Master de Mathématiques et Applications

Spécialité *Mathématiques de la Modélisation*

**Méthodes Numériques Probabilistes**

Tony Lelièvre et Julien Reygner

Version du 11 octobre 2021
Foreword

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

- 🌋 Can be made during class.
- 🏡 Can be made at home between two classes.
- ☕ Can be made in order to study before the exam.

**Some references.** For general lecture notes (in French) on probability theory, we refer to [8] for a first course (level L3), and to [11] for more advanced material, in particular on measure theory and Lebesgue integration. At several points in Part II, we follow the presentation of [13], which is very complete and pleasant to read. We also refer to [7, 2]. In Part III, we essentially follow [4] and refer to [12, 9] for more advanced material. Many interesting applications of stochastic processes to various fields are developed in [5, 3].
Contents

I The Monte Carlo method 1

1 Random number simulation 3
  1.1 Principle of the Monte Carlo method 3
  1.2 Random number generation 7

2 Confidence intervals 17
  2.1 Convergence in distribution and Central Limit Theorem 17
  2.2 Confidence intervals 20
  2.3 Variance reduction 23

II Markov chains 27

3 Markov chains in finite spaces 29
  3.1 General definitions 29
  3.2 Stationary distributions 33
  3.3 Extension to general state spaces 36

4 Ergodic theory of Markov chains 39
  4.1 Extensions of the Law of Large Numbers and the Central Limit Theorem 39
  4.2 Geometric convergence and spectral aspects 46
  4.3 Mixing time and coupling 48

5 The Markov Chain Monte Carlo method 53
  5.1 Reversible Markov chains 53
  5.2 Markov chain Monte Carlo 56

III Diffusion processes 63

References 65
Part I

The Monte Carlo method
Lecture 1

Random number simulation

The Monte Carlo method is a numerical method designed to approximate integrals, say on $\mathbb{R}^d$, of the form

$$I = \int_{x \in \mathbb{R}^d} f(x) p(x) dx,$$  \hspace{1cm} (1.1)

where $p: \mathbb{R}^d \to \mathbb{R}$ is a function such that

$$p(x) \geq 0, \text{ } dx\text{-almost everywhere}, \text{ and } \int_{x \in \mathbb{R}^d} p(x) dx = 1. \hspace{1cm} (1.2)$$

The method is based on a probabilistic interpretation of $p$ and $I$, which we first detail.

1.1 Principle of the Monte Carlo method

1.1.1 Random variables and expectation

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 on $\Omega$;
- $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{A})$, that is to say a nonnegative measure such that $\mathbb{P}(\Omega) = 1$.

Measurable sets $A \in \mathcal{A}$ are usually called events. An event $A$ such that $\mathbb{P}(A) = 1$ is called almost sure.

Definition 1.1.2 (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

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

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

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

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 $X$ is distributed according to $P$, that is to say that $P_X = P$. Notice that the triple $(E, \mathcal{E}, P_X)$ is a probability space itself.

The set of probability measures on $(E, \mathcal{E})$ is denoted by $\mathcal{P}(E)$. 

Example 1.1.3 (Throwing two dice). Assume that you throw two dice. The natural probability space to model the outcome of the experiment is \( \Omega = \{1, \ldots, 6\}^2 \) endowed with its power set and the probability measure \( \mathbb{P} \) defined by \( \mathbb{P}(\{\omega\}) = 1/36 \) for any \( \omega \in \Omega \). The random variable defined by \( X(\omega) = \omega_1 + \omega_2 \), for any \( \omega = (\omega_1, \omega_2) \in \Omega \), which designates the sum of the values of the two dice, takes its values in the space \( E = \{2, \ldots, 12\} \) endowed with its power set, and has a law \( P_X \) given by \( P_X(\{2\}) = \mathbb{P}(\{(1, 1)\}) = 1/36 \), \( P_X(\{3\}) = \mathbb{P}(\{(1, 2), (2, 1)\}) = 2/36 \), and so on.

Assume that \( E = \mathbb{R}^d \), endowed with its Borel \( \sigma \)-field \( \mathcal{E} = \mathcal{B}(\mathbb{R}^d) \).

Definition 1.1.4 (Probability density). A probability density is a measurable function \( p : \mathbb{R}^d \to \mathbb{R} \) which satisfies (1.2). A random variable \( X \) in \( \mathbb{R}^d \) is said to have density \( p \) with respect to the Lebesgue measure whenever

\[
\forall C \in \mathcal{B}(\mathbb{R}^d), \quad \mathbb{P}(X \in C) = \int_{x \in \mathbb{R}^d} 1_{\{x \in C\}} p(x) dx;
\]

in other words, when \( P_X(dx) = p(x) dx \).

We recall that by the Radon–Nikodym Theorem, the random variable \( X \) has a density with respect to the Lebesgue measure if and only if \( P_X(C) = 0 \) for all \( C \in \mathcal{B}(\mathbb{R}^d) \) with Lebesgue measure 0.

Example 1.1.5 (Dart game). Let \( E = \{x \in \mathbb{R}^2 : |x| \leq 1\} \) represent the target of a dart game and \( X \in E \) be the location of the dart that you just have thrown. If you are a beginner, it may be considered that the probability that \( X \) belongs to some subset \( C \) of \( E \) is proportional to the surface of \( C \), whatever the location of \( C \) in \( E \), so that \( X \) may be assumed to have the uniform density \( p_X(x) = \pi^{-1} 1_{\{x \in E\}} \). If you already are an expert, then it becomes more likely that your dart falls in areas of \( E \) which are close to the center, so that the law of \( X \) can be modeled by a density of the form \( p_X(x) = q(|x|) \), where \( q : [0, 1] \to [0, +\infty) \) is nonincreasing.

Remark 1.1.6 (* Atoms, discrete and continuous random variables). Assume that the measurable space \((E, \mathcal{E})\) is such that \( \{x\} \in \mathcal{E} \) for all \( x \in E \) (this the case of all measurable spaces considered in this course). An atom of a probability measure \( \mu \) on \((E, \mathcal{E})\) is an \( x \in E \) such that \( \mu(\{x\}) > 0 \).

A random variable is called continuous if its law does not have any atom. Obviously, random variables in \( \mathbb{R}^d \) which admit a density with respect to the Lebesgue measure, such as in Example 1.1.5, are continuous, but the converse is not true, see Exercise 1.2.19 below.

On the other hand, a random variable is called discrete if it takes its values in a finite or countably infinite space \( E \), as in Example 1.1.3. In this case, its law writes as a (weighted) sum of Dirac masses and is called purely atomic.

Obviously, there are random variables which are neither continuous nor discrete: coming back to Example 1.1.5, assume that we now want to take into account the fact that you sometimes miss the target, say with probability \( \theta \in (0, 1) \), and call \( \dagger \) the value of \( X \) in this case. Then the state space for \( X \) becomes \( E \cup \{\dagger\} \), and its law \( P_X = (1 - \theta) 1_{\{x \in E\}} p_X(x) dx + \theta 1_{\{x = \dagger\}} \) has an atom in \( \dagger \).

For all \( p \in [1, +\infty) \), we denote by \( L^p(\Omega, \mathcal{A}, \mathbb{P}) \), or more 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 \( \Omega \).

Definition 1.1.7 (Expectation). Let \( X \in L^1(\mathbb{P}) \). The expectation of \( X \) is the Lebesgue integral

\[
\mathbb{E}[X] := \int_{\omega \in \Omega} X(\omega) d\mathbb{P}(\omega).
\]
Remark 1.1.8. When $X$ is nonnegative but not necessarily in $L^1(\mathbb{P})$, the integral in Definition 1.1.7 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 $\mathbb{E}[X] < +\infty$.

Exercise 1.1.9. Check that if $1 \leq p \leq q$, then $L^q(\mathbb{P}) \subset L^p(\mathbb{P})$ and $\mathbb{E}[|X|^p]^{1/p} \leq \mathbb{E}[|X|^q]^{1/q}$.

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 $f : E \rightarrow \mathbb{R}$ 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 $E = \mathbb{R}^d$ and $X$ has density $p$ with respect to the Lebesgue measure, then

$$
\mathbb{E}[f(X)] = \int_{x \in \mathbb{R}^d} f(x) p(x) dx,
$$

which provides a probabilistic representation of the integral $\mathcal{I}$ defined in (1.1).

Sketch of proof. If $f$ is a simple function, that is to say $f(x) = \sum_{i=1}^L a_i \mathbb{1}_{\{x \in C_i\}}$ for $a_1, \ldots, a_L \in \mathbb{R}$ and $C_1, \ldots, C_L \in \mathcal{E}$, then the identities of Theorem 1.1.10 follow from the linearity of the Lebesgue integral. The generalisation to any measurable function $f$ relies on the following approximation argument. First write $f = f^+ - f^-$ with $f^+, f^- \geq 0$ the respective positive and negative parts of $f$. It is known [11, Proposition 2.1.2] that one may construct sequences $f^+_n$ of simple functions such that $0 \leq f^+_n \leq f^+$ and $f^+_n \uparrow f^+$, $dx$-almost everywhere, to which one may thus apply the Monotone Convergence Theorem [11, Théorème 2.1.1].

Remark 1.1.11. The proof of Theorem 1.1.10 shows in particular that

$$
\mathbb{E}\left[\mathbb{1}_{\{X \in C\}}\right] = \mathbb{P}(X \in C),
$$

for any $C \in \mathcal{E}$. This identity is constantly used in the sequel.

Up to renaming $f(X)$ into $X$, there is no loss of generality in writing directly

$$
\mathcal{I} = \mathbb{E}[X],
$$

for some random variable $X$ in $\mathbb{R}$. For the sake of simplicity we shall keep this notation in the sequel of this lecture.

1.1.2 The Law of Large Numbers

Probability theory not only provides a representation of $\mathcal{I}$ but also an implementable method to approximate its value, based on the Law of Large Numbers.

Definition 1.1.12 (Independence). Let $I$ be a set of indices and $(X_i)_{i \in I}$ be a family of random variables on $\Omega$, such that every $X_i$ takes its values in a measurable space $(E_i, \mathcal{E}_i)$. The variables $(X_i)_{i \in I}$ are called independent if, for any finite set of indices $\{i_1, \ldots, i_k\} \subset I$, for all $C_1 \in \mathcal{E}_{i_1}, \ldots, C_k \in \mathcal{E}_{i_k}$,

$$
\mathbb{P}(X_{i_1} \in C_1, \ldots, X_{i_k} \in C_k) = \mathbb{P}(X_{i_1} \in C_1) \cdots \mathbb{P}(X_{i_k} \in C_k).
$$
Using the same approximation argument as in the proof of Theorem 1.1.10, we obtain the following useful property of independent random variables.

**Lemma 1.1.13** (Independence and expectation). Let \( I \) be a set of indices and \((X_i)_{i \in I}\) be a family of independent random variables on \( \Omega \), such that every \( X_i \) takes its values in a measurable space \((E_i, \mathcal{E}_i)\). For all \( i \in I \), let \( f_i : E_i \to \mathbb{R} \) be a measurable function such that \( f_i(X_i) \in L^1(\mathbb{P}) \). Then, for any finite set of indices \( \{i_1, \ldots, i_k\} \subset I \), \( f_{i_1}(X_{i_1}) \cdots f_{i_k}(X_{i_k}) \in L^1(\mathbb{P}) \) and

\[
\mathbb{E} \left[ f_{i_1}(X_{i_1}) \cdots f_{i_k}(X_{i_k}) \right] = \mathbb{E} \left[ f_{i_1}(X_{i_1}) \right] \cdots \mathbb{E} \left[ f_{i_k}(X_{i_k}) \right].
\]

Lemma 1.1.13 actually provides a characterisation of independence since it is clear that if its conclusion holds then taking \( f_i(x) = \mathbb{1}_{\{x \in C_{i_1}\}} \), \( \ldots \), \( f_{i_k}(x) = \mathbb{1}_{\{x \in C_{i_k}\}} \), we recover the condition of Definition 1.1.12 by Remark 1.1.11.

A family of independent random variables \((X_i)_{i \in I}\) such that all spaces \( E_i \) are the same and all variables \( X_i \) have the same law is called **independent and identically distributed**, which we shall abbreviate to **iid**.

**Theorem 1.1.14** (Law of Large Numbers). Let \((X_i)_{i \geq 1}\) be a sequence of iid random variables in \( L^1(\mathbb{P}) \). Their empirical mean

\[
\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i
\]

satisfies

\[
\lim_{n \to +\infty} \bar{X}_n = \mathbb{E}[X_1], \quad \text{almost surely},
\]

that is to say that the event \( \{\omega \in \Omega : \lim_{n \to +\infty} \bar{X}_n(\omega) = \mathbb{E}[X_1]\} \) has probability 1.

**Remark 1.1.15.** Since the random variables \( X_i \) are identically distributed, they all have the same expectation and therefore the limit \( \mathbb{E}[X_1] \) is the expectation of any of the variables \( X_i \), \( i \geq 1 \). By the linearity of the Lebesgue integral, it is also the expectation of \( \bar{X}_n \).

The proof of Theorem 1.1.14 is difficult and we admit the result (see [8, Théorème 5.2.2] for an elementary proof). However it is easy to get an intuition of why such a statement should hold true, and also a rate of convergence, by stating a weak version of the Law of Large Numbers. In order to do so, we first introduce the notion of **variance**.

**Definition 1.1.16** (Variance). The variance of a random variable \( X \in L^2(\mathbb{P}) \) is defined by

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

Notice that, by Exercise 1.1.9, the assumption that \( X \in L^2(\mathbb{P}) \) ensures that \( \mathbb{E}[X] \) is well-defined.

**Exercise 1.1.17.** Prove the following properties of variance.

1. For any \( X \in L^2(\mathbb{P}) \), for any \( a, b \in \mathbb{R} \), \( \text{Var}(aX + b) = a^2 \text{Var}(X) \).
2. For any independent variables \( X, Y \in L^2(\mathbb{P}) \), \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \).

**Exercise 1.1.18.** If \( X, Y \in L^2(\mathbb{P}) \) are independent, what is the value of \( \text{Var}(X - Y) ? \)

We may now state and prove the weak Law of Large Numbers.
1.2 Random number generation

**Proposition 1.1.19 (Weak Law of Large Numbers).** Let \((X_i)_{i \geq 1}\) be a sequence of iid random variables in \(L^2(P)\). For all \(n \geq 1\),

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

so that \(\overline{X}_n\) converges to \(E[X_1]\) in \(L^2(P)\).

**Proof.** We first notice that, following Remark 1.1.15, 
\(E[X_1] = E[\overline{X}_n]\) and therefore

\[
E \left[ (\overline{X}_n - E[X_1])^2 \right] = E \left[ (\overline{X}_n - E[\overline{X}_n])^2 \right] = \text{Var}(\overline{X}_n).
\]

Now the definition of \(\overline{X}_n\) and the results of Exercise 1.1.17 allow to write

\[
\text{Var}(\overline{X}_n) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i),
\]

and since the variables \(X_i\) are identically distributed, all terms of the sum in the right-hand side are equal to \(\text{Var}(X_1)\), which completes the proof. \(\square\)

Even under the assumption that \(X_i \in L^2(P)\), one can however not deduce Theorem 1.1.14 directly from the conclusion of Proposition 1.1.19 since the convergence in \(L^2(P)\) does not necessarily imply the almost sure convergence.

Let us come back to the issue of estimating the integral \(I\) defined in Equation (1.3). As soon as \(X \in L^1(P)\), Theorem 1.1.14 suggests to approximate \(I\) by

\[
\hat{I}_n := \frac{1}{n} \sum_{i=1}^n X_i,
\]

(1.4)

where \(X_1, \ldots, X_n\) are independent random variables with the same distribution as \(X\).

This is the essence of the Monte Carlo method. It raises at least two important questions: how to draw iid samples with a given law? and how much should we take in order to ensure that \(\hat{I}_n\) is a good approximation of \(I\)? We start to discuss the former in Section 1.2 and the latter in Lecture 2, but both questions will actually remain at the heart of the whole course.

### 1.2 Random number generation

#### 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.’\(^1\). 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\(^2\), 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 1.1.12;
(ii) they look uniformly distributed in \( \{0, \ldots, m - 1\} \) in the sense that each integer \( x \in \{0, \ldots, m - 1\} \) appears in the sequence \( (x_n)_{n \geq 1} \) with equal frequency \( 1/m \).

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

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

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 \( U \sim \mathcal{U}[0,1] \).

More generally, the uniform distribution on the interval \( [a, b] \), denoted by \( \mathcal{U}[a,b] \), is the probability measure with density

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

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

\( \blacklozenge \) **Exercise 1.2.3.** Let \( U \sim \mathcal{U}[0,1] \). 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). If $U \sim U[0, 1]$, then $X := a + (b-a)U \sim U[a, b]$.

1.2.2 Classical discrete distributions

We first introduce several discrete distributions.

Definition 1.2.6 (Bernoulli, binomial and geometric distributions). Let $p \in [0, 1]$.

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

(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 $T \sim Geo(p)$.

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

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

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

Exercise 1.2.8. Let $(U_n)_{n \geq 1}$ be a sequence of independent uniform variables on $[0, 1]$.

1. Using an if test, how to draw a random variable $X \sim B(p)$?
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

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

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

$^3$See [11, Exercice p. 10].

$^4$It is a practical consequence of the Dynkin System Theorem [11, 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 $\mathcal{B}$.
Exercise 1.2.11 (Properties of CDFs). Let $F_X$ be the CDF of a random variable $X$. Show that:
\begin{itemize}
  \item [1.] $F_X$ is nondecreasing;
  \item [2.] $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F_X(x) = 1$;
  \item [3.] $F_X$ is right continuous and has left limits.
\end{itemize}

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 $F_X' = p$.

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 $U \sim \mathcal{U}[0, 1]$. The random variables $X$ and $F_X^{-1}(U)$ have the same distribution.

Proof. By Lemma 1.2.13 and Definition 1.2.1, for all $x \in \mathbb{R},$
\[ \mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(U \leq F_X(x)) = \int_{u=0}^{F_X(x)} du = F_X(x), \]
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
\[ p(x) = \mathbb{1}_{\{x>0\}} \lambda e^{-\lambda x}. \]

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

Exercise 1.2.16 (Properties of exponential distributions). Let $X \sim \mathcal{E}(\lambda)$.
\begin{itemize}
  \item [1.] Compute $\mathbb{E}[X]$ and $\text{Var}(X)$.
  \item [2.] If $a > 0$, what is the law of $aX$?
\end{itemize}
An immediate computation shows that the CDF of $X$ writes
\[
F_X(x) = \begin{cases} 
0 & \text{if } x \leq 0, \\
1 - e^{-\lambda x} & \text{otherwise.}
\end{cases}
\]

As a consequence, for all $u \in [0, 1]$,
\[
F_X^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u),
\]
with the obvious convention that $\ln 0 = -\infty$. Therefore, to draw a random variable $X \sim \mathcal{E}(\lambda)$, one may take a uniform variable $U$ on $[0, 1]$ and return $-\frac{1}{\lambda} \ln(1 - U)$. Notice that, by Exercise 1.2.3, it is also equivalent to return $-\frac{1}{\lambda} \ln(U)$.

