution and $Y_n \rightarrow y$ in probability, where $y$ is a constant. Then $(X_n,Y_n) \rightarrow (X,y)$ 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 $X_n \rightarrow X$ in probability, then $X_n \rightarrow X$ in distribution.

Conversely, for any $x \in \mathbb{R}^d$, if $X_n \rightarrow x$ in distribution then $X_n \rightarrow x$ 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 \( \epsilon > 0 \),

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

By Proposition 2.1.15 (i), \( \mathbb{P}(|f(X_n) - f(X)| \geq \epsilon) \) goes to 0 when \( n \to \infty \). Therefore

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

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 \( \psi : [0, +\infty) \to [0, 1] \) such that \( \psi(0) = 0 \) and \( \psi(r) \geq \mathbb{I}_{(r \geq \epsilon)} \) for any \( r \geq 0 \). Then the function \( f : \mathbb{R}^d \to \mathbb{R} \) is continuous and bounded, and it satisfies

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

Applying Definition 2.1.11 with this function, we get

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

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

**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 \( (X_n)_{n \geq 1} \) 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 for any \( M \geq 0 \),

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

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

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

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

1. Show that for any \(n \geq 1\),

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

2. Deduce that

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

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

\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 \(f : E \to \mathbb{R}\) 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) \(X_n \to X\) in distribution;
(ii) for any \(x \in E\), \(\mathbb{P}(X_n = x) \to \mathbb{P}(X = x)\);
(iii) \(\sum_{x \in E} |\mathbb{P}(X_n = x) - \mathbb{P}(X = x)| \to 0\).

\subsection*{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 \(f : \mathbb{R}^d \to \mathbb{R}\), we look for conditions under which \(\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]\). 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 \(X_n \to X\) in distribution and \(f : \mathbb{R}^d \to \mathbb{R}\) be bounded. Denote by \(C_f\) the set of \(x \in \mathbb{R}^d\) such that \(f\) is continuous at \(x\). If \(\mathbb{P}(X \in C_f) = 1\) then \(\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]\).

We leave the proof of Proposition 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 \(F(x^-) = F(x)\),

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

In particular,

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

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

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 $L^1(\mathbb{P})$ 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 $E[X_n] \to E[X]$.


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 statements 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. \qed

Theorem 2.2.3 (Strong Law of Large Numbers). If $X_1 \in L^1(\mathbb{P}; \mathbb{R}^d)$ then $X_n \to \mathbb{E}[X_1]$ 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 $L^1$). 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 $X_n \to \mathbb{E}[X_1]$ 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 $M \geq 1$,

$$\mathbb{E} \left[ |X_n|^p \mathbb{I}_{\{|X_n| \geq M\}} \right] \leq \mathbb{E} \left[ |X_1|^p \mathbb{I}_{\{|X_1| \geq M\}} \right].$$

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 $X'_i = X_i \mathbb{I}_{\{|X_i| \leq i\}}$ and proves that Theorem 2.2.3 is equivalent to the statement that

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

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

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^n Z_i = 0, \quad \text{almost surely.} \quad (2.2)$$

Besides, an explicit computation yields

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

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
\[
\lim_{N \to +\infty} \frac{1}{N} \sum_{n=1}^{N} U_n = T, \quad \text{almost surely}.
\]

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 \((\sum_{i=1}^{n} a_n u_i)_{N \geq 1}\) has a finite limit then \(a_n \sum_{i=1}^{n} u_i\) converges to 0.

4. Conclusion of the proof: combining (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 1981\(^4\). A recent preprint by Fitzsimmons\(^5\) 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 \(X_1 \in \mathbb{L}^2(\mathbb{P}; \mathbb{R}^d)\) 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
\[
\lim_{n \to +\infty} \frac{1}{\sqrt{n}} \left( \frac{X_n - \mathbb{E}[X_1]}{\sqrt{n}} \right) = \mathcal{N}(0, K).
\]

**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
\[
Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,
\]
so that the characteristic function of \(Z_n\) writes, for all \(u \in \mathbb{R}^d\),
\[
\Phi_{Z_n}(u) = \mathbb{E} \left[ \exp \left( i \left\langle u, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \right\rangle \right) \right] = \mathbb{E} \left[ \exp \left( i \left\langle \frac{u}{\sqrt{n}}, Y_i \right\rangle \right) \right]^n = \Phi_{Y_1}(u/\sqrt{n})^n,
\]
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
\[
\frac{\partial \Phi_{Y_1}}{\partial u_i}(0) = i \mathbb{E}[Y_{1,i}] = 0, \quad \frac{\partial^2 \Phi_{Y_1}}{\partial u_i \partial u_j}(0) = i^2 \mathbb{E}[Y_{1,i}Y_{1,j}] = -K_{i,j},
\]

\(^4\)https://link.springer.com/article/10.1007/BF01013465

\(^5\)https://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 $Z$. \qed

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. \qed

**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
\[ \lim_{n \to +\infty} \frac{\sqrt{n}}{\sigma} (\hat{I}_n - I) = N(0, 1), \]
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
\[ I'_n = \left[ \hat{I}_n - \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \hat{I}_n + \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \right] \]
Figure 2.2: Quantiles of the standard Gaussian distribution. The hatched area on the figure is equal to $1 - \alpha$.

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

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

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 $I_n$. 

### 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

$$
\mathbb{P} \left( J \in J_n \right) \geq 1 - \alpha. \tag{2.4}
$$

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

$$
\forall a > 0, \quad \mathbb{P} \left( |\hat{J}_n - J| \geq a \right) \leq \frac{\text{Var}(\hat{J}_n)}{a^2} = \frac{\sigma^2}{na^2},
$$

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

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

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 $\lambda \geq 0$, $\mathbb{E}[\exp(\lambda Z_1)] \leq \exp(\lambda^2/8)$. 

