UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Markov chains and Markov chain Monte Carlo methods

Autor: Ariadna Gómez del Pulgar Martínez

Director: Dr. Carles Rovira Escofet
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 13 de juny de 2022

Contents

Abstract

The aim of this project is to thoroughly study the main properties of discrete-time Markov chains with finite state spaces and one of its applications that finds greatest usage, Markov chain Monte Carlo (MCMC) methods, which are simulation tools to estimate integrals and sample from distributions. A brief description of regular Monte Carlo is included to introduce and understand MCMC. Aside from the theoretical description and algorithms, practical considerations to take into account when implementing MCMC, such as the thermalization of chains and determining the number of iterations, are included as well. A simple example of the calculation of $\Gamma(3/2)$ is executed so as to illustrate the functioning and performance of MCMC.

2020 Mathematics Subject Classification. 11K45, 60J10, 60J20, 65C05, 65C40.

Acknowledgements

I would like to thank my advisor, Carles Rovira, for conducting and doing an exhaustive follow-up of this bachelor thesis, as well as for his help and ideas in terms of both the content and form of this work.

I would also like to thank my friends and family for their support and encouragement during not only the writing process of this thesis, but the entirety of the duration of this degree.

Chapter 1

Introduction

This work presents a description of the properties and behavior of discrete-time Markov chains of finite state space, and, as an application of these stochastic processes, Markov chain Monte Carlo methods are presented.

Markov chains are stochastic processes described by an initial distribution and a matrix that encodes the probabilities of the transitions between different states. A distinguishing feature of these processes is that the probability of moving to a certain state only depends on the current state, and not in past states. In **Chapter 2**, a thorough study of the main results regarding Markov chains is presented, including all the necessary theorems that ensure the adequate and desired behavior of Markov chain Monte Carlo, such as the asymptotic convergence results and the ergodic theorem. Only discrete-time Markov chains of finite state space are described. This is due to the inner nature of simulations and computers. As simulated chains will be updated after each time unit according to a certain algorithm, they are discrete-time chains, and the finiteness of the quantity of numbers that a computer can generate and work with implies that the state space will be finite. However, this does not stop Markov chain Monte Carlo methods to be useful in infinite-state problems, as we will see.

In **Chapter 3**, Monte Carlo methods are described as a short introduction to understand Markov chain Monte Carlo. Monte Carlo methods are simulation approaches to compute integrals or to sample probability distributions that otherwise would be too difficult to compute or that would drive an unreliable result. Both regular Monte Carlo and importance sampling are presented, as well as the situations in which Monte Carlo fails and Markov chain Monte Carlo is the only feasible strategy to assess certain problems.

In **Chapter 4**, Markov chain Monte Carlo methods are presented. The goal of these methods is the same as regular Monte Carlo, although their nature makes them appropriate to handle certain problems, especially those that include multi-

dimensional distributions. This chapter includes a description of their theoretical background, the most used algorithms, practical considerations to take when running them and an illustrative case of the use of a Metropolis algorithm to compute $\Gamma(3/2)$ so as to exemplify the mentioned aspects. Multidimensional Markov chain Monte Carlo are also commented on.

Finally, in **Chapter 5** the conclusions of this thesis are exposed as a recapitulation of the most important results that have been previously seen and proven.

As it is in **Chapter 2** where the foundations of Markov chains are defined and the theme of **Chapters 3 and 4** is an application of Markov chains, the first chapter has a much more theoretical and academic sense, including exhaustive proofs of the results, whereas the last two chapters are notably descriptive.

Basic probability-related definitions, such as independent events, expectancy, probability space, random variable..., are omitted because of their general knowledge nature and so as to shorten the length of the work. Nonetheless, they can be found in [1].

Chapter 2

Discrete-time Markov chains

This first chapter is mainly based on the first chapter of [2], and was complemented with the second chapter of [3].

2.1 Basic definitions

Firstly, the basic definitions needed for the construction of the concept of discrete-time Markov chains will be introduced.

Throughout the entirety of this thesis, we will be working with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The first necessary definition to start this work is that of stochastic process.

Definition 2.1. *A stochastic process is a family $\{X_t, t \in T\}$ of random variables $X_t : \Omega \longrightarrow I$ defined in a common probability space and indexed by the set T.*

T is called the **index set**. It stands for the concept of time and describes the evolution of the process. If T is uncountable (for instance, $T = \mathbb{R}$ or $T = [a, b] \subset \mathbb{R}_+$), we say that the process evolves in continuous time. On the other hand, if T is discrete (for instance, $T = \mathbb{N}$), then we say that the process evolves in discrete time. In that case, when T is increased by one, a unit of time passes. Discrete-time Markov chains, which are the subject of interest of this work, are examples of the latter.

The set I is called **state-space**, and every $i \in I$ are called states. From now on, I will be considered to be a countable set. If, additionally, I is finite, its states will be labelled as $1, 2, \ldots, N$.

Definition 2.2. *$\lambda = (\lambda_i : i \in I)$ is a measure on I if $0 \leq \lambda_i < \infty$ for all $i \in I$. If, additionally, the total mass $\sum_{i \in I} \lambda_i$ equals 1, λ is a distribution.*

Distributions and measures can be thought of as row vectors whose components are indexed by I.

An example of a distribution that will be recurrently used throughout this work is the so-called *unit mass* at i, $\delta_i = (\delta_{ij} : j \in I)$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.3. *A matrix $P = (p_{ij} : i, j \in I)$ is a stochastic matrix if every row $(p_{ij} : j \in I)$ is a distribution:*

- $p_{ij} \in [0, 1]$ *for all* $i, j \in I$;
- $\sum_{j \in I} p_{ij} = 1$, *for all* $i \in I$.

If $A = (a_{ij} : i, j \in I)$ and $B = (b_{ij} : i, j \in I)$ are stochastic matrices, then $C = A \times B$, where $C = (c_{ij} : i, j \in I)$ is defined as $c_{ij} = \sum_{k \in I} a_{ik} b_{kj}$, is a stochastic matrix as well, since

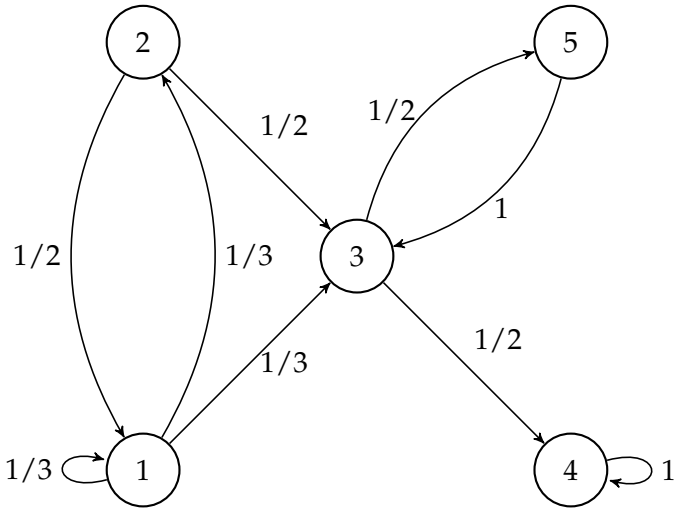
$$\sum_{j \in I} c_{ij} = \sum_{j \in I} \left(\sum_{k \in I} a_{ik} b_{kj} \right) = \sum_{k \in I} \left(\sum_{j \in I} a_{ik} b_{kj} \right) = \sum_{k \in I} a_{ik} \left(\sum_{j \in I} b_{kj} \right) = \sum_{k \in I} a_{ik} = 1.$$

Definition 2.4. *$(X_n)_{n \geq 0}$ is a homogenous Markov chain with initial distribution λ and transition matrix P, or Markov(λ, P), if*

1. *X_0 has distribution λ: $\mathbb{P}(X_0 = i_0) = \lambda_{i_0}$;*
2. *for $n \geq 0$, conditional on $X_n = i$, X_{n+1} has distribution $(p_{ij} : j \in I)$ and is independent of $X_0, \ldots, X_{n-1}$: $\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) = p_{i_n i_{n+1}}$.*

The last equality of the second property of **Definition 2.4** stands for the homogeneity of the Markov chain, as it implies that probability does not depend on time.

A simple way to describe discrete-time Markov chains and, especially, the concept of transition matrices, is by diagrams. Diagrams represent the different states of the state-space and the possible transitions between states after a unit of time, as well as the probabilities of those transitions. An illustrative case to describe this concept is depicted in **Example 2.5**.



Example 2.5.

This diagram corresponds to a Markov chain with a state-space of 5 states: I = {1, 2, 3, 4, 5}. According to the diagram, after a unit of time:

1. If the process is in state 1, the system will stay in state 1 or move to either states 2 or 3 with a probability of 1/3.

2. If the process is in state 2, the system will move to either states 1 or 3 with a probability of 1/2.

3. If the process is in state 3, the system will move to either states 4 or 5 with a probability of 1/2.

4. If the process is in state 4, the system will stay in that state.

5. If the process is in state 5, the system will move to state 3.

The stochastic matrix that corresponds to this diagram is the following one:

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Note that $p_{ij} \in [0,1]$ for all $i, j \in I$ and $\sum_{j \in I} p_{ij} = 1$, for all $i \in I$, as described before.

There exists a bijection between stochastic matrices and diagrams.

2.2 Defining properties of discrete-time Markov chains

In this section, the most defining properties of Markov chains are presented.

Theorem 2.6. *A discrete-time random process $(X_n)_{n\geq 0}$ is Markov(λ, P) if and only if for all $i_0, \ldots, i_{n+1} \in I$ and $n \geq 0$*

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \times \cdots \times p_{i_{n-1} i_n}. \tag{2.1}$$

Proof. Suppose that $(X_n)_{n\geq 0}$ is *Markov*(λ, P). Then

$$\begin{aligned}
&\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \\
&\quad = \mathbb{P}(X_n = i_n | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \mathbb{P}(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&\quad = \ldots = \mathbb{P}(X_n = i_n | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \times \\
&\quad \times \mathbb{P}(X_{n-1} = i_{n-1} | X_0 = i_0, \ldots, X_{n-2} = i_{n-2}) \times \cdots \times \mathbb{P}(X_0 = i_0) \\
&\quad = p_{i_{n-1} i_n} p_{i_{n-2} i_{n-1}} \cdots \lambda_{i_0}.
\end{aligned}$$

Here, the definition of conditional probability, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$, was used $n-1$ times, and in the last equality the properties (1) and (2) of **Definition 2.4** were used.

To prove the converse implication, we have to see that, if $(X_n)_{n\geq 0}$ satisfies (2.1), then it also satisfies properties (1) and (2) of **Definition 2.4**.

1. For $n = 1$, we have

   $$\mathbb{P}(X_0 = i_0, X_1 = i_1) = \lambda_{i_0} p_{i_0 i_1}.$$

   Therefore

   $$\mathbb{P}(X_0 = i_0) = \sum_{i_1 \in I} \mathbb{P}(X_0 = i_0, X_1 = i_1) = \sum_{i_1 \in I} \lambda_{i_0} p_{i_0 i_1} = \lambda_{i_0} \sum_{i_1 \in I} p_{i_0 i_1} = \lambda_{i_0},$$

   since P is a stochastic matrix and each of its rows is a distribution.

2. By the definition of conditional probability,

   $$\begin{aligned}
   &\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n) = \\
   &\quad = \frac{\mathbb{P}(X_0 = i_0, \ldots, X_{n+1} = i_{n+1})}{\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n)} \\
   &\quad = \frac{\lambda_{i_0} p_{i_0 i_1} \times \ldots \times p_{i_{n-1} i_n} p_{i_n i_{n+1}}}{\lambda_{i_0} p_{i_0 i_1} \times \ldots \times p_{i_{n-1} i_n}} \\
   &\quad = p_{i_n i_{n+1}}.
   \end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) = \\
&\quad = \frac{\mathbb{P}(X_n = i_n, X_{n+1} = i_{n+1})}{\mathbb{P}(X_n = i_n)} \\
&\quad = \frac{\sum_{i_0 \in I} \dots \sum_{i_{n-1} \in I} \mathbb{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i_n, X_{n+1} = i_{n+1})}{\mathbb{P}(X_n = i_n)} \\
&\quad = \frac{\sum_{i_0 \in I} \dots \sum_{i_{n-1} \in I} \lambda_{i_0} p_{i_0 i_1} \times \dots \times p_{i_{n-1} i_n} p_{i_n i_{n+1}}}{\mathbb{P}(X_n = i_n)} \\
&\quad = \frac{p_{i_n i_{n+1}} \sum_{i_0 \in I} \dots \sum_{i_{n-1} \in I} \lambda_{i_0} p_{i_0 i_1} \times \dots \times p_{i_{n-1} i_n}}{\mathbb{P}(X_n = i_n)} \\
&\quad = \frac{p_{i_n i_{n+1}} \sum_{i_0 \in I} \dots \sum_{i_{n-1} \in I} \mathbb{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i_n)}{\mathbb{P}(X_n = i_n)} \\
&\quad = \frac{p_{i_n i_{n+1}} \mathbb{P}(X_n = i_n)}{\mathbb{P}(X_n = i_n)} \\
&\quad = p_{i_n i_{n+1}}.
\end{aligned}$$

Therefore, $\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \dots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) = p_{i_n i_{n+1}}$, as we wanted to see.

□

The two following theorems reinforce the idea of the lack of memory of Markov chains. One of them, the strong Markov property, requires an extra definition: stopping times.

Theorem 2.7. (Markov property). *Let $(X_n)_{n \geq 0}$ be Markov(λ, P). Then, conditional on $X_m = i$, $(X_{m+n})_{n \geq 0}$ is Markov(δ_i, P) and is independent of the random variables $X_0, \dots, X_m$.*

Proof. Recall that, by definition, two events A and B are conditionally independent if, given a third event C, $\mathbb{P}(A \cap B | C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$. We have to prove that, for any event A determined by $X_0, \dots, X_m$,

$$\begin{aligned}
&\mathbb{P}(\{X_m = i_m, \dots, X_{m+n} = i_{m+n}\} \cap A | X_m = i) = \\
&\quad = \mathbb{P}(X_m = i_m, \dots, X_{m+n} = i_{m+n} | X_m = i) \mathbb{P}(A | X_m = i) \\
&\quad = \delta_{i i_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+1}} \mathbb{P}(A | X_m = i),
\end{aligned}$$

and then the result follows from **Theorem 2.6**.

First, consider the case of elementary events $A = \{X_0 = i_0, \ldots, X_m = i_m\}$. In that case

$$\begin{aligned}
&\mathbb{P}(\{X_m = i_m, \ldots, X_{m+n} = i_{m+n}\} \cap A | X_m = i) = \\
&\quad = \mathbb{P}(X_0 = i_0, \ldots, X_{m+n} = i_{m+n} | X_m = i) \\
&\quad = \frac{\mathbb{P}(X_0 = i_0, \ldots, X_{m+n} = i_{m+n}, X_m = i)}{\mathbb{P}(X_m = i)} \\
&\quad = \frac{\mathbb{P}(X_m = i_m, \ldots, X_{m+n} = i_{m+n}, i = i_m)\mathbb{P}(X_0 = i_0, \ldots, X_m = i_m, i = i_m)}{\mathbb{P}(X_m = i)} \\
&\quad = \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \mathbb{P}(X_0 = i_0, \ldots, X_m = i_m | X_m = i) \\
&\quad = \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \mathbb{P}(A | X_m = i).
\end{aligned}$$

Here, the independency of Markov chains expressed in the second property of **Definition 2.4** and the equation (2.1) were used.