**Exercise 1.2.17** (Other standard densities). Apply the inverse CDF method to the following standard probability densities.

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

**Exercise 1.2.18** (Poisson distribution). A random variable $N \in \mathbb{N}$ is distributed according to the Poisson distribution with parameter $\lambda > 0$ if, for any $k \in \mathbb{N}$,
\[
\mathbb{P}(N = k) = e^{-\lambda} \frac{\lambda^k}{k!}.
\]
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 $k \geq 0$, $\mathbb{P}(S_n = k) \to \mathbb{P}(N = k)$. Which interpretation of the Poisson distribution can you deduce?
3. Let $(X_i)_{i \geq 1}$ be a sequence of independent exponential random variables with parameter $\lambda$. 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 $[0, 1]$.

**Exercise 1.2.19** (An example of a continuous variable without density). The Cantor function is a function $F : [0, 1] \to [0, 1]$ which is continuous, nondecreasing, and such that $F(0) = 0$ and $F(1) = 1$, so that it is the CDF of a random variable $X \in [0, 1]$. Besides, it has the peculiarity that, $dx$-almost everywhere, $F'(x)$ exists and is equal to $0$.

1. Justify that the law of $X$ does not have any atom (see Remark 1.1.6).
2. Show that $X$ does not have a density with respect to the Lebesgue measure.

### 1.2.4 Gaussian random variables

We recall the value of the Gauss integral:
\[
\int_{x \in \mathbb{R}} \exp\left( -\frac{x^2}{2} \right) \, dx = \sqrt{2\pi}.
\]
**Definition 1.2.20** (Gaussian random variables). A random variable $X$ in $\mathbb{R}$ is Gaussian with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$ if it has the density

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

We denote $X \sim \mathcal{N}(\mu, \sigma^2)$.

In the particular case where $\mu = 0$ and $\sigma^2 = 1$, the law of $X$ is called the standard Gaussian distribution. Gaussian variables are also called normal.

**Exercise 1.2.21** (Properties of Gaussian random variables). Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

1. Compute $\mathbb{E}[X]$ and $\text{Var}(X)$.
2. For $a, b \in \mathbb{R}$, what is the law of $aX + b$? In particular, what is the law of $(X - \mu)/\sigma$?

It follows from Exercise 1.2.21 that if $G$ is a standard Gaussian variable, then $X = \mu + \sigma G \sim \mathcal{N}(\mu, \sigma^2)$. As a consequence, 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

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

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. We first need to give complementary results on the independence of random variables.

**Proposition 1.2.22** (Independence and density). Let $X, Y$ be two random variables in $\mathbb{R}$, with respective densities $p$ and $q$. These variables are independent if and only if the random vector $(X, Y) : \Omega \to \mathbb{R}^2$ has density $p(x)q(y)$.

**Proof.** Let $C, D \in \mathcal{B}(\mathbb{R})$, and let $P_{(X,Y)}$ denote the law of the random vector $(X, Y)$. By definition, we have

$$\mathbb{P}(X \in C, Y \in D) = \int_{(x,y) \in \mathbb{R}^2} \mathbb{1}_{\{x \in C, y \in D\}} P_{(X,Y)}(dx dy).$$

If $(X, Y)$ has density $p(x)q(y)$, then using the property that $\mathbb{1}_{\{x \in C, y \in D\}} = \mathbb{1}_{\{x \in C\}} \mathbb{1}_{\{y \in D\}}$ we may rewrite the right-hand side

$$\int_{(x,y) \in \mathbb{R}^2} \mathbb{1}_{\{x \in C, y \in D\}} p(x)q(y) dx dy = \int_{x \in \mathbb{R}} \mathbb{1}_{\{x \in C\}} p(x) dx \int_{y \in \mathbb{R}} \mathbb{1}_{\{y \in D\}} q(y) dy = \mathbb{P}(X \in C) \mathbb{P}(Y \in D),$$

so that $X$ and $Y$ are independent. Conversely, if $X$ and $Y$ are independent then

$$\mathbb{P}(X \in C, Y \in D) = \mathbb{P}(X \in C) \mathbb{P}(Y \in D) = \int_{(x,y) \in \mathbb{R}^2} \mathbb{1}_{\{x \in C, y \in D\}} p(x)q(y) dx dy,$$

so that the measures $P_{(X,Y)}(dx dy)$ and $p(x)q(y)dx dy$ coincide on the set of rectangles $\{C \times D : C, D \in \mathcal{B}(\mathbb{R})\}$. Since the Borel $\sigma$-field $\mathcal{B}(\mathbb{R}^2)$ is generated by this set\(^6\), we deduce that $P_{(X,Y)}(dx dy) = p(x)q(y)dx dy$.

\(^6\text{See [11, Définition 1.1.4, Exercice p. 10].}\)
The extension of Proposition 1.2.22 to a finite family of random variables \((X_i)_{1 \leq i \leq k}\) such that each \(X_i\) takes its values in \(\mathbb{R}^d\) is straightforward. In fact, more generally, a family of random variables \((X_i)_{i \in I}\) is independent if for any \(\{i_1, \ldots, i_k\} \subset I\), the law of \((X_{i_1}, \ldots, X_{i_k})\) is the product measure \(P_{X_{i_1}} \otimes \cdots \otimes P_{X_{i_k}}\).

\[\textbf{Exercise 1.2.23.} \text{ Let } X \text{ and } Y \text{ be two independent random variables in } \mathbb{R} \text{ with respective densities } p \text{ and } q. \text{ Show that } Z = X + Y \text{ has density } p \ast q(z) = \int_{y \in \mathbb{R}} p(z - y)q(y)\,dy.\]

We may now present the Box–Muller method.

**Proposition 1.2.24** (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.** Let \(C \in \mathcal{B}(\mathbb{R}^2)\). We compute the law of the random vector \((X, Y)\) by writing

\[
\mathbb{P}(\{(X, Y) \in C\}) = \mathbb{P}
\left(
\begin{array}{c}
\sqrt{R} \cos \Theta, \\
\sqrt{R} \sin \Theta
\end{array}
\right) \in C
\right)
\]

\[= \int_{\omega \in \Omega} \mathbb{1}_{\{(\sqrt{R(\omega)} \cos(\Theta(\omega))), \sqrt{R(\omega)} \sin(\Theta(\omega))) \in C\}} \, d\mathbb{P}(\omega)
\]

\[= \int_{r=0}^{+\infty} \int_{\theta=0}^{2\pi} \mathbb{1}_{\{(\sqrt{r} \cos \theta, \sqrt{r} \sin \theta) \in C\}} \frac{d\theta}{2\pi} \frac{1}{2} e^{-r/2} \, dr,
\]

where we have used Theorem 1.1.10 and Proposition 1.2.22 to identify the law of the pair \((R, \Theta)\).

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

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

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

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

so that Proposition 1.2.22 allows to conclude that \(X\) and \(Y\) are independent standard Gaussian variables. \(\square\)

Since both \(R\) and \(\Theta\) can be sampled using the inverse CDF method, Proposition 1.2.24 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 (think of Example 1.1.5)?

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

In the general case, an intuitive procedure can be formulated as follows (see also Figure 1.1):

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

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

\( \square \) Exercise 1.2.25. 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

\[
\mathbb{P}(X_1 \notin D, \ldots, X_{n-1} \notin D, X_n \in C \cap D) = \mathbb{P}(X_1 \notin D \cdot \cdots \cdot \mathbb{P}(X_{n-1} \notin D) \mathbb{P}(X_n \in C \cap D) \\
= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \int_{x \in R} 1_{\{x \in C \cap D\}} \frac{dx}{|R|} \\
= \left(1 - \frac{|D|}{|R|}\right)^{n-1} \frac{|D|}{|R|} \int_{x \in \mathbb{R}^d} 1_{\{x \in C\}} p(x) dx,
\]

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

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

which shows that \( X_N \) has density \( p \).
Exercise 1.2.26. Show that the random variables $X_N$ and $N$ are independent.

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

**Theorem 1.2.27** (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 1.2.27 follows from the same computation as in the example of uniform distributions, for which $q$ is the uniform distribution on the rectangle, and $k = |R|/|D|$.

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

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

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 $k$?
Random number simulation
Lecture 2

Confidence intervals

Let \((X_i)_{i \geq 1}\) be a sequence of iid random variables in \(\mathbb{R}\). If these variables are in \(L^2(\mathbb{P})\), the weak Law of Large Numbers stated in Proposition 1.1.19 shows that \(\bar{X}_n - \mathbb{E}[X_1]\) is of the order of magnitude of \(\sqrt{\text{Var}(X_1)/n}\). A more detailed description of the asymptotic law of \(\bar{X}_n - \mathbb{E}[X_1]\) is provided by the Central Limit Theorem, which then allows to construct confidence intervals for \(f = \mathbb{E}[X_1]\).

2.1 Convergence in distribution and Central Limit Theorem

The Central Limit Theorem relies on the notion of convergence in distribution. We introduce this notion for sequences of random variables in \(\mathbb{R}^d\).

Definition 2.1.1 (Convergence in distribution). A sequence of random variables \((X_n)_{n \geq 1}\) in \(\mathbb{R}^d\) converges in distribution to a random variable \(X\) if, for any continuous and bounded function \(\varphi : \mathbb{R}^d \to \mathbb{R}\), \(\mathbb{E}[\varphi(X_n)]\) converges to \(\mathbb{E}[\varphi(X)]\).

In other words, \(X_n\) converges to \(X\) in distribution if and only if \(P_{X_n}\) converges weakly\(^1\) to \(P_X\). Since convergence in distribution is actually merely a property of the laws of \(X_n\) and \(X\), we shall sometimes write \(X_n \to P\) to say that \(X_n\) converges in distribution to a random variable \(X\) distributed according to \(P\).

Remark 2.1.2. If \(X_n\) converges to \(X\) almost surely, then by the Dominated Convergence Theorem, \(X_n\) converges to \(X\) in distribution. The converse statement is not true, and does not even make a clear sense, because if \(X_n\) converges to \(X\) in distribution, it also converges in distribution to any random variable \(X'\) having the same law as \(X\), so that it is not clear what should be the almost sure limit of \(X_n\).

Theorem 2.1.3 (Central Limit Theorem). Let \((X_i)_{i \geq 1}\) be a sequence of iid random variables in \(L^2(\mathbb{P})\). We have

\[
\sqrt{n}(\bar{X}_n - \mathbb{E}[X_1]) \to \mathcal{N}(0, \sigma^2), \quad \text{in distribution,}
\]

where \(\sigma^2 = \text{Var}(X_1)\).

When \(\sigma^2 = 0\), the notation \(\mathcal{N}(0, 0)\) simply refers to the Dirac distribution in 0, that is to say the law of the constant random variable \(X = 0\).

Before detailing the proof of Theorem 2.1.3, we introduce the notion of characteristic function.

\(^1\)The weak topology on the space \(\mathcal{P}(\mathbb{R}^d)\) of probability measures on \(\mathbb{R}^d\) is usually defined with respect to the set \(C_b(\mathbb{R}^d)\) of continuous and bounded functions on \(\mathbb{R}^d\).
Definition 2.1.4 (Characteristic function). Let $X$ be a random variable in $\mathbb{R}^d$. The characteristic function of $X$ is the function $\Phi_X : \mathbb{R}^d \to \mathbb{C}$ defined by

$$\forall u \in \mathbb{R}^d, \quad \Phi_X(u) := \mathbb{E} [\exp(i \langle u, X \rangle)] = \mathbb{E} [\cos \langle u, X \rangle] + i \mathbb{E} [\sin \langle u, X \rangle].$$

By Theorem 1.1.10, we may rewrite

$$\Phi_X(u) = \int_{x \in \mathbb{R}^d} e^{i \langle u, x \rangle} P_X(dx),$$

which establishes a close link between the characteristic function of $X$ and the Fourier transform of $P_X$. From this link we may infer the following properties.

Proposition 2.1.5 (Properties of characteristic functions).

(i) If $\Phi_X(u) = \Phi_Y(u)$ for all $u \in \mathbb{R}^d$, then $X$ and $Y$ have the same distribution.

(ii) $X_n$ converges in distribution to $X$ if and only if $\Phi_{X_n}(u)$ converges to $\Phi_X(u)$ for all $u \in \mathbb{R}^d$.

The following result is given for random variables in $\mathbb{R}$, in which case $\Phi_X(u)$ simply writes $\mathbb{E}[e^{iuX}]$.

Lemma 2.1.6 (Taylor expansion). Let $p \geq 1$ and $X \in L^p(\mathbb{R})$.

(i) $\Phi_X$ is of class $C^p$ on $\mathbb{R}$ and

$$\Phi^{(p)}_X(u) = i^p \mathbb{E} [X^p e^{iuX}].$$

(ii) When $u \to 0$,

$$\Phi_X(u) = 1 + iu \mathbb{E}[X] - \frac{u^2}{2} \mathbb{E}[X^2] + \cdots + \frac{(iu)^p}{p!} \mathbb{E}[X^p] + o(u^p).$$

Proof. We prove the first point with $p = 1$. For all $u \in \mathbb{R}$,

$$\lim_{h \to 0} \frac{e^{i(u+h)X} - e^{iuX}}{h} = iXe^{iuX},$$

and since $|e^b - e^a| \leq |b - a|$, we have

$$\left| \frac{e^{i(u+h)X} - e^{iuX}}{h} \right| \leq |uX|.$$

As a consequence, the Dominated Convergence Theorem ensures that

$$\lim_{h \to 0} \mathbb{E} \left[ \frac{e^{i(u+h)X} - e^{iuX}}{h} \right] = \mathbb{E} [iXe^{iuX}],$$

that is to say that $\Phi_X$ is differentiable and $\Phi'_X(u) = i \mathbb{E}[X e^{iuX}]$. Using the dominated convergence again, it is easy to see that $\Phi'_X$ is continuous on $\mathbb{R}$. The extension to the argument to any $p \geq 1$ is straightforward, and the second point of the lemma is a direct application of Taylor’s formula.

Exercise 2.1.7 (Characteristic function of Gaussian random variables). Let $G \sim \mathcal{N}(0, 1)$. A reader familiar with the theory of distributions may check that a probability measure is a tempered distribution, which allows to define its Fourier transform.
1. Show that $\Phi_G$ is $C^1$ on $\mathbb{R}$, and that for all $u \in \mathbb{R}$, $\Phi'_G(u) + u\Phi_G(u) = 0$.
2. Deduce that $\Phi_G(u) = \exp(-u^2/2)$.
3. If $X \sim \mathcal{N}(\mu, \sigma^2)$, what is the expression of $\Phi_X(u)$?
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 are now ready to prove Theorem 2.1.3.

Proof of Theorem 2.1.3. For all $i \geq 1$, let $Y_i := X_i - \mathbb{E}[X_1]$, so that $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i^2] = \text{Var}(X_i) = \sigma^2$. We also denote

$$Z_n = \sqrt{n} (X_n - \mathbb{E}[X_1]) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,$$

so that, for all $u \in \mathbb{R}$,

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

where we have used the fact that the variables $Y_i$ are iid. By Lemma 2.1.6,

$$\Phi_{Y_1}(u/\sqrt{n}) = 1 - \frac{u^2 \sigma^2}{2n} + o \left( \frac{1}{n} \right)$$

when $n \to +\infty$. Using Lemma 2.1.8 below, we deduce that

$$\lim_{n \to +\infty} \Phi_{Z_n}(u) = \exp \left( -\frac{u^2 \sigma^2}{2} \right),$$

which by Exercise 2.1.7 is the characteristic function of $Z \sim \mathcal{N}(0, \sigma^2)$. As a consequence, Proposition 2.1.5 ensures that $Z_n$ converges in distribution to $Z$. \hfill \Box

In the proof of Theorem 2.1.3, we have used the following technical result.

Lemma 2.1.8 (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

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

We insist on the fact that this argument cannot be applied directly to $(1 + (\theta + \epsilon_n)/n)^n$ because $\epsilon_n$ is a complex number. We now compare both prelimits by writing,

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

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

and it is immediate that the sequence $\sup_{u \in [0,1]} |1 + (\theta + u\epsilon_n)/n|^{n-1}$ is bounded, which proves the claimed identity. \hfill \Box
Exercise 2.1.9 (Stronger convergence in the CLT). With the notation of the proof of Theorem 2.1.3, 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, by Remark 2.1.2, if such a variable exists, then necessarily $Z \sim \mathcal{N}(0, \sigma^2)$.

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?

### 2.2 Confidence intervals

Let us come back to the estimator $\hat{I}_n$, defined in (1.4), of the integral $I$ defined in (1.3). The Central Limit Theorem shows that, as soon as $X \in L^2(\mathbb{P})$, $\sqrt{n}(\hat{I}_n - I)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$, with $\sigma^2 = \text{Var}(X)$. We use this result to quantify the quality of the approximation of $I$ by $\hat{I}_n$ through the notion of (asymptotic) confidence interval.

**Definition 2.2.1** (Confidence interval). Let $\alpha \in (0, 1)$. A confidence interval for $I$ with level $1 - \alpha$ is an interval $[\hat{I}_n^-, \hat{I}_n^+]$ whose boundaries can be computed from the observation of $X_1, \ldots, X_n$ and such that

$$\lim_{n \to +\infty} \mathbb{P}(\hat{I}_n^- \leq I \leq \hat{I}_n^+) = 1 - \alpha.$$ 

The first step towards the construction of a confidence interval for $I$ is the following technical result on convergence in distribution. Given an interval $I$ of $\mathbb{R}$, we denote by $\partial I$ the boundaries of $I$; it is a set with at most two elements.