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

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

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 \( \mathbb{I} \).

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

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

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

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

Since \( f(X_i) \sim \mathcal{B}(\mathbb{I}) \), we have \( \sigma^2 = \text{Var}(f(X_1)) = \mathbb{I}(1 - \mathbb{I}) \simeq \mathbb{I} \), so that the length of the Monte Carlo confidence interval writes
\[
\ell_n \simeq 2\phi_{1-\alpha/2} \sqrt{\frac{\mathbb{I}}{n}}.
\]

Assume that we want this length to be smaller than \( \epsilon \mathbb{I} \), in order for the estimation of \( \mathbb{I} \) to have a relative precision of \( \epsilon \). Then we need to take \( n \) such that
\[
2\phi_{1-\alpha/2} \sqrt{\frac{\mathbb{I}}{n}} \leq \epsilon \mathbb{I},
\]
that is to say
\[ n \geq \left( \frac{2\phi_1 - \alpha/2}{\epsilon} \right)^2 \frac{1}{\mathbb{J}}. \]

For \( \epsilon = 0.01 \) and \( \alpha = 0.05 \), using the analytic bound on \( \mathbb{J} \) we obtain that \( n \) should be at least \( 5.6 \times 10^9 \), which is impossible to realise in practice.

In this section, we present variance reduction techniques which allow to construct estimators of \( \mathbb{J} \) 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 \( \beta \in \mathbb{R} \),
\[ \mathbb{J} = \mathbb{E}[f(X)] = \mathbb{E}[f(X) - \beta Y] + \beta \mathbb{E}[Y], \]

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

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

We may already remark that if \( \text{Cov}(f(X), Y) = 0 \) then \( (\sigma_{CV,\beta}^2)^2 \) 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 \( (\sigma_{CV,\beta}^2)^2 \) is minimal is then
\[ \beta^* = \frac{\text{Cov}(f(X), Y)}{\text{Var}(Y)}, \]

which yields the variance
\[ (\sigma_{CV,\beta^*}^2)^2 = \sigma^2 (1 - \rho^2), \]

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

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
\[ \hat{C}_n = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \hat{\mathbb{J}}_n)(Y_i - \bar{Y}_n). \]

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 \(f(X_i), Y_i \in L^2(\mathbb{P})\). For all \(n \geq 1\), let

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

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

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

where

\[
(\hat{\sigma}_{CV}^2)^2 = \hat{\sigma}_n^2 \left( 1 - \frac{\hat{C}_n^2 \text{Var}(Y)}{\hat{\sigma}_n^2} \right),
\]

satisfies

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

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

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

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 \(\hat{I}_{CV}^n\).

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

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

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

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

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

$$\left(\sigma_{Q}^{IS}\right)^{2} := \text{Var}(f(Y)w(Y))$$

as small as possible.

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

$$dQ^{*}(x) = \frac{|f(x)|}{\bar{J}} dP(x).$$

(i) $Q^{*}$ satisfies (2.5) and $\left(\sigma_{Q}^{IS}\right)^{2} = \bar{J}^{2} - J^{2}$.

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

(iii) If $f$ has constant sign $P$-almost everywhere, then $\left(\sigma_{Q}^{IS}\right)^{2} = 0$.

Proof. As a preliminary remark, we note that for any $Q$ satisfying (2.5),

$$\left(\sigma_{Q}^{IS}\right)^{2} = \mathbb{E} \left[ (f(Y)w(Y))^{2} \right] - J^{2}, \quad Y \sim Q. \tag{2.6}$$

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

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

with respect to $1_{\{f(x) \neq 0\}} dQ^{*}(x)$, therefore $Q^{*}$ satisfies (2.5) and besides, if $Y^{*} \sim Q^{*}$, then

$$\mathbb{E} \left[ (f(Y^{*})w^{*}(Y^{*}))^{2} \right] = \int_{x \in E} 1_{\{f(x) \neq 0\}} |f(x)|^{2} \left( \frac{\bar{J}}{|f(x)|} \right)^{2} dQ^{*}(x)$$

$$= \bar{J}^{2} \int_{x \in E} 1_{\{f(x) \neq 0\}} dQ^{*}(x)$$

$$= \bar{J}^{2},$$

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 $\bar{J}$ and $w$, and the Cauchy–Schwarz inequality,

$$\bar{J}^{2} = \left( \int_{x \in E} |f(x)| 1_{\{f(x) \neq 0\}} dP(x) \right)^{2}$$

$$= \left( \int_{x \in E} |f(x)|w(x) 1_{\{f(x) \neq 0\}} dQ(x) \right)^{2}$$

$$\leq \int_{x \in E} |f(x)|^{2}w(x)^{2} 1_{\{f(x) \neq 0\}} dQ(x)$$

$$= \mathbb{E} \left[ (f(Y)w(Y))^{2} \right],$$

with $Y \sim Q$. Combined with (2.6), this estimate completes the proof of (ii).
In practice it is impossible to implement the method with the optimal measure $Q^*$ since the latter depends explicitly on the quantity $\mathbb{I}$, which is likely to be unknown — and, in the case where $f$ is nonnegative $P$-almost everywhere, is exactly the quantity $\mathbb{I}$ 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 $\mathbb{P}(X \geq 20)$, for $X \sim \mathcal{N}(0, 1)$, given in the introduction of this section, the optimal density is proportional to $1_{\{y \geq 20\}}e^{-y^2/2}$. 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 $\mathbb{I} > |\mathbb{I}|$ (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 $Y \sim Q$,

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

A trivial example would be to assume that there exists $x_0 \in E$ such that $f(x_0) = \mathbb{I}$, 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
Part III

Diffusion processes and partial differential equations
Bibliography