Generally, any event A determined by $X_0, \ldots, X_m$ can be written as a countable disjoint union of elementary events of the previous form:

$$A = \bigcup_{k=1}^{\infty} A_k.$$

Because of the σ-additivity of $\mathbb{P}$, the expression holds:

$$\begin{aligned}
&\mathbb{P}(\{X_m = i_m, \ldots, X_{m+m} = i_{m+n}\} \cap A | X_m = i) = \\
&\quad = \mathbb{P}(\{X_m = i_m, \ldots, X_{m+n} = i_{m+n}\} \cap \bigcup_{k=1}^{\infty} A_k | X_m = i) \\
&\quad = \sum_{k=1}^{\infty} \mathbb{P}(\{X_m = i_m, \ldots, X_{m+n} = i_{m+n}\} \cap A_k | X_m = i) \\
&\quad = \sum_{k=1}^{\infty} \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \mathbb{P}(A_k | X_m = i) \\
&\quad = \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \sum_{k=1}^{\infty} \mathbb{P}(A_k | X_m = i) \\
&\quad = \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \mathbb{P}(\bigcup_{k=1}^{\infty} A_k | X_m = i) \\
&\quad = \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}} \mathbb{P}(A | X_m = i).
\end{aligned}$$

□

Definition 2.8. *A random variable $T : \Omega \to \{0, 1, \ldots\} \cup \{\infty\}$ is a stopping time if the event $\{T = n\}$ depends only on $X_0, \ldots, X_n$ for $n = 0, 1, 2, \ldots$.*

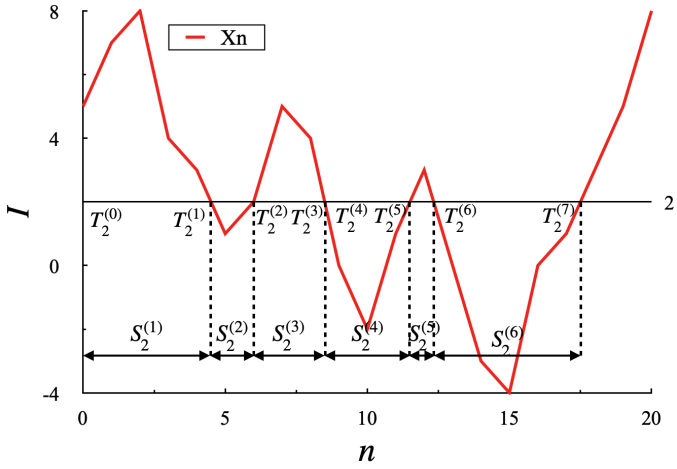


Figure 2.1: diagram that illustrates the concepts of the first rth passage times and excursions to state 2 of a Markov chain.

Some examples of stopping times that will be frequently dealt with during this work are the following:

- The first passage time, defined as

$$T_i(\omega) = \inf\{n \geq 1 : X_n(\omega) = i\},$$

where $\inf\emptyset = \infty$. The rth passage time $T_i^{(r)}$ to state i can be defined inductively by

$$\begin{aligned} &T_i^{(0)}(\omega) = 0, \\ &T_i^{(1)}(\omega) = T_i(\omega), \\ &T_i^{(r+1)}(\omega) = \inf\{n \geq T_i^{(r)}(\omega) + 1 : X_n(\omega) = i\}, \text{ for } r = 0, 1, 2, \ldots \end{aligned}$$

- The length of the rth excursion to i, $S_i^{(r)}$, is defined as

$$S_i^{(r)} = \begin{cases} T_i^{(r)} - T_i^{(r-1)} & \text{if } T_i^{(r-1)} < \infty \\ 0 & \text{otherwise.} \end{cases}$$

A simple illustration of these concepts is depicted in **Figure 2.1**.

From now on, the notation with subindex $\mathbb{P}_i(A)$ will be used to refer to the conditional probability $\mathbb{P}(A|X_0 = i)$. In a similar sense, when dealing with conditional expectancies, they will be written as $\mathbb{E}_i(A)$.

Theorem 2.9. (Strong Markov property). *Let $(X_n)_{n\geq 0}$ be Markov(λ, P) and let T be a stopping time of $(X_n)_{n\geq 0}$. Then, conditional on $T < \infty$ and $X_T = i$, $(X_{T+n})_{n\geq 0}$ is Markov(δ_i, P) and independent of $X_0, \ldots, X_T$.*

Proof. Let B be and event determined by $X_0, \ldots, X_T$. Once again, by the definition of conditional independency, it has to be proven that

$$\begin{aligned}\mathbb{P}(\{X_T = j_0, \ldots, X_{T+n} = j_n\} \cap B | T < \infty, X_T = i) = \\ = \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n)\mathbb{P}(B | T < \infty, X_T = i).\end{aligned}$$

Note that, as $T < \infty$, it can be said that $T = m$. Then, $B \cap \{T = m\}$ is determined by $X_0, \ldots, X_m$. Then by the Markov property at time T, $(X_{T+n})_{n\geq 0}$ is Markov(δ_i, P) and is independent of $X_0, \ldots, X_T$. Therefore, it is independent of B and T as well. Then,

$$\begin{aligned}\mathbb{P}(\{X_T = j_0, \ldots, X_{T+n} = j_{T+n}\} \cap B \cap \{T = m\} \cap \{X_T = i\}) = \\ = \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n)\mathbb{P}(B \cap \{T = m\} \cap \{X_T = i\}).\end{aligned} \tag{2.2}$$

If we sum over $m = 0, 1, 2, \ldots$ (that is, all the different possible values of T) and then use the σ-additivity of $\mathbb{P}$ on the left-hand side of (2.2), we get that

$$\begin{aligned}&\sum_{m\geq 0} \mathbb{P}(\{X_T = j_0, \ldots, X_{T+n} = j_{T+n}\} \cap B \cap \{T = m\} \cap \{X_T = i\}) = \\ &= \mathbb{P}(\{X_T = j_0, \ldots, X_{T+n} = j_{T+n}\} \cap B \cap \bigcup_{m\geq 0} \{T = m\} \cap \{X_T = i\}) \\ &= \mathbb{P}(\{X_T = j_0, \ldots, X_{T+n} = j_{T+n}\} \cap B \cap \{T < \infty\} \cap \{X_T = i\}).\end{aligned}$$

If we do so on the right-hand side of (2.2), we get

$$\begin{aligned}&\sum_{m\geq 0} \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n)\mathbb{P}(B \cap \{T = m\} \cap \{X_T = i\}) \\ &= \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n) \sum_{m\geq 0} \mathbb{P}(B \cap \{T = m\} \cap \{X_T = i\}) \\ &= \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n)\mathbb{P}(B \cap \bigcup_{m\geq 0} \{T = m\} \cap \{X_T = i\}) \\ &= \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n)\mathbb{P}(B \cap \{T < \infty\} \cap \{X_T = i\}).\end{aligned}$$

Thus, we get that

$$\begin{aligned}\mathbb{P}(\{X_T = j_0, \ldots, X_{T+n} = j_{T+n}\} \cap B \cap \{T < \infty\} \cap \{X_T = i\}) = \\ = \mathbb{P}_i(X_0 = j_0, \ldots, X_n = j_n)\mathbb{P}(B \cap \{T < \infty\} \cap \{X_T = i\}).\end{aligned} \tag{2.3}$$

Now, if we divide the left-hand side of (2.3) by $\mathbb{P}(T < \infty, X_T = i)$ and use the definition of conditional probability, we obtain that

$$\frac{\mathbb{P}(\{X_T = j_0, \dots, X_{T+n} = j_{T+n}\} \cap B \cap \{T < \infty\} \cap \{X_T = i\})}{\mathbb{P}(T < \infty, X_T = i)} =$$
$$= \mathbb{P}(\{X_T = j_0, \dots, X_{T+n} = j_{T+n}\} \cap B | \{T < \infty\} \cap \{X_T = i\}).$$

If we do so on the right-hand side of (2.3), we obtain

$$\frac{\mathbb{P}_i(X_0 = j_0, \dots, X_n = j_n)\mathbb{P}(B \cap \{T < \infty\} \cap \{X_T = i\})}{\mathbb{P}(T < \infty, X_T = i)}$$
$$= \mathbb{P}_i(X_0 = j_0, \dots, X_n = j_n)\mathbb{P}(B | \{T < \infty\} \cap \{X_T = i\}).$$

Therefore, we can conclude that

$$\mathbb{P}(\{X_T = j_0, \dots, X_{T+n} = j_{T+n}\} \cap B | \{T < \infty\} \cap \{X_T = i\}) =$$
$$= \mathbb{P}_i(X_0 = j_0, \dots, X_n = j_n)\mathbb{P}(B | \{T < \infty\} \cap \{X_T = i\}),$$

as wanted. □

2.3 n-step transition probabilities

In this section, we will talk about how to compute the probability of reaching any state in a certain amount of time or number of steps.

Recall that a distribution λ can be thought of as a row vector, and if the state-space I is finite then λ will be an N-vector. In that case, a stochastic matrix P will be an $N \times N$-matrix.

The multiplication of a row vector and a matrix, λP, is defined as

$$(\lambda P)_j = \sum_{i \in I} \lambda_i p_{ij},$$

and the matrix $P^2 = P \times P$ is defined as

$$(P^2)_{ij} = \sum_{j \in I} p_{ij} p_{jk}.$$

By recursion of this last expression, we can define the n-th power of a matrix P, P^n, as

$$(P^n)_{ij} = \sum_{i_2 \in I} \cdots \sum_{i_{n-1} \in I} p_{ii_2} p_{i_2 i_3} \cdots p_{i_{n-2} i_{n-1}} p_{i_{n-1} j}.$$

P^0 is defined as the identity matrix, $(Id)_{ij} = \delta_{ij}$.

From now on, the (i,j) entry in P^n will be referred to as $p_{ij}^{(n)} = (P^n)_{ij}$.

The probability that after n steps the Markov chain is in a given state j corresponds to $(\lambda P^n)_j$, and the probability to reach state j from state i in n states is given by $p_{ij}^{(n)}$. Both of this particularities are proven in the following theorem.

Theorem 2.10. *Let $(X_n)_{n\geq 0}$ be Markov(λ, P). Then, for all $n, m \geq 0$,*

1. $\mathbb{P}(X_n = j) = (\lambda P^n)_j$;

2. $\mathbb{P}_i(X_n = j) = \mathbb{P}(X_{n+m} = j | X_m = i) = p_{ij}^{(n)}$.

Proof. 1. We can compute the probability $\mathbb{P}(X_n = j)$ as

$$\begin{aligned}\mathbb{P}(X_n = j) &= \sum_{i_0 \in I} \dots \sum_{i_{n-1} \in I} \mathbb{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = j) \\ &= \sum_{i_0 \in I} \dots \sum_{i_{n-1} \in I} \lambda_{i_0} p_{i_0 i_1} \times \dots \times p_{i_{n-1} j} \\ &= \sum_{i_0 \in I} \lambda_{i_0} p_{i_0 j}^{(n)} \\ &= (\lambda P^n)_j.\end{aligned}$$

In the second equality, **Theorem 2.6** was used.

2. By the Markov property, conditional on $X_m = i$, $(X_{m+n})_{n\geq 0}$ is *Markov*(δ_i, P). If we take $\lambda = \delta_i$ in (1), we obtain

$$\mathbb{P}_i(X_n = j) = (\delta_i P)_j = \sum_{i \in I} \delta_{ij} p_{ij}^{(n)} = p_{ij}^{(n)}.$$

□

As $P^{n+m} = P^n P^m$, then

$$p_{ij}^{(n+m)} = \sum_{k \in i} p_{ik}^{(n)} p_{kj}^{(m)}. \tag{2.4}$$

This last result is known as the **Chapman-Kolmogorov equation**. It implies that the probability of reaching the state j in $n+m$ steps if we part from state i equals the sum of the probabilities of all the different possible trajectories that connect i and j in that same amount of steps.

2.4 Recurrence, transience and communicating classes

Definition 2.11. *Let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix P, and let i, j be states. We say that i leads to j, or $i \longrightarrow j$, if*

$$\mathbb{P}_i(X_n = j \text{ for some } n \geq 0) > 0.$$

We say that i communicates with j, or $i \longleftrightarrow j$, if both $i \longrightarrow j$ and $j \longrightarrow i$

In other words, a state i leads to state j if it is possible to reach state j parting from state i. If, additionally, it is possible to return to state i from state j, we say that i communicates with j.

The following theorem gives several ways to identify wether or not one state leads to another one.

Theorem 2.12. *For distinct states i and j the following are equivalent:*

- $i \longrightarrow j$;
- $p_{i_1 i_2} p_{i_2 i_3} \times \cdots \times p_{i_{n-1} i_n} > 0$ *for some states* $i_1, \ldots, i_n$ *with* $i_1 = i$, $i_n = j$;
- $p_{ij}^{(n)} > 0$ *for some* $n \geq 0$.

Proof. First, note that, since by **Theorem 2.10** $p_{ij}^{(n)} = \mathbb{P}_i(X_n = j)$,

$$p_{ij}^{(n)} \leq \mathbb{P}_i(X_n = j \text{ for some } n \geq 0).$$

Therefore, if $p_{ij}^{(n)} > 0$, then $\mathbb{P}_i(X_n = j \text{ for some } n \geq 0) > 0$, which implies that $i \longrightarrow j$. As a consequence, $(3) \implies (1)$. Additionally, if $i \longrightarrow j$, then $\mathbb{P}_i(X_n = j \text{ for some } n \geq 0)$, and therefore for that same n we have that $\mathbb{P}_i(X_n = j) = p_{ij}^{(n)} > 0$. Thus, $(1) \implies (3)$, and $(1) \iff (3)$.

On the other hand, by the definition of the n-th power of a matrix P, we have that

$$p_{ij}^{(n)} = \sum_{i_2 \in I} \cdots \sum_{i_{n-1} \in I} p_{i i_2} p_{i_2 i_3} \times \cdots \times p_{i_{n-1} j}.$$

As a consequence, if $p_{ij}^{(n)} > 0$, then $\sum_{i_2 \in I} \ldots \sum_{i_{n-1} \in I} p_{i i_2} p_{i_2 i_3} \times \cdots \times p_{i_{n-1} j} > 0$, and thus one of the summands has to be positive. Therefore, there exist some states $i_1, \ldots, i_n$ with $i_1 = i$, $i_n = j$ such that $p_{i_1 i_2} p_{i_2 i_3} \times \cdots \times p_{i_{n-1} i_n} > 0$. Since in this case we are dealing with an equality, the converse also holds. This proves that $(3) \iff (2)$.

□

In the light of this last theorem, $\longleftrightarrow$ defines an equivalence relation on I, since

- $\longleftrightarrow$ is transitive: if $i \longrightarrow j$, then $p_{i_1 i_2} \times \cdots \times p_{i_{n-1} i_n} > 0$ for some states $i_1, \ldots, i_n$ with $i_1 = i$, $i_n = j$, and if $j \longrightarrow k$, then $p_{j_1 j_2} \times \cdots \times p_{j_{m-1} j_m} > 0$ for some states $j_1, \ldots, j_m$ with $j_1 = j$, $j_n = k$. Therefore, $p_{i_1 i_2} \times \cdots \times p_{i_{n-1} i_n} p_{j_1 j_2} \times \cdots \times p_{j_{m-1} j_m} > 0$ for some states $i_1, \ldots, i_n, j_1, \ldots, j_m$ with $i_1 = i$ and $j_m = k$, and thus $i \longrightarrow k$. The same reasoning applies for $k \longrightarrow j \longrightarrow i$. As a consequence, if $i \longleftrightarrow j \longleftrightarrow k$, then $i \longleftrightarrow k$;
- $\longleftrightarrow$ is reflexive: since, by **Theorem 2.7**, conditional on $X_0 = i$, $(X_n)_{n \geq 0}$ is $Markov(\delta_i, P)$, $\mathbb{P}_i(X_0 = i) = 1$, and therefore $\mathbb{P}(X_n = i \text{ for some } n \geq 0) = 1 > 0$. Thus, $i \longrightarrow i$, and this implies that $i \longleftrightarrow i$;
- $\longleftrightarrow$ is symmetric: if $i \longleftrightarrow j$, then $i \longrightarrow j$ and $j \longrightarrow i$, and therefore $j \longleftrightarrow i$.