**Lemma 2.2.2** (Convergence in distribution and probabilities). Let $Z_n$ be a sequence of real-valued random variables which converges in distribution to $Z$. For any interval $I$ of $\mathbb{R}$ such that $\mathbb{P}(Z \in \partial I) = 0$,

$$\lim_{n \to +\infty} \mathbb{P}(Z_n \in I) = \mathbb{P}(Z \in I).$$

**Proof.** Let $a, b$ be the boundaries of $I$, we assume that $-\infty < a < b < +\infty$ in order not to overload the proof but the arguments can easily be adapted to other cases of interval. For all $M > 2/(b - a)$, let $\psi_M^-$ and $\psi_M^+$ be the functions defined by Figure 2.1. They satisfy

$$\forall z \in \mathbb{R}, \quad \psi_M^-(z) \leq 1_{\{z \leq I\}} \leq \psi_M^+(z),$$

so that, for all $n \geq 1$,

$$\mathbb{E}\left[\psi_M^-(Z_n)\right] \leq \mathbb{P}(Z_n \in I) \leq \mathbb{E}\left[\psi_M^+(Z_n)\right].$$

Since the functions $\psi_M^-$ and $\psi_M^+$ are continuous and bounded,

$$\mathbb{E}\left[\psi_M^-(Z)\right] \leq \liminf_{n \to +\infty} \mathbb{P}(Z_n \in I) \leq \limsup_{n \to +\infty} \mathbb{P}(Z_n \in I) \leq \mathbb{E}\left[\psi_M^+(Z)\right].$$

Besides, for any $z \in \mathbb{R} \setminus \{a, b\}$,

$$\lim_{M \to +\infty} \psi_M^-(z) = \lim_{M \to +\infty} \psi_M^+(z) = 1_{\{z \in I\}}.$$

Since $\mathbb{P}(Z \in \{a, b\}) = 0$, we deduce that

$$\lim_{M \to +\infty} \psi_M^-(Z) = \lim_{M \to +\infty} \psi_M^+(Z) = 1_{\{Z \in I\}}, \quad \text{almost surely},$$
and thus by the Dominated Convergence Theorem,
\[ P(Z \in I) \leq \lim \inf_{n \to +\infty} P(Z_n \in I) \leq \lim \sup_{n \to +\infty} P(Z_n \in I) \leq P(Z \in I), \]
from which we conclude that
\[ \lim_{n \to +\infty} P(Z_n \in I) = P(Z \in I). \]

Remark 2.2.3. If $Z$ has a density, then the condition that $P(Z \in \partial I) = 0$ is satisfied for any interval $I$ of $\mathbb{R}$.

Combining Lemma 2.2.2 and Remark 2.2.3 with the Central Limit Theorem, we deduce that for any $a < b$ in $\mathbb{R}$,
\[ \lim_{n \to +\infty} P \left( a \leq \frac{1}{\sqrt{n} \sigma^2} (\bar{X}_n - \hat{J}_n) \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_{a}^{b} \exp \left( -\frac{z^2}{2} \right) \, dz. \]
Recall that we denote by $\Phi$ the CDF of the $\mathcal{N}(0,1)$ distribution. For any $u \in (0,1)$, let us define the quantile of order $u$ of the law $\mathcal{N}(0,1)$ by
\[ \phi_u := \Phi^{-1}(u). \]

Exercise 2.2.4. Check that $\phi_{1-u} = -\phi_u$.

For $\alpha \in (0,1/2)$, we have
\[ \frac{1}{\sqrt{2\pi}} \int_{-\phi_{1-\alpha/2}}^{\phi_{1-\alpha/2}} \exp \left( -\frac{z^2}{2} \right) \, dz = 1 - \alpha, \]
see Figure 2.2.

As a consequence, the interval
\[ I_n = \left[ \hat{J}_n - \phi_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}}, \hat{J}_n + \phi_{1-\alpha/2} \sqrt{\frac{\sigma^2}{n}} \right] \]
satisfies the property that
\[ \lim_{n \to +\infty} P(J \in I_n) = 1 - \alpha, \]
but its boundaries depend on $\sigma^2 = \text{Var}(X)$. Yet, since we are trying to estimate the unknown quantity $\mathbb{E}[X]$, it is likely that $\text{Var}(X)$ be also unknown! A standard solution to this issue consists in replacing $\sigma^2$ with its estimator
\[ \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2, \]
which by Theorem 1.1.14 converges almost surely to $\sigma^2$. The validity of this approximation is ensured by the following result.

**Lemma 2.2.5** (Slutsky’s Theorem). Let $(Y_n)_{n \geq 1}$ and $(Z_n)_{n \geq 1}$ be two sequences of random variables in $\mathbb{R}$ such that:

(i) $Y_n$ converges in distribution to some deterministic value $y \in \mathbb{R}$;
(ii) $Z_n$ converges in distribution to some random variable $Z$ in $\mathbb{R}$.

Then the pair $(Y_n, Z_n)$ converges in distribution to $(y, Z)$.

**Proof.** We use the characterisation of the convergence in distribution by characteristic functions from Proposition 2.1.5. Let $X_n = (Y_n, Z_n)$. For all $u = (v, w) \in \mathbb{R}^2$,

$$
\Phi_{X_n}(u) = \mathbb{E} [\exp(i(uY_n + vZ_n))] \\
= \mathbb{E} [\exp(i(uY_n + vZ_n)) - \exp(i(uy + vZ_n))] + \mathbb{E} [\exp(i(uy + vZ_n))].
$$

Since the function $z \rightarrow \exp(i(uy + vz))$ is continuous and bounded, the convergence in distribution of $Z_n$ to $Z$ ensures that

$$
\lim_{n \to +\infty} \mathbb{E} [\exp(i(uy + vZ_n))] = \mathbb{E} [\exp(i(uy + vZ))] = \Phi_X(u), \quad X := (y, Z).
$$

On the other hand, for all $n \geq 1$,

$$
|\exp(i(uY_n + vZ_n)) - \exp(i(uy + vZ_n))| = |\exp(ivZ_n)(\exp(i(uY_n)) - \exp(i(uy)))| \\
\leq |\exp(i(uY_n)) - \exp(i(uy))|,
$$

and the right-hand side converges to 0 almost surely while being bounded from above by 2. By the Dominated Convergence Theorem, we deduce that

$$
\lim_{n \to +\infty} \mathbb{E} [\exp(i(uY_n + vZ_n)) - \exp(i(uy + vZ_n))] = 0,
$$

which shows that $\Phi_{X_n}(u)$ converges to $\Phi_X(u)$ and completes the proof. \qed

**Remark 2.2.6.** We formulate a few remarks about Lemma 2.2.5.

(i) By Remark 2.1.2, if $Y_n$ converges to $y$ almost surely (which will often be the case when we employ Slutsky’s Theorem), it also converges in distribution.

(ii) The convergence in distribution of $(Y_n, Z_n)$ implies the convergence in distribution of any continuous function of the pair $(Y_n, Z_n)$, in particular $Y_n + Z_n$ and $Y_n Z_n$. 

<table>
<thead>
<tr>
<th>$1 - \alpha$</th>
<th>$\phi_{1-\alpha/2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>1.65</td>
</tr>
<tr>
<td>95%</td>
<td>1.96</td>
</tr>
<tr>
<td>99%</td>
<td>2.58</td>
</tr>
</tbody>
</table>
(iii) Recall that such a statement as ‘\(X_n\) converges in distribution to \(X\)’ only means that the law of \(X_n\) converges to the law of \(X\). Therefore, if we take two sequences of random variables \(Y_n, Z_n\) which respectively converge to \(Y\) and \(Z\) in distribution, but we do not give any information on the joint law of the pair \((Y_n, Z_n)\), in general no statement can be made on the convergence in distribution of this pair. Slutsky’s Theorem is an exception to this rule. Another notable exception is the case where \(Y_n\) and \(Z_n\) are assumed to be independent for any \(n \geq 1\): in this case, if \(Y_n \rightarrow Y\) and \(Z_n \rightarrow Z\) in distribution, then \((Y_n, Z_n) \rightarrow (Y, Z)\) in distribution, where \(Y\) and \(Z\) are independent.

Since \(\sigma^2/\hat{\sigma}^2_n\) converges almost surely to 1 and \(\sqrt{n/\sigma^2(\hat{\sigma}^2_n - j)}\) converges in distribution to \(N(0, 1)\), we indeed deduce from Lemma 2.2.5 and Remark 2.2.6 that

\[
\sqrt{n \sigma^2_n} (\hat{\sigma}^2_n - j) = \sqrt{\frac{\sigma^2}{\hat{\sigma}^2_n}} \sqrt{n} (\hat{\sigma}^2_n - j) \rightarrow N(0, 1) \quad \text{in distribution},
\]

which leads to the following final statement.

**Corollary 2.2.7 (Confidence interval in the Monte Carlo method).** The interval

\[
[\hat{\sigma}^2_n - \phi_{1-\alpha/2} \sqrt{\frac{\sigma^2}{\hat{\sigma}^2_n}}, \hat{\sigma}^2_n + \phi_{1-\alpha/2} \sqrt{\frac{\sigma^2}{\hat{\sigma}^2_n}}]
\]

is a confidence interval with level \(1 - \alpha\) for \(\sigma^2\).

### 2.3 Variance reduction

Up to the error induced by the approximation of \(\sigma^2\) by \(\hat{\sigma}^2_n\), the length of the confidence interval obtained in Corollary 2.2.7 for \(\hat{\sigma}^2_n\) is

\[
\ell_n := 2\phi_{1-\alpha/2} \sqrt{\frac{\sigma^2}{\hat{\sigma}^2_n}}.
\]

Let \(G \sim N(0, 1)\). Assume that the quantity which we are trying to approximate is

\[
\hat{I} = \mathbb{P}(G \geq 20) = \mathbb{E}[X], \quad X := 1_{\{G \geq 20\}}.
\]

On the one hand, an upper bound on \(\hat{I}\) can be obtained analytically by writing

\[
\hat{I} = \frac{1}{\sqrt{2\pi}} \int_{y=20}^{+\infty} y^{-y^2/2} dy \leq \frac{1}{\sqrt{2\pi}} \int_{y=20}^{+\infty} \frac{y}{20} y^{-y^2/2} dy = \frac{e^{-20^2/2}}{20\sqrt{2\pi}} \approx 2.8 \times 10^{-89}.
\]

On the other hand, the Monte Carlo method suggests to draw iid realisations \(G_1, \ldots, G_n\) of \(N(0, 1)\), let \(X_i = 1_{\{G_i \geq 20\}}\) and approximate \(\hat{I}\) with

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

**Exercise 2.3.1.** What is the law of the random variable \(N = \inf\{n \geq 1 : \hat{I}_n \neq 0\}\)? What is its expectation?
Since $X_i \sim B(J)$, we have $\sigma^2 = \text{Var}(X_1) = J(1 - J) \simeq J$, so that the length of the Monte Carlo confidence interval writes

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

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

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

that is to say

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

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

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

### 2.3.1 Control variate

Assume that we want to estimate $J = E[X]$ thanks to iid samples $X_1, \ldots, X_n$. Assume also that we are able to sample iid random variables $Y_1, \ldots, Y_n$ whose common expectation $E[Y]$ is known analytically. Then, for all $\beta \in \mathbb{R}$,

$$J = E[X] = E[X - \beta Y] + \beta E[Y],$$

which suggests to approximate $J$ by the estimator

$$\hat{J}_{n}^{\text{CV,}\beta} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \beta Y_i) + \beta E[Y].$$

The variance of this estimator is $(\sigma^{\text{CV,}\beta})^2 / n$, where $(\sigma^{\text{CV,}\beta})^2 = \text{Var}(X - \beta Y)$. In order to study this quantity, we introduce the notion of covariance.

**Definition 2.3.2 (Covariance).** For all $X, Y \in L^2(P)$, the covariance between $X$ and $Y$ is defined by


**Remark 2.3.3.** The covariance has the following properties.

1. It is a symmetric, bilinear and nonnegative form on $L^2(P)$.
2. $\text{Cov}(X, X) = \text{Var}(X)$ and $\text{Var}(X + Y) = \text{Var}(X) + 2 \text{Cov}(X, Y) + \text{Var}(Y)$.
3. If $X$ and $Y$ are independent then $\text{Cov}(X, Y) = 0$, but the converse is not true in general$^3$.
4. By the Cauchy–Schwarz inequality, $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$.

When $\text{Var}(X)\text{Var}(Y) > 0$, we define the correlation coefficient between $X$ and $Y$ by

$$\rho_{X,Y} := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \in [-1, 1].$$

$^3$Take $X \sim \mathcal{N}(0, 1)$ and $Y = \epsilon X$ with $\epsilon$ independent from $X$ and $P(\epsilon = 1) = P(\epsilon = -1) = 1/2$. 
With these notions at hand, we may rewrite
\[(\sigma_{CV,\beta}^2)^2 = \text{Var}(X) - 2\beta \text{Cov}(X, Y) + \beta^2 \text{Var}(Y).\]

We may already remark that if \(\text{Cov}(X, Y) = 0\) then \((\sigma_{CV,\beta}^2)^2\) is always larger than the variance \(\text{Var}(X)\) associated with the original Monte Carlo estimator: for the control variate method to be efficient, it is thus necessary that \(X\) and \(Y\) be correlated. The choice of \(\beta\) for which \((\sigma_{CV,\beta}^2)^2\) is minimal is then
\[\beta^* = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)},\]

which yields the variance
\[(\sigma_{CV,\beta^*}^2)^2 = \text{Var}(X) \left(1 - \rho_{X,Y}^2\right).\]

As a consequence, the more \(X\) and \(Y\) are correlated, the better the variance reduction.

In practice, the optimal choice of \(\beta\) depends on the quantities \(\text{Cov}(X, Y)\) and \(\text{Var}(Y)\) which may need to be estimated. Let us introduce the following quantities:
\[
\tilde{\sigma}_{X,n}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \quad \tilde{\sigma}_{Y,n}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2, \quad \tilde{C}_{X,Y,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n).
\]

Theorem 1.1.14 shows that
\[\hat{\beta}^*_n := \frac{\tilde{C}_{X,Y,n}}{\tilde{\sigma}_{Y,n}^2}\]

converges to \(\beta^*\) almost surely, and Lemma 2.2.5 then allows to show the following result.

**Proposition 2.3.4** (Control variate method). Let \((X_i, Y_i)_{1 \leq i \leq n}\) be a sequence of iid pairs such that \(X_i, Y_i \in L^2(P)\) and \(\text{Var}(X_1) \cdot \text{Var}(Y_1) > 0\). For all \(n \geq 1\), let
\[\widehat{\mathcal{J}}_n := \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\beta}^*_n Y_i) + \hat{\beta}^*_n \mathbb{E}[Y],\]

with \(\hat{\beta}^*_n := \tilde{C}_{X,Y,n}/\tilde{\sigma}_{Y,n}^2\). The interval
\[
\left[\frac{\widehat{\mathcal{J}}_n}{1 - \frac{\sqrt{(\tilde{\sigma}_{n}^2 CV)^2}}{n}}, \frac{\widehat{\mathcal{J}}_n}{1 + \frac{\sqrt{(\tilde{\sigma}_{n}^2 CV)^2}}{n}}\right],
\]

where
\[(\tilde{\sigma}_{n}^2 CV)^2 = \tilde{\sigma}_{X,n}^2 \left(1 - \frac{\tilde{C}_{X,Y,n}}{\tilde{\sigma}_{X,n}^2 \tilde{\sigma}_{Y,n}^2}\right),\]

is a confidence interval with level \(1 - \alpha\) for \(\mathcal{J} = \mathbb{E}[X_1]\).

**Exercise 2.3.5.** Let \(G \sim \mathcal{N}(0, 1)\). For all \(\lambda > 0\), we define
\[I_\lambda = \mathbb{E} \left[\frac{1}{1 + \lambda G^2}\right] = \frac{1}{\sqrt{2\pi}} \int_{x \in \mathbb{R}} \frac{e^{-x^2/2}}{1 + \lambda x^2} \, dx.\]

Let \(G_1, \ldots, G_n\) be independent \(\mathcal{N}(0, 1)\) variables, and let \(X_i = 1/(1 + \lambda G_i^2), Y_i = 1 - \lambda G_i^2\).

1. Compute \(\mathbb{E}[Y_i]\).
2. Compare numerically the variances of the Monte Carlo estimator \(\widehat{\mathcal{J}}_n\) and of the control variate estimator \(\tilde{\sigma}_{n}^2 CV\).
3. How does this comparison vary with \(\lambda\)? What is your interpretation of this fact?
2.3.2 Importance sampling

In order to present the importance sampling method, it is more convenient to come back to the framework of Equations (1.1)–(1.2), where \( I \) writes \( \mathbb{E}[f(X)] \) with \( X \) a random variable in \( \mathbb{R}^d \) with density \( p \), and \( f \in L^1(P_X) \), rather than the simplified case of Equation (1.3).

Let \( q \) be a probability density on \( \mathbb{R}^d \) such that, \( dx \)-almost everywhere, if \( q(x) = 0 \) then \( p(x) = 0 \)— equivalently, \( q(x) > 0 \), \( P_X(dx) \)-almost everywhere. Then it is obvious that

\[
I = \int_{x \in \mathbb{R}^d} f(x) p(x) dx = \int_{y \in \mathbb{R}^d} f(y) \frac{p(y)}{q(y)} q(y) dy = \mathbb{E}[f(Y) w(Y)],
\]

where \( Y \) has density \( q \) and \( w \) is the likelihood ratio \( \frac{p}{q} \). As a consequence, the quantity

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

where \( Y_1, \ldots, Y_n \) are iid with density \( q \), converges almost surely to \( I \). The whole game of importance sampling then consists in choosing \( q \) in order to make the asymptotic variance

\[
(\sigma_{IS}^2) := \text{Var}(f(Y) w(Y))
\]
as small as possible.

**Lemma 2.3.6 (Optimal choice of \( q \)).** Assume that \( f \geq 0 \) and that \( I > 0 \). Let \( q \) be the probability density on \( \mathbb{R}^d \) defined by \( q(y) = f(y) p(y)/I \), and let \( w = p/q \). If \( Y \) is a random variable with density \( q \), then \( \text{Var}(f(Y) w(Y)) = 0 \).

**Proof.** This is trivial since for all \( y \in \mathbb{R}^d \), \( f(y) w(y) = I \). \(\square\)

Of course in practice it is impossible to implement the method with the optimal density \( q \) since the latter depends explicitly on the quantity \( I \) which is assumed to be unknown. Still, this lemma suggests that a ‘good’ choice of \( q \) would be one which gives a large probability to values of \( y \) which make the product \( f(y) p(y) \) large.

**Exercise 2.3.7.** For the example of the estimation of \( \mathbb{P}(G \geq 20) \), for \( G \sim \mathcal{N}(0, 1) \), 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 \( \epsilon = 0.01 \)?
Part II

Markov chains
Lecture 3

Markov chains in finite spaces

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

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

by the empirical mean of iid samples $f(X_1), \ldots, f(X_n)$. In Part I, the presentation was restricted to the case where $E = \mathbb{R}^d$ and $X$ has a density $p$, but it is clear that the principle of the method remains the same in general.

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

Throughout this part (except for Section 3.3), we fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let $E$ be a finite set, with cardinality $m \geq 1$, endowed with the discrete $\sigma$-field $\mathcal{E}$ which contains all $2^m$ subsets of $E$.

3.1 General definitions

3.1.1 Conditional probability

Definition 3.1.1 (Conditional probability). Let $A \in \mathcal{A}$ be an event such that $\mathbb{P}(A) > 0$. For all $B \in \mathcal{A}$, the conditional probability of $B$ given $A$ is

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

Notice that the mapping $\mathbb{P}(.|A)$ defines a new probability measure on $(\Omega, \mathcal{A})$, which can be seen as the ‘restriction’ (suitably renormalised) of $\mathbb{P}$ to $A$. The following formula is often useful.

Lemma 3.1.2 (Total probability formula). Let $I$ be a finite or countably infinite set of indices and let $(A_i)_{i \in I}$ be a partition of $\Omega$ into elements of $\mathcal{A}$. For any event $B$,

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i) \mathbb{P}(A_i),$$

where we take the convention that $\mathbb{P}(B|A_i) \mathbb{P}(A_i) = 0$ if $\mathbb{P}(A_i) = 0$. 
Proof. By \(\sigma\)-additivity,
\[
P(B) = P\left(\bigcup_{i \in I} B \cap A_i\right) = \sum_{i \in I} P(B \cap A_i),
\]
and either \(P(A_i) = 0\) in which case \(P(B \cap A_i) = 0 = P(B|A_i)P(A_i)\), or \(P(A_i) > 0\) in which case Definition 3.1.1 yields \(P(B \cap A_i) = P(B|A_i)P(A_i)\).

\(\square\)

Remark 3.1.3 (Conditional probability and independence). Two events \(A\) and \(B\) are called independent whenever the random variables \(\mathbb{1}_A\) and \(\mathbb{1}_B\) are independent in the sense of Definition 1.1.12. It is then easily checked that \(A\) and \(B\) are independent if and only if \(P(A \cap B) = P(A)P(B)\). In other words, when \(P(A) > 0\) and \(P(B) > 0\), the events \(A\) et \(B\) are independent if and only if \(P(A|B) = P(A)\) or equivalently \(P(B|A) = P(B)\): conditioning by one of the events does not modify the probability of the other event.

### 3.1.2 Markov chain and the Markov property

Loosely speaking, a Markov chain is a sequence of random variables \((X_n)_{n \geq 0}\) with values in the finite set \(E\), such that at each step \(n \geq 0\), if \(X_n = x\) then the next state \(y\) for \(X_{n+1}\) is chosen randomly, with probability \(P(x, y)\). This is translated in more formal terms in Definition 3.1.7 below. We first introduce the notion of stochastic matrix.

Definition 3.1.4 (Stochastic matrix). A stochastic matrix on \(E\) is an \(m \times m\) matrix \(P\) with coefficients \(P(x, y)\), \(x, y \in E\) which satisfy:

(i) for all \(x, y \in E\), \(P(x, y) \geq 0\);
(ii) for all \(x \in E\), \(\sum_{y \in E} P(x, y) = 1\).

In other words, each row of the matrix \(P\) represents a probability measure \(P(x, \cdot)\) on \(E\). For this reason, we shall take the convention to identify measures on \(E\) with row vectors of \(\mathbb{R}^m\), and dually, functions from \(E\) to \(\mathbb{R}\) will be identified with column vectors of \(\mathbb{R}^m\).

Exercise 3.1.5 (Expectation). Let \(\mu\) be a probability measure on \(E\), \(X\) be a random variable in \(E\) with distribution \(\mu\) and \(f : E \to \mathbb{R}\). Check that \(f \in L^1(\mu)\) and that \(\mathbb{E}[f(X)] = \mu f\).

We also denote by \(\mathbf{1}\) the column vector of which all coordinates are equal to 1, and recall that the set of probability measures on \(E\) is denoted by \(\mathcal{P}(E)\). Since \(E\) is a finite set with cardinality \(m\), \(\mathcal{P}(E)\) can be identified with the simplex \(\{\mu \in [0, 1]^m : \mu \mathbf{1} = 1\}\).

Exercise 3.1.6 (Properties of stochastic matrices). Let \(P\) be a stochastic matrix.

1. Show that \(P \mathbf{1} = \mathbf{1}\).
2. Show that, for any \(\mu \in \mathcal{P}(E)\), \(\mu P \in \mathcal{P}(E)\).
3. Show that, for any stochastic matrix \(Q\) on \(E\), \(PQ\) remains a stochastic matrix.

We may now introduce the notion of Markov chain.

Definition 3.1.7 (Markov chain). Let \(P\) be a stochastic matrix. A sequence of random variables \((X_n)_{n \geq 0}\) in \(E\) is called a Markov chain with transition matrix \(P\) if, for all \(n \geq 0\), for any \(x_0, \ldots, x_n \in E\) such that \(P(X_0 = x_0, \ldots, X_n = x_n) > 0\), for all \(x_{n+1} \in E\),

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

(3.1)
Equation (3.1) is called the Markov property. It expresses the fact that the law of the future value $X_{n+1}$ only depends on the past trajectory $X_0, \ldots, X_n$ through the current state $X_n$.

Example 3.1.8 (The Ehrenfest urn). Consider a box divided into two compartments, called A and B, and which contains $N$ particles, see Figure 3.1. At each step, one particle is chosen uniformly at random and moved to the other compartment. There are at least two ways to describe this dynamics.

![Figure 3.1: The Ehrenfest urn with $N = 10$ particles.](image)

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

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

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

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

and the other coefficients are 0.

From Definition 3.1.7 we deduce the following properties related with the law of the sequence $(X_n)_{n \geq 0}$.

Proposition 3.1.9 (Marginal distributions of a Markov chain). Let $(X_n)_{n \geq 0}$ be a Markov chain with transition matrix $P$. For all $n \geq 0$, let $\mu_n \in \mathcal{P}(E)$ denote the law of the random variable $X_n$.

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

$$\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \mu_0(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n). \quad (3.2)$$

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

Before detailing the proof of Proposition 3.1.9, we emphasise a few of its consequences.

Remark 3.1.10. (i) The first assertion shows that the law of any vector $(X_0, \ldots, X_n)$ is entirely characterised by two objects: the initial distribution $\mu_0$ and the transition matrix $P$.

(ii) The second assertion immediately yields the identity $\mu_n = \mu_0 P^n$. In particular, applying the result of Exercise 3.1.5 shows that for any $f : E \to \mathbb{R}$, $\mathbb{E}[f(X_n)] = \mu_0 P^n f$. 


Proof of Proposition 3.1.9. We prove the first assertion by induction on \( n \geq 0 \). For \( n = 0 \) this is immediate. Let \( n \geq 0 \) be such that (3.2) holds, and let \( x_0, \ldots, x_n, x_{n+1} \in E \). If \( \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = 0 \), then on the one hand the fact that \( \{X_0 = x_0, \ldots, X_{n+1} = x_{n+1}\} \subset \{X_0 = x_0, \ldots, X_n = x_n\} \) ensures that the former event has also probability 0, while on the other hand the identity (3.2) implies that \( \mu_0(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) = 0 \) and therefore this quantity remains 0 when multiplied by \( P(x_n, x_{n+1}) \). If \( \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0 \), Definition 3.1.11 allows to write

\[
\mathbb{P}(X_0 = x_0, \ldots, X_{n+1} = x_{n+1}) = \mathbb{P}(X_{n+1} = x_{n+1}|X_0 = x_0, \ldots, X_n = x_n)\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \mu_0(x_0)P(x_0, x_1) \cdots P(x_n, x_{n+1}),
\]
where we have used (3.1) and (3.2) at the last line.

The second assertion follows from the computation

\[
\mathbb{P}(X_{n+1} = y) = \sum_{x \in E} \mathbb{P}(X_{n+1} = y|X_n = x)\mathbb{P}(X_n = x) = \sum_{x \in E} P(x, y)\mu_n(x) = \mu_nP(y),
\]
in which we have used Lemma 3.1.2 and (3.1). \( \square \)

Remark 3.1.11 (* Kolmogorov and Fokker–Planck equations, Feynman–Kac formula). Under the assumptions of Proposition 3.1.9, let \( L = P - I \) where \( I \) the identity on \( E \).

(i) The identity

\[
\mu_{n+1} - \mu_n = \mu_nL
\]

is a discrete version of the Fokker–Planck equation which will be seen in Lecture ??.

(ii) Likewise, fixing \( N \geq 1 \), \( f : E \to \mathbb{R} \), and letting, for all \( n \in \{0, \ldots, N\} \) and \( x \in E \),

\[
u_n(x) := \mathbb{E}[f(X_N)|X_n = x] := \sum_{y \in E} f(y)\mathbb{P}(X_N = y|X_n = x),
\]

it is easily seen that

\[
\begin{cases}
u_{n+1} - \nu_n + Lu_{n+1} = 0, & 0 \leq n < N, \\
u_N = f,
\end{cases}
\]

which is a discrete version of the Feynman–Kac formula.

(iii) Last, recalling that \( P^n(x, y) = \mathbb{P}(X_n = y|X_0 = x) \), we get that the obvious identities

\[
P^{n+1} - P^n = LP^n = P^nL,
\]

respectively correspond to the backward and forward Kolmogorov equations.

Exercise 3.1.12 (Random mapping representation). Let \((F, \mathcal{F})\) be a measurable space (not necessarily finite, nor even countably infinite) and \( h : E \times F \to E \) be a measurable function. Let \( X_0 \) be a random variable in \( E \) with law \( \mu_0 \), and \( (U_n)_{n \geq 1} \) be a sequence of iid random variables in \( F \), independent from \( X_0 \). Consider the sequence \((X_n)_{n \geq 0}\) defined by the recursive relation

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

1. Check that \((X_n)_{n \geq 0}\) is a Markov chain and write its transition matrix.
2. Conversely, show that any Markov chain in $E$ admits such a representation\(^1\). Is this representation unique?

In practice we shall often compare Markov chains with the same transition matrix $P$ but different initial distributions. It will then be helpful to use the notation $P_\pi$, $E_\pi$, $\ldots$ to emphasise the fact that the initial distribution of the chain is $\mu$. When this initial distribution is a Dirac distribution at some $x \in E$ (that is to say that $X_0 = x$ almost surely), we shall write $P_x$, $E_x$, $\ldots$ rather than $P_\delta_x$, $E_\delta_x$, $\ldots$. As an example, we may observe from Proposition 3.1.9 and Remark 3.1.10 that when $X_0 = x$, the law of $X_n$ is related with the $n$-th power of $P$ by the identity

$$\forall x, y \in E, \quad P_x(X_n = y) = P^n(x, y).$$

Exercise 3.1.13 (A reformulation of the Markov property). Let $n, m \geq 0$, $F : E^n \to \mathbb{R}$ and $G : E^m \to \mathbb{R}$. For all $x, x' \in E$, show that

$$E_x \left[ F(X_1, \ldots, X_n) 1_{\{X_n = x'\}} G(X_{n+1}, \ldots, X_{n+m}) \right] = E_x \left[ F(X_1, \ldots, X_n) 1_{\{X_n = x\}} \right] E_{x'} \left[ G(X_1, \ldots, X_m) \right].$$

### 3.2 Stationary distributions

The first step to establish a connection between Markov chains and the Monte Carlo method is the notion of *stationary distribution*.

#### 3.2.1 Definition

**Definition 3.2.1** (Stationary distribution). Let $(X_n)_{n \geq 0}$ be a Markov chain in $E$ with transition matrix $P$. A probability measure $\pi$ on $E$ is called a stationary distribution for $(X_n)_{n \geq 0}$ if it satisfies

$$\pi P = \pi.$$

The denomination ‘stationary’ comes from the following result.

**Proposition 3.2.2** (Stationary distribution). Let $\pi$ be a stationary distribution for $(X_n)_{n \geq 0}$. For any $n \geq 0$,

$$\forall x \in E, \quad P_\pi(X_n = x) = \pi(x);$$

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

**Proof.** It is a straightforward consequence of the second assertion of Remark 3.1.10. \(\square\)

Exercise 3.2.3. Show that if $\pi$ is a stationary distribution for $(X_n)_{n \geq 0}$, then the whole sequence is actually stationary in the sense that for any $k \geq 0$ and $n \geq 0$, the vectors $(X_0, \ldots, X_n)$ and $(X_k, \ldots, X_{k+n})$ have the same distribution under $P_\pi$.

Exercise 3.2.4. Consider the Ehrenfest urn from Example 3.1.8.

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

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

---

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

\(^2\)Exercise 3.2.2 is a reformulation of the Markov property.
3.2.2 Existence of a stationary distribution

**Proposition 3.2.5** (Existence of stationary distribution). *Every Markov chain in* $E$ *admits at least one stationary distribution.*

We provide two different proofs of Proposition 3.2.5, both of which rely on the observation that the set $\mathcal{P}(E)$ is a convex and compact subset of $\mathbb{R}^m$.

*Proof by Brouwer’s Fixed Point Theorem.* By Exercise 3.1.6, the mapping $\mu \mapsto \mu P$ is continuous from $\mathcal{P}(E)$ to $\mathcal{P}(E)$. As a consequence, Brouwer’s Fixed Point Theorem ensures that it admits a fixed point in $\mathcal{P}(E)$.

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

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

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

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

By Exercise 3.1.6 and convexity, $\hat{\mu}_n \in \mathcal{P}(E)$ and by compactness, there exists an increasing sequence $(n_\ell)_{\ell \geq 1}$ such that $\hat{\mu}_{n_\ell}$ converges to some $\pi \in \mathcal{P}(E)$ when $\ell \to +\infty$. Since, for all $\ell \geq 1$,

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

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

*Remark 3.2.6.* The second proof shows that if the stationary distribution $\pi$ is unique, then the sequence $\hat{P}_n$ converges to $\pi$, whatever the initial distribution $\mu$.

3.2.3 Irreducibility and uniqueness of stationary distributions

What may prevent a stationary distribution from being unique? Let $E_1, E_2$ be two disjoint subsets of the finite space $E$, $P_1$ and $P_2$ be stochastic matrices respectively defined on $E_1$ and $E_2$, and let $\pi_1, \pi_2$ be some associated stationary distributions. On the space $E' = E_1 \cup E_2$, define the stochastic matrix $P'$ by the block decomposition

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

similarly, define the probability measures

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

Then it is clear that both $\pi'_1$ and $\pi'_2$ (and as a consequence, all their convex combinations) are stationary distributions for $P'$. Observe that in this situation, if the chain starts at some point $x \in E_1$, then $X_n$ will remain in $E_1$ for all $n \geq 1$, see also Figure 3.2.

This remark motivates the following definition.
3.2 Stationary distributions

Figure 3.2: On the space \( E = \{1, 2, 3, 4, 5\} \), arrows represent the possible moves of the Markov chain. Clearly, a chain started in \( E_1 = \{1, 2, 3\} \) can never go to \( E_2 = \{4, 5\} \).

**Definition 3.2.7** (Irreducibility). A Markov chain with transition matrix \( P \) on \( E \) is called irreducible if, for all \( x, y \in E \), there exist \( n \geq 1 \) and \( x = x_0, x_1, \ldots, x_n = y \) such that

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

We shall also say that the matrix \( P \) is irreducible.

**Exercise 3.2.8.** Check that the condition that there exist \( x = x_0, x_1, \ldots, x_n = y \) such that \( P(x_0, x_1) \cdots P(x_{n-1}, x_n) > 0 \) is equivalent to \( P_x(X_n = y) > 0 \).

**Proposition 3.2.9** (Uniqueness of a stationary distribution). If the stochastic matrix \( P \) is irreducible, then it possesses a unique stationary distribution.

We start the proof of Proposition 3.2.9 with the following exercise.

**Exercise 3.2.10.** Let \( P \) be an irreducible stochastic matrix, and let \( \pi \) be an associated stationary distribution. Show that for all \( x \in E \), \( \pi(x) > 0 \).

We now introduce a useful object.

**Definition 3.2.11** (Dirichlet form). Let \( P \) be a stochastic matrix, and let \( \pi \) be an associated stationary distribution. The Dirichlet form of \((P, \pi)\) is the quadratic form \( \mathcal{E}_\pi \) defined on \( \mathbb{R}^E \) by

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

**Lemma 3.2.12** (Another expression for \( \mathcal{E}_\pi \)). For all \( f \in \mathbb{R}^E \),

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

**Proof.** From Definition 3.2.11, we write

\[
\mathcal{E}_\pi(f) = \frac{1}{2} \sum_{x,y \in E} \left( f(y)^2 - 2f(y)f(x) + f(x)^2 \right) \pi(x) P(x, y)
= \frac{1}{2} \sum_{y \in E} f(y)^2 \pi P(y) - \sum_{x \in E} f(x) \pi(x) Pf(x) + \frac{1}{2} \sum_{x \in E} f(x)^2 \pi(x),
\]

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

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

which is the expected expression. \(\square\)
Lemma 3.2.13 (Dimension of $\ker(P - I)$ for irreducible matrices). Let $P$ be an irreducible stochastic matrix. For all $f \in \mathbb{R}^E$, if $Pf = f$ then there exists $c \in \mathbb{R}$ such that $f = cf$.

Proof. Let $\pi$ be a stationary distribution for $P$. If $Pf = f$, then Lemma 3.2.12 immediately shows that $E_\pi(f) = 0$, therefore by Definition 3.2.11 and Exercise 3.2.10, $f(x) = f(y)$ for all pairs $(x, y)$ such that $P(x, y) > 0$. We now take arbitrary $x, y \in E$ and let $x = x_0, x_1, \ldots, x_n = y$ be given by Definition 3.2.7. From this definition, $P(x_i, x_{i+1}) > 0$ for all $i = 0, \ldots, n - 1$, so that by the argument above, $f(x_0) = \cdots = f(x_n)$ and thus $f$ is a constant function on $E$.

To complete the proof of Proposition 3.2.9 we shall need the following elementary result from linear algebra.

Exercise 3.2.14. Let $A \in \mathbb{R}^{m \times m}$.

1. Show that $(\text{im} A)^ot = \ker A^\top$.
2. Deduce that the spaces $\{u \in \mathbb{R}^m : uA = 0\}$ and $\{v \in \mathbb{R}^m : Av = 0\}$ have the same dimension. Hint: think of the Rank-Nullity Theorem.