Since $\longleftrightarrow$ is an equivalence relation, it partitions I into equivalence classes, which are denominated **communicating classes**.

Definition 2.13. *Let C be a communicating class of a Markov chain. We say that C is closed if*

$$i \in C, i \longrightarrow j \Longrightarrow j \in C.$$

This last definition means that a Markov chain will not escape a closed class: once it visits a state in this communicating class, it will keep visiting states in that same class.

Definition 2.14. *We say that a state i is absorbing if $\{i\}$ is a closed class.*

Once a chain visits an absorbing state, it will remain in that state forever. In **Example 2.5**, state 4 is an absorbing state.

Definition 2.15. *An irreducible chain is a chain or transition matrix P where I is a single class.*

In an irreducible chain, each state is accessible from each one of the other states. In this type of chains, features that are class properties affect all the states in I.

Definition 2.16. *Let $(X_n)_{n \geq 0}$ be a Markov chain with transition matrix P. We say that a state i is recurrent if*

$$\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 1.$$

We say that a state i is transient if

$$\mathbb{P}_i(X_n = i \text{ for infinitely many } n) = 0.$$

A recurrent state is a state that the Markov chain will keep visiting, whereas a transient state is a state that the chain will eventually leave and not return to.

Each state in a Markov chain is either transient or recurrent, and, moreover, recurrence and transience are class properties. This will be seen in **Theorem 2.21** and **Theorem 2.22**. However, to prove that we first need two previous lemmas and various definitions. The proofs of these lemmas are omitted because of their simplicity. Among those, the rth passage time $T_i^{(r)}$ and the length of the rth excursion to state i, $S_i^{(r)}$, are used. Recall that they were presented as examples of stopping times earlier in this work.

Lemma 2.17. *For $r = 2, 3, \ldots$, conditional on $T_i^{(r-1)} < \infty$, $S_i^{(r)}$ is independent of $\{X_m : m \leq T_i^{(r-1)}\}$ and*

$$\mathbb{P}(S_i^{(r)} = n | T_i^{(r-1)} < \infty) = \mathbb{P}_i(T_i = n).$$

Proof. Can be found in [2].

□

Recall that the indicator function $1_{\{X_i=j\}}$ is the random variable defined as

$$1_{\{X_i=j\}} = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{otherwise.} \end{cases}$$

This variable will be recurrently used from now on, starting on the following definition.

Definition 2.18. *The number of visits to state i, V_i, is the random variable defined as*

$$V_i = \sum_{n=0}^{\infty} 1_{\{X_n=i\}}.$$

By definition of the indicator function, each time that the chain visits state i we will sum one unity to V_i, thus obtaining the total number of visits to that state.

Definition 2.19. *The return probability to state i, f_i, is defined as the probability, conditional on $X_0 = i$, that the first passage time is finite:*

$$f_i = \mathbb{P}_i(T_i < \infty).$$

Lemma 2.20. *For $r = 0, 1, 2, \ldots$, we have $\mathbb{P}_i(V_i > r) = f_i^r$.*

Proof. Can be found in [2].

□

Although **Lemma 2.17** and **Lemma 2.20** might seem slightly cryptic, their meaning is quite straightforward.

For **Lemma 2.17**, we have that, since $S_i^{(r)} = T_i^{(r)} - T_i^{(r-1)}$, then it is first required that $T_i^{(r-1)}$ is finite for $S_i^{(r)}$ to be finite. If that is fulfilled, the probability that the length of the rth excursion $S_i^{(r)}$ is n equals the probability that the next passage time to i after the $(r-1)$th passage, $T_i^{(r)}$, equals n. Because of the independency on the past of Markov chains, this last probability is the same as the probability that the first passage time happens at time n. This independency on the past also introduces the idea that $S_i^{(r)}$, which depends on $T_i^{(r)}$ and $T_i^{(r-1)}$, is independent on anything that happened before $T_i^{(r-1)}$.

For **Lemma 2.20**, we have that, in order to have a visit to state i, then we must have $T_i < \infty$. Thus, the probability that the number of visits to state i is, at least, r, equals the probability that at least the first r passage times to state i are finite, which, due to independency, corresponds to the multiplication, r times, of the probability that $T_i < \infty$.

Theorem 2.21. *The following dichotomy holds:*

1. *if* $\mathbb{P}_i(T_i < \infty) = 1$, *then* i *is recurrent and* $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$;
2. *if* $\mathbb{P}_i(T_i < \infty) < 1$, *then* i *is transient and* $\sum_{n=0}^{\infty} p_{ii}^{(n)} < \infty$.

Proof. 1. If $\mathbb{P}_i(T_i < \infty) = 1$, then, by definition of f_i, $f_i = 1$. As a consequence, by **Lemma 2.20**

$$\mathbb{P}_i(V_i = \infty) = \lim_{r\to\infty} \mathbb{P}_i(V_i > r) = \lim_{r\to\infty} f_i^r = \lim_{r\to\infty} 1^r = 1.$$

This implies that the state i is recurrent, as $(X_n)_{n\geq 0}$ will visit i infinitely many times with probability 1.

Now, note that the expected value of the number of visits V_i is

$$\mathbb{E}_i(V_i) = \mathbb{E}_i\left(\sum_{n=0}^{\infty} 1_{\{X_n=i\}}\right) = \sum_{n=0}^{\infty} \mathbb{E}_i\left(1_{\{X_n=i\}}\right) = \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = i) = \sum_{n=0}^{\infty} p_{ii}^{(n)}.$$

Here, **Theorem 2.10** was used to express $\mathbb{P}_i(X_n = i)$ as $p_{ii}^{(n)}$.

$\mathbb{E}_i(V_i)$ can also be computed as $\mathbb{E}_i(V_i) = \sum_{n=0}^{\infty} n\mathbb{P}_i(V_i = n)$, and since $\mathbb{P}_i(V_i = \infty) = 1$, $\mathbb{E}_i(V_i) = \infty$, and therefore $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$, as we wanted to see.

2. If $\mathbb{P}_i(T_i < \infty) < 1$, then $f_i < 1$ and

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \mathbb{E}_i(V_i) = \sum_{r=0}^{\infty} \mathbb{P}_i(V_i > r) = \sum_{r=0}^{\infty} f_i^r = \frac{1}{1-f_i} < \infty,$$

as we wanted to see. The fact that $f_i < 1$ ensures the convergence of the geometric progression.

Additionally, $\mathbb{E}_i(V_i) < \infty$ implies that $\mathbb{P}_i(V_i = \infty) = 0$, which means that $(X_n)_{n\geq 0}$ will not visit the state i infinitely many times, and therefore i is transient.

□

An important particularity arises from this last result: as $\mathbb{P}_i(T_i < \infty)$ is either 1 or smaller than 1, every state is either transient or recurrent.

Theorem 2.22. *Let C be a communicating class. Then either all states in C are transient or all are recurrent.*

Proof. Let i, j be any pair of states in a communicating class C, and suppose that i is transient. Since $i \longleftrightarrow j$, then there exist $n, m \geq 0$ such that $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$. For all r,

$$p_{ii}^{(n+r+m)} \geq p_{ij}^{(n)} p_{jj}^{(r)} p_{ji}^{(m)},$$

as $p_{ii}^{(n+r+m)}$ accounts for all the different routes that connect the state i with itself, and not only those that have an incursion in state j. Thus

$$p_{jj}^{(r)} \leq \frac{p_{ii}^{(n+r+m)}}{p_{ij}^{(n)} p_{ji}^{(m)}} \Longrightarrow \sum_{r=0}^{\infty} p_{jj}^{(r)} \leq \sum_{r=0}^{\infty} \frac{p_{ii}^{(n+r+m)}}{p_{ij}^{(n)} p_{ji}^{(m)}} = \frac{1}{p_{ij}^{(n)} p_{ji}^{(m)}} \sum_{r=0}^{\infty} p_{ii}^{(n+r+m)} < \infty,$$

where the second result of **Theorem 2.21** has been used. Once again by **Theorem 2.21**, this implies that j is also a transient state.

The converse implies the recurrence class property.

□

Since recurrence and transience are class properties, if a chain is irreducible, then it is entirely either transient or recurrent. If the second case holds, then, for each pair of states $i, j \in I$ $\mathbb{P}_i(X_n = j \text{ for some } n \geq 0) = 1$; that is, we can reach state j in n steps if we part from state i for a certain $n \geq 0$, which implies that there exists $n \geq 0$ such that $p_{ij}^{(n)} > 0$.

The following theorem is necessary to prove several forthcoming results, such as the ergodic theorem. The proof will be omitted for simplicity reasons. Nonetheless, it shows a intuitable result: that if a chain is irreducible and recurrent, as we will keep visiting each of the states, then the first passage time to all states from any of the other states must be finite.

Theorem 2.23. *Suppose P is irreducible and recurrent. Then, for all $j \in I$ we have $\mathbb{P}(T_j < \infty) = 1$.*

Proof. Can be found in [2].

□

2.5 Invariant distributions, detailed balance and convergence to equilibrium

Definition 2.24. *We say that a measure λ is invariant (also referred to as stationary or equilibrium) for a matrix P if*

$$\lambda P = \lambda.$$

Note that an invariant measure corresponds to an eigenvector of eigenvalue 1 of the matrix P, if there exists such eigenvalue.

Furthermore, if λ is invariant for P, then $\lambda P^n = \lambda$ for every $n \geq 2$, as

$$\lambda P^n = (\lambda P)P^{n-1} = \lambda P^{n-1} = (\lambda P)P^{n-2} = \lambda P^{n-2} = \ldots = \lambda.$$

This, together with **Theorem 2.10**, gives an idea of the meaning of an invariant distribution. If we reach an invariant distribution at some point, from that time on the distribution of states will be that same one: the probability of being in a certain state will remain the same regardless the number of steps. Moreover, if the initial distribution λ is invariant, after each step the probability distribution of the states is always the same, and equal to the initial distribution. This explains the term "invariant" used to refer to this type of distributions.

Definition 2.25. *A stochastic matrix P and a measure λ are said to be in detailed balance if*

$$\lambda_i p_{ij} = \lambda_j p_{ji} \text{ for all } i, j.$$

The distributions that are in detailed balance with P are invariant for P, as we will see in the following lemma. This will be crucial to ensure the convergence of Markov chain Monte Carlo methods.

Lemma 2.26. *If P and λ are in detailed balance, then λ is invariant for P.*

Proof. For every $i \in I$, we have

$$(\lambda P)_i = \sum_{j \in I} \lambda_j p_{ji} = \sum_{j \in I} \lambda_i p_{ij} = \lambda_i \sum_{j \in I} p_{ij} = \lambda_i,$$

as P is a stochastic matrix. Thus, $\lambda = \lambda P$, and therefore λ is invariant for P, as wanted. □

Definition 2.27. *If a state $i \in I$ is recurrent, its expected return time, m_i, is defined as the expected value, conditional on $X_0 = i$, of the first passage time to state i:*

$$m_i = \mathbb{E}_i(T_i).$$

A stronger property than recurrence related to the expected return time is positive recurrence.

Definition 2.28. *A state i is positive recurrent if its expected return time is finite:*

$$m_i < \infty.$$

Otherwise, it is called null recurrent.

Note that, if a state is positive recurrent, it is, in particular, recurrent.

As happens with recurrence and transience as well, positive recurrence is a class property. Furthermore, a relevant result about Markov chains is that, if a chain is irreducible and positive recurrent, it has an invariant distribution, π, which corresponds to $\pi_i = 1/m_i$. This result will be proven in **Theorem 2.32**. However, to prove that we first need other theorems related to the expected time spent in a certain state between visits to another state, which is presented in the following definition.

Definition 2.29. *For a fixed state k, the expected time spent in state i between visits to k is defined as*

$$\gamma_i^k = \mathbb{E}_k \left(\sum_{n=0}^{T_k - 1} 1_{\{X_n = i\}} \right).$$

Once again, in this last definition the indicator function accounts for the number of times that the chains visits a state. In this case, it counts the number of times that the chain visits a state i between times 0 and the first passage time to state k parting from k.

Theorem 2.30. *Let P be irreducible and recurrent. Then*

1. $\gamma_k^k = 1$;

2. $\gamma^k = (\gamma_i^k : i \in I)$ *satisfies* $\gamma^k P = \gamma^k$ *(that is,* γ^k *is invariant for* P*);*

3. $0 < \gamma_i^k < \infty$ *for all* $i \in I$.

Proof. 1. By definition, $\gamma_k^k = \mathbb{E}_k \sum_{n=0}^{T_k - 1} 1_{\{X_n = k\}}$. Since T_k is the first passage time to state k, the chain will not visit this state until that moment, and therefore $1_{\{X_n = k\}} = 0$ except for $n = 0$. Thus, $\gamma_k^k = 1$.

2. For $n = 1, 2, \ldots$ the event $\{n \leq T_k\}$ depends only on $X_0, \ldots X_{n+1}$, as it is only required that any of those variables is equal to k. Then, by the Markov property at $n - 1$,

$$\begin{aligned}\mathbb{P}_k(X_{n-1} = i, X_n = j \text{ and } n \leq T_k) = \\ = \mathbb{P}_k(X_{n-1} = i \text{ and } n \leq T_k)\mathbb{P}_i(X_1 = j) \\ = \mathbb{P}_k(X_{n-1} = i \text{ and } n \leq T_k) p_{ij}.\end{aligned}$$

As P is recurrent, by **Theorem 2.23** we have that $\mathbb{P}_k(T_k < \infty) = 1$. As a consequence, under $\mathbb{P}_k$, $T_k < \infty$ and, by definition of first passage time, $X_0 = X_{T_k} = k$ with probability 1. Therefore

$$\begin{aligned}\gamma_j^k &= \mathbb{E}_k \left(\sum_{m=0}^{T_{k-1}} 1_{\{X_m = j\}} \right) \\ &= \mathbb{E}_k \left(\sum_{n=1}^{T_k} 1_{\{X_n = j\}} \right) \\ &= \mathbb{E}_k \left(\sum_{n=1}^{\infty} 1_{\{X_n = j \text{ and } n \leq T_k\}} \right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}_k(X_n = j \text{ and } n \leq T_k) \\ &= \sum_{n=1}^{\infty} \sum_{i \in I} \mathbb{P}_k(X_{n-1} = i, X_n = j \text{ and } n \leq T_k) \\ &= \sum_{n=1}^{\infty} \sum_{i \in I} \mathbb{P}_k(X_{n-1} = i \text{ and } n \leq T_k) p_{ij}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in I} p_{ij} \sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = i \text{ and } n \leq T_k) \\
&= \sum_{i\in I} p_{ij} \sum_{m=0}^{\infty} \mathbb{P}_k(X_m = i \text{ and } m \leq T_k - 1) \\
&= \sum_{i\in I} p_{ij} \mathbb{E}_k \left(\sum_{m=0}^{\infty} 1_{\{X_m = i \text{ and } m \leq T_k - 1\}} \right) \\
&= \sum_{i\in I} p_{ij} \mathbb{E}_k \left(\sum_{m=0}^{T_k - 1} 1_{\{X_m = i\}} \right) \\
&= \sum_{i\in I} \gamma_i^k p_{ij}.
\end{aligned}
$$

This implies that, by the definition of the multiplication of a matrix by a vector, for every $j \in I$ $\gamma_j^k = (\gamma^k P)_j$, and therefore $\gamma^k = \gamma^k P$, as wanted.

In this development, two different changes of variable were used. Firstly, n was substituted by $m + 1$. However, this variation only affects the limits of the sum, and not the indicator function, since $X_0 = X_{T_k} = k \neq j$ with probability 1. The second substitution was, again, changing m by $n - 1$, which does affect both the limits of the sum and the considered states.

3. Since P is irreducible, for each state $i \in I$ there exist $n, m \geq 0$ such that $p_{ik}^{(n)}, p_{ki}^{(m)} > 0$. As it has just been proved, $\gamma^k = \gamma^k P$, and as a consequence $\gamma^k = \gamma^k P^m$. Therefore

$$
\gamma_i^k = (\gamma^k P^m)_i = \sum_{j\in I} \gamma_j^k p_{ji}^{(m)} = \gamma_k^k p_{ki}^{(m)} + \sum_{j\in I, j\neq k} \gamma_j^k p_{ji}^{(m)} = p_{ki}^{(m)} + \sum_{j\in I, j\neq k} \gamma_j^k p_{ji}^{(m)}.
$$

Since $\sum_{j\in I, j\neq k} \gamma_j^k p_{ji}^{(m)} \geq 0$, we have that $\gamma_i^k \geq p_{ki}^{(m)}$, which, at the same time, is greater than 0. This implies that $0 < \gamma_i^k$, as wanted.

Additionally, using a similar argument, we have that

$$
1 = \gamma_k^k = (\gamma^k P^n)_k = \sum_{j\in I} \gamma_j^k p_{jk}^{(n)} = \gamma_i^k p_{ik}^{(n)} + \sum_{j\in I, j\neq i} \gamma_j^k p_{jk}^{(n)}.
$$

Since $\sum_{j\in I, j\neq i} \gamma_j^k p_{jk}^{(n)} \geq 0$, we have that $1 \geq \gamma_i^k p_{ik}^{(n)}$. As $1 \geq p_{ik}^{(n)} > 0$, γ_i^k has to be finite, as wanted.

□

The third implication of this last theorem shows an interesting result: if a Markov chain with transition matrix P is irreducible and recurrent, then the expected visits to any other state in between visits to a fixed state k will, surely, not be 0.

Theorem 2.31. *Let P be irreducible and let λ be an invariant measure for P with $\lambda_k = 1$. Then $\lambda \geq \gamma^k$. The equality holds when, in addition, P is recurrent.*

Proof. For each $j \in I$, since λ is an invariant measure, we have

$$\begin{aligned}
\lambda_j &= \sum_{i_1 \in I} \lambda_{i_1} p_{i_1 j} \\
&= \sum_{i_1 \neq k} \lambda_{i_1} p_{i_1 j} + \lambda_k p_{kj} \\
&= \sum_{i_1 \neq k} \lambda_{i_1} p_{i_1 j} + p_{kj} \\
&= \sum_{i_1 \neq k} \left(\sum_{i_2 \in I} \lambda_{i_2} p_{i_2 i_1} \right) p_{1_1 j} + p_{kj} \\
&= \sum_{i_1, i_2 \neq k} \lambda_{i_2} p_{i_2 i_1} p_{1_1 j} + \sum_{i_1 \neq k} \lambda_k p_{k i_1} p_{i_1 j} + p_{kj} \\
&= \sum_{i_1, i_2 \neq k} \lambda_{i_2} p_{i_2 i_1} p_{1_1 j} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} + p_{kj} \\
&= \ldots = \sum_{i_1, \ldots, i_n = \neq k} \lambda_{i_n} p_{i_n i_{n-1}} \cdots p_{i_1 j} + \\
&+ \left(p_{kj} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} + \ldots + \sum_{i_1, \ldots, i_{n-1} \neq k} p_{k i_{n-1}} \cdots p_{i_2 i_1} p_{i_1 j} \right).
\end{aligned}$$

As $\sum_{i_1, \ldots, i_n \neq k} \lambda_{i_n} p_{i_n i_{n-1}} \cdots p_{i_1 j} \geq 0$, for $j \neq k$ we have

$$\lambda_j \geq p_{kj} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} + \ldots + \sum_{i_1, \ldots, i_{n-1} \neq k} p_{k i_{n-1}} \cdots p_{i_2 i_1} p_{i_1 j}.$$

Each of the summands of this last expressions stands for the probability that, parting from state j, the chain will not reach state k again until a certain step, as they take into account every possible route that connect states j and k for the first time in a certain amount of steps. In other words, for each $m \leq n-1$, $\sum_{i_1, \ldots, i_m \neq k} p_{k i_m} \cdots p_{i_1 j} = \mathbb{P}_k(X_m = j \text{ and } T_k \geq m)$. Therefore

$$\begin{aligned}
\lambda_j &\geq \sum_{m=1}^{n-1} \mathbb{P}_k(X_m = j \text{ and } T_k \geq m) \\
&\longrightarrow \sum_{m=1}^{\infty} \mathbb{P}_k(X_m = j \text{ and } T_k \geq m) = \gamma_j^k \text{ as } n \longrightarrow \infty.
\end{aligned}$$

Here, it was used that $\gamma_j^k = \sum_{n=1}^{\infty} \mathbb{P}_k(X_n = j \text{ and } n \leq T_k)$, as was observed in the proof of **Theorem 2.30**.

Therefore, $\lambda \geq \gamma^k$, as wanted.

Now, suppose that P is additionally recurrent. Then, by **Theorem 2.30**, we have that $\gamma^k P = \gamma^k$.

Now, let $\mu = \lambda - \gamma^k$. μ is invariant for P, as $\mu P = (\lambda - \gamma^k)P = \lambda P - \gamma^k P = \lambda - \gamma^k = \mu$. This also implies that $\mu P^n = \mu$ for $n \geq 2$. As it has just been proven, $\lambda \geq \gamma^k$, which implies that $\lambda - \gamma^k = \mu \geq 0$. Moreover, $\mu_k = \lambda_k - \gamma_k^k = 1 - 1 = 0$, by **Theorem 2.30**.

Since P is recurrent, then, given $i \in I$, there exists n such that $p_{ik}^{(n)} > 0$. Then, $0 = \mu_k = \sum_{j \in I} \mu_j p_{jk}^{(n)} = \mu_i p_{ik}^{(n)} + \sum_{j \neq i} \mu_j p_{jk}^{(n)} \geq \mu_i p_{ik}^{(n)}$. Since $p_{ik}^{(n)} > 0$, it must be $\mu_i = 0$. Thus, $\mu = 0$, which implies that $\lambda = \gamma^k$, as wanted. □

Theorem 2.32. *Let P be irreducible. Then the following are equivalent:*

1. *every state is positive recurrent;*
2. *some state i is positive recurrent;*
3. *P has an invariant distribution, π say.*

Moreover, when (3) holds we have $m_i = \frac{1}{\pi_i}$ for all i.

Proof. (1) $\implies$ (2) Obvious.

(2) $\implies$ (3) If a state i is positive recurrent, in particular it is recurrent. Therefore, as P is irreducible and recurrence is a class property, P is recurrent. Then, by **Theorem 2.30**, γ^i is invariant for P.

Note that $\sum_{j \in I} \gamma_j^i = \sum_{j \in I} \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} 1_{\{X_n = j\}} \right]$. Thus, $\sum_{j \in I} \gamma_j^i$ accounts for the expected visits to all states between visits to i. Therefore, it stands for the expected time between visits to i. Then

$$\sum_{j \in I} \gamma_j^i = \mathbb{E}_i(T_i) = m_i < \infty,$$

as i is positive recurrent.

Thus, $\pi_j = \gamma_j^i / m_i$ is an invariant distribution for P, as

$$\sum_{j \in I} \gamma_j^i = m_i \implies \frac{1}{m_i} \sum_{j \in I} \gamma_j^i = \sum_{j \in I} \frac{\gamma_j^i}{m_i} = \sum_{j \in I} \pi_j = 1$$

and

$$(\pi P)_j = \sum_{k \in I} \pi_k p_{kj} = \sum_{k \in I} \frac{\gamma_k^i}{m_i} p_{kj} = \frac{1}{m_i} \sum_{k \in I} \gamma_k^i p_{kj} = \frac{1}{m_i} \gamma_j^i = \pi_j,$$

since γ^i is invariant for P.

(3) $\Longrightarrow$ (1) Take any state k. As π is a distribution, $\sum_{i\in I}\pi_i = 1$, and therefore there exists some state i such that $\pi_i > 0$. For that same state i, as P is irreducible, there exists $n > 0$ such that $p_{ik}^{(n)} > 0$. Additionally, since π is invariant, $\pi P^n = \pi$. Therefore, $\pi_k = \sum_{i\in I}\pi_i p_{ik}^{(n)} > 0$ for some n.

Let $\lambda_i = \pi_i/\pi_k$. We have that

$$(\lambda P)_i = \sum_{j\in I}\lambda_j p_{ij} = \sum_{j\in I}\frac{\pi_j}{\pi_k}p_{ji} = \frac{1}{\pi_k}\sum_{j\in I}\pi_j p_{ji} = \frac{\pi_i}{\pi_k} = \lambda_i,$$

and therefore λ is invariant for P. Besides, $\lambda_k = \pi_k/\pi_k = 1$. As a consequence, by **Theorem 2.31**, $\lambda \geq \gamma^k$. Hence

$$m_k = \sum_{i\in I}\gamma_i^k \leq \sum_{i\in I}\lambda_i = \sum_{i\in I}\frac{\pi_i}{\pi_k} = \frac{1}{\pi_k}\sum_{i\in I}\pi_i = \frac{1}{\pi_k} < \infty,$$

since, by definition, $\pi_k = \gamma_k^i/m_i$, and, by **Theorem 2.30**, $0 < \gamma_i^k < \infty$ for all $i \in I$.

As $m_k < \infty$ for any state k, every state is positive recurrent, as wanted.

In addition to that, as P is positive recurrent, in particular it is recurrent, and therefore by **Theorem 2.31** $\lambda = \gamma^k$. Thus, for each $i \in I$, we have that

$$m_k = \sum_{i\in I}\gamma_i^k = \sum_{i\in I}\lambda_i = \sum_{i\in I}\frac{\pi_i}{\pi_k} = \frac{1}{\pi_k}\sum_{i\in I}\pi_i = \frac{1}{\pi_k},$$

which gives the desired invariant distribution. □

Definition 2.33. *A state i is called aperiodic if $p_{ii}^{(n)} > 0$ for all sufficiently large n, or, equivalently, if the set $\{n \geq 0 : p_{ii}^{(n)} > 0\}$ has no common divisor other than 1.*

An aperiodic state is a state that the chain can always visit for a sufficiently large n. If all states in a chain are aperiodic, then the chain can visit each one of them, for a sufficiently large n.

Aperiodicity is a class property, as well as happens with transiency, recurrence and positive recurrence.

Lemma 2.34. *Suppose P is irreducible and has an aperiodic state i. Then, for all states j and k, $p_{jk}^{(n)} > 0$ for all sufficiently large n. In particular, taking $j = k$, all states are aperiodic.*

Proof. Since P is irreducible, there exits $r, s \geq 0$ such that $p_{ji}^{(r)}, p_{ik}^{(s)} > 0$. Then, for all sufficiently large n

$$0 < p_{ji}^{(r)} p_{ii}^{(n)} p_{ik}^{(s)} \leq p_{jk}^{(r+n+s)},$$

as wanted. The inequalities hold because $p_{jk}^{(n+r+s)}$ accounts for all the possible routes that connect states j and k in $n + r + s$ steps, not only those that have an incursion in state i. □

If a Markov chain is irreducible, recurrent and aperiodic, then it is usually referred to as ergodic.

Generally, there is not a criterion to determine whether a stochastic matrix is aperiodic or not, and it must be checked in each particular case. Nonetheless, if all of its entries are positive, all transitions between states are permitted, and therefore it will be aperiodic. The higher the number of non-zero entries a matrix has, the higher the chances that it is aperiodic are.

The following theorem shows a key feature of irreducible and aperiodic chains that have an invariant distribution: the fact that this distribution coincides with the limiting distribution of the chain.

Theorem 2.35. (Convergence to equilibrium). *Let P be irreducible and aperiodic, and suppose that P has an invariant distribution π. Let λ be any distribution. Suppose that $(X_n)_{n\geq 0}$ is Markov(λ, P). Then*

$$\mathbb{P}(X_n = j) \longrightarrow \pi_j \text{ as } n \longrightarrow \infty \text{ for all } j.$$

In particular,

$$p_{ij}^{(n)} \longrightarrow \pi_j \text{ as } n \longrightarrow \infty \text{ for all } i, j.$$

Proof. The main argument of this proof is the coupling between two Markov chains.

First, let $(Y_n)_{n\geq 0}$ be *Markov*(π, P) and independent of $(X_n)_{n\geq 0}$. Fix a referent state b and let T be the first time such that both Y_n and X_n visit the state b:

$$T = \inf\{n \geq 1 : X_n = Y_n = b\}.$$

Step 1 We first show that $\mathbb{P}(T < \infty) = 1$.

Consider the process $W_n = (X_n, Y_n)$. It is a Markov chain on $I \times I$ with transition probabilities

$$\tilde{p}_{(i,k)(j,l)} = p_{ij}p_{kl}$$

and initial distribution

$$\mu_{(i,k)} = \lambda_i \pi_k.$$

Since P is aperiodic, for all states i, j, k, l we have

$$\tilde{p}_{(i,k)(j,l)}^{(n)} = p_{ij}^{(n)} p_{kl}^{(n)} > 0$$

for all sufficiently large n, so $\tilde{P}$ is aperiodic.

Also, $\tilde{P}$ has an invariant distribution

$$\tilde{\pi}_{(i,k)} = \pi_i \pi_k$$

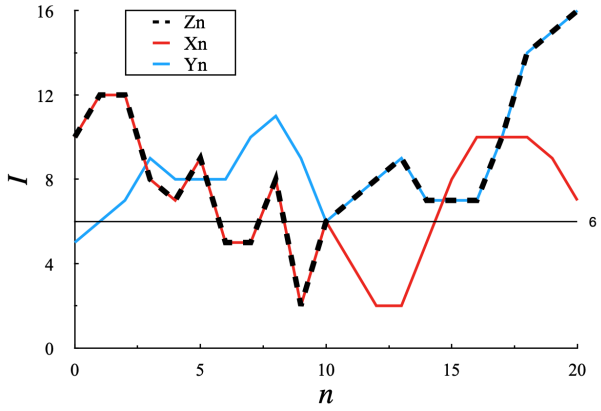


Figure 2.2: diagram that illustrates the concept of the process Z_n, with $b = 6$.

given that

$$(\tilde{\pi}\tilde{P})_{(i,k)} = \sum_{j,l\in I} \tilde{\pi}_{(j,l)}\tilde{p}_{(j,l),(i,k)} = \sum_{j,l\in I} \pi_j \pi_l p_{ji} p_{lk} = \left(\sum_{j\in I} \pi_j p_{ji}\right)\left(\sum_{l\in I} \pi_l p_{lk}\right) = \pi_i \pi_k = \tilde{\pi}_{(i,k)}.$$

As $\tilde{P}$ has an invariant distribution, by **Theorem 2.32** it is positive recurrent, and in particular it is recurrent. Note that T is the first passage time of the process W_n to the state (b, b). Then, by **Theorem 2.23**, $\mathbb{P}(T < \infty) = 1$.

Step 2 Consider the process

$$Z_n = \begin{cases} X_n & \text{if } n < T \\ Y_n & \text{if } n \geq T \end{cases}$$

A diagram that exemplifies this process is presented in **Figure 2.2**.

Intuitively, Z_n must be $Markov(\lambda, P)$, as we part from X_n, which has initial distribution λ and transition matrix P, and at time $n = T$ it shifts to Y_n, which has transition matrix P as well. This aspect is going to be proven by using the Markov chain W_n.