Proof of Proposition 3.2.9. Lemma 3.2.13 shows that the kernel (for the right multiplication) of $P - I$ has dimension 1. Since any stationary distribution for $P$ is in the kernel (for the left multiplication) of $P - I$, and that any two distinct probability measures cannot be collinear, we deduce that there is at most one such distribution.

Exercise 3.2.15 (The coupon collector). A brand of chocolate eggs hides surprise gifts in each egg. There are $N$ different models of gifts, each of which is equally likely to be hidden in a given egg. We denote by $X_n \in \{0, \ldots, N\}$ the number of different gifts that you have collected after eating $n$ eggs, and $\tau_N = \inf\{n \geq 0 : X_n = N\}$ the time at which you have found all eggs.

1. Show that $(X_n)_{n \geq 0}$ is a Markov chain and write its transition matrix.
2. Is this chain irreducible?
3. Describe the set of its stationary distributions.
4. Compute $E_0[\tau_N]$ and give an equivalent of this quantity when $N \to +\infty$. Hint: define $\eta_0 = 0$ and, for $i \in \{1, \ldots, N\}$, $\eta_i = \inf\{n \geq 1 : X_{n-1} + n = i\}$. How to express $\tau_N$ in terms of $\eta_1, \ldots, \eta_N$? What is the law of each $\eta_i$?
5. Show that, for any $c > 0$, $P(\tau_N > [N \ln N + cN]) \leq e^{-c}$. Hint: for $i \in \{1, \ldots, N\}$ and $k \geq 1$, introduce the event $A^k_i = \{\text{no gift of the } i\text{-th type has been found in the first } k \text{ eggs}\}$.

3.3 Extension to general state spaces

3.3.1 Countably infinite state spaces

Let $E$ be a countably infinite space, endowed with the discrete $\sigma$-field $\mathcal{E}$. We shall still call stochastic matrix a collection of numbers $\{P(x, y) : x, y \in E\}$ satisfying the conditions of Definition 3.1.4 and keep using matrix-product-like notation such as $\mu P$, $Pf$. The definition of a Markov chain remains unchanged.

Example 3.3.1 (Simple random walk on $\mathbb{Z}^d$). The simple random walk on $\mathbb{Z}^d$ is the random sequence $(X_n)_{n \geq 0}$ which at each step picks up its next state uniformly among its neighbours. More precisely, it is the Markov chain in $\mathbb{Z}^d$ with transition matrix

$$P(x, y) = \begin{cases} \frac{1}{2d} & \text{if } |x - y| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

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

The major difference with the theory of Markov chains in finite spaces is that a stationary distribution need not exist.

**Exercise 3.3.2.** Show that a stationary distribution \( \pi \) for the simple random walk on \( \mathbb{Z} \) necessarily satisfies \( \pi(x + 1) = \pi(x) \) for all \( x \in \mathbb{Z} \) and therefore cannot exist. Generalise the argument to \( \mathbb{Z}^d \) for any \( d \geq 1 \).

**Remark 3.3.3 (Invariant unbounded measure).** Let \( P \) denote the transition matrix of the simple random walk on \( \mathbb{Z}^d \). It is obvious that the unbounded measure \( \pi \) defined on \( \mathbb{Z}^d \) by \( \pi(x) = 1 \) for all \( x \in \mathbb{Z}^d \) satisfies \( \pi P = \pi \), but this measure cannot be normalised to a probability distribution. Still, it is sometimes useful to call \( \pi \) an invariant (unbounded) measure for \( P \).

The existence of stationary distributions is related with the tail of the return times

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

For a given stochastic matrix \( P \) on \( E \), a state \( x \in E \) is called **transient** if \( P_x(\tau_x = +\infty) > 0 \) and **recurrent** if \( P_x(\tau_x = +\infty) = 0 \). Furthermore, recurrent states are called **null** if \( \mathbb{E}_x[\tau_x] = +\infty \) and **positive** if \( \mathbb{E}_x[\tau_x] < +\infty \). If \( P \) is irreducible, then all states have the same nature and being transient, null recurrent or positive recurrent becomes a property of the chain. Anticipating on the ergodic theory of Markov chains which will be detailed in Lecture 4 for the case where \( E \) is finite (in which all irreducible Markov chains are necessarily positive recurrent), let us assert that an irreducible Markov chain always satisfies the property that

\[ \forall x \in E, \quad \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1\{X_i = x\} = \frac{1}{\mathbb{E}_x[\tau_x]}, \quad \text{almost surely}. \]

If the chain is transient or null recurrent, the right-hand side is 0. In contrast, if the chain is positive recurrent, then setting

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

defines a probability measure \( \pi \) on \( E \) such that \( \pi P = \pi \), and it is the unique stationary distribution for \( P \). We refer to [2, 7] for more detail.

**Exercise 3.3.4 (Recurrence and transience of the simple random walk).** The purpose of this exercise is to show that for \( d \in \{1, 2\} \), the random walk on \( \mathbb{Z}^d \) is null recurrent, while for \( d \geq 3 \) it is transient.

1. Show that an irreducible Markov chain \( (X_n)_{n \geq 0} \) is transient if and only if \( \sum_{n=1}^{+\infty} P_x(X_n = x) < +\infty \) for any \( x \in E \).
2. For the simple random walk in dimension \( d = 1 \), compute \( P_0(X_n = 0) \) and conclude.
3. We let \( d = 2 \) and denote by \( (X^1_n, X^2_n) \) the coordinates of the simple random walk.
   (a) Let \( U_n = (X^1_{n+1} - X^1_n) + (X^2_{n+1} - X^2_n) \) and \( V_n = (X^1_{n+1} - X^1_n) - (X^2_{n+1} - X^2_n) \).
   Write the law of the pair \((U_n, V_n)\) and show that the sequence \((U_n, V_n)_{n \geq 0}\) is iid.
   (b) For \( n \) odd, what is the value of \( P_0(X_n = 0) \)?
   (c) For \( n \) even, express the event \( \{X_n = 0\} \) in terms of \( \sum_{k=0}^{n-1} U_k + \sum_{k=0}^{n-1} V_k \) and deduce the value of \( P_0(X_n = 0) \).
   (d) Conclude.
4. We now assume that \( d \geq 3 \) and denote by \( \varphi \) the characteristic function of \( X_1 \) under \( P_0 \), defined by

\[ \forall u \in \mathbb{R}^d, \quad \varphi(u) := \mathbb{E}_0\left[e^{i(u \cdot X_1)}\right]. \]
(a) Show that, for all \( u = (u_1, \ldots, u_d) \in \mathbb{R}^d \),
\[ \varphi(u) = \frac{1}{d} (\cos u_1 + \cdots + \cos u_d) . \]

(b) Show that
\[ \sum_{k=0}^{+\infty} \mathbb{P}_0(X_{2k} = 0) = \frac{1}{(2\pi)^d} \int_{u \in (-\pi, \pi)^d} \frac{du}{1 - \varphi^2(u)} . \]

(c) Conclude.

### 3.3.2 Continuous state spaces

For continuous state spaces, things become more delicate already at the level of the definition of a Markov chain. Indeed, since we expect to be able to address random variables with densities, events of the form \( \{ X_n = x_n \} \) will generally have probability 0 and will therefore not be appropriate to define Markov chains.

Let \((E, \mathcal{E})\) be a measurable space. The starting point of the theory is the following suitable extension of the notion of stochastic matrix.

**Definition 3.3.5 (Stochastic kernel).** A stochastic kernel on \( E \) is a function \( P : E \times E \to [0, 1] \) such that:

(i) for all \( x \in E \), the function \( C \in E \mapsto P(x, C) \) is a probability measure;

(ii) for all \( C \in \mathcal{E} \), the function \( x \in E \mapsto P(x, C) \) is measurable.

A sequence of random variables \( (X_n)_{n \geq 0} \) in \( E \) is then called a Markov chain with transition kernel \( P \) and initial distribution \( \mu \) if, for all \( n \geq 0 \) and \( C_0, \ldots, C_n \in \mathcal{E} \),
\[ \mathbb{P}(X_0 \in C_0, \ldots, X_n \in C_n) = \int x_0 \in C_0, \ldots, x_n \in C_n \mu(dx_0)P(x_0, dx_1) \cdots P(x_{n-1}, dx_n) . \]

It then becomes a conventional notation to write
\[ \mathbb{P}(X_{n+1} \in C|X_n = x) = P(x, C) , \]
even when the event \( \{X_n = x\} \) has probability 0, so that the conditional probability in the left-hand side is not well-defined in the sense of Subsection 3.1.1. Hence, the stochastic kernel \( P \) has the straightforward interpretation that, once in the state \( x \in E \), the chain picks up its next state according to the probability measure \( P(x, \cdot) \) on \( E \).

A stationary distribution for a stochastic kernel \( P \) is a probability measure \( \mu \) on \( E \) such that, for any \( C \in \mathcal{E} \),
\[ \mu(C) = \int x \in E \mu(dx)P(x, C) . \]

**Exercise 3.3.6.** Let \( (\xi_n)_{n \geq 1} \) be a sequence of iid random variables with distribution \( \mathcal{N}(0, 1) \). Let \( X_0 \) be a random variable in \( \mathbb{R} \) with distribution \( \mu \), independent from the sequence \( (\xi_n)_{n \geq 1} \). Let us define the sequence \( (X_n)_{n \geq 0} \) by
\[ X_n = \alpha X_{n-1} + \xi_n , \]
where \( \alpha \in (0, 1) \).

1. Show that \( (X_n)_{n \geq 0} \) is a Markov chain and compute its transition kernel.
2. Find \( \sigma^2 > 0 \) such that \( \mathcal{N}(0, \sigma^2) \) is a stationary distribution for this chain.

In general, both the notions of irreducibility and recurrence, which are respectively related with the uniqueness and existence of stationary distributions, need to be adapted to the continuous state space framework. We do not enter into the details of this theory here and refer the interested reader to \([7, 6, 14]\).
Lecture 4

Ergodic theory of Markov chains

Let \((X_n)_{n \geq 0}\) be an irreducible Markov chain on the finite space \(E\). We saw in Lecture 3 that this chain possesses a unique stationary distribution \(\pi\), and for all \(f : E \to \mathbb{R}\), the expectation of the random variable \(\frac{1}{n} \sum_{k=0}^{n-1} f(X_k)\) converges to \(\pi f\), which establishes a first connection between \(\pi\) and the long time behaviour of \((X_n)_{n \geq 0}\). In this lecture, this connection is investigated in more detail.

Throughout this lecture, the space \(E\) is still assumed to be finite, with cardinality \(m\).

4.1 Extensions of the Law of Large Numbers and the Central Limit Theorem

In this section, we generalise the Law of Large Numbers and the Central Limit Theorem to Markov chains.

4.1.1 Law of Large Numbers for Markov chains

Let \((X_n)_{n \geq 0}\) be a Markov chain in \(E\). For all \(x \in E\), let us define the return time to \(x\) by

\[\tau_x := \inf\{n \geq 1 : X_n = x\}\]

Lemma 4.1.1 (Finiteness of the expected return time). If the chain \((X_n)_{n \geq 0}\) is irreducible, then for any \(x \in E\), \(\mathbb{E}_x[\tau_x] < +\infty\).

With the vocabulary introduced in Subsection 3.3.1, the chain \((X_n)_{n \geq 0}\) is positive recurrent.

Proof. Let \(x \in E\). For all \(x' \in E\), Definition 3.2.7 implies that there exists \(n_{x'} \geq 1\) such that \(P^{n_{x'}}(x', x) > 0\). Let

\[\kappa := \min_{x' \in E} P_{x'}^{n_{x'}}(x', x) > 0, \quad m := \max_{x' \in E} n_{x'} \geq 1,\]

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

\[\kappa \leq P_{x'}^{n_{x'}}(x', x) = P_{x'}(X_{n_{x'}} = x) \leq P_{x'}(\tau_x \leq n_{x'}) \leq P_{x'}(\tau_x \leq m).\]

Hence, for any \(\ell \geq 1\),

\[\mathbb{P}_x(\tau_x > \ell m) = \mathbb{P}(X_1 \neq x, \ldots, X_{\ell m} \neq x) = \sum_{x' \neq x} \mathbb{E}_x \left[ F(X_1, \ldots, X_{(\ell-1)m}) \mathbb{1}_{\{X_{(\ell-1)m} = x'\}} G(X_{(\ell-1)m+1}, \ldots, X_{\ell m}) \right],\]
where

\[ F(x_1, \ldots, x_{(\ell-1)m}) = \mathbb{I}_{\{x_1 \neq \ldots, x_{(\ell-1)m} \neq x\}}, \quad G(x_1, \ldots, x_m) = \mathbb{I}_{\{x_1 \neq \ldots, x_m \neq x\}}. \]

Using Exercise 3.1.13, we get

\[
\mathbb{E}_x \left[ F(X_1, \ldots, X_{(\ell-1)m}) \mathbb{I}_{\{X_{(\ell-1)m} = x'\}} G(X_{(\ell-1)m+1}, \ldots, X_{\ell m}) \right] \\
= \mathbb{E}_x \left[ F(X_1, \ldots, X_{(\ell-1)m}) \mathbb{I}_{\{X_{(\ell-1)m} = x'\}} \mathbb{E}_{x'} [G(X_1, \ldots, X_m)] \right] \\
= \mathbb{P}_x (X_1 \neq \ldots, X_{(\ell-1)m-1} \neq x, X_{(\ell-1)m} = x') \mathbb{P}_x (X_1 \neq \ldots, X_m \neq x) \\
= \mathbb{P}_x (\tau_x > (\ell - 1)m, X_{(\ell-1)m} = x') \mathbb{P}_x' (\tau_x > m) \\
\leq (1 - \kappa) \mathbb{P}_x (\tau_x > (\ell - 1)m, X_{(\ell-1)m} = x'),
\]

so that

\[ \mathbb{P}_x (\tau_x > \ell m) \leq (1 - \kappa) \mathbb{P}_x (\tau > (\ell - 1)m), \]

and thus

\[ \mathbb{P}_x (\tau_x > \ell m) \leq (1 - \kappa)^{\ell}. \]

We complete the proof by remarking that, by the Fubini–Tonelli Theorem,

\[
\mathbb{E}_x [\tau_x] = \mathbb{E}_x \left[ \sum_{n=0}^{+\infty} \mathbb{I}_{\{n < \tau_x\}} \right] \\
= \sum_{n=0}^{+\infty} \mathbb{P}_x (\tau_x > n) \\
= \sum_{\ell=0}^{+\infty} \sum_{k=0}^{m-1} \mathbb{P}_x (\tau_x > \ell m + k) \\
\leq \sum_{\ell=0}^{+\infty} m \mathbb{P}_x (\tau_x > \ell m) \\
\leq \sum_{\ell=0}^{+\infty} m (1 - \kappa)^{\ell} \\
= \frac{m}{\kappa}. \quad \Box
\]

\[ \textbf{Exercise 4.1.2 (Exponential moments).} \] Under the assumptions of Lemma 4.1.1, show that there exists \( \epsilon > 0 \) such that \( \mathbb{E}_x [\exp(\epsilon \tau_x)] < +\infty \). Deduce that for all \( p \geq 1 \), \( \mathbb{E}_x [(\tau_x)^p] < +\infty \).

Let \( x \) be the starting point of the chain \( (X_n)_{n \geq 0} \). If the latter is irreducible, then Lemma 4.1.1 allows to define the sequence \( (\tau_x^\ell)_{\ell \geq 0} \) by \( \tau_x^0 = 0 \) and

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

Then we shall see in Lemma 4.1.4 that this sequence divides the trajectory \( (X_n)_{n \geq 0} \) into excursions \( (X_{\ell}^0, \ldots, X_{\ell+1}^0), (X_{\ell}^1, \ldots, X_{\ell+1}^1), \ldots \), which are independent and identically distributed and to which we may thus apply the Law of Large Numbers to obtain the following result. Notice that it is obvious that \( \tau_x^\ell \to +\infty \), almost surely, when \( \ell \to +\infty \).
4.1 Extensions of the Law of Large Numbers and the Central Limit Theorem

**Theorem 4.1.3 (Ergodic Theorem).** Let \((X_n)_{n \geq 0}\) be an irreducible Markov chain, the unique stationary distribution of which is denoted by \(\pi\). For all \(f : E \rightarrow \mathbb{R}\),

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

Let us emphasise the fact that the result of Theorem 4.1.3 holds whatever the initial distribution \(\mu_0\) of \(X_0\).

The essential ingredient in the proof of Theorem 4.1.3 is the following result.

**Lemma 4.1.4 (Strong Markov property I).** Under the assumptions of Theorem 4.1.3, for any \(f : E \rightarrow \mathbb{R}\) the random variables \((Z_\ell)_{\ell \geq 0}\) defined by

\[
Z_\ell := \sum_{i=\tau_\ell}^{\tau_{\ell+1}-1} f(X_i)
\]

are iid under \(\mathbb{P}_x\).

**Proof.** For the sake of simplicity, we only prove that \(Z_0\) and \(Z_1\) are independent and have the same law, but the ideas remain the same for an arbitrary number of variables. Let \(C_0, C_1\) be Borel sets of \(\mathbb{R}\). We first write

\[
\mathbb{P}_x(Z_0 \in C_0, Z_1 \in C_1) = \sum_{n,m=1}^{+\infty} \mathbb{P}_x(Z_0 \in C_0, Z_1 \in C_1, \tau_x^1 = n, \tau_x^2 = n + m).
\]

For given values of \(n, m\),

\[
\mathbb{P}_x(Z_0 \in C_0, Z_1 \in C_1, \tau_x^1 = n, \tau_x^2 = m) = \mathbb{E}_x \left[ F(X_1, \ldots, X_n) 1_{\{X_n = x\}} G(X_{n+1}, \ldots, X_{n+m}) \right],
\]

where

\[
F(x_1, \ldots, x_n) = 1_{\{\sum_{i=0}^{n-1} f(x_i) \in C_0, x_1, \ldots, x_{n-1} \neq x\}},
\]

\[
G(x_1, \ldots, x_m) = 1_{\{\sum_{i=0}^{m-1} f(x_i) \in C_1, x_1, \ldots, x_{m-1} \neq x, x_m = x\}},
\]

and we denote \(x_0 = x\) in both sums defining \(F\) and \(G\). By Exercise 3.1.13,

\[
\mathbb{E}_x \left[ F(X_1, \ldots, X_n) 1_{\{X_n = x\}} G(X_{n+1}, \ldots, X_{n+m}) \right] = \mathbb{E}_x \left[ F(X_1, \ldots, X_n) 1_{\{X_n = x\}} \right] \mathbb{E}_x \left[ G(X_{n+1}, \ldots, X_{n+m}) \right]
\]

\[
= \mathbb{P}_x(Z_0 \in C_0, \tau_x^1 = n) \mathbb{P}_x(Z_0 \in C_1, \tau_x^2 = m),
\]

so that

\[
\mathbb{P}_x(Z_0 \in C_0, Z_1 \in C_1) = \mathbb{P}_x(Z_0 \in C_0) \mathbb{P}_x(Z_0 \in C_1),
\]

which shows that the variables \(Z_0\) and \(Z_1\) are independent and have the same law. \(\square\)

**Proof of Theorem 4.1.3.** The proof is divided into three steps.

**Step 1.** We assume that \(X_0 = x\) almost surely and fix \(y \in E\). In this step, we show that there exists a deterministic number \(a(x, y)\) such that

\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{\{X_i = y\}} = a(x, y), \quad \text{almost surely.}
\]
For all $n \geq 1$, there exists $L \geq 0$ such that
\[
\tau_x^L \leq n < \tau_x^{L+1},
\]
and thus
\[
\sum_{\ell=0}^{L-1} \sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1} \mathbb{1}_{\{X_i=y\}} \leq \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=y\}} \leq \sum_{\ell=0}^{L} \sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1} \mathbb{1}_{\{X_i=y\}},
\]
where the denominator in the right-hand side is not 0 for $n \geq \tau_x^1$. Letting
\[
Z_x^y = \sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1} \mathbb{1}_{\{X_i=y\}}, \quad \mathcal{Z}_\ell = \tau_x^{\ell+1} - \tau_x^\ell,
\]
we deduce from Lemma 4.1.4 that the sequences $(Z_x^y)_{\ell \geq 0}$ and $(\mathcal{Z}_\ell)_{\ell \geq 0}$ are iid under $\mathbb{P}_x$. Besides, by Lemma 4.1.1,
\[
0 \leq \mathbb{E}_x[Z_0^y] \leq \mathbb{E}_x[\mathcal{Z}_0] < +\infty.
\]
As a consequence, the Law of Large Numbers shows that
\[
\lim_{L \to +\infty} \frac{1}{L} \sum_{\ell=0}^{L-1} Z_x^y = \mathbb{E}_x[Z_0^y], \quad \lim_{L \to +\infty} \frac{1}{L} \sum_{\ell=0}^{L-1} \mathcal{Z}_\ell = \mathbb{E}_x[\mathcal{Z}_0], \quad \text{almost surely},
\]
and thus
\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=y\}} = \frac{\mathbb{E}_x[Z_0^y]}{\mathbb{E}_x[\mathcal{Z}_0]} =: a(x,y), \quad \text{almost surely}.
\]

**Step 2.** In this step, we identify $a(x,y)$. Since the sequence $\frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=y\}}$ is bounded, the Dominated Convergence Theorem shows that
\[
\lim_{n \to +\infty} \mathbb{E}_x \left[ \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=y\}} \right] = a(x,y).
\]
But for all $n \geq 1$,
\[
\mathbb{E}_x \left[ \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=y\}} \right] = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{P}_x(X_i = y) = \frac{1}{n} \sum_{i=0}^{n-1} P^i(x,y),
\]
and by Remark 3.2.6, the right-hand side converges to $\pi(y)$. Thus, $a(x,y) = \pi(y)$ for any $x \in E$.

**Step 3.** We now conclude the proof. Let $\mu_0$ denote the initial distribution of the chain and let $f : E \to \mathbb{R}$. By Lemma 3.1.2,
\[
\mathbb{P} \left( \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \pi f \right) = \sum_{x \in E} \mathbb{P}_x \left( \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \pi f \right) \mu_0(x).
\]
By Steps 1 and 2, for all $x \in E$,
\[
\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \sum_{y \in E} f(y) \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{\{X_i=y\}}.
\]
converges almost surely under $P_x$ to
\[ \sum_{y \in E} f(y)\pi(y) = \pi f. \]

Therefore
\[ \sum_{x \in E} P_x \left( \lim_{n \to +\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \pi f \right) \mu_0(x) = \sum_{x \in E} \mu_0(x) = 1, \]
and the proof is completed. \(\square\)

**Remark 4.1.5** (A representation of $\pi$). Taking $y = x$ in the proof above, we observe that, by the definition of $\tau_x$,
\[ Z_0^x = \sum_{i=0}^{\tau_x-1} \mathbb{1}_{\{X_i=x\}} = 1, \quad \text{almost surely under } P_x, \]
so that the identity established in Step 2 rewrites
\[ \pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]}, \]
which is a useful representation formula for the stationary distribution.

**Exercise 4.1.6** (Other representation formulæ). Show that, for any $y \in E$,
\[ \mathbb{E}_x \left[ \sum_{i=0}^{\tau_y-1} \mathbb{1}_{\{X_i=y\}} \right] = \frac{\pi(y)}{\pi(x)}, \]
and that for any $f : E \to \mathbb{R}$,
\[ \mathbb{E}_x \left[ \sum_{i=0}^{\tau_x-1} f(X_i) \right] = \frac{\pi f}{\pi(x)}. \]

**Exercise 4.1.7.** Consider the Ehrenfest urn model with an even number of particles $N$.
1. If all particles are initially in the same compartment, how long does it take, in average, for all particles to come back to this compartment?
2. If each compartment initially contains half the total number of particles, how long does it take in average for the chain to return to this state?

With similar arguments to the proof of Lemma 4.1.4, the following slightly different version of the strong Markov property, which will be used in the proof of Theorem 4.1.9, can be stated. The proof is left as an exercise.

**Lemma 4.1.8** (Strong Markov property II). Let $(X_n)_{n \geq 0}$ be an irreducible Markov chain with initial distribution $\mu_0$ and transition matrix $P$. Let $x \in E$.

(i) $\tau_x < +\infty$ almost surely.

(ii) The sequence $(Y_m)_{m \geq 0}$ defined by $Y_m = X_{m+\tau_x}$ is a Markov chain with initial condition $x$ and transition matrix $P$. 
4.1.2 Central Limit Theorem

Theorem 4.1.3 allows to extend the idea of the Monte Carlo method to Markov chains as follows: if \((X_n)_{n \geq 0}\) is an irreducible Markov chain with (unique) stationary distribution \(\pi\), then for any \(f : E \to \mathbb{R}\), the expectation
\[
\mathbb{E} = \pi f = \sum_{x \in E} f(x) \pi(x)
\]
can be approximated by
\[
\hat{\mathbb{E}}_n = \frac{1}{n} \sum_{i=0}^{n-1} f(X_i).
\]

For a given probability measure \(\pi\), how to construct an irreducible Markov chain with stationary distribution \(\pi\) will be the subject of Lecture 5. In the present subsection, we show that the Law of Large Numbers of Theorem 4.1.3 can be complemented with a Central Limit Theorem, which thus provides confidence intervals for \(\mathbb{E}\).

**Theorem 4.1.9** (Markov chain Central Limit Theorem). Under the assumptions of Theorem 4.1.3, for all \(f : E \to \mathbb{R}\) there exists \(\sigma^2(f) \in [0, +\infty)\) such that
\[
\lim_{n \to +\infty} \sqrt{n} \left( \hat{\mathbb{E}}_n - \mathbb{E} \right) = N(0, \sigma^2(f)), \quad \text{in distribution.}
\]

**Proof.** We follow the same idea as for the proof of Theorem 4.1.3. We first define
\[
\forall x \in E, \quad \tilde{f}(x) = f(x) - \pi f,
\]
and let
\[
\bar{Z}_\ell = \sum_{i=\tau_x^\ell}^{\tau_x^{\ell+1}-1} \tilde{f}(X_i), \quad \ell \geq 0.
\]

By Lemma 4.1.4, the variables \((\bar{Z}_\ell)_{\ell \geq 0}\) are iid under \(\mathbb{P}_x\); by Exercise 4.1.2 and the boundedness of \(f\), they are in \(L^2(\mathbb{P}_x)\); by Exercise 4.1.6, their expectation under \(\mathbb{P}_x\) writes
\[
\mathbb{E}_x \left[ \bar{Z}_0 \right] = \frac{\pi \tilde{f}}{\pi(x)} = 0.
\]

Therefore, the Central Limit Theorem shows that
\[
\lim_{L \to +\infty} \frac{1}{\sqrt{L}} \sum_{\ell=0}^{L-1} \bar{Z}_\ell = N(0, v(x, f)), \quad \text{in distribution},
\]
where
\[
v(x, f) := \mathbb{E}_x \left[ \left( \bar{Z}_0 \right)^2 \right] = \mathbb{E}_x \left[ \left( \sum_{i=0}^{\tau_x^L} (f(X_i) - \pi f) \right)^2 \right].
\]

By Slutsky’s Theorem (see Lemma 2.2.5), we deduce that
\[
\frac{1}{\sqrt{\tau_x^L}} \sum_{i=0}^{\tau_x^L-1} (f(X_i) - \pi f) = \sqrt{\tau_x^L} \frac{1}{\sqrt{L}} \sum_{\ell=0}^{L-1} \bar{Z}_\ell
\]
4.1 Extensions of the Law of Large Numbers and the Central Limit Theorem

converges in distribution to the law \( N(0, \pi(x)v(x, f)) \). Using the same ‘squeezing’ argument of \( n \) between \( \tau_x^L \) and \( \tau_x^{L+1} \) as in Theorem 4.1.3, we deduce the intermediary result that, for all \( x \in E \), 
\[
\sqrt{n} \left( \bar{J}_n - J \right) = \frac{1}{\sqrt{n}} \sum_{i=0}^{\tau_x-1} (f(X_i) - J) + \sqrt{\frac{n - \tau_x}{n}} \sqrt{n - \tau_x} \left( \hat{\delta}_{n-\tau_x} - J \right),
\]
where
\[
\hat{\delta}_m := \frac{1}{m} \sum_{j=0}^{m-1} f(Y_j).
\]
By the first part of the proof,
\[
\lim_{n \to +\infty} \sqrt{n - \tau_x} \left( \hat{\delta}_{n-\tau_x} - J \right) = \lim_{m \to +\infty} \sqrt{m} \left( \hat{\delta}_m - J \right) = N(0, \pi(x)v(x, f)), \quad \text{in distribution,}
\]
while
\[
\lim_{n \to +\infty} \frac{1}{\sqrt{n}} \sum_{i=0}^{\tau_x-1} (f(X_i) - J) = 0, \quad \lim_{n \to +\infty} \sqrt{\frac{n - \tau_x}{n}} = 1, \quad \text{almost surely.}
\]
As a consequence, it follows from Slutsky’s Theorem that
\[
\lim_{n \to +\infty} \sqrt{n} \left( \bar{J}_n - J \right) = N(0, \pi(x)v(x, f)), \quad \text{in distribution,}
\]
which completes the proof since the prelimit does not depend on the choice of \( x \in E \). \( \square \)

In Proposition 4.1.10 and Exercise 4.1.11, we give two alternative expressions for \( \sigma^2(f) \).

**Proposition 4.1.10** (Asymptotic variance). Under the assumptions of Theorem 4.1.3, we have
\[
\sigma^2(f) = \text{Var}_\pi(f(X_0)) + 2 \sum_{k=1}^{+\infty} \text{Cov}_\pi(f(X_0), f(X_k)).
\]
We do not prove Proposition 4.1.10 here, and rather refer to [5, Section 1.6].

**Exercise 4.1.11** (Poisson equation). Let \( f : E \to \mathbb{R} \). We call Poisson equation the equation
\[
\forall x \in E, \quad -(P-I)g(x) = f(x) - \pi f.
\]
We assume that \( P \) is irreducible and write \( \pi^\bot = \{ g \in \mathbb{R}^E : \pi g = 0 \} \).

1. Show that \( \pi^\bot \) is stable by \( P-I \), and that the restriction \( (P-I)|_{\pi^\bot} \) defines an automorphism of \( \pi^\bot \).
2. Deduce that the Poisson equation possesses a unique solution \( g \in \pi^\bot \), and describe the set of all solutions in \( \mathbb{R}^E \).
3. Show that, for all \( x \in E \), \( g(x) = \sum_{n=0}^{+\infty} \mathbb{E} \left[ f(X_n) \right] - \pi f \).
4. Show that \( \sigma^2(f) = \pi(g^2) - \pi((Pg)^2) \).
4.2 Geometric convergence and spectral aspects

In this section, we develop a somewhat different point of view on the long time behaviour of an irreducible Markov chain and its stationary distribution by investigating the conditions under which \( X_n \) converges in distribution to \( \pi \), and if it does, at which rate.

Exercise 4.2.1. Let \( (X_n)_{n \geq 0} \) be a Markov chain on the finite space \( E \) and \( f : E \to \mathbb{R} \). Using the results from Lecture 3, show that if \( X_n \) converges in distribution to some probability measure \( \mu \) on \( E \), then necessarily \( \mu \) is a stationary distribution for \( (X_n)_{n \geq 0} \).

4.2.1 Spectrum of stochastic matrices

General remarks

Let \( (X_n)_{n \geq 0} \) be a Markov chain with transition matrix \( P \) and initial distribution \( \mu \). Since the law of \( X_n \) writes \( \mu P^n \), it is natural to expect the spectrum of \( P \) to play a role in the description of its \( n \to +\infty \) limit.

Exercise 4.2.2. Show that for any (complex) eigenvalue \( \lambda \) of \( P \), \( |\lambda| \leq 1 \).

If \( P \) is irreducible, then 1 is a single eigenvalue, with associated eigenvectors \( 1 \) (to the right) and \( \pi \) (to the left). Let us denote by \( R \) the subset of row vectors \( \rho \in \mathbb{R}^m \) such that \( \sum_{x \in E} \rho(x) = 0 \).

Exercise 4.2.3. 1. Show that the linear application \( \rho \mapsto \rho P \) maps \( R \) to itself. We denote it by \( P_{|R} \).

2. Show that \( \lambda \) is an eigenvalue of \( P_{|R} \) if and only if it is an eigenvalue of \( P \) and \( \lambda \neq 1 \).

Period of a Markov chain

Consider the sequence \( (X_n)_{n \geq 0} \) defined in the two-point space \( E = \{-1, 1\} \) by \( X_n = (-1)^n X_0 \). It is an irreducible Markov chain with transition matrix

\[
P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},
\]

and unique stationary distribution \( \pi \) such that \( \pi(-1) = \pi(1) = 1/2 \).

Exercise 4.2.4. Without using the results of Section 4.1, check directly that the chain \( (X_n)_{n \geq 0} \) satisfies the Law of Large Numbers of Theorem 4.1.3. What do you think of the Central Limit Theorem?

Yet, for any initial distribution \( \mu_0 \) on \( E \) which is not \( \pi \), the law \( \mu_n \) of \( X_n \) under \( \mathbb{P}_{\mu_0} \) does not converge to \( \pi \). This is related to the phenomenon of periodicity.

We recall that the greatest common divisor of a set of nonnegative integers \( \mathcal{N} \) is defined by

\[
gcd(\mathcal{N}) = \max\{ k \geq 1 : \forall n \in \mathcal{N}, \exists \ell \in \mathbb{N} : n = k\ell \}.
\]

Definition 4.2.5 (Period). Let \( (X_n)_{n \geq 0} \) be a Markov chain with transition matrix \( P \). For all \( x \in E \), the period of \( x \) is \( \gcd\{n \geq 1 : P^n(x, x) > 0\} \).

Exercise 4.2.6. Compute the period of the states \(-1 \) and \( 1 \) in the example of the two-point space described above.
Lemma 4.2.7 (Period of an irreducible chain). If the chain \((X_n)_{n \geq 0}\) is irreducible, then all states have the same period, which is thus called the period of the chain. The chain is called aperiodic if its period is 1 and periodic otherwise.

Proof. For all \(x \in E\), we denote \(\mathcal{N}(x) = \{n \geq 1 : P^n(x, x) > 0\}\). Let \(x, y \in E\). By Definition 3.2.7, there exist \(r \geq 1\) and \(\ell \geq 1\) such that \(P^r(x, y) > 0\) and \(P^\ell(y, x) > 0\). Let \(m = r + \ell\). Then \(m \in \mathcal{N}(x) \cap \mathcal{N}(y)\) and \(\mathcal{N}(x) + m \subset \mathcal{N}(y)\), so that \(\gcd \mathcal{N}(y)\) divides all elements of \(\mathcal{N}(x)\) and thus \(\gcd \mathcal{N}(y) \leq \gcd \mathcal{N}(x)\). By the same arguments, \(\gcd \mathcal{N}(x) \leq \gcd \mathcal{N}(y)\) and the proof is completed. \(\square\)

The Perron–Frobenius Theorem and geometric convergence

Proposition 4.2.8 (Perron–Frobenius Theorem). Let \(P\) be the transition matrix of an irreducible Markov chain \((X_n)_{n \geq 0}\), with period \(k \geq 1\). The eigenvalues \(\lambda\) of \(P\) such that \(|\lambda| = 1\) are the \(k\)-th roots of unity, and they are all simple.

We refer to [7, Theorem 3.11] for the proof. Define
\[
\lambda_* := \max\{|\lambda|, \lambda \neq 1\} \text{ is an eigenvalue of } P = \max\{|\lambda|, \lambda \text{ is an eigenvalue of } P|_R\},
\]
so that if \(P\) is aperiodic, \(\lambda_* < 1\). We shall call the quantity \(1 - \lambda_*\) the spectral gap of \(P\). For any \(\mu \in \mathcal{P}(\mathbb{R})\),
\[
\mu P^n - \pi = (\mu - \pi)P^n = (\mu - \pi)P^n|_R;
\]
since \(\mu - \pi \in \mathbb{R}\). This yields the following geometric convergence result.

Theorem 4.2.9 (Geometric convergence). Let \(P\) be an irreducible and aperiodic stochastic matrix with stationary distribution \(\pi\) and spectral gap \(1 - \lambda_* \in (0, 1]\). For any norm \(\| \cdot \|\) on \(\mathbb{R}^m\) and for any \(\alpha \in (\lambda_*, 1]\), there exists a constant \(C_\alpha\) such that, for all \(n \geq 0\),
\[
\sup_{\mu \in \mathcal{P}(E)} \|\mu P^n - \pi\| \leq C_\alpha \alpha^n.
\]

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

\[ \square \]

**Remark 4.2.10.** Owing to the fact that the binomial terms in the Dunford decomposition of \( P|_\mathcal{R} \) grow polynomially in \( n \), one cannot take \( \alpha = \lambda_\star \) in the proof above, except if \( N = 0 \), that is to say if \( P|_\mathcal{R} \) is diagonalisable. We shall discuss this case in more detail in Lecture 5.