By applying the strong Markov property to $(W_n)_{n\geq 0}$ at the stopping time T, $(X_{T+n}, Y_{T+n})_{n\geq 0}$ is $Markov(\delta_{(b,b)}, \tilde{P})$. and independent of $(X_0, Y_0), (X_1, Y_1), \ldots, (X_T, Y_T)$. Analogously, the process $(Y_{T+n}, X_{T+n})_{n\geq 0}$ is $Markov(\delta_{(b,b)}, \tilde{P})$ and is independent of $(X_0, Y_0), (X_1, Y_1), \ldots, (X_T, Y_T)$. Hence $W'_n = (Z_n, Z'_n)$ is $Markov(\mu, \tilde{P})$, where

$$Z'_n = \begin{cases} Y_n & \text{if } n < T \\ X_n & \text{if } n \geq T \end{cases}$$

as W'_n corresponds to W_n until $n = T$ and to $(Y_{T+n}, X_{T+n})_{n\geq 0}$ after $n = T$.

This implies, in particular, that $(Z_n)_{n\geq 0}$ is $Markov(\lambda, P)$.

Step 3. We have

$$\mathbb{P}(Z_n = j) = \mathbb{P}(X_n = j \text{ and } n < T) + \mathbb{P}(Y_n = j \text{ and } n \geq T),$$

that is, the process Z_n will have an incursion in state j if either the chain X_n visits that state in a time shorter than T or the chain Y_n visits that state in a time longer than T.

Recall that $(Y_n)_{n\geq 0}$ is $Markov(\pi, P)$, and π is invariant for P, which implies that $\pi P^n = \pi$. Then, by **Theorem 2.10**, $\mathbb{P}(Y_n = j) = (\pi P^n)_j = \pi_j$. Hence

$$\begin{aligned}
|\mathbb{P}(X_n = j) - \pi_j| &= |\mathbb{P}(Z_n = j) - \mathbb{P}(Y_n = j)| \\
&= |\mathbb{P}(X_n = j \text{ and } n < T) - \mathbb{P}(Y_n = j \text{ and } n < T)| \\
&= |\mathbb{P}(X_n = j)\mathbb{P}(n < T) - \mathbb{P}(Y_n = j)\mathbb{P}(n < T)| \\
&= |\left[\mathbb{P}(X_n = j) - \mathbb{P}(Y_n = j)\right]\mathbb{P}(n < T)| \\
&= |\mathbb{P}(X_n = j) - \mathbb{P}(Y_n = j)||\mathbb{P}(n < T)| \\
&\leq \mathbb{P}(n < T) \longrightarrow 0 \text{ as } n \longrightarrow \infty,
\end{aligned}$$

since as time increases and $\mathbb{P}(T < \infty) = 1$ it will be less likely that the chains X_n and Y_n have not coincided.

Thus, it has been proven that $|\mathbb{P}(X_n = j) - \pi_j| \longrightarrow 0$ as $n \longrightarrow \infty$, and therefore $\mathbb{P}(X_n = j) \longrightarrow \pi_j$ as $n \longrightarrow \infty$, as wanted. □

To exemplify the concept of periodicity and its relevance to the previous theorem and its proof, consider the following transition matrix:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Note that

$$P^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = Id.$$

Thus, $P^{2n} = Id$, and $P^{2n+1} = P$, for all $n \geq 0$. This implies that any Markov chain with this transition probability will alternately stay in one state and shift to the other state. Therefore, it has no limiting distribution, since $p_{ij}^{(n)}$ does not converge. Furthermore, this chain is not aperiodic: it returns to each state after every 2 time units, thus having period 2.

P has an invariant distribution $\pi = (1/2 \;\; 1/2)$, as

$$(1/2 \;\; 1/2)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (1/2 \;\; 1/2).$$

However, this invariant distribution does not coincide with the limiting distribution, which does not exist. The reason for that is that, if we consider the processes $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ that start, respectively, from states 0 and 1, the chains will never coincide, which makes the proof of **Theorem 2.35** fail.

An important remark about **Theorem 2.35** is that the uniqueness of the limit implies that the invariant distribution is unique.

Let us recapitulate and link the most important results seen up to this point regarding invariant distributions. If a Markov chain with transition matrix P is irreducible and aperiodic and it has an invariant distribution, the latter coincides with the limiting distribution, thus guaranteeing the uniqueness of the invariant distribution. Having an invariant distribution is ensured if P is, additionally, positive recurrent, which is a class property. Therefore, if P is irreducible, aperiodic and positive recurrent, it has a unique invariant distribution $\pi_i = 1/\mathbb{E}_i(T_i)$, which also corresponds to the limiting distribution. Moreover, if under those conditions there exists a distribution λ such that λ and P are in detailed balance, λ will coincide as well with the invariant distribution and the limiting distribution. This last feature, alongside the ergodic theorem, constitutes the basis of the Markov chain Monte Carlo algorithms.

2.6 Ergodic theorem

The most relevant result concerning Markov chains for Markov chain Monte Carlo methods is the ergodic theorem, which explains the behavior of averages and, specifically, the average time spent in every state in the long run.

This theorem is a version of the strong law of large numbers for Markov chains. To understand this law, the definitions of the expectancy of a random variable and almost sure convergence are essential. Recall that the latter occurs when, given a sequence of random variables $\{X_n, n \geq 1\}$, there exists a random variable X such that

$$\lim_{n\to\infty} X_n(\omega) = X(\omega),$$

except, possibly, for a subset $N \in \mathcal{F}$ of probability 0. We will write this as $\mathbb{P}(X_n \longrightarrow X \text{ as } n \longrightarrow \infty) = 1$.

Theorem 2.36. (Strong law of large numbers). *Let $Y_1, Y_2, \ldots$ be a sequence of independent, identically distributed, non-negative random variables with $\mathbb{E}(Y_1) = \mu$. Then*

$$\mathbb{P}\left(\frac{Y_1 + \ldots + Y_n}{n} \longrightarrow \mu \text{ as } n \longrightarrow \infty\right) = 1.$$

The proof of this theorem can be found in [1].

The ergodic theorem involves the number of visits to a certain state before a certain moment in time, which is included in the following definition.

Definition 2.37. *The number of visits to i before n is defined as*

$$V_i(n) = \sum_{k=0}^{n-1} 1_{\{X_k=i\}}.$$

It is important to remark the difference between $V_i(n)$ and V_i. While the former counts the visits to state i only up to the moment n, the latter accounts for all the visits to state i, up to ∞.

Additionally, observe that $V_i(n)/n$ stands for the proportion of time before n spent in state i. Consequently, since the ergodic theorem regards the average time spent in every state in the long run, it must relate to $V_i(n)$.

Theorem 2.38. (Ergodic theorem). *Let P be irreducible and let λ be any distribution. If $(X_n)_{n\geq 0}$ is Markov(λ, P), then*

$$\mathbb{P}\left[\frac{V_i(n)}{n} \longrightarrow \frac{1}{m_i} \text{ as } n \longrightarrow \infty\right] = 1$$

Moreover, in the positive recurrent case, for any bounded function $f : I \longrightarrow \mathbb{R}$ we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \longrightarrow \overline{f} \text{ as } n \longrightarrow \infty\right] = 1$$

where

$$\overline{f} = \sum_{i\in I} \pi_i f_i$$

and where $(\pi_i : i \in I)$ is the unique invariant distribution.

Proof. Firstly, let us consider the case in which P is transient. In that case, for every state $i \in I$, $\mathbb{P}_i(X_n = i \text{ for infinitely many} n) = 0$, and consequently the number of visits to state i will be finite. Thus, $V_i(n) \leq V_i$, which accounts for all the visits to state i, not only those up to the moment n. Additionally, by **Theorem 2.21**, we have that $\mathbb{P}_i(T_i < \infty) < 1$, which implies that $\mathbb{P}_i(T_i = \infty) > 0$. This suggests that $m_i = \mathbb{E}_i(T_i) = \sum_{n=0}^{\infty} n\mathbb{P}_i(T_i = n) = \infty$, and therefore $1/m_i = 0$. On the other hand,

$$0 \leq \frac{V_i(n)}{n} \leq \frac{V_i}{n} \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Thus, $V_i(n)/n \longrightarrow 1/m_i$, as wanted.

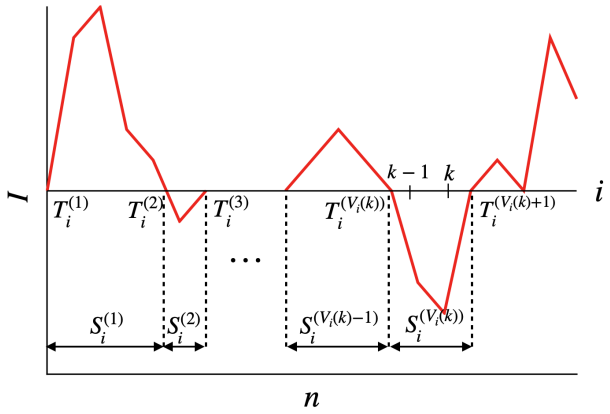


Figure 2.3: diagram that illustrates the last passage time and excursion to a state i before k and the first passage time and excursion to state i after $k-1$, for an initial distribution δ_i.

Now, suppose that P is irreducible. Fix a state i and let $T = T_i$. By **Theorem 2.21**, we have that, since i is recurrent, $\mathbb{P}_i(T_i < \infty) = 1$. Therefore, by the strong Markov property at T, $(X_{T+n})_{n\geq 0}$ is $Markov(\delta_i, P)$ and is independent of $X_0, X_1, \ldots, X_T$. In the long run, as we will visit the state i infinitely many times with probability 1, the time spent in state i will be the same for $(X_n)_{n\geq 0}$ and for $(X_{T+n})_{n\geq 0}$. Thus, it suffices to prove the desired convergence for the initial distribution $\lambda = \delta_i$.

Recall that, by **Lemma 2.17**, the lengths of the rth excursions to state i, $S_i^{(1)}, S_i^{(2)}, \ldots$ are independent and identically distributed with $\mathbb{E}_i[S_i^{(r)}] = \mathbb{E}_i(T_i) = m_i$. Additionally,

$$S_i^{(1)} + \ldots + S_i^{(V_i^{(n)}-1)} \leq n-1,$$

the left-hand side of the inequality being the time of the last visit to state i before n, and

$$S_i^{(1)} + \ldots + S_i^{(V_i^{(n)})} \geq n,$$

the left-hand side of the inequality being the time of the first visit to state i after $n-1$. These concepts are illustrated in **Figure 2.3**.

Therefore,

$$\frac{S_i^{(1)} + \ldots + S_i^{(V_i^{(n)}-1)}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{S_i^{(1)} + \ldots + S_i^{(V_i^{(n)})}}{V_i(n)}. \tag{2.5}$$

By the strong law of large numbers,

$$\mathbb{P}\left[\frac{S_i^{(1)} + \ldots + S_i^{(n)}}{n} \longrightarrow m_i \text{ as } n \longrightarrow \infty\right] = 1,$$

and, since P is recurrent,

$$\mathbb{P}[V_i(n) \longrightarrow \infty \text{ as } n \longrightarrow \infty] = 1.$$

Then, letting $n \longrightarrow \infty$ in (2.5), the left-hand side and the right-hand side of the inequality will both converge to m_i with probability 1, and therefore so will do the term in the middle:

$$\mathbb{P}\left[\frac{n}{V_i(n)} \longrightarrow m_i \text{ as } n \longrightarrow \infty\right] = 1,$$

which implies that

$$\mathbb{P}\left[\frac{V_i(n)}{n} \longrightarrow \frac{1}{m_i} \text{ as } n \longrightarrow \infty\right] = 1,$$

as wanted.

Now, suppose that $(X_n)_{n\geq 0}$ is positive recurrent, which, by **Theorem 2.32**, is equivalent to having an invariant distribution $\pi_i = 1/m_i$. Let $(\pi_i : i \in I)$ be the invariant distribution of $(X_n)_{n\geq 0}$ and let $f : I \longrightarrow \mathbb{R}$ be a bounded function; that is, $|f| \leq M$, for some $M < \infty$. Without loss of generality, by dividing f by M, we can assume that $|f| \leq 1$. For any $J \subseteq I$, we have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \overline{f}\right| &= \left|\sum_{i\in I}\frac{V_i(n)}{n}f_i - \sum_{i\in I}\pi_i f_i\right| \\
&= \left|\sum_{i\in I}\left[\frac{V_i(n)}{n} - \pi_i\right]f_i\right| \\
&= \left|\sum_{i\in J}\left[\frac{V_i(n)}{n} - \pi_i\right]f_i + \sum_{i\notin J}\left[\frac{V_i(n)}{n} - \pi_i\right]f_i\right| \\
&\leq \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right|\left|f_i\right| + \sum_{i\notin J}\left|\frac{V_i(n)}{n} - \pi_i\right|\left|f_i\right| \\
&\leq \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\left|\frac{V_i(n)}{n} - \pi_i\right| \\
&\leq \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\frac{V_i(n)}{n} + \sum_{i\notin J}\pi_i \\
&\leq 2\sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + 2\sum_{i\notin J}\pi_i.
\end{aligned}$$

In this development, it was used that $\sum_{k=0}^{n-1} f(X_k) = \sum_{i\in I} V_i(n) f_i$, where $f_i = f(i)$, as $V_i(n)$ accounts for all the times before n that $X_k = i$. Additionally, the triangular

inequality, $|a \pm b| \leq |a| + |b|$, alongside $|a \cdot b| = |a| \cdot |b|$, was used. Furthermore, in the last inequality it was used that, since if we count the visits to all states in I before n we will get that same n and π is a distribution, we have

$$\sum_{i \in I} \frac{V_i(n)}{n} = 1 \Longrightarrow$$

$$\sum_{i \notin J} \frac{V_i(n)}{n} = 1 - \sum_{i \in J} \frac{V_i(n)}{n} = \sum_{i \in I} \pi_i - \sum_{i \in J} \frac{V_i(n)}{n} = \sum_{i \in J} \left[\pi_i - \frac{V_i(n)}{n} \right] + \sum_{i \notin J} \pi_i \leq$$

$$\leq \sum_{i \in J} \left| \pi_i - \frac{V_i(n)}{n} \right| + \sum_{i \notin J} \pi_i.$$

We proved above that

$$\mathbb{P}\left[\frac{V_i(n)}{n} \longrightarrow \pi_i \text{ as } n \longrightarrow \infty \right] = 1.$$

Given $\varepsilon > 0$, choose J finite so that

$$\sum_{i \notin J} \pi_i < \frac{\varepsilon}{4}$$

and $N = N(\omega)$ such that, for $n \geq N(\omega)$,

$$\sum_{i \in I} \left| \frac{V_i(n)}{n} - \pi_i \right| < \frac{\varepsilon}{4}.$$

Then, for $n \geq N(\omega)$

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \overline{f} \right| < 2\frac{\varepsilon}{4} + 2\frac{\varepsilon}{4} = \varepsilon,$$

thus ensuring the desired convergence.

□

Chapter 3

Monte Carlo methods

This chapter has been written following [4].

Monte Carlo methods are simulation tools to calculate integrals and estimate probability distributions, especially those in higher dimensions or that do not have an analytical solution, when numerical methods such as the Simpson rule fail or have a substantially big variance.

The basic idea behind Monte Carlo is the following. Suppose that we want to calculate a certain integral

$$I = \int_a^b f(x)dx. \tag{3.1}$$

By multiplying and dividing (3.1) by $(b-a)$, we get that

$$I = (b-a)\int_a^b f(x)\frac{1}{(b-a)}dx \iff \frac{I}{(b-a)} = \int_a^b f(x)\frac{1}{(b-a)}dx. \tag{3.2}$$

The right-hand side of this last equality can be understood as the expectation of $f(x)$ under a uniform distribution $X \sim Unif(a,b)$.

By the strong law of large numbers, presented in **Theorem 2.36**, we have that, given n samples of the distribution X, then

$$\mathbb{P}\left[\frac{f(X_1)+\ldots+f(X_n)}{n} \longrightarrow \mu \text{ as } n \longrightarrow \infty\right] = 1,$$

where $\mu = \mathbb{E}[f(X)]$, which coincides with the right-hand side of (3.2).

By this method, (3.1) can be estimated as the arithmetic mean of $f(x)$ given a sufficiently big sample of a uniform random variable, and, following that, (3.2) can be obtained by multiplying this result by $(b-a)$.

3.1 Regular Monte Carlo

More generally, following this last reasoning, in order to solve an integral of the form

$$J = \int f(x)h(x)dx = \mathbb{E}_h[f(x)], \tag{3.3}$$

one can use a sample $(X_1, \ldots, X_n)$ generated from the density h and, by the almost sure convergence that the strong law of large numbers guarantees, approximate (3.3) as

$$\mathbb{E}_h[f(x)] = \overline{f(X)} = \frac{1}{n}\sum_{i=1}^{n} f(X_i). \tag{3.4}$$

Note that, in (3.3) and (3.4), the notation has a different meaning than before: now, the subindex does not refer to an initial condition, but to the distribution according to which the expectation is being calculated.

The error in Monte Carlo integration can be assessed through the variance, for the variance of a quantity accounts for how far away the obtained outcome might be from the factual result. The variance can be estimated either as

$$\begin{aligned}
\mathrm{var}[\overline{f(X)}] &= \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}[f(X_i)] \\
&= \frac{1}{n^2} n \cdot \mathrm{var}[f(X_1)] \\
&= \frac{1}{n}\mathbb{E}_h[(f(X) - \mathbb{E}_h[f(X)])^2] \\
&= \frac{1}{n}\int \{f(x) - \mathbb{E}_h[f(X)]\}^2 h(x)dx
\end{aligned}$$

or as the variance of the sample

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[f(X_i) - \overline{f(X)}\right]^2,$$

which accounts for how dispersed the samples are from its mean value.

By taking the definition based on the sample, the central limit theorem can be applied to construct bounds and confidence intervals of the variance when the variance is finite, and to estimate the order of the variance.

The central limit theorem involves the definition of convergence in distribution. Recall that this convergence happens when, given sequence of random variables $\{X_n, n \geq 1\}$ and the sequence of respective probability density functions $\{F_n, n \geq$

$1\}$, there exists a random variable X with probability density function F such that

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

for all $x \in \mathbb{R}$ where F is continuous.

Theorem 3.1. (Multidimensional central limit theorem). *Let $\{X_n, n \geq 1\}$ be an independent and identically distributed sequence of k-dimensional random vectors. Let $S_n = X_1 + \ldots + X_n$. Suppose that each one of the components of X_1 is square integrable and let $\mathbb{E}(X_1) = m$, $\mathbb{E}[(X_1 - m)(X_1 - m)^T] = \Lambda$. Then*

$$\frac{S_n - nm}{\sqrt{n}}$$

converges in distribution to a multidimensional normal random variable $N(0, \Lambda)$.

The proof of this theorem can be found in [5].

In the particular case of one dimension, this result implies that, if $\{X_i, i = 1, \ldots, n\}$ is an iid sequence with mean $\mu = \mathbb{E}(X_i)$ and variance σ^2, then $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$ converges in distribution to a standard normal variable $N(0, 1)$; that is,

$$\lim_{n\to\infty} \mathbb{P}\left[a \leq \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \leq b\right] = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

As a consequence, the exact result of (3.3) will be within the interval $\overline{f(X)} \pm \sigma/\sqrt{n}$ with a probability of 68%, or within the interval $\overline{f(X)} \pm 1.96\sigma/\sqrt{n}$ with a probability of 95%.

Theorem 3.1 also implies that, for n large, the error in the Monte Carlo estimate is of order $1/\sqrt{n}$. This might seem apparently large, as other numerical integration techniques such as the trapezoidal rule and the Simpson's rule have an error of order $1/n^2$ and $1/n^4$, respectively. However, the usefulness of the Monte Carlo method relies in multidimensional integration, when these numerical approximation methods suffer the so-called curse of dimensionality: they have an error of order $1/n^{2/d}$ and $1/n^{4/d}$, respectively, where d is the number of dimensions, whereas the multidimensional central limit theorem still guarantees an error of order $1/\sqrt{n}$ in the variance of Monte Carlo estimates.

Nonetheless, to obtain more accurate results, variance must be reduced. There are several variance reduction techniques. The one that might arise naturally is to increase the size of the sample. This, however, can be computationally and time expensive, and therefore other methods may be preferably implemented, such as using importance sampling.

3.2 Importance sampling

The idea of importance sampling is to calculate (3.3) by generating samples from a distribution that is easier to simulate or that represents a smaller variance than h, since a certain integral is not tied to only one distribution. That is done by considering

$$J = \mathbb{E}_h[f(x)] = \int f(x)h(x)dx = \int f(x)\frac{h(x)}{g(x)}g(x)dx. \tag{3.5}$$

This equality is known as the **importance sampling fundamental identity**.

Then, a sample X_i, $i = 1, \ldots, n$, can be generated from the distribution g and $\mathbb{E}_h[f(x)]$ can be approximated as

$$\mathbb{E}_h[f(x)] = \frac{1}{n}\sum_{i=1}^{n}\frac{h(X_i)}{g(X_i)}f(X_i), \tag{3.6}$$

as the strong law of large numbers guarantees the convergence of this estimate.

3.3 When does Monte Carlo fail?

Apart from when the variance is extremely large, Monte Carlo fails when it is difficult to sample from the distribution of interest.

Either by regular Monte Carlo or by importance sampling, generating samples from a certain distribution is mandatory. Nonetheless, this process can be difficult in several dimensions unless the distribution can be expressed in a product form

$$h(X_1, \ldots, X_n) = \prod_{i \in n} h_i(X_i),$$

for a computer can only generate pseudo-random numbers and the state space becomes exceptionally large. Thus, when multidimensional distributions cannot be expressed as a product of univariate distributions, sampling from it can be remarkably difficult. Therefore, an alternative approach must be chosen. A straightforward, prompt and low-variance process that resolves this difficulty is Markov chain Monte Carlo. Often, it is the only way to address the product-form issue, and, additionally, generating a Markov chain instead of simulating a distribution can have computational and timing advantages.

Chapter 4

Markov chain Monte Carlo methods

One of the most important applications of Markov chains to statistics and physics are Markov chain Monte Carlo (MCMC) methods.

The aim of MCMC is to approximate integrals of the form of (3.3) by generating a Markov chain $(X_n)_{n\geq 0}$ with stationary distribution h. If the chain is irreducible and positive recurrent, then by the ergodic theorem we have that, for any bounded function f,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \longrightarrow \sum_{i\in I} h_i f_i = \mathbb{E}_h[f(X)] \text{ as } n \longrightarrow \infty$$

with probability 1.

By using this method, direct sampling from the distribution h is avoided, which, as mentioned, can have enormous advantages.

This chapter has been written following the articles [6] and [7], as well as the books [4] and [8] for the different versions of the algorithms and the first and eighth chapters of [9] for how to determine the total number of iterations and the burn-in iterations.

4.1 General basis of the algorithms

An irreducible and positive recurrent Markov chain $(X_n)_{n\geq 0}$ with a desired stationary distribution π has to be constructed. The irreducibility of the chain needs to be assessed in each particular case. This can be done by drawing the associated diagram if the state-space is not too large, or by checking that the transition probabilities between all states are positive for some n. Positive recurrence is guaranteed by the fact that the chain has a stationary distribution, as was seen in

Theorem 2.32. On the other hand, the convergence to the stationary distribution is ensured by fulfilling the detailed balance equation

$$\pi_i p_{ij} = \pi_j p_{ji},$$

where P is the probability transition matrix of the chain. This is accomplished by defining P as

$$p_{ij} = q_{ij}\alpha_{ij}, \text{ for } i \neq j$$
$$p_{ii} = 1 - \sum_{j \neq i} p_{ij},$$

where $Q = (q_{ij} : i, j \in I)$ is the transition matrix of an arbitrary irreducible Markov chain with state space I and α_{ij} is defined as

$$\alpha_{ij} = \frac{s_{ij}}{1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}},$$

where s_{ij} is a symmetric function of i and j such that $0 \leq \alpha_{ij} \leq 1$ for all i, j.

Note that, with this transition matrix, we have that

$$\pi_i p_{ij} = \pi_i q_{ij} \frac{s_{ij}}{1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}} = \frac{\pi_i q_{ij} s_{ij}}{\frac{\pi_j q_{ji} + \pi_i q_{ij}}{\pi_j q_{ji}}} = \frac{\pi_j q_{ji} s_{ji}}{\frac{\pi_j q_{ji} + \pi_i q_{ij}}{\pi_i q_{ij}}} = \pi_j q_{ji} \frac{s_{ji}}{1 + \frac{\pi_j}{\pi_i}\frac{q_{ji}}{q_{ij}}} = \pi_j p_{ji}.$$

Thus, the detailed balance equation is fulfilled and therefore, by **Lemma 2.26**, π is invariant for P. Then, if P is irreducible, the ergodic theorem ensures the desired expectation convergence.

The meaning of this definition of transition matrix P is that, if the chain is in state i, a candidate to the next state j is chosen according to the transition probabilities q_{ij}. Then, this candidate is accepted with probability α_{ij}. If j is not accepted, then the following state is, once again, i.

To fulfill the condition $0 \leq \alpha_{ij} \leq 1$ for all i, j, s_{ij} can generally be defined as

$$s_{ij} = g\left[\min\left(\frac{\pi_i q_{ij}}{\pi_j q_{ji}}, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right)\right],$$

where $g(x)$ is a function such that $0 \leq g(x) \leq 1 + x$ for $0 \leq x \leq 1$. Thus, $0 \leq s_{ij} \leq 1 + \min(\pi_i q_{ij}/\pi_j q_{ji}, \pi_j q_{ji}/\pi_i q_{ij})$, ensuring that $0 \leq \alpha_{ij} \leq 1$. Generally, only choices of s_{ij} that solely involve the quantity $\pi_j q_{ji}/\pi_i q_{ij}$ and its inverse are used. This quantity is known as the **test ratio** or **acceptance ratio**. The test ratio encodes how probable the new proposed sample value is with respect to the current sample value, according to π.

Since only the test ratio enters the simulation, the distribution π needs to be known only up to a multiplicative constant beforehand. This is another advantage

of MCMC methods in front of regular Monte Carlo. However, if the multiplicative constant is different to 1 (that is, if we use a chain with stationary distribution π and $\pi_0 + \ldots + \pi_S \neq 1$, where S is the number of states in the state space), we are, indeed, estimating the normalized expectation

$$E(f) = \frac{\sum_{i=0}^{S} f(X_i)\pi_i}{\sum_{i=0}^{S} \pi_i}. \tag{4.1}$$

If we want to estimate a normalized integral, as usually happens in statistical physics, this comes handy.

Another advantage of only using the test ratio is that, then, continuous distributions can be used, and not only discrete ones. If we consider $\pi(x_i)d\mu(x_i)$, $p(x_i, x_j)d\mu(x_j)$ and $q(x_i, x_j)d\mu(x_j)$ to be the probability elements for the continuous distributions π, p and q, when entering them in the definition of α and p we have that

$$\alpha_{ij} = \frac{s_{ij}}{1 + \frac{\pi(x_i)d\mu(x_i)q(x_i,x_j)d\mu(x_j)}{\pi(x_j)d\mu(x_j)q(x_j,x_i)d\mu(x_i)}} = \frac{s_{ij}}{1 + \frac{\pi(x_i)q(x_i,x_j)}{\pi(x_j)q(x_j,x_i)}},$$

$$p(x_i, x_j)d\mu(x_j) = q(x_i, x_j)d\mu(x_j)\alpha_{ij} \Longrightarrow p(x_i, x_j) = q(x_i, x_j)\alpha_{ij},$$

thus obtaining the same working mechanism that we had for discrete distributions.

However, it is notable that, because of working with computers, which can only generate a finite quantity of numbers, we will never work with continuous distributions, only with discrete approximations of such distributions.

Different variants of this generic algorithm can be implemented, depending on the requirements and characteristics of the different problems to solve. The most relevant of them are the Metropolis algorithm, the Metropolis-Hastings algorithm and the Gibbs sampler.

4.2 The Metropolis algorithm

Nicholas Metropolis [6] was the first mathematician to think of and use MCMC in the context of statistical physics, and Wilfred Hastings [7] later described the mathematical reasoning behind the general algorithm.

For the Metropolis algorithm, the transition probabilities are symmetric, that is, $q(x|y) = q(y|x)$, and the function s_{ij} is

$$s_{ij} = \begin{cases} 1 + \frac{\pi_i}{\pi_j} & \text{if } \frac{\pi_j}{\pi_i} \geq 1 \\ 1 + \frac{\pi_i}{\pi_j} & \text{otherwise.} \end{cases}$$

This implies that the probability of acceptance will have the form

$$\alpha = \min\left[1, \frac{\pi(Y_t)}{\pi(X_t)}\right].$$

This procedure can be implemented in a conventional computer as follows:

1. Initialize the Markov chain by choosing any state X_0 for $t = 0$.

2. Given X_t, repeat:

   (a) Generate $Y_t \sim q(\cdot|X_t)$.

   (b) Generate $U \sim Unif(0,1)$.

   (c) Accept Y_t with probability $\alpha(X_t, Y_t)$, where

   $$\alpha(X_t, Y_t) = \min\left[1, \frac{\pi(Y_t)}{\pi(X_t)}\right];$$

   that is, if $u \leq \alpha(X_t, Y_t)$, then $X_{t+1} = Y_t$; otherwise, $X_{t+1} = X_t$.

According to this algorithm, whenever we try to move to a point in a higher-density region (that is, a more probable point than the current state), the move will always be accepted. On the other hand, moving to lower density regions will not always be accepted. In this way, the algorithm makes the chain remain in higher-density regions of π, thus obtaining larger samples from these parts, while only sporadically visiting the lower-density regions, performing an adequate sampling of the distribution.

The choice of q_{ij} should be so that the transition probabilities are as easy to simulate as possible. In that sense, normal distributions or uniform distributions centered in X_t are commonly used. Different methods to simulate probability distributions can be found in [1] and [8].

4.3 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a generalization of the Metropolis algorithm for the case in which the transition probabilities are not symmetric. Thus, the acceptance probability is

$$\alpha(X_t, Y_t) = \min\left[1, \frac{\pi(Y_t)q(X_t|Y_t)}{\pi(X_t)q(Y_t|X_t)}\right].$$

4.4 The Gibbs sampler

Gibbs sampling is a special case of the Metropolis-Hastings algorithm. It concerns multidimensional random variables of the form $X = (x_i, \ldots, x_n)$, with a joint distribution $p(x_1, \ldots, x_n)$, and is especially useful when it is difficult to simulate the joint distribution of them, but not the conditional distribution.

The algorithm works as follows:

1. A starting value $X^{(0)}$ is defined.
2. To generate $X^{(i+1)} = (x_1^{(i+1)}, \ldots, x_n^{(i+1)})$, we sample each component from the conditional distribution $p(x_j^{(i+1)} | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

In this algorithm, all samples are used, without discarding any of them, contrary to what happened in the Metropolis and Metropolis-Hastings algorithms.

4.5 Importance sampling

As happened with regular Monte Carlo, importance sampling can be implemented. This can be done by considering another stationary distribution π'. Note that (4.1) can be rewritten as

$$E(f) = \frac{\frac{\sum_{i=0}^{S}[f(X_i)/\pi_i']\pi_i}{\sum_{i=0}^{S}\pi_i'}}{\frac{\sum_{i=0}^{S}(\pi_i/\pi_i')\pi_i'}{\sum_{i=0}^{S}\pi_i'}}.$$

Thus, the integral of interest can be computed as

$$E_\pi(f) = \frac{\sum_{i=1}^{n}[f(X_i)\pi_i/\pi_i']/n}{\sum_{i=1}^{n}(\pi_i/\pi_i')/n}.$$

4.6 Determining the total number of iterations

As it has been previously mentioned, MCMC methods rely on asymptotic results, and therefore they require a large number of iterations. Instead of running an unnecessarily overly long chain, the required number of iterations to ensure the convergence of the chain to its stationary distribution can be assessed while running the simulation.

A method to determine whether the chain has converged or not is to run several chains with different starting points, which need to be sufficiently spaced from each other. When the chains have "forgotten" the initial states and present a similar behavior, convergence can be assumed. That can be evaluated by comparing

the variance between the different sequences with the variance of each individual sequence. When the former is smaller than the latter, approximate convergence can be presumed, according to the own variability of the chains. This method is known as the **Gelman-Rubin convergence diagnostic**. To see a thorough description of the method and the theoretical basis behind it in terms of underestimates and unbiased overestimates, see [10].

Let us consider m Markov chains, $m \geq 2$, each one with a different initial state, the same transition matrix and the same number of iterations, n. In this development, the subindex i will refer to the labelling of the different chains, $i = 1, \ldots, m$, whereas the subindex j will refer to the position of the states in a single chain, $j = 1, \ldots, n$.

With this notation, the variance between sequences can be computed as

$$B = \frac{n}{m-1} \sum_{i=1}^{m} (\overline{X_i} - \overline{X}) = \frac{n}{m-1} \sum_{i=1}^{m} \left[\frac{1}{n} \sum_{j=1}^{n} X_{ij} - \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{n} \sum_{j=1}^{n} X_{ij} \right) \right],$$

where $\overline{X_i} = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$ is the average of a certain chain and $\overline{X} = \frac{1}{m} \sum_{i=1}^{m} \overline{X_i}$ is the average of all chain averages.

Additionally, the variance within sequences is

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2 = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \overline{X_i})^2 \right],$$

where $s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \overline{X_i})^2$ is the variance of a single sequence. Thus, W corresponds to the mean variance of all the different chains.

Incorporating these two variance components, we can compute the following estimate of the variance of X:

$$\text{var}(X) = \frac{n-1}{n} W + \frac{1}{n} B.$$

Now, let R be defined as

$$R = \frac{\text{var}(X)}{W} = 1 - \frac{1}{n} + \frac{1}{n} \frac{B}{W}.$$

As $n \longrightarrow \infty$, if B and W are indistinguishable, R will tend to 1. Note that the fact that the fact that the between and within sequences variances become similar implies that all the chains present such a resemblant behavior that they cannot be discerned from each other. This happens they have reached their invariant distribution.

Generally, values of R between 0.97 and 1.03 imply that the convergence of the chains has been achieved, thus setting the number of required iterations.

Furthermore, testing convergence is not only convenient to address the total number of iterations. It is important to check for undiagnosed slow convergence. If, for instance, we generated a single chain and it converged slowly or got trapped in a region of high probability that does not correspond to the stationary distribution, it would be unnoticeable. Assessing convergence by running different chains can help identify these difficulties too.

4.7 Thermalization

As it has already been discussed, when using MCMC we are referring to asymptotic results: only the limiting distribution is of interest. For that reason, it is common to discard the first iterations of the chain where it explores the state space but still has not reached the stationary distribution. The consequence of this procedure is that the average of $f(X_n)$ will be a much more precise estimate of the expectancy $\mathbb{E}_h[f(x)]$. This process is called **thermalization** or **burn-in**.

By discarding these first M iterations, if the total number of iterations is N, then the average of f will be calculated as

$$\frac{1}{N-M}\sum_{k=M+1}^{N} f(X_k).$$

To determine the number of burn-in iterations, the Gelman-Rubin convergence diagnostic can be used as well, since it points out when the chains converge and after that point all the samples will be of interest to us. Generally, the burn-in iterations should represent 1% or 2% of the total iterations of the chain.

4.8 Multidimensional MCMC

As mentioned, avoiding sampling from a certain distribution finds its greatest advantage when dealing with the multidimensional case. In that case, we have a product state space of the form

$$I = \prod_{m\in\Lambda} S_m,$$

where λ is a finite set and S_m are state spaces that can be arbitrarily large, thus making I immensely large. When the distribution of interest cannot be expressed as a product of univariate distributions, Markov chain Monte Carlo becomes the only feasible strategy.

Suppose that the invariant distribution π is d-dimensional. Then, the simulated Markov chain will have the form $X(t) = (X_1(t), \ldots, X_d(t))$. The usefulness of

Markov chain Monte Carlo relies on the fact that every coordinate can evolve separately from the others, even if the transition probabilities cannot be expressed as a product form. Thus, one can either change all coordinates at a time, change one coordinate selected at random, or change one coordinate selected following a determinate sequence. In these last two cases, measures of the process need to be made at time $0, d, 2d, \ldots$.

4.9 Worked example: calculation of $\Gamma(3/2)$

As an example of the implementation of Markov chain Monte Carlo, we used a Metropolis algorithm to calculate the integral

$$\Gamma(3/2) = \int_0^\infty x^{1/2} e^{-x} dx = \frac{\sqrt{\pi}}{2} \approx 0.886227. \tag{4.2}$$

The graphics that are referred to in this section can be found in **Appendix A**. The code that was run to obtain the results is included in **Appendix B**.

A plot of the function to integrate is presented in **Figure 1**. From this plot, it is notable that the regions with a higher concentration of volume are for $x < 5$, and therefore a Markov chain Monte Carlo method should concentrate the sampling in this region.

By using an integral whose value is already known, we can compare the obtained simulation result with the theoretical result and calculate the corresponding relative error, which gives us an idea of how far the simulated result is from the factual value. This is recommended for all Monte Carlo methods: to use the algorithm first on a known result before confronting an unsolved problem to test its accuracy.

To implement the Markov chain Monte Carlo method, the function $f(x) = \sqrt{x}$ was thought of as the function of interest, whereas $h(x) = e^{-x}$ was used as the probability distribution according to which the expectation of $f(x)$ is being calculated. Thus, $h(x)$ will correspond to the invariant distribution of the chain. Note that, in this case, $h(x)$ corresponds to an exponential distribution of parameter $\lambda = 1$, and therefore it is already normalized. On the contrary, if we were to use an unnormalized distribution, the result of the algorithm would be normalized by the normalization constant of the distribution.

To implement the algorithm, the transition probabilities q where defined as $q(Y_t|X_t) = X_t + 0.25u$, where u is a random number generated from a uniform distribution $Unif(-0.5, 0.5)$. Thus, q is symmetric. Furthermore, q is easily implementable in a computer, which is its greatest attractive. Nonetheless, since only positive values of x are of interest to us, we will only accept values of Y_t that

are positive. u was generated by using the `drand48()` function of *C*, which has a period of 10^{14}. This means that the sequence of pseudo-random numbers generated from this function will not repeat itself until the 10^{14}th time that the function is used. This should be large enough to support the generation of several long Markov chains.

Note that, in our case, the transition probabilities correspond to a continuous distribution and the state space is, theoretically, infinite. To use discrete transition probabilities and a transition probability as the ones presented in **Chapter 2**, a discretization of the state space would be needed. However, it is not necessary, since a computer can only simulate a finite number of states and therefore, in practice, we are dealing with a finite state space and discrete transition probabilities.

In order to implement the Gelman-Rubin criterion, three different chains were run, with initial states 0, 2 and 5, respectively. The condition $0.97 < R < 1.03$ was fulfilled at the iteration 6332. Since the burn-in iterations should represent around 1% of the total number of iterations, the latter was set to be 622300, and the first 6332 iterations were discarded. A representation of the between-chains variance, the within-chains variance and the parameter R is depicted in **Figure 2**. From this plot, it is notable that as the number of iterations increases the different variances of the chains decrease, which indicates the progressive convergence to the stationary distribution.

The first 15000 states of the chains are represented in **Figure 3**. The totality of the chains was not represented, as the number of iterations is extremely large and for a number of iterations larger than 6332 the chains have already converged. It is notable that, generally, many of the values of the chains are below 3, and the majority of them is below 1.5. This indicates that the chains do a more thorough exploration and take more values from that region, which is the region of higher probability of the function $\sqrt{x}e^{-x}$. To further exemplify this phenomenon, an example of a chain with initial state $X_0 = 50$ and 50000 iterations is plotted in **Figure 4**. Here, it is clear that, although the chain started at a value of low probability, it rapidly reaches lower values of x, where it remains exploring the space state as it is a region of higher probability.

To prove the convergence of the chain to the desired stationary distribution, e^{-x}, a histogram of the values of the chain after the burn-in is presented in **Figure 5**. The chain had starting value $X_0 = 2$ and runs for 50000 iterations. However, the first 6332 iterations, corresponding to the burn-in, were discarded, since in that interval of time the chain still had not converged to the stationary distribution. Thus, only 43668 values were used. As can be undoubtedly seen in the plot, the distribution of values in intervals adequately matches the expected exponential distribution. This manifests the usefulness of Markov chain Monte Carlo methods

not only to calculate integrals, but also to sample from probability distributions.

The evolution of $\overline{\sqrt{X_k}} = \frac{1}{k}\sum_{i=0}^{i=k}\sqrt{X_i}$, which, by the ergodic theorem, converges to the value of (4.2), is presented in **Figure 6**. The convergence to the expected value of the integral of each one of the different chains as the number of iterations increases is clear in the graphic, especially after the thermalization has ended.

The result of the integral by using the Markov chains is calculated as

$$\overline{\sqrt{X}} = \frac{1}{633200 - 6332}\sum_{i=6332}^{i=633200}\sqrt{X_i}.$$

By using this, the obtained values are 0.942024 for the chain with starting value $X_0 = 0$, 0.883861 for the chain with starting value $X_0 = 2$ and 0.892217 for the chain with starting value $X_0 = 5$. These results are strongly similar, which confirms the independence of the result and the initial value of the chain. In addition, they represent a relative error of 6.3%, 0.27% and 0.68%, respectively. Since the relative errors are small, especially those of the starting values 2 and 5, we can conclude that the implemented Markov chain Monte Carlo method has successfully computed (4.2), an otherwise not analytically solvable integral.

Chapter 5

Conclusions

The main goal of this work was to study the properties and behavior of discrete-time Markov chains $(X_n)_{n\geq 0}$ with a finite state-space I, and later, as an application of these, to describe Markov chain Monte Carlo methods.

As it has been seen, discrete-time Markov chains with a finite state space can be described by an initial distribution λ and a stochastic matrix P that stands for the transition probabilities between states after a unit of time. A distinctive trait of this type of stochastic processes is that the transitions from a state to the following state only depend on the current state of the chain, and not on the states previous to that. Moreover, the powers of the transition matrix represent the transition probabilities between states in the number of states corresponding to the power, and if we first multiply that by the initial distribution we obtain the probability distribution of states.

The fact that the probability of transition between states in any number of steps is positive establishes a recurrence relation in the state space, thus partitioning the latter into communicating classes. Regarding communicating classes, the chains of most interest are irreducible chains, which have only a communicating class that englobes the entirety of the state-space, and thus all states are accessible from any other state. Many properties of the states are class properties, such as recurrence, transience, positive recurrence and aperiodicity. A recurrent state is a state that the chain will always keep visiting, while a transient state is a state that the chain will eventually not visit anymore. Each state in a chain is either recurrent or transient. A positive recurrent state is a recurrent state that, additionally, presents a finite expected return time. On the other hand, an aperiodic state is a state that can be visited for all sufficiently large number of states.

Markov chains may have stationary distributions, that is, probability distributions of the state space that do not vary after a unit of time: $\lambda P = \lambda$. If a chain is irreducible, having a stationary distribution and being positive recurrent are

equivalent. Additionally, if a distribution is in detailed balance with P, that is, $\lambda_i p_{ij} = \lambda_j p_{ji}$ for all i, j, then λ is stationary. One of the most important results concerning stationary distributions regards the asymptotic convergence of chains: if they are irreducible and aperiodic and have a stationary distribution, then the probability distribution of the states tends to the stationary distribution when time tends to ∞.

Another result regarding the asymptotic behavior of Markov chains is the ergodic theorem, which ensures, amongst others, that, given any bounded function $f : i \longrightarrow \mathbb{R}$ and an irreducible and positive recurrent chain, then $\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$ converges to $\sum_{i \in I} \pi_i f_i$ as time tends to ∞. This theorem represent the basis for Markov chain Monte Carlo methods.

In regular Monte Carlo methods, integrals of the form $\int f(x)h(x)dh$ are estimated as expectancies of functions according to a certain distribution, $\mathbb{E}_h[f(X)]$. A number of points are sampled from the distribution, and then the strong law of large numbers ensures the convergence of the mean of the values to the desired expectancy. This procedure, however, can become difficult in several dimensions if the probability distribution from which we need to sample cannot be expressed as a product of unidimensional distributions. It is then when Markov chain Monte Carlo methods are most useful.

In Markov chain Monte Carlo methods, an irreducible and positive recurrent chain with stationary distribution h is generated, and the ergodic theorem then ensures the convergence of the mean value of $f(X)$ to $\mathbb{E}_h[f(X)]$, which estimates the integral. In order to generate a chain with a certain probability distribution, algorithms such as Metropolis-Hastings, Metropolis and the Gibbs sampler are used. In this algorithms, the states of the Markov chain are generated so that the chain does a thorough exploration of higher probability density regions of h and sporadically visits regions of low probability. This procedure avoids direct sampling from h, thus preventing the downsides of regular Monte Carlo. The number of iterations needed for the chains can be determined by running several chains at the same time and comparing the within-chains variance and the between-chains variance: when they are indistinguishable, the chains have converged.

The calculation of $\Gamma(3/2)$ via a Metropolis algorithm with transition probabilities $q(Y_t|X_t) = X_t + 0.25u$ has shown the efficiency of the method, as it has arisen satisfactory results in both computing the integral and sampling the distribution.

Because of the characteristics, reliability and computational advantages of Markov chain Monte Carlo methods, they find their greatest application domains in Bayesian statistics and statistical physics. Both of these fields usually involve integrals of over hundreds of parameters: in the former, to estimate the posterior distributions of parameters, and in the latter, to compute expectations of physical quantities.

Bibliography

[1] Sanz-Solé, M., *Probabilitats*. Edicions de la Universitat de Barcelona, Barcelona, 1999.

[2] Norris, J. M., *Markov chains*. Cambridge University Press, Cambridge, 1997.

[3] Resnick, S. I., *Adventures in stochastic processes*. Birkhäuser, Boston, 1992.

[4] Robert, C. P. and Casella, G., *Monte Carlo statistical methods*. Springer, New York, 2004.

[5] Nualart, and Sanz-Solé. M., *Curs de probabilitats*. Promociones y Publicaciones Universitarias, Barcelona, 1990.

[6] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H., *Equation of state calculations by fast computing machines*. The Journal of Chemical Physics, volume 21, number 6, June 1953.

[7] Hastings, W. K., *Monte Carlo sampling methods using Markov chains and their applications*. Biometrika, volume 57, number 1, April 1970.

[8] Rubinstein, R. Y. and Kroese, D. P., *Simulation and the Monte Carlo method*. Wiley, New Jersey, 2017.

[9] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

[10] Gelman, A. and Rubin, D. B., *Inference from iterative simulation using multiple sequences*. Statistical Science, volume 7, number 4, 1992.

Appendix A: plots to illustrate the simulation of $\Gamma(3/2)$

In this appendix, the different plots and graphics corresponding to the obtained data from the calculation of $\Gamma(3/2)$ by a Metropolis algorithm are included.

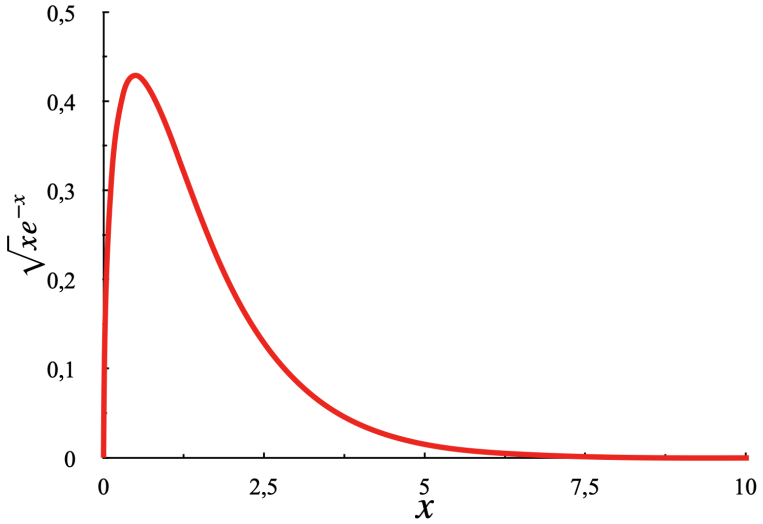


Figure 1: plot of the function $\sqrt{x}e^{-x}$, for positive values of x.

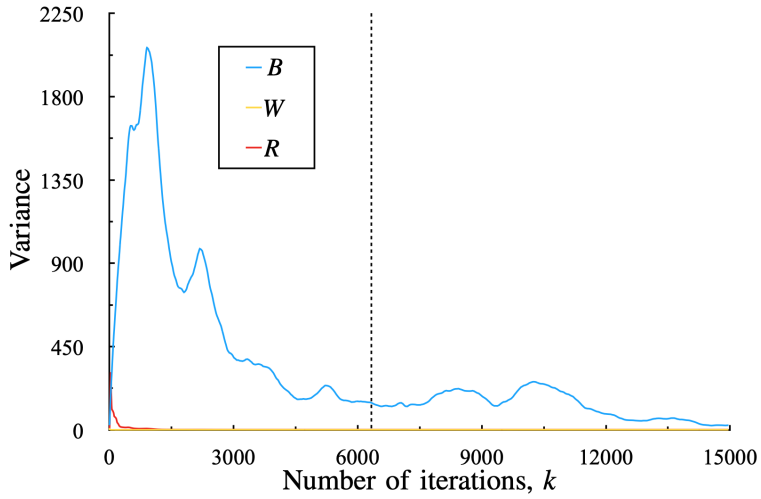


Figure 2: evolution of the different variances of interest with the number of iterations. The iteration 6332, corresponding to the end of the burn-in, is highlighted.

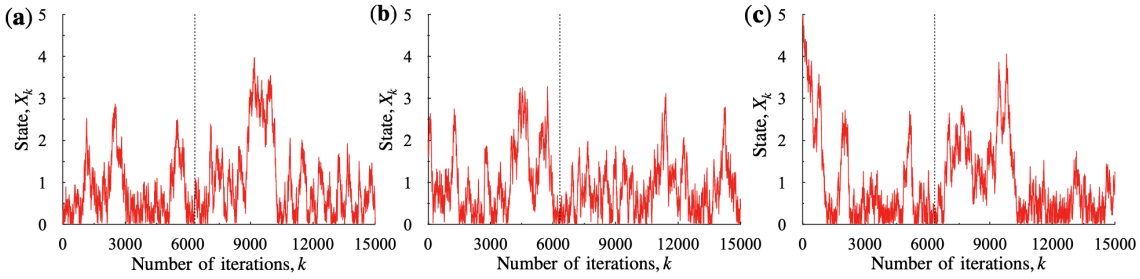


Figure 3: Markov chains with transition probabilities $q(Y_t|X_t) = X_t + 0.25u$, stationary distribution e^{-x} and initial states **(a)** $X_0 = 0$, **(b)** $X_0 = 2$ and **(c)** $X_0 = 5$. The iteration 6332, corresponding to the end of the burn-in, is highlighted.

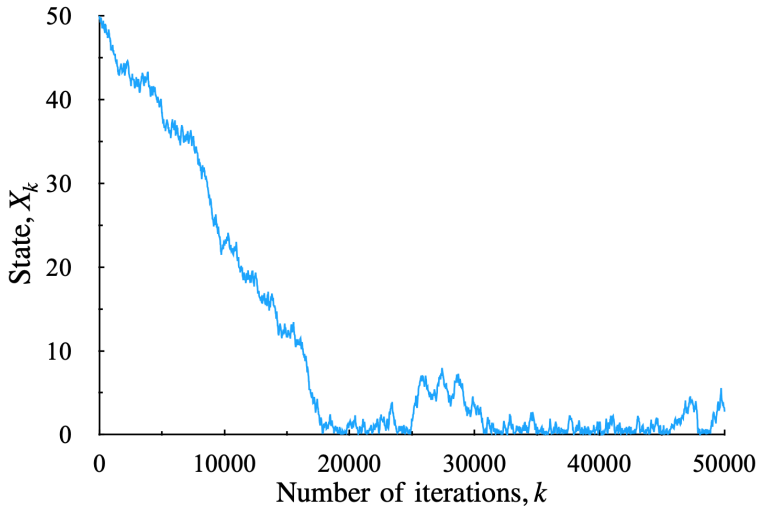


Figure 4: Markov chain with transition probabilities $q(Y_t|X_t) = X_t + 0.25u$, stationary distribution e^{-x} and initial state $X_0 = 50$.

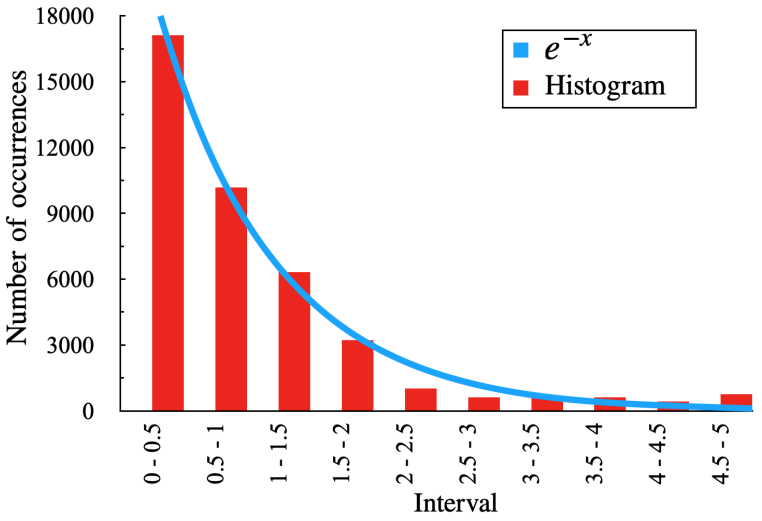


Figure 5: histogram of the different values of a Markov chain with starting value $X_0 = 2$, after the burn-in has ended, for an amount of 43668 iterations. A plot of the function e^{-x} is included for comparison.

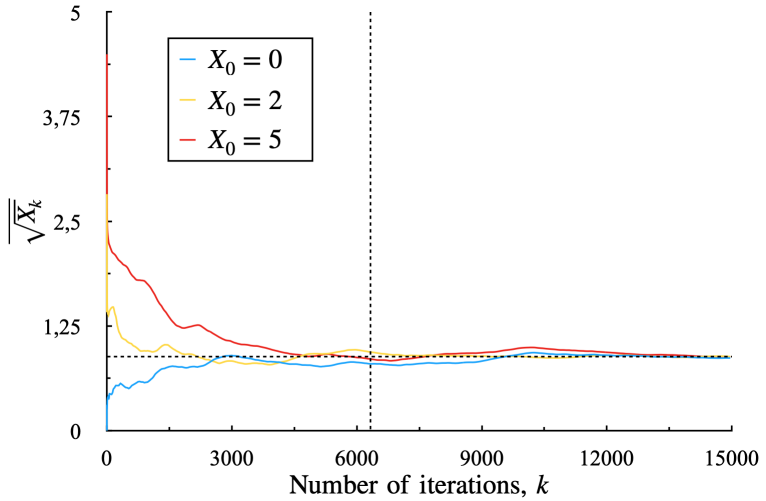


Figure 6: evolution of $\overline{\sqrt{X_k}} = \sum_{i=0}^{i=k} \sqrt{X_i}$ with the number of iterations. The iteration 6332, corresponding to the end of the burn-in, is highlighted, as well as the theoretical value of (4.2).

Appendix B: code of the simulation

In this appendix, the code for the Metropolis algorithm that was used to calculate the value of $\Gamma(3/2)$ is presented. This code is written in C language. Slight variations of this code, such as eliminating the part corresponding to the Gelman-Rubin criterion and/or imposing a certain number of iterations, were used to represent the different figures presented in **Appendix A**.

```
#include <math.h>
#include <time.h>
#include <stdio.h>
#include <stdlib.h>

#define PI 4*atan(1)

double fun(double);

double generate(double);

double alpha(double, double);

int main(void) {
    int i, j, k, conv, num;
    double aux, u, B, W, **cad, *mitja, *mean1, mean2, *si2;
    FILE *cadenes, *mitjanes, *variances;
    char nomcad[20], nommit[20], nomvar[20];

/*Opening of the different files used to draw the plots*/
    printf("Name of the chains' file: \n");
    scanf("%s", nomcad);
```

```
    cadenes = fopen(nomcad, "w");

    if(cadenes == NULL) {
        printf("Error. \n");
        exit(1);
    }

    printf("Name of the means' file: \n");
    scanf("%s", nommit);

    mitjanes = fopen(nommit, "w");

    if(mitjanes == NULL) {
        printf("Error. \n");
        exit(1);
    }

    printf("Name of the variances' file: \n");
    scanf("%s", nomvar);

    variances = fopen(nomvar, "w");

    if(variances == NULL) {
        printf("Error. \n");
        exit(1);
    }

/*Reading the number of chains to simulate and their initial states*/
    printf("Number of chais to simulate: \n");
    scanf("%d", &num);


    cad = (double **)malloc(num*sizeof(double *));

    if(cad == NULL) {
        printf("Error. \n");
        exit(1);
    }
```

```
    for(i = 0; i < num; i++) {
        cad[i] = (double *)malloc(5000000*sizeof(double));
        if(cad[i] == NULL) {
            printf("Error. \n");
            exit(1);
        }
    }

    mitja = (double *)malloc(num*sizeof(double));

    if(mitja == NULL) {
        printf("Error. \n");
        exit(1);
    }

    mean1 = (double *)malloc(num*sizeof(double));

    if(mean1 == NULL) {
        printf("Error. \n");
        exit(1);
    }

    si2 = (double *)malloc(num*sizeof(double));

    if(si2 == NULL) {
        printf("Error. \n");
        exit(1);
    }

    for(i = 0; i < num; i++) {
        printf("Initial state of the chain n. %d : \n", i + 1);
        scanf("%le", &cad[i][0]);
    }

/*Initializing the different variables*/
    for(i = 0; i < num; i++) {
        for(j = 1; j < 5000000; j++) {
            cad[i][j] = 0;
        }
```

```
    }

    for(i = 0; i < num; i++) {
        mitja[i] = 0;
        fprintf(mitjanes, "%le \t", mitja[i]);
    }

    fprintf(mitjanes, "\n");

    for(i = 0; i < num; i++) {
        mean1[i] = cad[i][0];
    }

    i = 0;

/*Gelman–Rubin convergence diagnostic*/
    do {
        B = 0;

        W = 0;

        mean2 = 0;

        for(j = 0; j < num; j++) {
            si2[j] = 0;
        }

        for(j = 0; j < num; j++) {
            fprintf(cadenes, "%le \t", cad[j][i]);

/*A candidate is generated*/
            aux = generate(cad[j][i]);

            u = drand48();

/*The candidate is accepted or rejected*/
            if(u < alpha(cad[j][i], aux)) {
                cad[j][i + 1] = aux;
            } else {
```

```
                cad[j][i + 1] = cad[j][i];
            }
        }

/*Mean value of every chain*/
        for(j = 0; j < num; j++) {
            mean1[j] = mean1[j] + cad[j][i + 1];
        }

/*Mean of the means of the chains*/
        for(j = 0; j < num; j++) {
            mean2 = mean2 + mean1[j]/(i + 1);
        }

        mean2 = mean2/num;

/*Variance between chains*/
        for(j = 0; j < num; j++) {
            B = B + (mean1[j]/(i + 1) - mean2)*
                (mean1[j]/(i + 1) - mean2);
        }

        B = B*(i + 1)/(num - 1);

/*Variance of every chain*/
        for(j = 0; j < num; j++) {
            for(k = 0; k < i + 1; k++) {
                si2[j] = si2[j] + (cad[j][k] - mean1[j]/(i + 1))*
                         (cad[j][k] - mean1[j]/(i + 1));
            }
            si2[j] = si2[j]/(i + 1);
        }

/*Variance within chains*/
       for(j = 0; j < num; j++) {
            W = W + si2[j];
        }

        W = W/num;
```

```
        fprintf(variances, "%le \t %le \t %le \n", B, W,
                        1 - 1/(i + 1) + B/(W*(i + 1)));

        i = i + 1;

    } while((1 - 1/i + B/(W*i)) > 1.03 || (1 - 1/i + B/(W*i)) < 0.97);

    printf("The chains have converged at the iteration %d \n", i);

    conv = i*100;

/*Proper chain simulation*/
    for(i = conv/100; i < conv; i++) {
        for(j = 0; j < num; j++) {
            fprintf(cadenes, "%le \t", cad[j][i]);

/*A candidate is generated*/
            aux = generate(cad[j][i]);

            u = drand48();

/*The candidate is accpented with probability alpha*/
            if(u < alpha(cad[j][i], aux)) {
                cad[j][i + 1] = aux;
            } else {
                cad[j][i + 1] = cad[j][i];
            }

/*Updating of the mean value*/
            mitja[j] = mitja[j] + sqrt(cad[j][i + 1]);

            fprintf(mitjanes, "%le \t", mitja[j]/(i + 1 - conv/100));
        }

        fprintf(cadenes, "\n");

        fprintf(mitjanes, "\n");
```

```
    }

    printf("Expected result: %le \n", sqrt(PI)/2);

    for(i = 0; i < num; i++) {
        printf("Mean of the chain n. %d : %le \n",
                i + 1, mitja[i]/(conv - conv/100));
    }

    for(i = 0; i < num; i++) {
        printf("Relative error of the chain n. %d : %le per cent \n",
         i + 1, fabs(mitja[i]/(conv - conv/100) -
         sqrt(PI)/2)/(sqrt(PI)/2)*100);
    }


    fclose(cadenes);

    fclose(mitjanes);

    fclose(variances);


    for(i = 0; i < num; i++) {
        free(cad[i]);
    }

    free(cad);

    free(mitja);

    free(mean1);

    free(si2);

    return 0;

}
```

```
/*Function that calculates the exponential of a value*/
double fun(double x) {

    return exp(-x);

}


/*Function that generates positive candidates according to
the transition probabilities*/
double generate(double xt) {
    double au;

    au = -1;

    while (au < 0) {
        au = xt + 0.25*(drand48() - 0.5);
    }

    return au;

}


/*Function that calculates alpha*/
double alpha(double xt, double yt) {

    if(1 < fun(yt)/fun(xt)) {
        return 1;
    } else {
        return fun(yt)/fun(xt);
    }
}
```