### 4.3 Mixing time and coupling

Theorem 4.2.9 provides another point of view on the use of Markov chains in random numerical simulation: if one wants to sample a random variable \( X \) with a given distribution \( \pi \) on \( E \), a possible solution consists in constructing an irreducible and aperiodic Markov chain \( (X_n)_n \geq 0 \) with stationary distribution \( \pi \), and generate a trajectory \( (X_n)_{n \leq n \leq \infty} \) long enough for the law of \( X_\pi \) to be close to \( \pi \). The geometric decay given in Theorem 4.2.9 provides an upper bound on how large \( n \) should be taken, but in general it may be difficult to estimate the spectral radius \( \lambda_\star \) accurately.

#### 4.3.1 Definition of the mixing time

A sometimes more tractable measure is given by the notion of **mixing time**, which relies on the choice of the following specific distance on \( \mathcal{P}(E) \).

**Definition 4.3.1 (Total variation distance).** The total variation distance between two probability measures \( \mu \) and \( \nu \) on \( E \) is defined by
\[ \| \mu - \nu \|_{TV} = \frac{1}{2} \sum_{x \in E} |\mu(x) - \nu(x)|. \]

**Remark 4.3.2.** The factor \( 1/2 \) allows to ensure that \( \| \mu - \nu \|_{TV} \in [0, 1] \).

\[ \square \]

**Exercise 4.3.3.** Show the duality formula
\[ \| \mu - \nu \|_{TV} = \frac{1}{2} \sup_{\| f \|_\infty \leq 1} (\mu f - \nu f), \]
where \( \| f \|_\infty = \max_{x \in E} |f(x)| \).

**Definition 4.3.4 (Mixing time).** Let \( (X_n)_n \geq 0 \) be an irreducible and aperiodic Markov chain with transition matrix \( P \) and stationary distribution \( \pi \). For all \( n \geq 0 \), let us define
\[ d_{TV}(n) = \max_{x \in E} \| P^n(x, \cdot) - \pi \|_{TV}. \]

For all \( \epsilon > 0 \), the mixing time (at level \( \epsilon \)) of the Markov chain is defined by
\[ n_{\text{mix}}(\epsilon) = \inf \{ n \geq 0 : d_{TV}(n) \leq \epsilon \}. \]

\[ \square \]

**Exercise 4.3.5.** Show that the sequence \( (d_{TV}(n))_n \geq 0 \) is nonincreasing, so that \( d_{TV}(n) \leq \epsilon \) for all \( n \geq n_{\text{mix}}(\epsilon) \).
4.3 Mixing time and coupling

4.3.2 The coupling technique

In practical applications, it is often the case that the size of the state space $E$ depends on a parameter $N$ (think of the number of particles in the Ehrenfest urn), with which $\tau_{\text{mix}}(\epsilon)$ increases. To quote from [13], obtaining quantitative estimates on how the mixing behaviour of a Markov chain depends on $N$ is an important topic in the ‘modern theory’ of Markov chains. An often convenient tool is coupling. It is based on the following reformulation of the total variation distance.

**Lemma 4.3.6** (Total variation distance and coupling). For all $\mu, \nu \in \mathcal{P}(E)$,

$$\|\mu - \nu\|_{\text{TV}} = \inf_{(X,Y) \in \mathcal{C}(\mu,\nu)} \mathbb{P}(X \neq Y),$$

where $\mathcal{C}(\mu,\nu)$ denotes the set of random pairs $(X,Y)$ in $E \times E$ such that the respective marginal distributions of $X$ and $Y$ are $\mu$ and $\nu$.

A pair $(X,Y) \in \mathcal{C}(\mu,\nu)$ is called a *coupling* of $\mu$ and $\nu$. Before detailing the proof of Lemma 4.3.6, we present yet a few other expressions for the total variation distance.

**Exercise 4.3.7.** Let $\mu, \nu \in \mathcal{P}(E)$. Recall the notation $[a]_+ = \max\{a,0\}$, $[a]_- = \max\{-a,0\}$ and $a \land b = \min\{a,b\}$.

1. Show that $\|\mu - \nu\|_{\text{TV}} = \sum_{x \in E} [\mu(x) - \nu(x)]_+ = \sum_{x \in E} [\mu(x) - \nu(x)]_-.$
2. Show that $\|\mu - \nu\|_{\text{TV}} = 1 - \sum_{x \in E} \mu(x) \land \nu(x)$.

For both questions you may draw a picture.

**Proof of Lemma 4.3.6.** We first note that for any $(X,Y) \in \mathcal{C}(\mu,\nu),

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sum_{x \in E} |\mathbb{P}(X = x) - \mathbb{P}(Y = x)|$$

$$= \frac{1}{2} \sum_{x \in E} |\mathbb{P}(X = x, Y \neq x) - \mathbb{P}(Y = x, X \neq x)|$$

$$\leq \frac{1}{2} \sum_{x \in E} (\mathbb{P}(X \neq Y, X = x) + \mathbb{P}(X \neq Y, Y = x))$$

$$= \mathbb{P}(X \neq Y),$$

so that

$$\|\mu - \nu\|_{\text{TV}} \leq \inf_{(X,Y) \in \mathcal{C}(\mu,\nu)} \mathbb{P}(X \neq Y).$$

We now prove the converse inequality by constructing an *optimal coupling*\(^1\), that is to say a pair $(X,Y) \in \mathcal{C}(\mu,\nu)$ such that $\|\mu - \nu\|_{\text{TV}} = \mathbb{P}(X \neq Y)$. Using the result of Exercise 4.3.7, let us define

$$p = 1 - \|\mu - \nu\|_{\text{TV}} = \sum_{x \in E} \mu(x) \land \nu(x) \in [0,1],$$

and draw a random variable $U$ uniformly in $[0,1]$. If $U \leq p$, we draw $Z$ according to the probability measure $((\mu(x) \land \nu(x))/p$ and let $X = Y = Z$. If $U > p$, then necessarily $\|\mu - \nu\|_{\text{TV}} > 0$ and we draw $X$ and $Y$ independently according to the respective probability measures

$$\tilde{\mu}(x) = \frac{[\mu(x) - \nu(x)]_+}{\|\mu - \nu\|_{\text{TV}}}, \quad \tilde{\nu}(x) = \frac{[\mu(x) - \nu(x)]_-}{\|\mu - \nu\|_{\text{TV}}}.$$

\(^1\)This coupling is also sometimes called *maximal*, because it maximises the probability that $X = Y$. 


Since \( \tilde{\mu} \) and \( \tilde{\nu} \) have disjoint supports, it is straightforward that \( X = Y \) if and only if \( U \leq p \), so that
\[
P(X \neq Y) = P(U > p) = 1 - p = \|\mu - \nu\|_{TV}.
\]
We now check that \( (X,Y) \) is a coupling of \( \mu \) and \( \nu \). For all \( x \in E \),
\[
P(X = x) = P(X = x|U \leq p)P(U \leq p) + P(X = x|U > p)P(U > p)
\]
\[=
P(Z = x)p + \tilde{\mu}(x)(1 - p)
\]
\[= \mu(x) \land \nu(x) + [\mu(x) - \nu(x)]_+
\]
\[= \mu(x),
\]
and a similar computation shows that \( P(Y = x) = \nu(x) \).

Couplings can be employed to obtain upper bounds on mixing times as follows. Let \( P \) be an irreducible and aperiodic stochastic matrix on \( E \), with unique stationary distribution denoted by \( \pi \), and assume that you are able to construct a Markov chain \( (X_n, Y_n)_{n \geq 0} \) on \( E \times E \) such that:

(i) each sequence \( (X_n)_{n \geq 0} \) and \( (Y_n)_{n \geq 0} \) is a Markov chain with transition matrix \( P \);
(ii) if \( X_n = Y_n \) then \( X_m = Y_m \) for all \( m \geq n \).

Let \( \tau := \inf \{ n \geq 0 : X_n = Y_n \} \) be the coupling time of this pair. Then by Lemma 4.3.6, for any \( x, y \in E \),
\[
\|P^n(x, \cdot) - P^n(y, \cdot)\|_{TV} \leq \mathbb{P}_{x,y}(\tau > n),
\]
where the notation \( \mathbb{P}_{x,y} \) means that \( X_0 = x \) and \( Y_0 = y \).

Exercise 4.3.8. Check that this inequality implies that
\[
d_{TV}(n) \leq \max_{x,y} \mathbb{P}_{x,y}(\tau > n),
\]
where we recall that the notation \( d_{TV}(n) \) is introduced in Definition 4.3.4.

We deduce from Exercise 4.3.8 that as soon as the right-hand side is smaller than \( \epsilon \), then \( n \) is larger than the mixing time \( n_{mix}(\epsilon) \) of the single chain \( (X_n)_{n \geq 0} \). It is interesting to point out the fact that it is the construction of the random trajectory of the pair \( (X_n, Y_n)_{n \geq 0} \) which provides information on the rate of convergence of the deterministic quantity \( P^n(x, \cdot) \).

Example 4.3.9 (Upper bound on the mixing time for the lazy Ehrenfest urn). We consider the microscopic description of the Ehrenfest urn as introduced in Example 3.1.8, with state space \( E_{\text{micro}} = \{A, B\}^N \). Let \( P \) denote the associated transition matrix. To avoid periodicity, we consider the lazy version of this dynamics, that is to say the Markov chain on \( E_{\text{micro}} \) with transition matrix \( (P + I)/2 \). The corresponding dynamics is defined as follows: at each step, one particle \( i \in \{1, \ldots, N\} \) is chosen uniformly and kept in the same compartment with probability 1/2, or moved to the other compartment with probability 1/2.

We now construct a coupling of two Ehrenfest urns by imposing that, at each step, the same particle \( i \in \{1, \ldots, N\} \) is chosen, and in both urns, the particle is moved to the same compartment \( A \) or \( B \), drawn randomly and independently from the current location of the particles. It is quickly observed that this construction satisfies the conditions above; besides, the coupling time is the time at which all \( N \) particles have been selected. In other words, it has the law of the coupon collection time introduced in Exercise 3.2.15. As a consequence, for any pair of initial configurations, for any \( c > 0 \),
\[
d_{TV}(N \ln N + cN) \leq \mathbb{P}(\tau > N \ln N + cN) \leq e^{-c},
\]
so that
\[ n_{\text{mix}}(\epsilon) \leq N \ln N + N \ln(1/\epsilon). \]
Notice that the leading order in \( N \) does not depend on \( \epsilon \), which is often the case for mixing times.

\textbf{Exercise 4.3.10.} Show that the mixing time of the macroscopic description of the lazy Ehrenfest urn dynamics is necessarily smaller than the mixing time of the microscopic description, so that it is also in \( O(N \ln N) \).
Lecture 5

The Markov Chain Monte Carlo method

Let \( \pi \) be a probability measure on the finite space \( E \) with cardinality \( m \). Assume that we want to either compute an expectation of the form

\[
I = \sum_{x \in E} f(x) \pi(x),
\]

for some function \( f : E \to \mathbb{R} \), or generate iid random variables \( X_1, X_2, \ldots \) distributed according to \( \pi \). Both tasks are virtually elementary because the finiteness of \( E \) allows them to be handled by a simple enumeration procedure. However when \( m \) is large, this procedure may have a computational cost which makes it impractical.

An alternative approach, called the Markov chain Monte Carlo (MCMC) method, consists in constructing a Markov chain \((X_n)_{n \geq 0}\) of which \( \pi \) is a stationary distribution, and using either the Law of Large Numbers and Central Limit Theorem from Section 4.1 to compute an estimator and a confidence interval\(^1\) for \( I \), or the convergence theorem from Section 4.2 to sample independent random variables \( \widetilde{X}_1, \widetilde{X}_2, \ldots \) which are approximately distributed according to \( \pi \) by running independent realisations of the chain \((X_n)_{n \geq 0}\) on long enough times.

Often, MCMC algorithms return Markov chains which have the particularity of being reversible with respect to \( \pi \). As we first detail in this lecture, this property makes the analysis of the long time behaviour of the Markov chain somehow easier than in the general case discussed in Lecture 4.

5.1 Reversible Markov chains

5.1.1 Definition and general remarks

**Definition 5.1.1** (Reversibility). A Markov chain \((X_n)_{n \geq 0}\) with transition matrix \( P \) is said to be reversible with respect to \( \pi \in \mathcal{P}(E) \) if, for any \( x, y \in E \),

\[
\pi(x)P(x, y) = \pi(y)P(y, x).
\]

Equation (5.1) is called the detailed balance equation. The denomination ‘reversibility’ is explained by the following result.

\(^1\)For the latter purpose, we mention here that the asymptotic variance \( \sigma^2(f) \) in Theorem 4.1.9 is generally difficult to estimate, but we shall not discuss this issue further in these notes.
Proposition 5.1.2 (Reversibility). Let \((X_n)_{n \geq 0}\) be a Markov chain with transition matrix \(P\), reversible with respect to \(\pi\). For any \(n \geq 0\), the vectors \((X_0, \ldots, X_n)\) and \((X_n, \ldots, X_0)\) have the same distribution under \(\mathbb{P}_\pi\).

Proof. For any \(x_0, \ldots, x_n \in E\), we deduce from Proposition 3.1.9 that
\[
\mathbb{P}_\pi(X_0 = x_0, \ldots, X_n = x_n) = \pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n).
\]
Applying Definition 5.1.1 once shows that \(\pi(x_0)P(x_0, x_1) = P(x_1, x_0)\pi(x_1)\), and iterating this procedure leads to the identity
\[
\pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1}) \cdots P(x_1, x_0),
\]
the right-hand side of which is \(\mathbb{P}_\pi(X_0 = x_n, \ldots, X_n = x_0)\) by Proposition 3.1.9 again. 

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

Corollary 5.1.3 (Reversibility and stationary distribution). If \((X_n)_{n \geq 0}\) is reversible with respect to \(\pi\), then \(\pi\) is a stationary distribution for \((X_n)_{n \geq 0}\).

Notice that this result can also be obtained by the direct computation
\[
\pi P(y) = \sum_{x \in E} \pi(x)P(x, y) = \sum_{x \in E} P(y, x)\pi(y) = \pi(y),
\]
which uses the fact that \(\sum_{x \in E} P(y, x) = 1\).

Exercise 5.1.4. Show that for the Ehrenfest urn, both the microscopic and the macroscopic descriptions are reversible with respect to the stationary distributions from Exercise 3.2.4.

Exercise 5.1.5. Find an elementary example of a Markov chain which is not reversible.

5.1.2 Spectral characterisation of reversibility

For any \(\pi \in \mathcal{P}(E)\), we define the symmetric and bilinear form \(\langle \cdot, \cdot \rangle_\pi\) on \(\mathbb{R}^E\) by
\[
\forall f, g \in \mathbb{R}^E, \quad \langle f, g \rangle_\pi = \sum_{x \in E} f(x)g(x)\pi(x).
\]
An operator \(A\) is called symmetric in \(\ell^2(\pi)\) if \(\langle Af, g \rangle_\pi = \langle f, Ag \rangle_\pi\), for all \(f, g \in \mathbb{R}^E\).

Remark 5.1.6. If \(\pi\) is the stationary distribution of an irreducible stochastic matrix \(P\), then \(\pi(x) > 0\) for all \(x \in E\) and thus \(\langle \cdot, \cdot \rangle_\pi\) is a scalar product.

Proposition 5.1.7 (Spectral characterisation of reversibility). Let \(\pi \in \mathcal{P}(E)\). A Markov chain with transition matrix \(P\) is reversible with respect to \(\pi\) if and only if \(P\) is symmetric in \(\ell^2(\pi)\).
Proof. Assume that a Markov chain with transition matrix \( P \) is reversible with respect to \( \pi \). Then for all \( f, g \in \mathbb{R}^E \),

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

\[
= \sum_{x, y \in E} f(y)g(x)\pi(x)P(x, y)
\]

\[
= \sum_{x, y \in E} f(y)g(x)\pi(y)P(y, x)
\]

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

\[
= \langle f, P^2 \rangle_\pi.
\]

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

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

so that Equation (5.1) is satisfied. \( \square \)

The spectral study of reversible chains is facilitated by the following result.

**Proposition 5.1.8 (Eigenvalues of reversible chains).** Let \( P \) be the transition matrix of an irreducible chain which is reversible with respect to its invariant measure \( \pi \). The eigenvalues of \( P \) are real and can be labelled \( 1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_m \geq -1 \), and there exists an orthonormal basis \( \{f_1, \ldots, f_m\} \) of \( \ell^2(\pi) \) such that \( Pf_i = \lambda_i f_i \) for all \( i \).

**Remark 5.1.9.** If the chain is aperiodic, then by Proposition 4.2.8, \( \lambda_m > -1 \). As a consequence, \( \lambda_\star = \max\{|\lambda_2|, |\lambda_m|\} < 1 \).

**Exercise 5.1.10.** Use the Spectral Theorem to prove Proposition 5.1.8.

### 5.1.3 Long time behaviour of reversible Markov chains

From Proposition 5.1.8, we deduce another statement for the geometric convergence of \( P^nf \) to \( \pi f \).

**Proposition 5.1.11 (Rate of convergence for reversible chains).** Under the assumptions of Proposition 5.1.8, for all \( f : E \to \mathbb{R} \), for all \( n \geq 0 \),

\[
\|P^nf - \pi f\|_{\ell^2(\pi)}^2 \leq \lambda_{max}^{2n}\|f - \pi f\|_{\ell^2(\pi)}^2.
\]

Notice that this result only provides the geometric convergence of \( P^nf \) to \( \pi f \) if \( \lambda_\star < 1 \), that is to say if the chain is aperiodic (see Remark 5.1.9), which is of course in accordance with Theorem 4.2.9.

**Proof.** Notice that in Proposition 5.1.8 we may take \( f_1 = 1 \), in which case \( \langle f, f_1 \rangle_{\pi} = \pi f \), so that writing the orthogonal decomposition

\[
P^nf = \sum_{i=1}^m \langle P^nf, f_i \rangle_{\pi} f_i = \sum_{i=1}^m \langle f, P^n f_i \rangle_{\pi} f_i = \sum_{i=1}^m \lambda_i^n \langle f, f_i \rangle_{\pi} f_i
\]
yields

\[ P^n f - \pi f = \sum_{i=2}^{m} \lambda_i^n \langle f, f_i \rangle \pi f_i. \]

As a consequence,

\[ \| P^n f - \pi f \|_2^2(\pi) = \sum_{i=2}^{m} (\lambda_i^n \langle f, f_i \rangle \pi)^2 \leq \lambda_I^2 \sum_{i=2}^{m} (\langle f, f_i \rangle \pi)^2 \leq \lambda_I^2 \| f - \pi f \|_2^2(\pi). \]

\[ \square \]

Exercise 5.1.12. The chi-square distance on \( \mathcal{P}(E) \) is defined by

\[ \chi_2(\mu|\pi) = \begin{cases} \sum_{x \in E} (\frac{\mu(x)}{\pi(x)} - 1)^2 \pi(x) & \text{if } \mu(x) = 0 \text{ as soon as } \pi(x) = 0, \\ +\infty & \text{otherwise}. \end{cases} \]

Note that it is not a distance, because it is not symmetric in \( \mu \) and \( \pi \). Show that under the assumptions of Proposition 5.1.8, for any initial distribution \( \mu \in \mathcal{P}(E) \),

\[ \chi_2(\mu P^n|\pi) \leq \lambda_I^2 \chi_2(\mu|\pi). \]

Exercise 5.1.13. For a reversible chain, the inverse \( 1/(1 - \lambda^\ast) \) of the spectral gap is called the relaxation time. How can you compare the relaxation time and the mixing time?

Exercise 5.1.14. Under the assumptions of Proposition 5.1.8, shows that the asymptotic variance in the Central Limit Theorem for Markov chains (Theorem 4.1.9) rewrites

\[ \sigma^2(f) = \sum_{i=2}^{m} \frac{1 + \lambda_i}{1 - \lambda_i} \langle f, e_i \rangle^2 \pi. \]

5.2 Markov chain Monte Carlo

5.2.1 Gibbs measures

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

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

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

\[ Z_\beta = \sum_{x \in E} e^{-\beta V(x)} \]

is called the partition function. Obviously, any probability measure \( \pi \) on \( E \) writes under this form, since it suffices to set \( \beta = 1 \) and \( V(x) = -\ln \pi(x) \). In the sequel, we shall work under the following two assumptions which are related with the computational complexity of the underlying model:

(i) the function \( V \) is easy to evaluate;

(ii) the constant \( Z_\beta \) is not easy to compute.

These assumptions make the enumeration procedure discussed in the introduction impossible to implement.
Example 5.2.1 (Product measure in the canonical ensemble). Let \( d \geq 1 \) and \( \mu_1, \ldots, \mu_d \) denote probability measures on \( \mathbb{N} \). Let \( \overline{\mu} \) be the product measure of \( \mu_1, \ldots, \mu_d \), defined on \( \mathbb{N}^d \) by

\[
\overline{\mu}(x_1, \ldots, x_d) = \mu_1(x_1) \cdots \mu_d(x_d).
\]

This probability measure can be seen as an instance of a grandcanonical ensemble in statistical physics, in which particles can be in \( d \) different energy levels, and a configuration \( x \in \mathbb{N}^d \) represents the number of particles in each energy level. Under the measure \( \overline{\mu} \), the law of the number of particles in the \( i \)-th energy level is \( \mu_i \), and the number of particles in each energy level are independent.

Fix \( N \geq 1 \), define the canonical ensemble with \( N \) particles by

\[
E = \{ x \in \mathbb{N}^d : x_1 + \cdots + x_d = N \},
\]

and if \( \overline{\mu}(E) > 0 \), consider the conditional distribution \( \pi \) defined on \( E \) by

\[
\pi(x) = \frac{\overline{\mu}(x)}{\overline{\mu}(E)}.
\]

The probability measure \( \pi \) can be written under the form of a Gibbs measure with

\[
\beta = 1, \quad V(x) = -\sum_{i=1}^{d} \ln \mu_i(x_i), \quad Z_\beta = \overline{\mu}(E).
\]

Exercise 5.2.2. For Example 5.2.1, write a recursive formula (in terms of \( d \) and \( N \)) for the cardinality \( m \) of \( E \), and deduce the bound \( m \geq N^{d-1}/(d-1)! \).

Exercise 5.2.3 (Multinomial distribution). In Example 5.2.1, assume that for all \( i \in \{1, \ldots, d\} \), \( \mu_i \) is the Poisson distribution with parameter \( \lambda_i > 0 \), that is to say the probability measure on \( \mathbb{N} \) defined by

\[
\mu_i(x) = e^{-\lambda_i} \frac{\lambda_i^x}{x!},
\]

for all \( x \in \mathbb{N} \).

1. Fix \( N \geq 0 \) and compute the associated conditional product measure \( \pi \) on \( E \).
2. When \( d = 2 \), how do you interpret \( \pi \)?
3. Extrapolating from this example, describe an algorithm allowing to sample a random variable \( X \in E \) exactly distributed according to \( \pi \) in \( N \) steps.

5.2.2 The Metropolis algorithm

The Metropolis algorithm provides a method to construct a Markov chain which is reversible with respect to a given probability measure \( \pi \) on a finite space \( E \). In the sequel, we shall still assume that \( \pi(x) > 0 \) for all \( x \in E \). This is not a restrictive assumption since if \( \pi(x) = 0 \) then one may simply remove \( x \) from \( E \).

The basic ingredients of the construction of the Metropolis chain \( (X_n)_{n \geq 0} \) are:

- an irreducible stochastic matrix \( Q \) on \( E \) such that \( Q(x, y) > 0 \) if and only if \( Q(y, x) > 0 \), called the proposal matrix;
- an acceptance function \( F : (0, +\infty) \to (0, 1] \) such that \( F(\rho) = \rho F(1/\rho) \) for all \( \rho > 0 \).

---

Common acceptance functions are \( F(\rho) = \min(\rho, 1) \) (the Metropolis–Hastings rule) and \( F(\rho) = \rho/(1 + \rho) \) (the Barker rule).

When the chain is in the state \( x \in E \), the next state is computed as follows:

(i) draw a state \( y \) with probability \( Q(x, y) \),

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

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

otherwise remain at \( x \).

The condition on \( Q \) ensures that, almost surely, the ratio \( r(x, y) \) takes its values in \((0, +\infty)\). Besides, it can be computed without knowing the partition function \( Z_\beta \).

**Proposition 5.2.4** (Reversibility of the Metropolis chain). The Metropolis chain \((X_n)_{n \geq 0}\) is irreducible and reversible with respect to \( \pi \).

As a consequence, \( \pi \) is the unique stationary distribution of \((X_n)_{n \geq 0}\) and all convergence results from Lecture 4 and Section 5.1 can be applied to this chain.

**Proof.** Let \( P \) denote the transition matrix of the Metropolis chain. It follows from the description of this chain that for all \( x, y \in E \),

\[
P(x, y) = \begin{cases} Q(x, y)a(x, y) & \text{if } x \neq y, \\ 1 - \sum_{z \neq x} Q(x, z)a(x, z) & \text{if } x = y. \end{cases}
\]

We first check irreducibility. Let \( x, y \in E \). Since \( Q \) is irreducible, there exist \( n \geq 1 \) and \( x = x_0, \ldots, x_n = y \in E \) such that \( Q(x_i, x_{i+1}) > 0 \) for all \( i \in \{0, \ldots, n - 1\} \). Clearly, there is no loss of generality in assuming that \( x_i \neq x_{i+1} \). Then, as a consequence of the assumption on \( Q \), we also have \( Q(x_{i+1}, x_i) > 0 \). Therefore all ratios \( r(x_i, x_{i+1}) \) are positive, and so are their images by \( F \), so that

\[
P^n(x, y) \geq \mathbb{P}_x(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=0}^{n-1} Q(x_i, x_{i+1})a(x_i, x_{i+1}) > 0.
\]

We now check reversibility. For all \( x, y \in E \) such that \( x \neq y \), the property \( F(\rho) = \rho F(1/\rho) \) yields

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

which ensures that the detailed balance equation holds. \(\square\)

**Exercise 5.2.5.** Show that if \( Q \) is aperiodic, then \( P \) is also aperiodic.

**Remark 5.2.6.** If \( Q \) is already reversible with respect to \( \pi \), the chain constructed with the Metropolis–Hastings rule has transition matrix \( P = Q \), while the chain constructed with the Barker rule is the lazy version of \( Q \) with transition matrix \( P = (Q + I)/2 \). In the latter case, \( P \) is aperiodic even if \( Q \) is not, which shows that the converse statement to Exercise 5.2.5 does not hold.
Assume that the proposal matrix $Q$ is symmetric in the sense that $Q(x, y) = Q(y, x)$, and that the acceptance function is given by the Metropolis–Hastings rule. Then a move from $x$ to $y$ is accepted with probability

$$a(x, y) = e^{-\beta[V(y) - V(x)]_+},$$

where $V$ is the potential of the Gibbs measure $\pi$. As a consequence, moves which make the potential decrease are accepted with probability 1, while the probability of moves which increase the potential is exponentially small.

### 5.2.3 Two other applications of the Metropolis algorithm

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

#### Simulated annealing for optimisation

Consider the Gibbs measure

$$\pi(x) = \frac{1}{Z_\beta}e^{-\beta V(x)},$$

for some function $V : E \to \mathbb{R}$.

**Exercise 5.2.7.** Show that, when $\beta \to +\infty$, $\pi$ converges to the uniform distribution on the set $\arg\min V := \{x \in E : \forall y \in E, V(y) \geq V(x)\}$.

If one is interested in finding the global minima of $V$, then a first approach may consist in taking a ‘large’ value of $\beta$, constructing a Metropolis chain $(X_n)_{n \geq 0}$ reversible with respect to the Gibbs measure $\pi$ and running it on a long enough time for $X_n$ to be essentially concentrated on the global minima of $V$.

Observe that if the algorithm uses the Metropolis–Hastings rule with a symmetric proposal matrix $Q$, then according to the discussion in Subsection 5.2.2, the following two phenomena occur.

- Moves that make the value of $V(X_n)$ decrease are always accepted. This brings the chain toward ‘local minima’ of $V$ on a short time scale, in accordance with the idea of gradient descent algorithms. Here, the notion of a ‘local’ minimum has to be understood with respect to the graph structure induced on $E$ by the pairs $(x, y)$ such that $Q(x, y) > 0$.
- Moves that make the value of $V(X_n)$ increase are accepted with an exponentially small (but nonetheless positive) probability. This allows the chain to ‘escape’ local minima on long time scales and go exploring other local minima. This behaviour is an essential feature of stochastic algorithms.

The idea of **simulated annealing** is a refinement of the Metropolis algorithm, in which the parameter $\beta$ increases with time. Given a proposal matrix $Q$ on $E$, an acceptance function $F$, and a deterministic sequence $(\beta_n)_{n \geq 1}$ growing to $+\infty$ which we call a *cooling scheme*, it can be described as follows: for all $n \geq 0$, given the current state $X_n = x \in E$,

1. select a state $y$ with probability $Q(x, y)$;
2. set $X_{n+1} = y$ with probability

$$a_n(x, y) = F\left(\frac{\pi_{\beta_n+1}(y)Q(y, x)}{\pi_{\beta_n+1}(x)Q(x, y)}\right),$$

where for all $\beta > 0$, we denote by $\pi_\beta$ the Gibbs measure with potential $V$ and inverse temperature $\beta$, otherwise set $X_{n+1} = x$. 

Then it can be shown that under some assumptions on \( V \), there exist cooling schemes for which \( V(X_n) \) converges to \( \min_{x \in E} V(x) \). We refer to [5, Chapitre 2], [3, Chapitre 5.3] for details.

**Subset simulation for rare events**

We come back to the issue, already discussed in Lecture 2, of estimating a small probability \( p \) of the form \( \mathbb{P}(f(X) \geq M) \), where \( X \) is a random variable in \( E \) with distribution \( \mu, f : E \to [0, +\infty) \) and \( M > 0 \). For the sake of simplicity, we remain here in the case where \( E \) is finite, but with a large cardinality \( m \) which prevents any enumerative procedure from being implemented.

In contrast with the remainder of this section, here we assume that we are able to generate iid samples \( X_1, X_2, \ldots \) from \( \mu \). However, the Monte Carlo estimator

\[
\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f(X_i) \geq M\}}
\]

requires too many samples to be accurate — for the sequel of the argument, we retain from Lecture 2 that it takes at least \( 1/p \) samples to estimate \( p \) accurately.

The idea of **subset simulation**\(^3\), also called **multilevel splitting**\(^4\), consists in introducing \( L \geq 1 \) levels \( 0 = M_0 < M_1 < \cdots < M_L = M \), writing

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

and estimating each term of this product separately.

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

**Exercise 5.2.8.** For a fixed value of \( p \), what is the value \( L^* \) which minimises the complexity \( L/p^{1/L} \)? Deduce that the resulting optimal complexity is of order of magnitude \( -\ln p \).

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

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

and define, for all \( x \in E_\ell \),

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

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


5.2 Markov chain Monte Carlo

5.2.4 Gibbs sampler

The Gibbs sampler algorithm, sometimes called Glauber dynamics in statistical physics, is a MCMC method which provides an alternative to the Metropolis algorithm. It is designed for probability measures \( \pi \) on state spaces \( E \) which have the specific form \( E = S^\Lambda \), where \( S \) and \( \Lambda \) are finite spaces.

Example 5.2.9 (The Ising model). The Ising model is a simple (but extremely famous\(^5\)) model for ferromagnetism. Fix \( N \geq 1, d \geq 1 \) and let \( \Lambda_{N,d} = \mathbb{Z}^d \cap [-N,N]^d \). Each vertex \( u \) of the lattice \( \Lambda_{N,d} \) has a spin \( x_u \in \{-1, 1\} \), so that a configuration \( x = (x_u)_{u \in \Lambda_{N,d}} \) is an element of the space \( E = \{-1, 1\}^{\Lambda_{N,d}} \), the cardinality of which is \( m = 2^{(2N+1)^d} \). The potential of a configuration \( x \) is defined by

\[
V(x) = -\sum_{u,v \in \Lambda_{N,d}, \|u-v\|_1 = 1} x_u x_v,
\]

where the notation \( u \sim v \) means that \( \|u-v\|_1 = 1 \), where \( \| \cdot \|_1 \) is the Euclidean distance on \( \mathbb{R}^d \). For a fixed value of \( \beta \), configurations with the lowest potential are those where all spins are either equal to \(-1\) or equal to \(+1\). Since the associated Gibbs measure \( \pi(x) \) gives a larger weight to configurations with low potential, this model represents the fact that configurations where neighbouring vertices have the same spin are more likely.

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

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

and let

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

the set of configurations which can be obtained by changing the value of the spin \( x_u \).

Definition 5.2.10 (Gibbs sampler). Let \( \pi \) be a probability measure on \( E = S^\Lambda \), such that \( \pi(x) > 0 \) for all \( x \in E \). The Gibbs sampler of \( \pi \) is the Markov chain in \( E \) defined by, at each step:

(i) picking a vertex \( u \in \Lambda \) uniformly;

(ii) selecting the new value of the spin \( x_u \) according to the conditional probability \( \pi(\cdot|E_{x,u}) \), where \( x \) is the current configuration.

Let us provide more detail on the update of the spin \( x_u \). For all \( s \in S \), the spin \( x_u \) is updated to the value \( s \) with probability

\[
\pi(x^{u,s}|E_{x,u}) = \frac{\pi(x^{u,s})}{\sum_{s' \in S} \pi(x^{u,s'})} = \frac{e^{-\beta V(x^{u,s})}/Z_\beta}{\sum_{s' \in S} e^{-\beta V(x^{u,s'})}/Z_\beta} = \frac{1}{1 + \sum_{s' \in S \setminus \{s\}} e^{-\beta (V(x^{u,s'}) - V(x^{u,s}))}}.
\]

\(^5\) A lot of references can be found in [13, Chapter 15].
This identity shows that it is not necessary to know the value of the partition function $Z_\beta$ to compute the conditional probability $\pi(\cdot|E_{x,u})$; instead, only $|S|$ evaluations of the potential $V$ are used\(^6\).

**Remark 5.2.11** (Graph structure for $\Lambda$). *This evaluation is particularly fast when the potential $V$ depends on a geometrical structure of the set $\Lambda$. Assume indeed that the latter is the set of vertices of an undirected graph, and the the former writes under the form*

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

*for some symmetric function $w : S \times S \to \mathbb{R}$, and $u \sim v$ means that the pair $(u, v)$ is an edge of the graph. Then for all $x \in E$, $u \in \Lambda$, and $s, s' \in S$,

$$V(x^{u,s}) - V(x^{u,s'}) = 2 \sum_{v \in \Lambda} \sum_{u \sim v} w(s, x_v) - w(s', x_v),$$

*which makes the computation of the conditional probability $\pi(\cdot|E_{x,u})$ local in the sense that it only depends on the spins of the neighbouring vertices $v$ of $u$ in the configuration $x$.\(^7\)

\(^6\)Here and in the sequel, we denote by $|\cdot|$ the cardinality of a finite set.

\(^7\)Exercise 5.2.12. Write explicitly the update probabilities for the Ising model.

The interest of the Gibbs sampler is given by the following result.

**Proposition 5.2.13** (Reversibility). *Under the assumptions of Definition 5.2.10, the Gibbs sampler of $\pi$ is irreducible, aperiodic and reversible with respect to $\pi$.\(^8\)

**Exercise 5.2.14.** Prove the irreducibility and aperiodicity properties.

**Proof of reversibility.** From Definition 5.2.10, we deduce that the transition matrix $P$ of the Gibbs sampler writes, for all $x, y \in E$ with $x \neq y$,

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

Clearly, for all $x, y \in E$ and $u \in \Lambda$, we have $y \in E_{x,u}$ if and only if $x \in E_{y,u}$, and in this case $E_{x,u} = E_{y,u}$. As a consequence, in such a case,

$$\pi(x)P(x,y) = \pi(x) \frac{1}{|\Lambda|} \pi(y|E_{x,u}) = \frac{\pi(x)\pi(y)}{|\Lambda|\pi(E_{x,u})} = \frac{\pi(x)\pi(y)}{|\Lambda|\pi(E_{y,u})} = \pi(y)P(y,x),$$

which is the detailed balance condition needed to prove reversibility. If there is no $u \in \Lambda$ such that $y \in E_{x,u}$, or equivalently $x \in E_{y,u}$, then $P(x,y) = 0 = P(y,x)$ which makes the detailed balance also hold in this case, and thus completes the proof.\(^{\square}\)
Part III

Diffusion processes
Bibliography


---

7 https://hal.archives-ouvertes.fr/hal-01897577v1
8 http://www.hairer.org/notes/Markov.pdf
9 http://cermics.enpc.fr/~jourdain/probastat/poly.pdf
10 https://www.math.u-psud.fr/~jflegall/IPPA2.pdf
11 https://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf