National and Kapodistrian University of Athens

MSc Thesis

The Bootstrap Method for Discrete-Time Markov Chains and Applications

Author:
Panagiotis Andreou

Supervisor:
Dr. Samis Trevezas

A thesis submitted in partial fulfillment of the requirements for the degree of M.Sc. in Statistics and Operations Research

in the

Faculty of Science
Department of Mathematics

June 30, 2023

"The great thing about Statistics is that you get to play in everyone's back yard."

John Tukey

The Bootstrap Method for Discrete-Time Markov Chains and Applications

National & Kapodistrian University of Athens
Department of Mathematics

Panagiotis Andreou

Abstract

The present thesis aims at presenting the application of the Bootstrap method on dependent data. The Bootstrap method was introduced in the late '70s by the eminent statistician Bradley Efron, bringing a revolution in Statistics and many other related fields. In its initial formulation, this method considered independent data. Here we are dealing with data that exhibit a particular type of dependence, that of a discrete-time Markov chain. In the first part of the thesis, we provide a theorem-proof type of presentation of the basic Markov chain theory, both with discrete and arbitrary state space. In the second part, we focus on the problem of estimating the transition matrix of a Markov chain based on an observed path of the chain. We first examine how we can use asymptotic methods to tackle this problem, presenting both the classical and the Bayesian framework. Then, we show how we can exploit the Bootstrap method to approach this problem. We delve into both the frequentist and the Bayesian frameworks of tackling this problem, and we give detailed proofs of the main asymptotic results that validate these procedures. Finally, we apply the above theoretical methods in simulated and real data.

Η Μέθοδος Bootstrap στις Μαρκοβιανές Αλυσίδες Διακριτού Χρόνου και Εφαρμογές

Εθνικό & Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

Παναγιώτης Ανδρέου

Περίληψη

Η παρούσα διπλωματική εργασία αποσκοπεί στο να παρουσιάσει την εφαρμογή της μεθόδου Bootstrap σε εξαρτημένα δεδομένα. Η μέθοδος Bootstrap εισήχθη στα τέλη της δεκαετίας του 1970, φέρνοντας επανάσταση στη Στατιστική και σε πολλές επιστήμες που κάνουν χρήση αυτής. Ωστόσο, στην αρχική της μορφή η μέθοδος αυτή αφορούσε ανεξάρτητα δεδομένα. Εδώ, ασχολούμαστε με δεδομένα που έχουν μία συγκεκριμένη δομή εξάρτησης, αυτήν της Μαρκοβιανής αλυσίδας διακριτού χρόνου. Στο πρώτο μέρος της εργασίας, κάνουμε μία παρουσίαση της βασικής θεωρίας των Μαρκοβιανών αλυσίδων διακριτού χρόνου, τόσο με διακριτό όσο και με γενικό χώρο καταστάσεων. Στο δεύτερο μέρος της εργασίας, εξετάζουμε πώς η μέθοδος Bootstrap μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε τον πίνακα πιθανοτήτων μετάβασης μίας Μαρκοβιανής αλυσίδας έχοντας παρατηρήσει ένα μονοπάτι της. Εξετάζουμε και την κλασική και την Μπεϋζιανή αντιμετώπιση αυτού του προβλήματος και παρουσιάζουμε αναλυτικά τις αποδείξεις βασικών αποτελεσμάτων που αναδεικνύουν τις ασυμπτωτικές ιδιότητες των εκτιμητριών Bootstrap. Στο τελευταίο κεφάλαιο της διπλωματικής εργασίας, παρουσιάζονται κάποιες εφαρμογές σε προσομοιωμένα και πραγματικά δεδομένα.

Acknowledgments

The origins of this thesis lie in the sudden outburst of a global pandemic and my eventual inability to travel abroad and start my path as a graduate student. Despite the relative spontaneity of my decision to attend this MSc program, the current thesis is the result of the long-term help from numerous people.

First, I would like to thank my advisor, Dr. Samis Trevezas, for spending so much valuable time with me over the past few years, caring to help me improve both in academic and non-academic aspects. In every course I attended with Dr. Trevezas, he was trying his best to expose us to several concepts, always rooting for deep understanding and creativity rather than a constant replication of standard ideas. It should also be noted that this thesis would have never occurred if it weren't for Dr. Trevezas' crucial contribution in my acceptance to this MSc program after I had missed the application's deadline. As soon as he found out that I was no longer able to travel in the coming year, he jumped in and gave me a spot in the candidates' interviews. Later on, he came up with my thesis' topic, always showing great flexibility with me completing this thesis while being abroad.

Over the past few years, several other professors stood out for their outstanding teaching skills and personalities, teaching me plenty of things, both at a human and a mathematical level. These names include Professors Antonis Economou, Apostolos Giannopoulos, Loukia Meligkotsidou and Apostolos Burnetas from the National & Kapodistrian University of Athens, and Professors Amarjit Budhiraja, Sayan Banerjee, Jan Hannig and Mariana Olvera-Cravioto from the University of North Carolina at Chapel Hill. I thank them all for showing me how teaching should be done.

I would also like to thank my dear friends Giorgos, Apollonas, Kostis, Achilleas, George, Maria, Anna, Danai and Panagiota for the wonderful experiences we shared during our undergraduate years. No matter how many times I declined their invitations, being an antisocial geek, they kept inviting me!

None of this would have been possible without the constant support of my family over all these years. They stood by my side in every small step I was about to take. I am truly grateful to them for their selfless patience and devotion.

Lastly, there is a person that deserves a whole paragraph for himself, and even that will not be adequate. It is very likely that I would have never gotten into Statistics if it weren't for my dear friend Vasileios Katsianos, currently a Ph.D. student at the University of Chicago. First as a teaching assistant in the computer labs and later on as a friend, Vasilis kept helping me evolve as a person, showing tremendous amounts of patience and sanity every time I was not able to think for myself. This thesis is devoted to him.

Chapel Hill, NC

June, 2023

Contents

List of Figures

Part I

Probability Theory for Markov Chains

Chapter 1

Discrete-Time Markov Chains

Markov chains are some of the simplest and most useful stochastic processes. Roughly speaking, by a stochastic process we shall think of a collection of random phenomena whose evolution is examined over the passing of time.

1.1 Discrete state-space

1.1.1 Basic definitions and properties

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and S a countable[1] set in which the random variables take values. By random variable we mean a measurable function $X : \Omega \to S$ such that

$$p_i = \mathbb{P}(X = i) = \mathbb{P}(\{\omega \in \Omega : \, X(\omega) = i\}).$$

An $S \times S$ matrix $P = (p_{ij} : i, j \in S)$ will be called *stochastic* if every row is a distribution on S, i.e., if for every $i \in S$, $\sum_{j \in S} p_{ij} = 1$. We are now able to define a discrete-time Markov chain.

Definition 1.1.1 (Markov Chain). Let λ be a distribution and P a stochastic matrix. The family of random variables $(X_n)_{n \geq 0} = \{X_n : n \geq 0\}$ is called a *Markov Chain* with *initial distribution* λ and *transition matrix* P, if

A. $X_0 \sim \lambda$, i.e., $\mathbb{P}(X_0 = k) = \lambda_k, \ \forall k \in S$,

B. the next state of the chain depends only on its present state, i.e.,

$$\begin{aligned} &\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \\ &\qquad = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) = p_{i_n i_{n+1}}. \end{aligned} \tag{1.1.1}$$

[1] since we have assumed a discrete state-space

We will denote by Markov(λ, P) a Markov chain with initial distribution λ and transition matrix P. If the transition probabilities do not depend on the time at which we are examining the process, the Markov chain will be called *time-homogeneous*. From now on, whenever we say Markov chain, we will mean a time-homogeneous one, unless otherwise stated.

We will now prove a very useful Theorem about Markov chains. In particular, we will show that a Markov chain is entirely determined by its initial distribution and its transition matrix.

Theorem 1.1.1. *A stochastic process $(X_n)_{n=0}^{\infty}$ is Markov(λ, P) if, and only if, for all $N \in \mathbb{N}$ and all $i_0, i_1, \ldots, i_N \in S$,*

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{N-1} i_N}. \tag{1.1.2}$$

Proof. We consider a known result that a stochastic process is characterized by its finite-dimensional distributions[2]. Suppose that $(X_n)_{n=0}^{\infty}$ is Markov(λ, P) and let $N \in \mathbb{N}$. Then, by the multiplicative law of probability, we receive

$$\begin{aligned}
&\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N) = \\
&\quad = \mathbb{P}(X_0 = i_0) \cdot \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) \cdot \ldots \cdot \mathbb{P}(X_N \mid X_0 = i_0, \ldots, X_{N-1} = i_{N-1}) = \\
&\quad = \mathbb{P}(X_0 = i_0) \cdot \mathbb{P}(X_1 = i_1 \mid X_0 = i_0) \cdot \ldots \cdot \mathbb{P}(X_N \mid X_{N-1} = i_{N-1}) = \\
&\quad = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{N-1} i_N},
\end{aligned}$$

as equation (1.1.2) indicates. Conversely, assume that equation (1.1.2) holds for every $N \in \mathbb{N}$. The idea now is to sum over every possible state $i_N \in S$, so that the last term of the intersection will drop and then, using induction, we will get the equation (1.1.1), indicating that $(X_n)_{0 \leq n \leq N}$ is Markov(λ, P). Indeed,

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{N-1} i_N} \Rightarrow$$

$$\sum_{i_N \in S} \mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N) = \sum_{i_N \in S} \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{N-1} i_N} \Rightarrow$$

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{N-1} = i_{N-1}) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{N-2} i_{N-1}},$$

and thus equation (1.1.2) holds for $N - 1$ too. A simple inductive argument shows that equation (1.1.2) holds for every $n = 0, 1, \ldots, N$. Hence, the initial distribution satisfies $\mathbb{P}(X_0 = i_0) = \lambda_{i_0}$ and for every $n = 1, \ldots, N$ we have

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) =$$

[2] this is an application of Dynkin's $\pi - \lambda$ theorem

$$= \frac{\mathbb{P}(X_{n+1} = i_{n+1}, X_0 = i_0, X_1 = i_1, \dots, X_n = i_n)}{\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n)} =$$

$$= \frac{\lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} p_{i_n i_{n+1}}}{\lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}} = p_{i_n i_{n+1}},$$

and the proof is complete. □

Let

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

denote the Kronecker delta. Markov chains have the very useful property that at any point they start all over again, forming a new Markov chain that is independent of the past. This property is called *memorylessness*.

Theorem 1.1.2 (Markov property). *Let $(X_n)_{n\geq 0}$ be Markov(λ, P) and $X_m = i$ for some $m \geq 0$ and $i \in S$. Then, given that $X_m = i$, the process $(X_k)_{k\geq m}$ is Markov(δ_i, P) and is independent of $X_0, X_1, \dots, X_{m-1}$.*

Proof. We want to show that $(\{(X_0, \dots, X_{m-1}), (X_m, X_{m+1}, \dots)\} \mid X_m = i)$ are independent, and the second process is Markov(δ_i, P). The distribution of the finite vector $(X_0, \dots, X_{m-1})$ is determined by a probability function, while the distribution of the (infinite) vector $(X_m, X_{m+1}, \dots)$ is determined by the distributions of all the finite vectors $(X_m, X_{m+1}, \dots, X_{m+n})$, for $n \in \mathbb{N}$. Thus, it suffices to show that

$$\mathbb{P}(X_{0:m+n} = i_{0:m+n} | X_m = i) = \mathbb{P}(X_{0:m-1} = i_{0:m-1} | X_m = i) \cdot \\ \cdot \mathbb{P}(X_{m:m+n} = i_{m:m+n} | X_m = i),$$

for every $n \in \mathbb{N}$, and $\mathbb{P}(X_{m:m+n} = i_{m:m+n} | X_m = i)$ is given by the transition probability matrix P. The latter is immediate since $(X_n)_{n\geq 0}$ is by assumption a Markov chain. For the former, let $n \in \mathbb{N}$. Using the statement of Theorem 1.1.1, we get

$$\mathbb{P}(X_{0:m+n} = i_{0:m+n} | X_m = i) = \delta_{ii_m} \frac{\mathbb{P}(X_0 = i_0, \dots, X_m = i_m, \dots, X_{m+n} = i_{m+n})}{\mathbb{P}(X_m = i)}$$

$$= \frac{\mathbb{P}(X_0 = i_0, \dots, X_m = i_m)}{\mathbb{P}(X_m = i)} \delta_{ii_m} p_{i_m i_{m+1}} \cdots p_{i_{m+n-1} i_{m+n}}$$

$$= \mathbb{P}(X_{0:m-1} = i_{0:m-1} | X_m = i) \cdot \mathbb{P}(X_{m:m+n} = i_{m:m+n} | X_m = i),$$

as we wanted to show. □

Suppose that instead of a one-step transition, we want to examine the state in which the Markov chain will be after n-steps, i.e., for every $i, j \in S$ we are interested in the

probabilities

$$p_{ij}^{(n)} = \mathbb{P}(X_{m+n} = j \mid X_m = i) = \mathbb{P}(X_n = j \mid X_0 = i),$$

for every $n \geq 0$, where by definition we set $p_{ij}^{(0)} = \delta_{ij}$. Let

$$P^{(n)} = (p_{ij}^{(n)} : i, j \in S), \quad n \geq 0$$

denote the n-th order transition matrix. A surprisingly simple result tells us that $P^{(n)}$ is equal to the algebraic n-th power of the transition matrix, i.e., P^n.

Proposition 1.1.1. *For every natural number $n \geq 0$, we have $P^{(n)} = P^n$.*

Proof. The proof consists of the Total Law of Probability and a simple induction argument. For every $i, j \in S$ and $n \geq 0$, we have

$$\begin{aligned} p_{ij}^{(n)} &= \mathbb{P}(X_n = j \mid X_0 = i) = \sum_{k \in S} \mathbb{P}(X_n = j, X_{n-1} = k \mid X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}(X_{n-1} = k \mid X_0 = i) \mathbb{P}(X_n = j \mid X_{n-1} = k, X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}(X_{n-1} = k \mid X_0 = i) \mathbb{P}(X_n = j \mid X_{n-1} = k) = \sum_{k \in S} p_{ik}^{(n-1)} p_{kj}. \end{aligned}$$

In matrix form, the above result is expressed as $P^{(n)} = P^{(n-1)} \cdot P$. The desired result follows by induction. □

From the above proof we can actually infer the following very important identity, which is referred to as the *Chapman-Kolmogorov equation:*

$$P^{(n+m)} = P^{(n)} \cdot P^{(m)}, \quad \forall n, m \in \mathbb{N}. \tag{1.1.3}$$

1.1.2 Stopping times and the Strong Markov property

In the previous paragraph we began by assuming a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where Ω is a set, $\mathscr{F}$ is a σ-algebra (or σ-field) and $\mathbb{P}$ is a probability measure. But one could argue that these technical terms do not reflect one's intuitive ideas about randomness; at least not in an obvious way. In a setting of randomness, Ω is the set of all possible outcomes of an experiment, $\mathscr{F}$ is a collection of subsets of Ω in which an outcome $\omega \in \Omega$ will be, and $\mathbb{P}$ assigns a number to each set in $\mathscr{F}$ that shows how likely it is for this ω to lie in this set. The key interpretation of the σ-algebra $\mathscr{F}$ is *information.* It is the specific σ-algebra that tells us in which subsets of Ω will the observation ω be found.

Definition 1.1.2. Let $\{\mathscr{F}_n : n \in \mathbb{N}\}$ be a *filtration*, i.e., an increasing sequence of σ-algebras such that $\mathscr{F}_n \subseteq \mathscr{F}_{n+1} \subseteq \mathscr{F}$ for every $n \in \mathbb{N}$. A function $T : \Omega \to \mathbb{N} \cup \{\infty\}$ is called a

stopping time according to the filtration $(\mathscr{F}_n)_{n\geq 0}$, if

$$\{\omega \in \Omega : T(\omega) \leqslant n\} := \{T \leqslant n\} \in \mathscr{F}_n, \quad \forall n \in \mathbb{N}.$$

It is immediate that the above condition is equivalent to $\{T = n\} \in \mathscr{F}_n, \forall n \in \mathbb{N}$. One simply has to write $\{T = n\} = \{T \leqslant n\} \setminus \{T \leqslant n-1\}$ and use the properties of a σ-algebra. Intuitively, T is a stopping time if the event that T takes a specific value can only be determined by the information we have up to this stage, i.e., the event $\{T = n\}$ depends solely on $X_0, \ldots, X_n$. In other words, if someone is watching the stochastic process, he will know at the time when T takes place. There are some classic and very useful examples of stopping times in the case of Markov chains.

Example 1.1.1. Let $j \in S$. The *first passage time* $T_j : \Omega \to \mathbb{N} \cup \{\infty\}$ defined by

$$T_j(\omega) = \inf\{k \geqslant 1 : X_k(\omega) = j\}, \tag{1.1.4}$$

is a stopping time, since for every $n \in \mathbb{N}$ we have

$$\{T_j = n\} = \{\inf\{k \geqslant 1 : X_k = j\} = n\} = \{X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j\} \in \mathscr{F}_n.$$

This stopping time tells us the first time that the chain goes at the state j.

Example 1.1.2. Let $j \in S$. The *first hitting time* $H^A : \Omega \to \mathbb{N} \cup \{\infty\}$ defined by

$$H^A(\omega) = \inf\{k \geqslant 0 : X_k(\omega) \in A\}, \tag{1.1.5}$$

is a stopping time, since for every $n \in \mathbb{N}$ we have

$$\{H^A = n\} = \{\inf\{k \geqslant 0 : X_k \in A\} = n\} = \{X_0 \notin A, \ldots, X_{n-1} \notin A, X_n \in A\} \in \mathscr{F}_n.$$

This stopping time tells us the first time that the chain falls into the set A.

Example 1.1.3. Let $j \in S$. The *last exit time* $L^A : \Omega \to \mathbb{N} \cup \{\infty\}$ defined by

$$L^A(\omega) = \sup\{k \geqslant 0 : X_k(\omega) \in A\}, \tag{1.1.6}$$

is NOT a stopping time, since for every $n \in \mathbb{N}$ the event $\{L^A = n\}$ cannot be determined from the random variables $X_0, \ldots, X_n$ (it involves the future evolution of the chain as well). □

We are now ready to prove a very important property of Markov chains: the *Strong Markov property*. This Theorem generalizes Theorem (1.1.2) in the sense that one can, instead of conditioning on the state of the chain at a specific time, condition on a random time (specifically a stopping time) and the memoryless property will still hold.

Theorem 1.1.3 (Strong Markov property). *Let $(X_n)_{n\geqslant 0}$ be Markov(λ, P) and T a stopping time of $(X_n)_{n\geqslant 0}$. Then, conditional on $T < \infty$ and $X_T = i$, the process $(X_k)_{k\geqslant T}$ is Markov(δ_i, P) and independent of $X_0, \ldots, X_{T-1}$.*

Proof. Since T is a stopping time, it is determined by $X_0, X_1, \ldots, X_T$. If B is an event depending on $X_0, X_1, \ldots, X_T$, then for every $m \in \mathbb{N}$ the event $B \cap \{T = m\}$ is determined by $X_0, X_1, \ldots, X_m$. This observation is crucial, since we can now use Theorem 1.1.2. The Markov property at time m yields

$$\begin{aligned}\mathbb{P}(\{X_T = j_0, X_{T+1} = j_1, \ldots, X_{T+n} = j_n\} \cap B \cap \{T = m\} \cap \{X_T = i\}) = \\ = \mathbb{P}(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n \mid X_0 = i)\mathbb{P}(B \cap \{T = m\} \cap \{X_T = i\}) \Rightarrow\end{aligned}$$

$$\begin{aligned}\sum_{m=0}^{\infty} \mathbb{P}(\{X_T = j_0, X_{T+1} = j_1, \ldots, X_{T+n} = j_n\} \cap B \cap \{T = m\} \cap \{X_T = i\}) = \\ = \sum_{m=0}^{\infty} \mathbb{P}(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n \mid X_0 = i)\mathbb{P}(B \cap \{T = m\} \cap \{X_T = i\}) \Rightarrow\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\{X_T = j_0, X_{T+1} = j_1, \ldots, X_{T+n} = j_n\} \cap B \cap \{T < \infty\} \cap \{X_T = i\}) = \\ = \mathbb{P}(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n \mid X_0 = i)\mathbb{P}(B \cap \{T < \infty\} \cap \{X_T = i\}).\end{aligned}$$

Dividing by $\mathbb{P}(T < \infty, X_T = i)$, we get

$$\begin{aligned}\mathbb{P}(\{X_T = j_0, X_{T+1} = j_1, \ldots, X_{T+n} = j_n\} \cap B \mid T < \infty, X_T = i) = \\ = \mathbb{P}(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n \mid X_0 = i)\mathbb{P}(B \mid T < \infty, X_T = i),\end{aligned}$$

which proves the desired independence. □

1.1.3 Recurrence and Transience

If a Markov chain starts from a certain state, then how many times will the chain revisit this state? In response to this question, we classify a state according to the following definition.

Definition 1.1.3. Let $(X_n)_{n\geqslant 0}$ be a Markov chain with transition matrix P. We say that a state i is *recurrent* if

$$\mathbb{P}(X_n = i \text{ for infinitely many } n \mid X_0 = i) = 1.$$

If the *expected return time* $m_i := \mathbb{E}_i[T_i]$ is finite, then we say that i is *positive recurrent.*

We say that a state i is *transient* if

$$\mathbb{P}(X_n = i \text{ for infinitely many } n \mid X_0 = i) = 0.$$

Our goal is to show that these two notions are mutually exclusive, i.e., each state is either recurrent or transient. In order to do that, we will need more machinery. First, we give some definitions and state two useful Lemmas.

Definition 1.1.4. Let $j \in S$. We defined the first passage time $T_j : \Omega \to \mathbb{N} \cup \{\infty\}$ as $T_j(\omega) = \inf\{k \geqslant 1 : X_k(\omega) = j\}$. We define recursively the *n-th passage time* $T_j^{(n)}$ to the state j as a function $T_j^{(n)} : \Omega \to \mathbb{N} \cup \{\infty\}$ that satisfies

$$T_j^{(0)} = 0, \quad T_j^{(1)} = T_j, \quad T_j^{(n+1)}(\omega) = \inf\left\{k \geqslant T_j^{(n)}(\omega) + 1 : \ X_k(\omega) = j\right\},$$

for $n \in \mathbb{N}$. The *length of the n-th excursion* is then defined as

$$S_j^{(n)} = \begin{cases} T_j^{(n)} - T_j^{(n-1)}, & \text{if } T_j^{(n-1)} < \infty \\ 0, & \text{otherwise} \end{cases}.$$

Lemma 1.1.1. *Let $n \geq 2$ and $i \in S$. Then, conditional on $T_i^{(n-1)} < \infty$, $S_i^{(n)}$ is independent of X_k, $0 \leq k \leq T_i^{(n-1)}$ and*

$$\mathbb{P}\left(S_i^{(n)} = k \mid T_i^{(n-1)} < \infty\right) = \mathbb{P}(T_i = k \mid X_0 = i).$$

The proof is an application of the Strong Markov property on the stopping time $T = T_i^{(n-1)}$. The details can be found in [34]. Intuitively, the above lemma tells us that the time between two consecutive visits on the state $i \in S$ has the same distribution as that of the chain starting from i and revisiting it for the first time, which seems reasonable from the Strong Markov property.

Let V_i denote the total number of the Markov chain's visits to a state i. With the use of an indicator function, V_i can simply be expressed as

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

Since the functions $1(X_n = i)$ are non-negative and measurable, Beppo Levi's Theorem immediately yields

$$\mathbb{E}_i[V_i] = \int V_i \, d\mathbb{P} = \int \sum_{n=0}^{\infty} 1(X_n = i) \, d\mathbb{P} = \sum_{n=0}^{\infty} \int 1(X_n = i) \, d\mathbb{P}$$

$$= \sum_{n=0}^{\infty} \mathbb{E}_i \left[1(X_n = i)\right] = \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = i) = \sum_{n=0}^{\infty} p_{ii}^{(n)},$$

where we defined $\mathbb{P}_i(X_n = i) := \mathbb{P}(X_n = i \mid X_0 = i)$. Let $f_i := \mathbb{P}(T_i < \infty \mid X_0 = i)$ denote the *return probability* to i. There is a useful connection between V_i and f_i, as the next lemma indicates.

Lemma 1.1.2. *For every $n \in \mathbb{N}$ we have $\mathbb{P}_i(V_i > n) = f_i^n$.*

Proof. Let $n \in \mathbb{N}$, $\omega \in \Omega$ and $X_0 = i$. If $V_i(\omega) > n$, then $T_i^{(n)}(\omega) < \infty$. On the other hand, if $T_i^{(n)}(\omega) < \infty$, then $V_i(\omega) > n$. Hence, $\{V_i > n\} = \{T_i^{(n)} < \infty\}$. In order to prove the lemma, we will use induction on $\mathbb{N}$. For $n = 0$ the result is true. Assuming it is true for n, we have

$$\begin{aligned}\mathbb{P}_i(V_i > n+1) &= \mathbb{P}_i\left(T_i^{(n+1)} < \infty\right) = \mathbb{P}_i\left(T_i^{(n)} < \infty \text{ and } S_i^{(n+1)} < \infty\right) \\ &= \mathbb{P}_i\left(S_i^{(n+1)} < \infty \,\middle|\, T_i^{(n)} < \infty\right) \mathbb{P}_i\left(T_i^{(n)} < \infty\right) = f_i f_i^n = f_i^{n+1},\end{aligned}$$

where in the last equality we used Lemma 1.1.1, so the result is also true for $n+1$. The induction is complete. □

We are now ready to prove the main theorem of this paragraph, which confirms that each state is either recurrent or transient. The idea is to transfer the question in a more concrete one, i.e., whether the probability $\mathbb{P}_i(T_i < \infty)$ is equal to 1 or strictly less than 1. Another way to think about it is to study the series $\sum_{n=0}^{\infty} p_{ii}^{(n)}$. Intuitively, if the state i is recurrent, then the chain will return infinitely many times to it and, thus, the above series will diverge. If the state i is transient, then from a specific point the chain will never revisit this state, so the series is actually a finite sum and it converges.

Theorem 1.1.4. *Let $i \in S$ and T_i be the first passage time. Then,*

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

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

In particular, each state is either recurrent or transient.

Proof. (a) If $f_i := \mathbb{P}_i(T_i < \infty) = 1$, then we have

$$\begin{aligned}\mathbb{P}(X_n = i \text{ for infinitely many } n \mid X_0 = i) &= \mathbb{P}_i(V_i = \infty) \\ &= \lim_{n\to\infty} \mathbb{P}_i(V_i > n) = \lim_{n\to\infty} f_i^n = \lim_{n\to\infty} 1^n = 1,\end{aligned}$$

where in the third equality we used Lemma 1.1.2. Hence, i is recurrent and

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \mathbb{E}_i[V_i] = \infty.$$

(b) If $f_i := \mathbb{P}_i(T_i < \infty) < 1$, then Lemma 1.1.2 yields

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

thus we receive

$$\mathbb{P}(X_n = i \text{ for infinitely many } n \mid X_0 = i) = \mathbb{P}_i(V_i = \infty) = \lim_{n\to\infty} \mathbb{P}_i(V_i > n) = 0$$

and i is transient. □

1.1.4 Class structure

A very common idea in Mathematics is to classify objects according to a particular property. Two objects might not be "equal", but can be viewed as "equivalent" if certain common properties are shared. This leads us to the idea of an *equivalence relation*. An equivalence relation on a set S splits the set into subsets, each of which contains elements that can be considered "the same" according to a specified property. We give the technical definition.

Definition 1.1.5. Let S be a set and $S \times S = \{(i, j) : i, j \in S\}$ be its Cartesian product. A subset $R \subseteq S \times S$ is an *equivalence relation* on S if it satisfies the following properties:

(i) *Reflexive*: $(i, i) \in R, \forall i \in S$

(ii) *Symmetric*: if $(i, j) \in R$, then $(j, i) \in R, \forall i, j \in S$

(iii) *Transitive*: if $(i, j) \in R$ and $(j, \ell) \in R$, then $(i, \ell) \in R, \forall i, j, \ell \in S$.

The advantage of tracking an equivalence relation on a particular set is that we can then decompose the set into smaller sets and work with each set separately, simplifying our work. We will use this idea in our studying of Markov chains. We will define a binary relation on the state-space S that partitions it into smaller sets, the *equivalence classes*.

Definition 1.1.6. We say that a state $j \in S$ is *accessible* from a state $i \in S$, and write $i \to j$, if there exists an $n \in \mathbb{N}$ such that

$$p_{ij}^{(n)} := \mathbb{P}(X_n = j \mid X_0 = i) > 0.$$

Definition 1.1.7. We say that i *communicates* with j, and write $i \leftrightarrow j$, if both $i \to j$ and $j \to i$.

Proposition 1.1.2. *Let $i, j \in S$, $i \neq j$. The following equivalence holds:*

$$i \to j \quad \Leftrightarrow \quad \exists\, n \in \mathbb{N},\, \exists\, i_1, i_2, \ldots, i_{n-1} \in S : p_{ii_1} p_{i_1 i_2} \ldots p_{i_{n-1} j} > 0. \tag{1.1.7}$$

Proof. First, suppose that $i \to j$. Then there exists $n \in \mathbb{N}$ such that

$$0 < \mathbb{P}(X_n = j \mid X_0 = i) = \sum_{i_1,\ldots,i_{n-1}} p_{ii_1} p_{i_1 i_2} \ldots p_{i_{n-1} j},$$

where the idea was to sum over all possible paths that lead from i to j. Thus there exist $i_1, i_2, \ldots, i_{n-1} \in S$ such that $p_{ii_1}p_{i_1i_2}\ldots p_{i_{n-1}j} > 0$.

For the opposite direction, notice that

$$0 < p_{ii_1}p_{i_1i_2}\ldots p_{i_{n-1}j} \leqslant \mathbb{P}(X_n = j \mid X_0 = i)$$

and, thus, $i \to j$. □

Proposition 1.1.3. *The binary relation "$\leftrightarrow$" is an equivalence relation on the state-space S.*

Proof. Let $i, j, \ell \in S$. Clearly $i \leftrightarrow i$, since $p_{ii}^{(0)} = 1 > 0$. If $i \leftrightarrow j$, then by definition $j \leftrightarrow i$. Finally, let $i \leftrightarrow j$ and $j \leftrightarrow \ell$. Since we know from the above proposition that $i \to j$ and $j \to \ell$ implies $i \to \ell$, we get that $i \leftrightarrow j$ and $j \leftrightarrow \ell$ implies $i \leftrightarrow \ell$. Thus the binary relation "$\leftrightarrow$" is an equivalence relation on the state-space S. □

Definition 1.1.8. We say that "$\leftrightarrow$" partitions S into *communication classes.* If a Markov chain has only one communication class, then it is called *irreducible.*

We stated in the beginning of the paragraph that an equivalence relation divides S into subsets the objects of which exhibit similar properties. In the case of Markov chains, these properties are recurrence and transience. We can strengthen Theorem 1.1.4 and prove the following.

Theorem 1.1.5. *Let C be a communication class. Then, either all of its states are recurrent or all are transient.*

Proof. Let C be a communication class and $i, j \in C$. Assume that i is transient. Since $i, j \in C$, they communicate with each other, i.e.,

$$\exists\, n, m \in \mathbb{N} : \; p_{ij}^{(n)} > 0 \quad \text{and} \quad p_{ji}^{(m)} > 0.$$

We have for every $k \in \mathbb{N}$ that

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

since there are many possible paths leading from i back to i and we specify a certain one that passes through j. We thus have

$$p_{jj}^{(k)} \leqslant \frac{p_{ii}^{(n+k+m)}}{p_{ij}^{(n)} p_{ji}^{(m)}} \Rightarrow \sum_{k=0}^{\infty} p_{jj}^{(k)} \leqslant \frac{1}{p_{ij}^{(n)} p_{ji}^{(m)}} \sum_{k=0}^{\infty} p_{ii}^{(n+k+m)} < \infty,$$

where we know that the last series is finite by Theorem 1.1.4. We showed that if one state of C is transient, then every other state of C is transient. By duality, if one state of C is recurrent, then every other state of C is recurrent. We thus say that transience/recurrence is a *class property.* □

An immediate corollary of the previous theorem is that an irreducible Markov chain has either only transient or only recurrent states. If the latter is the case, one would expect that the chain would hit each state in finite time. The probabilistic analogue of this intuitive belief is stated in the following theorem, a proof of which can be found in [9].

Theorem 1.1.6. *If a Markov chain is irreducible and recurrent, then for every state $i \in S$ we have that $\mathbb{P}(T_i < \infty) = 1$.*

1.1.5 Stationarity

We will introduce the concept of a stationary distribution and use it to study the behavior of a discrete-time discrete state-space Markov chain as time goes to infinity. Intuitively speaking, a probability distribution is stationary with respect to a Markov chain if it is left invariant by the chain's transition matrix, i.e., a Markov chain that would start from a stationary distribution would never be able to "escape" from this distribution.

Definition 1.1.9. Let $(X_n)_{n\geq 0}$ be a Markov chain and $j \in S$ a state. Let

$$d_j = \gcd\Big\{n \in \mathbb{N} : p_{jj}^{(n)} > 0\Big\},$$

assuming that $\{n \in \mathbb{N} : p_{jj}^{(n)} > 0\} \neq \emptyset$, where gcd denotes the greatest common divisor.

(i) If $d_j = 1$, then the state j is called *aperiodic*.

(ii) If $d_j > 1$, then the state j is called *periodic* with *period* d_j.

If the set $\{n \in \mathbb{N} : p_{jj}^{(n)} > 0\}$ is empty, then we define $d_j = \infty$.

Let $(X_n)_{n\in\mathbb{N}}$ be Markov(π_0, P), i.e., the chain has initial distribution π_0 and transition matrix $P = (p_{ij})_{i,j\in S}$. We will denote by π_n the probability distribution of the random variable X_n, i.e., $\pi_n(j) := \mathbb{P}(X_n = j), j \in S$. We wish to study the limit of π_n as n approaches infinity. Since for every $n \in \mathbb{N}$ we have a distribution π_n, the object of study $(\pi_n)_{n\in\mathbb{N}}$ is a sequence of distributions and, thus, its limit needs more clarification. From now on, distribution means probability distribution, unless otherwise stated.

Definition 1.1.10. Let $(\pi_n)_{n\in\mathbb{N}}$ be a sequence of distributions and π be a distribution. We say that $(\pi_n)_{n\in\mathbb{N}}$ *converges* to π, and write $\pi_n \xrightarrow[n\to\infty]{} \pi$, if

$$\lim_{n\to\infty} \pi_n(j) = \pi(j) \quad \forall j \in S,$$

i.e., if pointwise convergence is satisfied.

Scheffé's Lemma[3] in the discrete setting yields the following:

$$\pi_n \to \pi \Rightarrow \sum_{j\in S}|\pi_n(j) - \pi(j)| \to 0. \tag{1.1.8}$$

Definition 1.1.11. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition matrix $P = (p_{ij})_{i,j\in S}$. A probability distribution $\pi = \{\pi(j) : j \in S\}$ on S is a *stationary distribution* for $(X_n)_{n\in\mathbb{N}}$ if it satisfies the following conditions:

$$\begin{cases} \pi = \pi P \\ \sum_{i\in S} \pi(i) = 1 \end{cases} \tag{1.1.9}$$

or, in analytical form,

$$\begin{cases} \pi(j) = \sum_{i\in S} \pi(i)p_{ij} \quad \forall j \in S \\ \sum_{i\in S} \pi(i) = 1 \end{cases} . \tag{1.1.10}$$

We will show that if the sequence $(\pi_n)_n$ converges to a distribution π, then π is stationary. That means that the only candidate limits for the limit distribution are distributions that satisfy the equations (1.1.9). Hence, we observe that already there is a connection between stationary and limit distributions of a Markov chain.

Theorem 1.1.7. *Let $(X_n)_{n\in\mathbb{N}}$ be Markov(π_0, P) and $(\pi_n)_n$ be the corresponding sequence of distributions. If $\pi_n \to \pi$, then π is a stationary distribution.*

Proof. A consequence of the Chapman-Kolmogorov equation (1.1.3) is that

$$\pi_n = \pi_0 P^n, \quad \forall n \in \mathbb{N}, \tag{1.1.11}$$

so for every $n \in \mathbb{N}$ we get recursively that

$$\pi_n = \pi_0 P^{n-1} P = (\pi_0 P^{n-1})P = \pi_{n-1}P \Leftrightarrow \pi_n(j) = \sum_{i\in S} \pi_{n-1}(i)p_{ij} \quad \forall j \in S.$$

Sending n to infinity, the left-hand side becomes $\pi(j)$ (by assumption) and for the right-hand side we get

$$\left| \sum_{i\in S} \pi_{n-1}(i)p_{ij} - \sum_{i\in S} \pi(i)p_{ij} \right| \leqslant \sum_{i\in S} |\pi_{n-1}(i) - \pi(i)|p_{ij} \leqslant \sum_{i\in S} |\pi_{n-1}(i) - \pi(i)| \xrightarrow[n\to\infty]{} 0,$$

[3] If μ is a σ-finite measure on the measurable space $(S, \mathscr{S})$, and f_n, f are measurable functions satisfying $\int_S f_n \, d\mu = \int_S f \, d\mu = 1$ and $f_n \to f$ a.s., then $\int_S |f_n(s) - f(s)| \, \mu(ds) \to 0$.
The proof follows upon noting that $\int_S |f_n(s) - f(s)| \, \mu(ds) = 2\int_S (f - f_n)^+ \, \mu(ds)$, $(f - f_n)^+ \to 0$ a.s., $0 \leq (f - f_n)^+ \leq f$, and using the Dominated Convergence Theorem.

where in the last limit we used (1.1.8). It follows that

$$\sum_{i\in S} \pi_{n-1}(i)p_{ij} \xrightarrow[n\to\infty]{} \sum_{i\in S} \pi(i)p_{ij}$$

We conclude that

$$\pi(j) = \sum_{i\in S} \pi(i)p_{ij} \quad \forall j \in S \Leftrightarrow \pi = \pi P,$$

i.e., π is a stationary distribution. □

The idea to name as *stationary* a distribution that satisfies (1.1.9) is better understood through the next theorem, which tells us that if a Markov chain starts from a stationary distribution, then its stochastic behavior will always be described by that. Furthermore, its joint distribution remains the same regardless of the time at which we are examining the case. In other words, the chain exhibits constant, or *stationary*, behavior over time.

Theorem 1.1.8. *Let $(X_n)_{n\in\mathbb{N}}$ be Markov(π_0, P). If π_0 is a stationary distribution, then $\pi_n = \pi_0$ for all $n \in \mathbb{N}$ and*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_k = x_k) = \mathbb{P}(X_n = x_0, X_{n+1} = x_1, \ldots, X_{n+k} = x_k),$$

for every $n, k \in \mathbb{N}$.

Proof. We will use induction to show that $\pi_n = \pi_0$ for all $n \in \mathbb{N}$. For $n = 0$ it is obviously true. For $n = 1$, we have from (1.1.11) that $\pi_1 = \pi_0 P = \pi_0$, thus it is also true. Assume that it is true for n. For $n + 1$ we have

$$\pi_{n+1} = \pi_n P = \pi_0 P = \pi_0,$$

and the induction is complete. Hence, $\pi_n = \pi_0$ for all $n \in \mathbb{N}$. Now, let $n, k \in \mathbb{N}$. From (1.1.1) and (1.1.2), we have that

$$\begin{aligned}
&\mathbb{P}(X_n = x_0, X_{n+1} = x_1, \ldots, X_{n+k} = x_k) = \\
&\quad = \mathbb{P}(X_n = x_0)\mathbb{P}(X_{n+1} = x_1 \mid X_n = x_0) \ldots \mathbb{P}(X_{n+k} = x_k \mid X_{n+k-1} = x_{k-1}) \\
&\quad = \pi_n(x_0)p_{x_0x_1} \ldots p_{x_kx_{k-1}} = \pi_0(x_0)p_{x_0x_1} \ldots p_{x_{k-1}x_k} \\
&\quad = \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_k = x_k).
\end{aligned}$$

□

1.1.6 Coupling

Coupling is an extremely useful technique that is often used in Probability Theory in order to prove statements that include random variables defined in different probability spaces. The goal is to be able to *compare* these random variables (notice that in general we cannot

write $\mathbb{P}(X \neq Y)$, unless X and Y are defined on the same probability space). The idea is to embed these random variables in a larger probability space in a way that does not alter their distributions. The idea of coupling can be used in very simple settings such as showing quickly that $X \sim \text{Bin}(n, 1/2)$ is stochastically larger than $Y \sim \text{Bin}(n, 1/3)$, as well as in much more complicated settings such as coupling evolving random graphs with multi-type branching processes to study centrality measures on networks. A typical example is the PageRank algorithm, originally developed by Brin & Page at Google, in 1996. Detailed expositions of how couplings between graphs are used along with the concept of local weak convergence, can be found in [25], [35], [36], [37]).

First we will give some technical definitions and theorems and then we will use a coupling argument in order to prove one of the main theorems of the chapter, regarding the limit behavior of a Markov chain. This paragraph is to be viewed mostly as an expository. For a detailed approach and rigorous proofs of the following statements, the reader is referred to [23].

Definition 1.1.12. Let $(X, \mathscr{A})$ be a measurable space. A *signed measure* on $(X, \mathscr{A})$ is a function $\mu : \mathscr{A} \to \overline{\mathbb{R}}$ such that

(i) μ takes on at most one of the values $-\infty$ or ∞

(ii) $\mu(\varnothing) = 0$

(iii) if $(B_n)_{n\geqslant 1}$ is a sequence of pairwise disjoint sets, then

$$\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n).$$

Definition 1.1.13. Let μ be a signed measure on the measurable space $(X, \mathscr{A})$ and let $P, N \in \mathscr{A}$. We say that

(i) P is *positive*, if $\mu(E \cap P) \geqslant 0$ for every $E \in \mathscr{A}$.

(ii) N is *negative*, if $\mu(E \cap N) \leqslant 0$ for every $E \in \mathscr{A}$.

Theorem 1.1.9 (Hahn decomposition). *Let μ be a signed measure on the measurable space $(X, \mathscr{A})$. Then there exist a positive set $P \in \mathscr{A}$ and a negative set $N \in \mathscr{A}$ such that $P \cap N = \varnothing$ and $X = P \cup N$.*

Notice that the Hahn decomposition need not be unique.

Definition 1.1.14. If $\{P, N\}$ is a Hahn decomposition of X, then we define the measures $\mu^+, \mu^- : \mathscr{A} \to [0, \infty]$ with $\mu^+(A) = \mu(P \cap A)$ and $\mu^-(A) = -\mu(N \cap A)$ for every $A \in \mathscr{A}$.

Definition 1.1.15. Let $(X, \mathscr{A})$ be a measurable space and μ, ν two measures on it. These measures are called *mutually singular* if there are disjoint sets $A, B \in \mathscr{A}$ such that $\mu(A) = 0$, $\nu(B) = 0$ and $X = A \cup B$.

Theorem 1.1.10 (Jordan decomposition). *Let μ be a signed measure on the measurable space $(X, \mathscr{A})$. Then there exist two mutually singular positive measures μ^+ and μ^- such that $\mu = \mu^+ + (-\mu^-)$.*

Definition 1.1.16. Let $(E, \mathscr{E})$ be a measurable space, where E is a Polish space (i.e., homeomorphic to a complete, separable metric space). If μ is a bounded signed measure on $(E, \mathscr{E})$ such that $\mu(E) = 0$, we define the *total variation norm* of μ as

$$\|\mu\|_{TV} := \sup_{\|f\|_\infty \leqslant 1} \left| \int_E f \, d\mu \right|.$$

Using the Jordan-Hahn Decomposition, we can prove that $\|\mu\|_{TV} = 2 \sup_{A \in \mathscr{E}} \mu(A)$.

Definition 1.1.17. Let $(E, \mathscr{E})$ be a measurable space and $\mathscr{E} \otimes \mathscr{E}$ denote the smallest σ-algebra containing $\mathscr{E} \times \mathscr{E}$. Let X and Y be two random variables defined on the probability spaces $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$, respectively, and taking values on the measurable space $(E, \mathscr{E})$. A *coupling* of the random variables X and Y is any pair of random variables $(\hat{X}, \hat{Y})$ taking values on $(E \times E, \mathscr{E} \otimes \mathscr{E})$ whose marginals have the same distribution as X and Y, i.e.,

$$X \overset{d}{=} \hat{X} \quad \text{and} \quad Y \overset{d}{=} \hat{Y}.$$

Our goal generally is to find a coupling that makes the total variation norm $\|\mathbb{P}_1 - \mathbb{P}_2\|_{TV}$ as small as possible. We state without proof the basic coupling inequality.

Theorem 1.1.11. *Given two random variables X and Y with respective probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, then any coupling $\hat{\mathbb{P}}$ of $\mathbb{P}_1$ and $\mathbb{P}_2$ satisfies*

$$\|\mathbb{P}_1 - \mathbb{P}_2\|_{TV} \leqslant 2\hat{\mathbb{P}}(\hat{X} \neq \hat{Y}). \tag{1.1.12}$$

Remark 1. In practice, if we have a random variable X defined on a probability space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$ and a random variable Y defined on a probability space $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$, the coupling argument allows us to define random variables $\hat{X}$ and $\hat{Y}$ on the space $\Omega = \Omega_1 \times \Omega_2$ such that

$$\hat{X}(\omega_1, \omega_2) = X(\omega_1) \quad \text{and} \quad \hat{Y}(\omega_1, \omega_2) = Y(\omega_2).$$

We can then define a probability measure $\mathbb{P}$ such that

$$\mathbb{P}(A_1 \times \Omega_2) = \mathbb{P}_1(A_1) \quad \text{and} \quad \mathbb{P}(\Omega_1 \times A_2) = \mathbb{P}_2(A_2),$$

for every $A_1 \in \mathscr{F}_1$, $A_2 \in \mathscr{F}_2$. Then the random variables $\hat{X}$ and $\hat{Y}$ are defined in Ω and have the same distribution with X and Y, respectively. Indeed,

$$\mathbb{P}(\hat{X} \in C) = \mathbb{P}(\{X \in C\} \times \Omega_2) = \mathbb{P}_1(X \in C)$$

and

$$\mathbb{P}(\hat{Y} \in C) = \mathbb{P}(\Omega_1 \times \{Y \in C\}) = \mathbb{P}_2(Y \in C),$$

for every $C \in \mathscr{F}_1 \otimes \mathscr{F}_2$. Such a measure $\mathbb{P}$ is called a *coupling measure*.

1.1.7 Limit behavior

We will use a coupling argument in order to prove a very important theorem, that connects the limit and the stationary distribution of a Markov chain if some "good" properties are satisfied. First, we state two very useful Lemmas, proofs of which can be found in [34].

Lemma 1.1.3. *Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible Markov chain with transition matrix P. Then, the following are equivalent:*

(i) every state is positive recurrent;

(ii) some state $i \in S$ is positive recurrent;

(iii) the chain has a stationary distribution π.

If (iii) holds, then $\pi_i = \dfrac{1}{m_i}$ for every $i \in S$, where $m_i = \mathbb{E}_i[T_i]$ is the expected return time to state i.

Lemma 1.1.4. *Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible Markov chain with transition matrix P and suppose there exists at least one aperiodic state. Then, for all sufficiently large n, we have $p_{jk}^{(n)} > 0$ for all $j, k \in S$ and all states are aperiodic.*

We are now ready to prove the main theorem of the paragraph. The idea is to create a second Markov chain (i.e., a *coupling* of the one we have) that has the desired stationary distribution as its initial distribution. From the theorems that we have proved so far, we understand the stochastic behavior of the second, constructed, Markov chain. The key-idea of coupling here is to let both Markov chains run simultaneously until the time they meet and then construct a new Markov chain that has the initial behavior of our desired chain and the limit behavior of the second, well-understood, chain. Using the Strong Markov property, we infer that the third chain is actually a copy of the first one, so we can show the result for the third one, something that is easier.

Theorem 1.1.12 (Convergence to Equilibrium). *Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible and aperiodic Markov chain with transition matrix P, arbitrary initial distribution λ, countable state-space S and suppose that the chain has a stationary distribution $\pi = (\pi(j) : j \in S)$. Then, we have that $\pi_n \xrightarrow[n\to\infty]{} \pi$ or, equivalently,*

$$\lim_{n\to\infty} \pi_n(j) = \pi(j) \quad \forall j \in S \Leftrightarrow \lim_{n\to\infty} \mathbb{P}(X_n = j) = \pi(j) \quad \forall j \in S$$

In addition, for all states $i, j \in S$ we have that $\lim_{n\to\infty} p_{ij}^{(n)} = \pi(j)$.

Proof. By assumption, $(X_n)_{n\in\mathbb{N}}$ is Markov(λ, P). Let $(Y_n)_{n\in\mathbb{N}}$ be Markov(π, P) and independent of $(X_n)_{n\in\mathbb{N}}$, where π is a stationary distribution for $(X_n)_{n\in\mathbb{N}}$. Let[4]

$$T := \inf\{n \in \mathbb{N} : X_n = Y_n\}$$

denote their first meeting time. For every $n \in \mathbb{N}$, the event $\{T = n\}$ is an element of the σ-algebra generated by $(X_k)_{0\leqslant k\leqslant n}$ and $(Y_k)_{0\leqslant k\leqslant n}$, so T is a stopping time. We will show that T is finite with probability 1. Let $W_n = (X_n, Y_n)$ be a Markov chain with state-space the Cartesian product $S \times S$. From the multiplicative behavior of the coupling measure, we have that $(W_n)_{n\in\mathbb{N}}$ has transition probabilities

$$\tilde{p}_{(i,k)(j,\ell)} = p_{ij}\, p_{k\ell} \quad \forall i, k, j, \ell \in S$$

and initial distribution

$$\mu_{ij} = \lambda_i\, \pi_j \quad \forall i, j \in S.$$

By assumption, $(X_n)_{n\in\mathbb{N}}$ is aperiodic, so from Lemma (1.1.4) we have that

$$\tilde{p}^{(n)}_{(i,k)(j,\ell)} = p^{(n)}_{ij}\, p^{(n)}_{k\ell} > 0 \quad \forall i, k, j, \ell \in S,$$

for sufficiently large n. Thus, the 2-dimensional Markov chain $(W_n)_{n\in\mathbb{N}}$ is irreducible. Furthermore, $(W_n)_{n\in\mathbb{N}}$ has a stationary distribution given by

$$\tilde{\pi}_{i,j} = \pi_i\, \pi_j, \quad \forall i, j \in S,$$

so by Lemma (1.1.3) we get that $(W_n)_{n\in\mathbb{N}}$ is positive recurrent, so from Theorem (1.1.6) we conclude that $\mathbb{P}(T < \infty) = 1$.

Remark 2. Since $\mathbb{P}(T < \infty) = 1$ and T is a positive random variable, we infer that T is a proper (non defective) random variable.

Now we want to create a Markov chain that has the same initial behavior as that of $(X_n)_{n\in\mathbb{N}}$ and the same limit behavior as that of $(Y_n)_{n\in\mathbb{N}}$, since we can study that via Theorem (1.1.7). So we begin by constructing a stochastic process $\{Z_n : n \in \mathbb{N}\}$, such that

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

A graphic illustration is showed below. The continuous line depicts the Markov chain $(X_n)_{n\in\mathbb{N}}$, the dotted line depicts the Markov chain $(Y_n)_{n\in\mathbb{N}}$ and the red line depicts the constructed stochastic process $(Z_n)_{n\in\mathbb{N}}$.

[4] here we use a coupling argument to define a sample space Ω on which both $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ are defined and have the same distributions as before.

We want to show that $(Z_n)_{n\in\mathbb{N}}$ is Markov(λ, P). Since $\mathbb{P}(T \geqslant 0) = 1$, by the definition of Z_n we have that $\mathbb{P}(Z_0 = X_0) = 1$, so $Z_0 \sim \lambda$. Now we have to show that

$$\mathbb{P}(Z_{n+1} = z_{n+1} \mid Z_n = z_n, \ldots, Z_0 = z_0) = p_{z_n, z_{n+1}}, \tag{1.1.13}$$

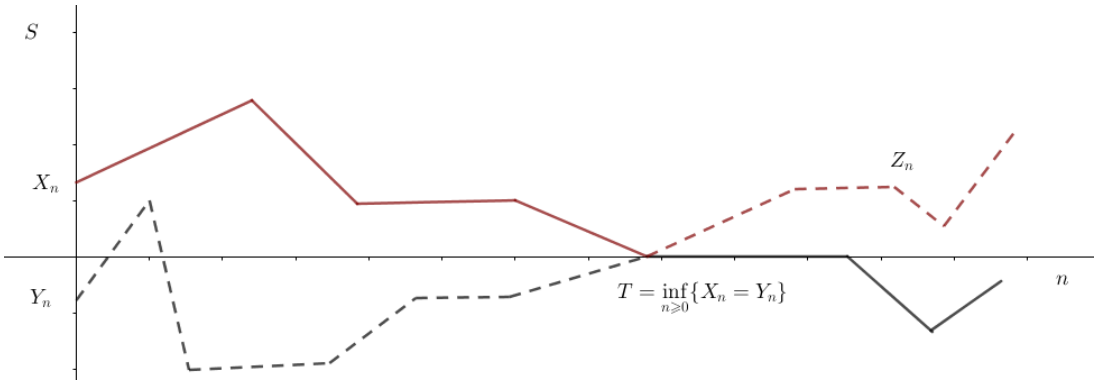

FIGURE 1.1: The coupling argument for Markov chains.

and then $(Z_n)_{n\in\mathbb{N}}$ will indeed be a Markov chain with initial distribution λ and transition matrix P. The idea is to partition the coupled sample space Ω into smaller subsets that occur for the different values that the stopping time T takes. Let $n \in \mathbb{N}$. We have that

$$\Omega = \bigcup_{k=0}^{n} \{T = k\} \cup \{T > n\},$$

so the joint probability becomes

$$\begin{aligned}\mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0) = \\ = \sum_{k=0}^{n} \mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0, T = k) + \\ + \mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0, T > n). \end{aligned} \tag{1.1.14}$$

We will examine these events separately. For every $k = 0, 1, \ldots, n$ we have that

$$\{Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0, T = k\} = \bigcap_{i=0}^{k} \{X_i = z_i\} \bigcap_{i=0}^{k-1} \{Y_i \neq z_i\} \bigcap_{i=k}^{n+1} \{Y_i = z_i\},$$

since T is the first meeting time of $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$, thus

$$\mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0, T = k) =$$

$$= \mathbb{P}\left(\bigcap_{i=0}^{k}\{X_i = z_i\}\right)\mathbb{P}\left(\bigcap_{i=0}^{k-1}\{Y_i \neq z_i\}\bigcap_{i=k}^{n+1}\{Y_i = z_i\}\right)$$

$$= \mathbb{P}\left(\bigcap_{i=0}^{k}\{X_i = z_i\}\right)\mathbb{P}\left(\bigcap_{i=0}^{k-1}\{Y_i \neq z_i\}\bigcap_{i=k}^{n}\{Y_i = z_i\}\right)p_{z_n z_{n+1}}$$

$$= \mathbb{P}\left(\bigcap_{i=0}^{k}\{X_i = z_i\}\bigcap_{i=0}^{k-1}\{Y_i \neq z_i\}\bigcap_{i=k}^{n}\{Y_i = z_i\}\right)p_{z_n z_{n+1}}$$

$$= \mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0, T = k)\, p_{z_n z_{n+1}},$$

where in the third equality we used the independence of $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$. Similarly, for the event $\{T > n\}$ we get

$$\{Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0, T > n\} = \bigcap_{i=0}^{n+1}\{X_i = z_i\}\bigcap_{i=0}^{n}\{Y_i \neq z_i\},$$

so the joint probability becomes

$$\mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0, T > n) = \mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0, T > n)\, p_{z_n z_{n+1}}$$

and, thus, equation (1.1.14) becomes

$$\mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0) =$$

$$= \sum_{k=0}^{n}\mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0, T = k)\, p_{z_n z_{n+1}} +$$

$$+ \mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0, T > n)\, p_{z_n z_{n+1}} =$$

$$= p_{z_n z_{n+1}}\left(\sum_{k=0}^{n}\mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0, T = k) + \mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0, T > n)\right) \Rightarrow$$

$$\mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0) = \mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0)\, p_{z_n z_{n+1}} \Rightarrow$$

$$p_{z_n z_{n+1}} = \frac{\mathbb{P}(Z_{n+1} = z_{n+1}, Z_n = z_n, \ldots, Z_0 = z_0)}{\mathbb{P}(Z_n = z_n, \ldots, Z_0 = z_0)} \Rightarrow$$

$$p_{z_n z_{n+1}} = \mathbb{P}(Z_{n+1} = z_{n+1} \mid Z_n = z_n, \ldots, Z_0 = z_0).$$

We conclude that $(Z_n)_{n\in\mathbb{N}}$ is Markov(λ, P). Hence, $(X_n)_{n\in\mathbb{N}}$ and $(Z_n)_{n\in\mathbb{N}}$ have the same stochastic behavior, so for every $j \in S$ we have that

$$\pi_n(j) = \mathbb{P}(X_n = j) = \mathbb{P}(Z_n = j) \quad \forall n \in \mathbb{N}.$$

Furthermore, since π is stationary and $Y_0 \sim \pi$, we know that $\pi(j) = \mathbb{P}(Y_n = j)$ for all

$j \in S, n \in \mathbb{N}$. Let $j \in S$. We have that

$$\begin{aligned}
|\pi_n(j) - \pi(j)| &= |\mathbb{P}(Z_n = j) - \mathbb{P}(Y_n = j)| \\
&= |\mathbb{P}(Z_n = j, T \leqslant n) + \mathbb{P}(Z_n = j, T > n) - \mathbb{P}(Y_n = j, T \leqslant n) - \mathbb{P}(Y_n = j, T > n)| \\
&= |\mathbb{P}(X_n = j, T > n) - \mathbb{P}(Y_n = j, T > n)| \\
&\leqslant \max\{\mathbb{P}(X_n = j, T > n), \mathbb{P}(Y_n = j, T > n)\} \leqslant \mathbb{P}(T > n).
\end{aligned}$$

We have, thus, shown that

$$|\pi_n(j) - \pi(j)| \leqslant \mathbb{P}(T > n) \xrightarrow[n\to\infty]{} 0, \quad \forall j \in S,$$

since T is proper by Remark 2 and consequently $\pi_n \xrightarrow[n\to\infty]{} \pi$. In order to show that $\lim_{n\to\infty} p_{ij}^{(n)} = \pi(j)$, we can simply start the chain $(X_n)_{n\in\mathbb{N}}$ from the state $i \in S$, i.e., take $\mathbb{P}(X_0 = i) = 1$, since in the preceding proof we started from an arbitrary initial distribution λ. Then, $p_{ij}^{(n)} = \mathbb{P}(X_n = j) \xrightarrow[n\to\infty]{} \pi(j)$. □

1.1.8 Ergodic Theorem

We would like to know more about the asymptotic behavior of the averages over different paths of a Markov chain, relating statistical and probabilistic properties of the chain. Until now, we have seen the Laws of Large Numbers as the most common way to study such averages. The problem is that the Laws of Large Numbers concern independent random variables, while the random variables that build a Markov chain are exhibiting dependence. The necessary modifications lead to a new class of extremely useful theorems, called *Ergodic Theorems*.

Ergodic theorems give us information about the asymptotic behavior of the time average of several orbits in a dynamical system, relating the so called *time average* and *space average* together. In this paragraph we will prove the Ergodic Theorem for Markov chains. We will state two versions of it and prove the first one, which is a special but very instructive case.

First, we state the Weak and the Strong Law of Large Numbers. The latter will be used in our proof of the Ergodic Theorem.

Theorem 1.1.13 (Weak Law of Large Numbers). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of iid[5] random variables, with $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}[X_1] = \mu$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu,$$

[5] independent and identically distributed

where the convergence in probability is defined by

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) \xrightarrow[n\to\infty]{} 0.$$

Theorem 1.1.14 (Strong Law of Large Numbers). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of iid random variables, with $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}[X_1] = \mu$. Then,*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu,$$

where the almost sure convergence is defined by

$$\mathbb{P}\left(\left\{ \omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) = \mu \right\} \right) = 1.$$

If, in addition, the random variables (X_n) are non-negative, then the Theorem holds even in the case $\mu = \infty$.

We are now ready to state the two cases of the Ergodic Theorem for Markov chains. First, recall that for a given state $i \in S$ we define

$$V_i(n) := \sum_{k=0}^{n-1} 1(X_k = i) \quad \text{and} \quad V_i := \sum_{k=0}^{\infty} 1(X_k = i)$$

to be the number of visits in i before the n-th step and the total number of visits in i, respectively. We also refer to Definition 1.1.4 for the n-th passage time, $T_i^{(n)}$, and the length of the n-th excursion, $S_i^{(n)}$. These two concepts are better illustrated in the following figure.

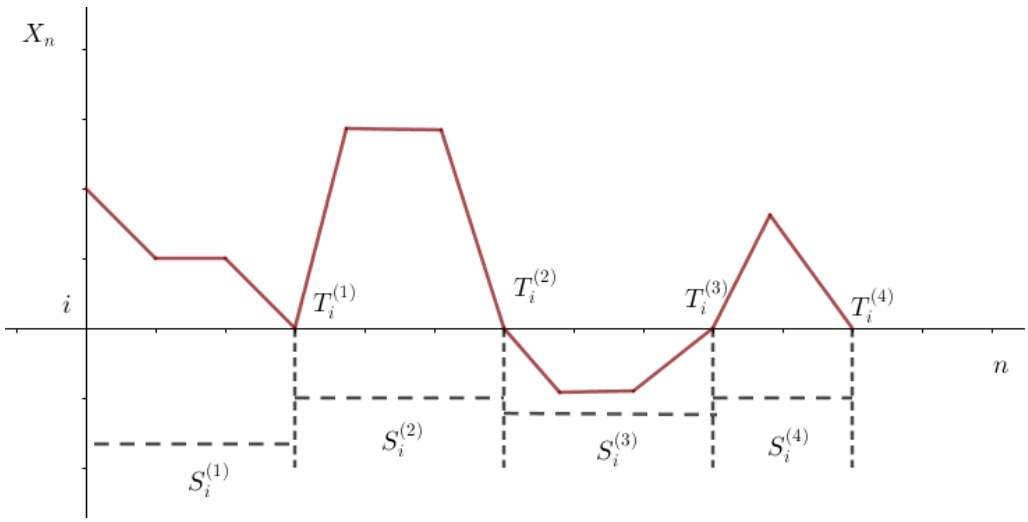

FIGURE 1.2: Passage times & lengths of excursions in Markov chains.

Theorem 1.1.15 (Ergodic Theorem I). *Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible Markov(λ, P), where λ is an arbitrary initial distribution. Then, for every $i \in S$,*

$$\frac{V_i(n)}{n} := \frac{1}{n}\sum_{k=0}^{n-1} 1(X_k = i) \overset{a.s.}{\to} \frac{1}{m_i}.$$

Proof. The Markov chain $(X_n)_{n\in\mathbb{N}}$ is irreducible, so it is either transient or recurrent. If it is transient, then we know that $\mathbb{P}(V_i < \infty) = 1$, and consequently

$$\frac{V_i(n)}{n} \leqslant \frac{V_i}{n} \xrightarrow[n\to\infty]{a.s.} 0 = \frac{1}{m_i},$$

since $m_i = \infty$. Suppose now that the chain is recurrent. Let $i \in S$. We know that $\mathbb{P}(T_i < \infty) = 1$. By the Strong Markov property (1.1.3) we have that the process $(X_{T_i+n})_{n\in\mathbb{N}}$ is Markov(δ_i, P) and is independent of $X_0, X_1, \ldots, X_{T_i}$, so it suffices to prove the theorem for δ_i as the initial distribution. By Lemma (1.1.1), we get that $S_i^{(1)}, S_i^{(2)}, \ldots$ are iid random variables with $\mathbb{E}_i[S_i^{(k)}] = m_i$ for all $k \in \mathbb{N}$. Thus, the Strong Law of Large Numbers can be used and we have that

$$\frac{S_i^{(1)} + \cdots + S_i^{(n)}}{n} \overset{a.s.}{\to} m_i. \tag{1.1.15}$$

Also, since $(X_n)_{n\in\mathbb{N}}$ is recurrent, we know that

$$V_i(n) \overset{a.s.}{\to} \infty. \tag{1.1.16}$$

Since we have taken $\lambda = \delta_i$, the following two inequalities hold:

$$T_i^{(V_i(n)-1)} = S_i^{(1)} + \cdots + S_i^{(V_i(n)-1)} \leqslant n - 1$$

and

$$T_i^{(V_i(n))} = S_i^{(1)} + \cdots + S_i^{(V_i(n))} \geqslant n,$$

so we take

$$\frac{S_i^{(1)} + \cdots + S_i^{(V_i(n)-1)}}{V_i(n)} \leqslant \frac{n}{V_i(n)} \leqslant \frac{S_i^{(1)} + \cdots + S_i^{(V_i(n))}}{V_i(n)}. \tag{1.1.17}$$

Combining (1.1.15), (1.1.16) and (1.1.17), we get that

$$\frac{n}{V_i(n)} \overset{a.s.}{\to} m_i$$

or, equivalently,

$$\frac{V_i(n)}{n} \overset{a.s.}{\to} \frac{1}{m_i}.$$

□

Theorem 1.1.16 (Ergodic Theorem II). *Let $(X_n)_{n\in\mathbb{N}}$ be an irreducible, positive recurrent*

Markov(λ, P) *and let* $(\pi(j) : j \in S)$ *be its unique stationary distribution. Then, for every bounded function* $f : S \to \mathbb{R}$*, we have that*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{a.s.}{\to} \sum_{j\in S} \pi(j) f(j).$$

1.2 General state-space

In the preceding section we assumed that the state-space S was discrete, i.e., either finite or countable. However, many interesting applications concern continuous state-spaces. For instance, Hamiltonian Monte Carlo ([3], [11], [21]) and many other MCMC methods ([2], [26]) operate on general spaces (mainly measurable topological spaces or metric spaces). This led to a need for a much more general and richer theory for Markov Chains. When dealing with general state-spaces, the idea is to think about *sets* instead of *points*. The dynamics of the chain are now described by a mapping called *kernel* instead of a matrix. The surprising thing is that most of the results that we stated and proved for the countable state-space case will still hold without assuming any specific structure for the new, possibly uncountable, state-space.

1.2.1 Kernels

A kernel generalizes the notion of the transition matrix that we saw in the countable case. Intuitively, a kernel is a mapping that tells us how possible it is to move to a specific set given the position at which we are now.

Definition 1.2.1. Let $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ be measurable spaces. We say that a mapping $P : X \times \mathscr{Y} \to [0, \infty]$ is a *kernel* on $X \times \mathscr{Y}$ if it satisfies the following:

(i) for every $A \in \mathscr{Y}$, the mapping $P(\cdot, A) : (X, \mathscr{X}) \to ([0, \infty], \mathscr{B}([0, \infty]))$, described by $x \mapsto P(x, A)$, is a Borel-measurable function.

(ii) for every $x \in X$, the mapping $P(x, \cdot) : \mathscr{Y} \to [0, \infty]$, described by $A \mapsto P(x, A)$, is a measure on $\mathscr{Y}$;

We say that the kernel P is

- *bounded*, if $\sup_{x\in X} P(x, Y) < \infty$;
- a *normalized kernel*, if $P(x, Y) = 1$ for all $x \in X$;
- a *Markov kernel*, if $(X, \mathscr{X}) = (Y, \mathscr{Y})$ and $P(x, X) = 1$ for all $x \in X$.

Definition 1.2.2. Let $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ be measurable spaces. Let ν be a positive σ-finite measure[6] on $(Y, \mathscr{Y})$ and $\xi : X \times Y \to [0, \infty]$ be an $\mathscr{X} \otimes \mathscr{Y}$-measurable function, where

[6] there exists a countable set I and measurable sets $(A_i)_{i\in I}$ of finite measure, such that $Y = \bigcup_{i\in I} A_i$.

$\mathscr{X} \otimes \mathscr{Y}$ is the product σ-algebra of $\mathscr{X}$ and $\mathscr{Y}$. We say that the kernel $P : X \times \mathscr{Y} \to [0, \infty]$ has *density* ξ, if

$$P(x, A) = \int_A \xi(x, y)\, \nu(dy),$$

for all $x \in X$, $A \in \mathscr{Y}$.

Remark 3. The notion of a kernel strictly generalizes that of a transition matrix in the countable case by taking ν in definition 1.2.2 to be the counting measure. In particular, let $(S, \mathscr{P}(S))$ be a measurable space, where S is a countable set and $\mathscr{P}(S)$ denotes the powerset of S, i.e., the set of all subsets of S. A Markov kernel P on $S \times \mathscr{P}(S)$ is a matrix $P = (p_{ij} : i, j \in S)$ such that each row $\{p_{ij} : j \in S\}$ is a probability function, so a Radon-Nikodym derivative. The kernel P is formally described by $P(i, \{j\}) = P(i, j) = p_{ij}$ for all $i, j \in S$, thus indeed generalizing the countable case.

1.2.1.1 Kernels and Integral Operators

If μ is a measure and f a measurable function, then we will be using the following notation interchangeably:

$$\mu f = \int f\, d\mu = \int \mu(dx) f(x),$$

assuming the integral exists. This last notation serves us better if we observe that a kernel gives rise to two *integral operators*. A kernel *acts* on measures from the right and on functions from the left, as the next proposition shows[7].

Proposition 1.2.1 (Kernels and Operators). *Let $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ be measurable spaces and $P : X \times \mathscr{Y} \to [0, \infty]$ be a kernel on $X \times \mathscr{Y}$.*

(i) If μ is a positive measure on $(X, \mathscr{X})$, then μP defined as

$$\mu P(A) = \int_X \mu(dx) P(x, A), \quad A \in \mathscr{Y},$$

is a positive measure on $(Y, \mathscr{Y})$.

(ii) If $f : Y \to \mathbb{R}$ is a measurable function, then $Pf : X \to \mathbb{R}$ defined as

$$Pf(x) = \int_Y P(x, dy) f(y), \quad x \in X,$$

is a measurable function.

Proof. (i) We have that $\mu P(A) \geqslant 0$ for every $A \in \mathscr{Y}$. Let $(A_i)_{i \in I}$ be a countable selection of disjoint elements of $\mathscr{Y}$. Using the Beppo Levi Theorem, we get that

$$\mu P\left(\bigcup_{i \in I} A_i\right) = \int_X \mu(dx) P\left(x, \bigcup_{i \in I} A_i\right) = \int_X \mu(dx) \sum_{i \in I} P(x, A_i)$$

[7] We assume that all the integrals that appear in the text exist, unless otherwise stated.

$$= \sum_{i \in I} \int_X \mu(dx) P(x, A_i) = \sum_{i \in I} \mu P(A_i),$$

thus μP is a positive measure on $(Y, \mathscr{Y})$.

(ii) The measurability of Pf is a direct consequence of the measurability of f and of the mapping $x \mapsto P(x, A)$ and of basic properties of the integral. □

Definition 1.2.3 (Product Kernel). If $(X, \mathscr{X})$, $(Y, \mathscr{Y})$, $(Z, \mathscr{Z})$ are measurable spaces and P_1, P_2 are kernels from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$ and from $(Y, \mathscr{Y})$ to $(Z, \mathscr{Z})$ respectively, then it can be proved that we can define a new kernel $P_1 P_2 : X \times \mathscr{Z} \to [0, \infty]$ by

$$P_1 P_2 := \int P_1(x, dy) P_2(y, A), \quad x \in X, A \in \mathscr{Z}.$$

The kernel $P_1 P_2$ is referred to as the *product kernel*. Inductively, we can also define the *n-th product kernel* P^n by

$$P^n(x, A) = \int_X P(x, dy) P^{n-1}(y, A),$$

from which we can take the Chapman-Kolmogorov equation

$$P^{n+m}(x, A) = \int_X P^n(x, dy) P^m(y, A). \tag{1.2.1}$$

Notice that if the state-space is discrete, then the kernel P is a matrix (namely, the transition matrix) and its n-th power is just the n-th power of the matrix. Thus, equation (1.2.1) generalizes equation (1.1.3), providing a useful connection of the two cases. In essence, the Chapman-Kolmogorov equation gives us information about the intermediate states from which the Markov chain will pass in order to go from a particular state (or set) to another.

Definition 1.2.4 (Tensor Products). We define the tensor product of n kernels and the tensor product of a measure with a kernel.

(i) If P is a kernel on $X \times \mathscr{Y}$, then we can define a kernel $P^{\otimes n}$ on $(X^n, \mathscr{X}^{\otimes n})$ such that

$$P^{\otimes n} f(x) = \int_{X^n} f(x_1, \ldots, x_n) P(x, dx_1) P(x_1, dx_2) \ldots P(x_{n-1}, dx_n).$$

This kernel is called the *n-th tensorial product* of P.

(ii) If ν is a σ-finite measure on $(X, \mathscr{X})$ and P is a kernel on $(X, \mathscr{Y})$, we define the *tensor product* of ν and P to be a measure on $(X \times Y, \mathscr{X} \otimes \mathscr{Y})$ such that

$$\nu \otimes P(A \times B) = \int_A \nu(dx) P(x, B),$$

for every $A \in \mathscr{X}$, $B \in \mathscr{Y}$.

1.2.1.2 Kernels and Random Variables

The notion of a kernel is strictly related to the concepts of conditional expectation and conditional probability.

Definition 1.2.5 (Conditional Expectation & Probability). Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}|X| < \infty$, and $\mathscr{G} \subseteq \mathscr{F}$ be a σ-algebra. A random variable $Y : \Omega \to \mathbb{R}$ is a *conditional expectation* for X with respect to $\mathscr{G}$, if two conditions are satisfied:

(i) Y is $\mathscr{G}$-measurable;

(ii) for every $A \in \mathscr{G}$, we have that $\int_A X \, d\mathbb{P} = \int_A Y \, d\mathbb{P}$.

It can be proved that such a random variable exists and is $a.s.$-unique (a proof is given in [10]). We write $Y = \mathbb{E}[X \mid \mathscr{G}]$. If we take X to be the A-indicator function, i.e., $X = 1_A$, we get the *conditional probability* of A with respect to $\mathscr{G}$, i.e., $\mathbb{P}(A \mid \mathscr{G}) = \mathbb{E}[1_A \mid \mathscr{G}]$.

We now give two propositions that clarify the connection between kernels and random variables.

Proposition 1.2.2. *Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ be measurable spaces and consider the mapping $P : X \times \mathscr{Y} \to [0, \infty]$ defined by*

$$P(x, A) = \mathbb{P}(Y \in A \mid X = x), \quad x \in X, A \in \mathscr{Y}.$$

Then, P is a normalized kernel on $X \times \mathscr{Y}$. We call it the conditional probability kernel of Y given X.

Proof. Let $A \in \mathscr{Y}$. The conditional probability $\mathbb{P}(Y \in A \mid X)$ is a $\sigma(X)$-measurable random variable, so the function $P(\cdot, A) : X \to [0, \infty]$, defined by $x \mapsto P(x, A)$, is Borel-measurable. Let $x \in X$. Then the function $P(x, \cdot) : \mathscr{Y} \to [0, \infty]$, defined by $A \mapsto P(x, A)$ is by definition a positive measure. Obviously, $P(x, Y) = 1$ and thus P is a normalized kernel. □

Proposition 1.2.3. *Let P be a conditional probability kernel of Y given X.*

(i) If the function $f : Y \to \mathbb{R}$ is measurable, then $Pf(x) = \mathbb{E}[f(Y) \mid X = x]$ for all $x \in X$, for Pf as in Proposition 1.2.1.

(ii) If μ is the probability distribution of X, then μP is the probability distribution of Y.

Proof. The proof contains ideas that have already been discussed; only the notation changes now.

(i) Since $P(x, A)$ expresses the conditional distribution of Y given that $X = x$, we have that

$$\mathbb{E}[f(Y) \mid X = x] = \int_X P(x, dy) f(y) = Pf(x).$$

(ii) For every $A \in \mathscr{Y}$, we have that

$$\begin{aligned}\mathbb{P}(Y \in A) = \mathbb{E}[\mathbb{P}(Y \in A \mid X)] &= \int_X \mu(dx) P(Y \in A \mid X = x) \\ &= \int_X \mu(dx) P(x, A) = \mu P(A),\end{aligned}$$

so μP is the probability distribution of Y. □

1.2.2 Homogeneous Markov Chains

Now that we have developed the useful theoretical machinery for kernels, we can define a Markov chain in the general case, where the state-space might be either countable or uncountable. First, we will need to give some measure-theoretic notation and terminology.

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(S, \mathscr{A})$ be a measurable space, where S is the state-space. Notice that, since we will start examining the chain's behavior in terms of sets instead of single points, we need a σ-algebra of the state-space's subsets; this is the role of $\mathscr{A}$. Let T be a set that denotes time. Throughout this paragraph we will take $T = \mathbb{N}$, unless otherwise stated. A *stochastic process* is a family $\{X_n : n \in T\}$ of random variables $X_n : \Omega \to S$. In this case, the stochastic process can be viewed and treated as a function $X : T \times \Omega \to S$ with $X(n, \omega) = X_n(\omega)$. Alternatively, one can define a stochastic process by examining its trajectory for each $\omega \in \Omega$. In this case, the stochastic process can be viewed and treated as a function $\hat{X} : \Omega \to S^T$, where for every $\omega \in \Omega$ the value $\hat{X}(\omega)$ is a function from T to S, satisfying $\hat{X}(\omega)(n) = X(\omega, n)$.

A sequence of σ-algebras $(\mathscr{F}_n)_{n \in T}$ is a *filtration* in $(\Omega, \mathscr{F})$ if for all $n \in T$ we have $\mathscr{F}_n \subseteq \mathscr{F}_{n+1} \subseteq \mathscr{F}$. If we endow a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with a filtration $(\mathscr{F}_n)_{n \in T}$, we get the *filtered probability space* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n \in T}, \mathbb{P})$. A stochastic process $\{X_n : n \in T\}$ is *adapted* to the filtration $(\mathscr{F}_n)_{n \in T}$ if for every $n \in T$ the random variable X_n is $\mathscr{F}_n$-measurable. We will write $\{(X_n, \mathscr{F}_n) : n \in T\}$ to denote an adapted stochastic process. The *natural filtration* for a stochastic process is the one given by $\mathscr{F}_n^X = \sigma(X_0, X_1, \ldots, X_n)$, $n \in T$. Notice that every stochastic process is adapted to its natural filtration.

We will now give some definitions for Markov chains with general state-space. The key-idea is that we will be referring to sets instead of single points. Intuitively, we shall think of the σ-algebra $\mathscr{F}_n$ as the *information* available at the n-th step.

Definition 1.2.6 (Markov chain). Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n \in T}, \mathbb{P})$ be a filtered probability space and $(S, \mathscr{A})$ be a measurable space. We will call an adapted stochastic process $\{(X_n, \mathscr{F}_n) :$

$n \in T\}$ a *Markov chain* with *state-space* S if

$$\mathbb{P}(X_{n+1} \in A \mid \mathscr{F}_n) \overset{a.s.}{=} \mathbb{P}(X_{n+1} \in A \mid X_n) \tag{1.2.2}$$

for all $n \in T$, $A \in \mathscr{A}$.

Note that we used the $a.s.$-notation, since the conditional probabilities are random variables, as we discussed in (1.2.5).

From now on, the time-set T will be the set $\mathbb{N}$ of natural numbers, unless otherwise stated. We now give the definition of a homogeneous Markov chain, making use of the notion of kernels that we introduced above. Note that, since we care about transitions within the same state-space, we will use a kernel on $S \times \mathscr{A}$, i.e., the Cartesian product of the set with a set of its own subsets, instead of involving a σ-algebra of another set.

Definition 1.2.7. Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ be a filtered probability space, $(S, \mathscr{A})$ be a measurable space and P be a kernel on $S \times \mathscr{A}$. We call the adapted stochastic process $\{(X_n, \mathscr{F}_n) : n \in \mathbb{N}\}$ a *homogeneous Markov chain* with state-space S and initial distribution the distribution of X_0, if

$$\mathbb{P}(X_{n+1} \in A \mid \mathscr{F}_n) \overset{a.s.}{=} P(X_n, A) \tag{1.2.3}$$

for all $n \in \mathbb{N}$, $A \in \mathscr{A}$.

Intuitively we shall think of the previous definition as follows: the left-hand side is a random variable that describes the next state of the Markov chain given the information that we have so far, while the right-hand side involves a kernel and, thus, is telling us how likely it is, given that the chain is in the state X_n at the n-th step, to go to the set A afterwards. The fact that the stage itself does not play any role in the above equality implies time-homogeneity.

In the countable state-space case, we showed that the Markov chain is entirely determined by its initial distribution and its transition matrix. It comes as no surprise that this property still holds if we replace the transition matrix with a Markov kernel. In particular, we have the following important analogue of Theorem (1.1.1)[8], connecting the countable and the uncountable case.

Theorem 1.2.1. *Let P be a Markov kernel on $S \times \mathscr{A}$ and λ an arbitrary distribution on S. An S-valued stochastic process $(X_n)_{n\in\mathbb{N}}$ is a homogeneous Markov chain with kernel P and initial distribution λ, if for every $n \in \mathbb{N}$ the joint distribution of the vector $(X_0, X_1, \ldots, X_n)$ is given by $\lambda \otimes P^{\otimes n}$.*

[8] a proof can be found in [12]

1.2.3 The Canonical Chain

A question that arises naturally is the following: given a probability distribution λ and a Markov kernel P, can one find a Markov chain that has λ as its initial distribution and P as its transition kernel? We will see that the answer is positive. Given a general state-space S, the key-idea is to consider a new filtered probability space, whose sample elements will be sequences of elements of S. This way, we will also be able to establish a nice connection with the shift transformation from Ergodic Theory and, thus, use its results in our studying of Markov chains' limit behavior.

Definition 1.2.8 (Coordinate Process). Let $(S, \mathscr{A})$ be a measurable space. Let $\Omega = S^{\mathbb{N}}$ be the set of sequences that take values in S, i.e.,

$$\Omega = S^{\mathbb{N}} = \{\omega = (\omega_0, \omega_1, \omega_2, \dots) : \omega_i \in S \ \forall i \in \mathbb{N}\},$$

endowed with the product σ-algebra $\mathscr{A}^{\otimes\mathbb{N}}$. The stochastic process $\{X_n : n \in \mathbb{N}\}$ defined by $X_n(\omega) = \omega_n$ for every $\omega = (\omega_0, \omega_1, \dots, \omega_n, \dots) \in \Omega$, is called the *coordinate process* on S. Every $\omega \in \Omega$ is called a *path* of the process. We endow the measurable space $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$ with the *canonical filtration* $\{\mathscr{F}_n : n \in \mathbb{N}\}$, where $\mathscr{F}_n = \sigma(X_0, \dots, X_n)$ for every $n \in \mathbb{N}$.

One can prove that, using an appropriate filtered probability space, we are always able to find a Markov chain with given initial distribution and transition kernel. The proof of this important result is highly technical, with emphasis mostly on measure-theoretical tools, and lies outside the scope of the current thesis. The interested reader is referred to [12] for a detailed proof.

Theorem 1.2.2. *Let $(S, \mathscr{A})$ be a measurable space and P a Markov kernel on $S \times \mathscr{A}$. For every probability measure λ on $(S, \mathscr{A})$, there exists a unique probability measure $\mathbb{P}_\lambda$ on the measurable space $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$ such that, for $\{\mathscr{F}_n : n \in \mathbb{N}\}$ as above, the adapted coordinate process $\{(X_n, \mathscr{F}_n) : n \in \mathbb{N}\}$ is a Markov chain with initial distribution λ and transition kernel P.*

Definition 1.2.9 (Canonical Markov Chain). Let $(S, \mathscr{A})$ be a measurable space. Consider the coordinate process $\{X_n : n \in \mathbb{N}\}$ and the canonical filtration $\{\mathscr{F}_n : n \in \mathbb{N}\}$ on the measurable space $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$. If P is a Markov kernel on $S \times \mathscr{A}$ and $\{\mathbb{P}_\lambda\}$ the family of probability measures on $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$ introduced by Theorem 1.2.2, then the coordinate process $\{X_n : n \in \mathbb{N}\}$ will be referred to as the *canonical Markov chain.*

1.2.4 Ergodic Theory and Markov Chains

Ergodic Theory is an independent branch of Mathematics that was introduced in the 20th century and, ever since, has had an immense effect on several scientific fields, such as Number Theory, Probability & Statistics, Riemannian Geometry, Statistical Mechanics etc.

Roughly speaking, Ergodic Theory deals with the study of the long-term statistical behavior of certain dynamical systems, especially those which exhibit *ergodicity*[9]. In this section, we will give some very introductory definitions and state the celebrated Birkhoff's Ergodic Theorem, a theorem that has had tremendous applications in the theory of Markov chains and MCMC methods. For a detailed presentation of Ergodic Theory, we refer the reader to [7], [12], [16] and [32].

Definition 1.2.10 (Dynamical System). Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $T : \Omega \to \Omega$ a measurable and measure preserving map, i.e., $T^{-1}(\mathscr{F}) \subseteq \mathscr{F}$ and $\mathbb{P}(T^{-1}(A)) = \mathbb{P}(A)$ for every $A \in \mathscr{F}$. Then, the quartet $(\Omega, \mathscr{F}, \mathbb{P}, T)$ is referred to as a *dynamical system*. We say that T is a *measure preserving transformation* and $\mathbb{P}$ is *invariant* under T. If T is invertible and T^{-1} is also measurable, we say that T is an *invertible measure preserving transformation*.

We will now use the space $S^{\mathbb{N}}$ that we used previously in order to describe the canonical chain.

Definition 1.2.11 (Shift operator). Let $(S, \mathscr{A})$ be a measurable space and take the associated measurable space $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$. The mapping $T : S^{\mathbb{N}} \to S^{\mathbb{N}}$ defined by

$$\omega = (\omega_0, \omega_1, \omega_2, \dots) \mapsto T(\omega) = (\omega_1, \omega_2, \omega_3, \dots),$$

is referred to as the *shift operator*. It can be shown that it is $\mathscr{A}^{\otimes\mathbb{N}}$-measurable.

Definition 1.2.12 (Stationary Process). A stochastic process $\{X_n : n \in \mathbb{N}\}$ is *stationary* if the joint distribution of $(X_{n_1}, \dots, X_{n_k})$ is the same as that of $(X_{n_1+m}, \dots, X_{n_k+m})$ for every $k, m, n_1, \dots, n_k \in \mathbb{N}$.

Definition 1.2.13. Let $(\Omega, \mathscr{F})$ be a measurable space and $T : \Omega \to \Omega$ a measurable map. A random variable $Y : \Omega \to \overline{\mathbb{R}}$ is called *invariant* for T if $Y \circ T = Y$. An event A is called *invariant* for T if $A = T^{-1}(A)$.

It is easy to check that the family

$$\mathscr{G} = \{A \in \mathscr{F} : A \text{ is invariant for } T\}$$

is a sub-σ-algebra of $\mathscr{F}$.

Definition 1.2.14 (Ergodic Dynamical System). A dynamical system $(\Omega, \mathscr{F}, \mathbb{P}, T)$ is *ergodic* if $\mathbb{P}(A) \in \{0, 1\}$ for every $A \in \mathscr{G}$.

We have now developed the machinery that is needed in order to connect Ergodic Theory to Markov chains. The connection involves the result from Theorem 1.2.2 and the shift operator, and can be summarized in the following proposition.

[9] intuitively, ergodicity expresses the idea that a point of a system will visit all the space in a way that is random and uniform

Proposition 1.2.4. *Let $(S, \mathscr{A})$ be a measurable space. A probability measure $\mathbb{P}$ on the canonical space $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$ is invariant under the shift operator $T : S^{\mathbb{N}} \to S^{\mathbb{N}}$ if, and only if, the coordinate process $\{X_n : n \in \mathbb{N}\}$ is stationary with respect to $\mathbb{P}$.*

Now we are ready to state and prove one of the most important theorems in Ergodic Theory and the theory of Markov chains; the Ergodic Theorem. There are several ergodic theorems, such as Birkhoff's pointwise ergodic theorem, von Neumann's mean ergodic theorem, the maximal ergodic theorem etc. Here we will state and prove the first one, namely the pointwise ergodic theorem, as introduced by George Birkhoff in 1932 (for historical notes the reader is referred to [49]). The proof of this result is in Appendix A.

Theorem 1.2.3 (Pointwise Ergodic Theorem). *Let $(\Omega, \mathscr{F}, \mathbb{P}, T)$ be a dynamical system, $Y \in L^1(\mathbb{P})$ a random variable and $\mathscr{G}$ the σ-algebra of T-invariant sets. Then, we have that*

$$\frac{1}{n}\sum_{k=0}^{n-1} Y \circ T^k \xrightarrow{a.s.} \mathbb{E}[Y \mid \mathscr{G}]. \tag{1.2.4}$$

If the dynamical system $(\Omega, \mathscr{F}, \mathbb{P}, T)$ is ergodic, the σ-algebra $\mathscr{G}$ is trivial and thus the ergodic theorem becomes

$$\frac{1}{n}\sum_{k=0}^{n-1} Y \circ T^k \xrightarrow{a.s.} \mathbb{E}[Y]. \tag{1.2.5}$$

We can now use Theorem 1.2.2 in order to study Markov chains via dynamical systems and get the Ergodic Theorem for Markov chains as a corollary of Theorem 1.2.3. Given the state-space $(S, \mathscr{A})$, we will consider the canonical space $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}})$ and the coordinate process $\{X_n : n \in \mathbb{N}\}$. For a Markov kernel P on $S \times \mathscr{A}$, we endow the canonical space with a family $(\mathbb{P}_\lambda)$ of probability measures such that the coordinate process is a Markov chain with initial distribution λ and kernel P. Let $T : S \to S$ denote the shift operator.

Theorem 1.2.4 (Ergodic Theorem for Markov chains). *Let P be a Markov kernel on $S \times \mathscr{A}$ and π a stationary distribution for P. Assume that the dynamical system $(S^{\mathbb{N}}, \mathscr{A}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, T)$ is ergodic and $Y \in L^1(\mathbb{P}_\pi)$. Then,*

$$\frac{1}{n}\sum_{k=0}^{n-1} Y \circ T_k \xrightarrow{a.s.} \mathbb{E}_\pi[Y]. \tag{1.2.6}$$

Part II

Statistics for Markov Chains

Chapter 2

Statistical Inference for Finite Markov Chains

2.1 Introduction

In Chapter 1 we studied the probabilistic behavior of a Markov chain. That is, we assumed that its transition matrix (or kernel) is known and gave a plethora of results that one can get. Now we turn our interest to a problem of statistical nature: *given* a path of Markov chain, what can we say about its transition probabilities? We will provide estimations in two frameworks: the Classical (or Frequentist) and the Bayesian one. For that purpose, we will be constrained in finite state-space Markov chains, although analogous results hold for more general state-spaces as well (extended presentations can be found in [1], [5], [6], [46]).

Let $\{X_n : n \in \mathbb{N}\}$ be a time-homogeneous Markov chain with finite state-space S. Without loss of generality, we can take $S = \{1, \dots, m\}$. For two states $i, j \in S$, we define the transition probability

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_1 = j \mid X_0 = i),$$

so the transition matrix P is the $m \times m$ stochastic matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}.$$

We will address the problem of estimating P on the basis of a path $\{x_0, x_1, \dots, x_n\}$ of our Markov chain. In the frequentist approach this is done by finding the Maximum Likelihood Estimator (MLE), while the Bayesian approach requires that we determine the posterior distribution.

2.2 Frequentist approach

Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with unknown transition matrix $P = (p_{ij})_{i,j\in S}$. The goal is to estimate P. In order to exploit the classical framework, we will make parametric assumptions; that is, we will assume that the unknown quantities form a subset of a Euclidean space. We will define two possible parametric spaces, find the MLE of P and provide asymptotic results for it.

2.2.1 Parametrization

Unless P has a special structure (zeros in several positions etc), all m^2 elements are considered unknown and, thus, the parametric space can be viewed as a subset of $\mathbb{R}^{m^2}$. If $p_i = (p_{i1}, \ldots, p_{im}) \in \mathbb{R}^m$ denotes the i-th row of P, then

$$P = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix} = (p_1 \cdots p_m)^T \in \mathbb{R}^{m^2}.$$

The two main ways to parametrize P are via the:

- Natural Parametric Space,
- Minimal Parametric Space.

Definition 2.2.1. The *natural parametric space* is the Cartesian product

$$\Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_m,$$

where each Θ_i, $1 \leqslant i \leqslant m$, is a probability simplex defined by

$$\Theta_i = \left\{ (p_{i1}, \ldots, p_{im}) : \ 0 \leqslant p_{i1}, \ldots, p_{im} \leqslant 1, \sum_{k=1}^{m} p_{ik} = 1 \right\} \subseteq [0,1]^m.$$

The idea for the use of another parametric space derives from the fact that, since the probabilities of a row add up to 1, we only need $m-1$ and not all m of them. A question that arises is which of the m elements of each row we will discard. If one-step transitions from a state to itself are possible (i.e., the diagonal elements of P are not identically equal to 0), we generally discard the diagonal elements.

Definition 2.2.2. The *minimal parametric space* is the Cartesian product

$$\Theta^* = \Theta_1^* \times \Theta_2^* \times \ldots \times \Theta_m^*,$$

where each Θ_i^*, $1 \leqslant i \leqslant m$, is defined by

$$\Theta_i^* = \left\{ (p_{i1}, \ldots, p_{i,i-1}, p_{i,i+1}, \ldots, p_{im}) : \sum_{k \in S_i} p_{ik} \leqslant 1 \right\} \subseteq [0,1]^{m-1},$$

where $S_i := \{1, \ldots, i-1, i+1, \ldots, m\}$.

We provide a graphical illustration of Θ_i and Θ_i^* in the case $m = 3$.

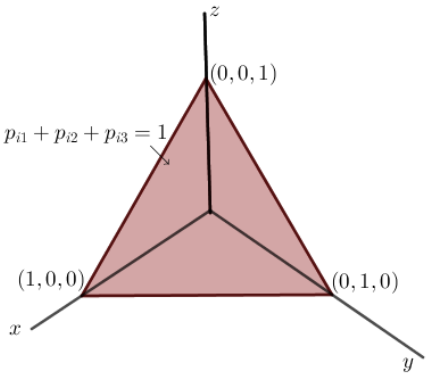

FIGURE 2.1: Natural Parametric Space

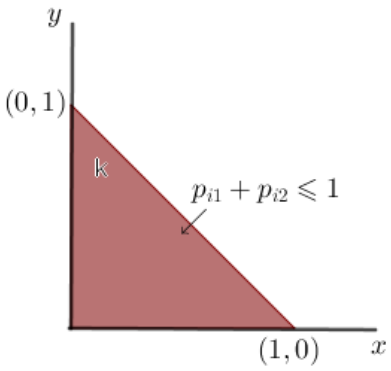

FIGURE 2.2: Minimal Parametric Space

2.2.2 Maximum Likelihood Estimation

Let λ be the initial distribution of $(X_n)_{n \in \mathbb{N}}$ and $\{x_0, x_1, \ldots, x_n\}$ a realization of the chain at time n. We define the *counting processes* $(n_{ij}(n))_{n \in \mathbb{N}}$ and $(n_i(n))_{n \in \mathbb{N}}$ by

$$n_{ij}(n) = \sum_{k=0}^{n-1} 1(X_k = i, X_{k+1} = j) \quad \text{and} \quad n_i(n) = \sum_{k=0}^{n-1} 1(X_k = i), \tag{2.2.1}$$

the total number of $i \to j$ transitions up to time n and the total number of visits in the state before time n, respectively. The likelihood $L(\lambda, P)$ is given by

$$\begin{aligned} L(\lambda, P) &= \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\ &= \lambda_{x_0} \cdot \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \ldots \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\ &= \lambda_{x_0} \cdot \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \ldots \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \\ &= \lambda_{x_0} \cdot \prod_{t=1}^{n} p_{x_{t-1} x_t} = \lambda_{x_0} \cdot \prod_{i \in S} \prod_{j \in S} p_{ij}^{n_{ij}(n)}. \end{aligned}$$

Since we cannot estimate the initial distribution λ based solely on one realization of the chain, we will work conditioning on X_0, so the likelihood becomes

$$L_n(P) = \mathbb{P}(X_1, \dots, X_n; P \mid x_0) = \prod_{i \in S} \prod_{j \in S} p_{ij}^{n_{ij}(n)} = \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{n_{ij}(n)}, \tag{2.2.2}$$

thus the log-likelihood is given by

$$\ell_n(P) := \log L_n(P) = \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij}(n) \log p_{ij}. \tag{2.2.3}$$

We wish to find the matrix $P = (p_{ij})_{i,j \in S}$ that maximizes the likelihood (2.2.2) or, equivalently, the log-likelihood (2.2.3) under the m constraint equations

$$\sum_{j=1}^{m} p_{1j} = 1, \sum_{j=1}^{m} p_{2j} = 1, \dots, \sum_{j=1}^{m} p_{mj} = 1. \tag{2.2.4}$$

The fact that we have m constraint equations reduces the degrees of freedom from m^2 (the number of the unknown parameters) to $m^2 - m$. We can thus maximize (2.2.3) either by using Lagrange multipliers or by eliminating parameters. We will proceed with the former. Since we have m constraints, we will take m Lagrange multipliers, namely $\lambda_1, \dots, \lambda_m$. Let $\underline{\lambda} = (\lambda_1, \dots, \lambda_m)$. The objective function then is

$$\begin{aligned} \Lambda(P, \underline{\lambda}) &= \ell_n(P) - \lambda_1 \left(\sum_{j=1}^{m} p_{ij} - 1 \right) - \dots - \lambda_n \left(\sum_{j=1}^{m} p_{ij} - 1 \right) \\ &= \ell_n(P) - \sum_{i=1}^{m} \lambda_i \left(\sum_{j=1}^{m} p_{ij} - 1 \right) = \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij}(n) \log p_{ij} - \sum_{i=1}^{m} \lambda_i \left(\sum_{j=1}^{m} p_{ij} - 1 \right). \end{aligned}$$

Using the constraint equations (2.2.4), we get

$$\frac{\partial \Lambda}{\partial \lambda_1} = \frac{\partial \Lambda}{\partial \lambda_2} = \dots = \frac{\partial \Lambda}{\partial \lambda_m} = 0.$$

For every $i, j \in S$, we have that

$$\frac{\partial \Lambda}{\partial p_{ij}} = 0 \Leftrightarrow \frac{n_{ij}(n)}{p_{ij}} - \lambda_i = 0 \Leftrightarrow p_{ij} = \frac{n_{ij}(n)}{\lambda_i}$$

and from (2.2.4) we receive

$$\sum_{j=1}^{m} p_{ij} = 1 \Leftrightarrow \sum_{j=1}^{m} \frac{n_{ij}(n)}{\lambda_i} = 1 \Leftrightarrow \lambda_i = \sum_{j=1}^{m} n_{ij}(n) = n_i(n),$$

where the last equality follows from a simple computation using (2.2.1). If we denote by $\hat{p}_{ij}(n)$ the maximum likelihood estimator of p_{ij}, the above arguments constitute (having omitted a few technical details) a proof of the following.

Proposition 2.2.1 (M.L.E. of Transition Matrix). *Let $(X_n)_{n\in\mathbb{N}}$ be a discrete-time homogeneous Markov chain with finite state-space S, initial distribution λ and transition matrix P. Let $\{x_0, x_1, \dots, x_n\}$ be an observed realization of the chain. If $\hat{P}_n$ denotes the maximum likelihood estimator of the transition matrix based on the given path at time n, then we have that $\hat{P}_n = (\hat{p}_{ij}(n))_{i,j\in S}$, where*

$$\hat{p}_{ij}(n) = \begin{cases} \dfrac{n_{ij}(n)}{n_i(n)}, & \text{if } n_i(n) > 0 \\ 0, & \text{if } n_i(n) = 0 \end{cases}.$$

2.2.3 Asymptotic Behavior

Now that we have an estimator, we would like to examine its asymptotic performance. We will see that the MLE is strongly consistent and asymptotically normal. We will prove the first (possibly avoiding a few technicalities) and state the second result. More concise explanations can be found in [1], [5], [6] and [46].

Theorem 2.2.1 (Consistency of the M.L.E.). *For every $i, j \in S$, we have that*

$$\hat{p}_{ij}(n) \xrightarrow{a.s.} p_{ij},$$

i.e., the maximum likelihood estimator $\hat{P}_n = (\hat{p}_{ij})_{i,j\in S}$ is strongly consistent for the transition matrix P.

Proof. Recall from definition (2.2.1) that

$$n_{ij}(n) = \sum_{k=0}^{n-1} 1(X_k = i, X_{k+1} = j) = \sum_{k=0}^{n-1} 1(X_k = i)1(X_{k+1} = j).$$

Applying Birkhoff's pointwise ergodic theorem, we get that

$$\begin{aligned} \frac{n_{ij}(n)}{n-1} &= \frac{1}{n-1}\sum_{k=0}^{n-1} 1(X_k = i, X_{k+1} = j) \\ &\xrightarrow{a.s.} \mathbb{E}\left[1(X_k = i, X_{k+1} = j)\right] \\ &= \mathbb{P}\left(X_k = i, X_{k+1} = j\right) \\ &= \mathbb{P}(X_k = i)\mathbb{P}(X_{k+1} = j \mid X_k = i) = \mathbb{P}(X_k = i)\, p_{ij}. \end{aligned} \tag{2.2.5}$$

Using the same idea for $n_i(n) = \sum_{k=0}^{n-1} 1(X_k = i)$, we have that

$$\frac{n_i(n)}{n-1} = \frac{1}{n-1}\sum_{k=0}^{n-1} 1(X_k = i) \xrightarrow{a.s.} \mathbb{E}\left[1(X_k = i)\right] = \mathbb{P}(X_k = i). \tag{2.2.6}$$

Combining (2.2.5) with (2.2.6), Slutsky's theorem yields

$$\hat{p}_{ij}(n) = \frac{n_{ij}(n)}{n_i(n)} = \frac{\frac{n_{ij}(n)}{n-1}}{\frac{n_i(n)}{n-1}} \xrightarrow{a.s.} \frac{\mathbb{P}(X_k = i)\, p_{ij}}{\mathbb{P}(X_k = i)} = p_{ij}$$

and the proof is complete. □

Theorem 2.2.2 (Asymptotic Normality of the M.L.E.). *Let $\pi = (\pi_1, \ldots, \pi_m)$ denote the stationary distribution of $(X_n)_{n\in\mathbb{N}}$. The maximum likelihood estimator matrix $\hat{P}_n$ is asymptotically normal, i.e.,*

$$\sqrt{n}(\hat{P}_n - P) \xrightarrow{d} \mathscr{N}_{m^2}(0, \Sigma), \tag{2.2.7}$$

where Σ is an $m^2 \times m^2$ block-diagonal matrix given by

$$\Sigma = \begin{pmatrix} \frac{1}{\pi_1}\Lambda_1 & 0 & \cdots & 0 \\ 0 & \frac{1}{\pi_2}\Lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\pi_m}\Lambda_m \end{pmatrix}$$

and Λ_i, $i = 1, 2, \ldots, m$, are $m \times m$ covariance matrices defined by

$$\Lambda_i = \begin{pmatrix} p_{i1}(1-p_{i1}) & -p_{i1}p_{i2} & \cdots & -p_{i1}p_{im} \\ -p_{i2}p_{i1} & p_{i2}(1-p_{i2}) & \cdots & -p_{i2}p_{im} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{im}p_{i1} & -p_{im}p_{i2} & \cdots & p_{im}(1-p_{im}) \end{pmatrix}.$$

In particular, for every $i, j \in S$ we have that

$$\sqrt{n\,\pi_i}\cdot\frac{\hat{p}_{ij}(n) - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})}} \xrightarrow{d} N(0,1). \tag{2.2.8}$$

Remark 4. Notice that there is an interesting similarity with the Central Limit Theorem applied to a multinomial distribution, that is

$$\sqrt{n}\cdot\frac{\hat{p}_i(n) - p_i}{\sqrt{p_i(1-p_i)}} \xrightarrow{d} N(0,1).$$

This does not come as a surprise, if we take into consideration the multinomial nature of the problem (remember that the likelihood contains terms of the form $p_{ij}^{n_{ij}(n)}$). The basic difference now is that the sample size n becomes $n\,\pi_i$, which is reasonable since we are now interested in the $i \to j$ transitions, so we want to examine how much time the Markov chain has spent in the i state; and that is exactly what the stationary distribution tells us.

Remark 5. The above theorem, along with Slutsky's theorem and the Continuous Mapping Theorem, can be used to derive asymptotic confidence intervals for p_{ij}, $i, j \in S$. In particular, for $a \in (0,1)$ we have that

$$I_{1-a}(p_{ij}) = \left(\hat{p}_{ij}(n) - z_{a/2}\sqrt{\frac{\hat{p}_{ij}(n)(1-\hat{p}_{ij}(n))}{n_i(n)}}\,,\, \hat{p}_{ij}(n) + z_{a/2}\sqrt{\frac{\hat{p}_{ij}(n)(1-\hat{p}_{ij}(n))}{n_i(n)}} \right)$$

is a $100(1-a)\%$ confidence interval for p_{ij}, where $z_{a/2}$ is such that $\mathbb{P}(Z > z_{a/2}) = \frac{a}{2}$ for $Z \sim N(0,1)$.

2.3 Bayesian approach

2.3.1 Introduction

In Classical Statistics, we view parameters as constants. This assumption might seem innocuous at first glance, but is crucial in the development and philosophy of this framework. It turns out that this is not the only way to carry out statistical inference. One can view an unknown quantity as a random variable and enter a quite different realm of inference, the realm of Bayesian statistics. Apart from all the technical differences that arise with such a change, the essence of the two approaches is quite distinct. In the heart of Bayesian inference lies the use of *probability* in order to quantify beliefs and uncertainty. The main tool of Bayesian statistics is Bayes' rule, which can intuitively be written as

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})\, P(\text{model})}{P(\text{data})}.$$

There are four important quantities in the above formula:

(1) the *likelihood*: $P(\text{data} \mid \text{model})$,

(2) the *prior* distribution: $P(\text{model})$,

(3) the *posterior* distribution: $P(\text{model} \mid \text{data})$,

(4) the *evidence* (*marginal likelihood*): $P(\text{data})$.

A detailed exposition of the Bayesian approach to inference is outside the scope of this thesis. For an in-depth presentation of this framework, the reader is referred to [20]. In essence, suppose we want to carry out inference about a (possibly multidimensional) parameter θ

on the basis of the observations x. If, by any means, we have prior information for θ, we can use it in our analysis and construct a prior distribution $p(\theta)$. We then compute the likelihood $f(x \mid \theta)$, whose importance is already understood from the classical framework. The ultimate goal of Bayesian statistics is to compute the posterior distribution $\pi(\theta \mid x)$. Unlike classical statistics, in which there are several ways of carrying out inference (pointwise estimation, confidence intervals etc), in Bayesian Statistics the posterior distribution *is* the inference. Any quantity of interest can be computed using probabilistic arguments on the posterior distribution. All the above characteristics are combined in Bayes' rule as follows:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, p(\theta)}{f(x)} = \frac{f(x \mid \theta)\, p(\theta)}{\int_\Theta f(x \mid \theta)\, p(\theta)\, d\theta} \propto f(x \mid \theta) p(\theta). \tag{2.3.1}$$

Notice that the denominator is a constant, so one can write $\pi(\theta \mid x) \propto f(x \mid \theta) p(\theta)$, where the symbol $\propto$ means "proportional to". Such writing is valid, since $\pi(\theta \mid x)$ is a probability distribution, so its integral is equal to 1 and, thus, the normalizing constant can be uniquely determined.

A key element of the Bayesian approach is the notion of *updating*: one has some prior beliefs and updates them on the basis of the observed data, in order to get the posterior distribution. In order to carry out Bayesian inference, we first need to specify a suitable prior distribution. A comfortable situation occurs when the prior and the posterior are in the same distribution family. In such a case, the prior distribution is said to be *conjugate* for the particular likelihood model.

2.3.2 Dirichlet Distribution

A likelihood that arises very often in Statistics is the binomial, since it measures the number of successes in a given number of Bernoulli trials. The conjugate family for a binomial likelihood model is the family of Beta distributions and this is easy to check. The multidimensional analogue of the binomial distribution is the multinomial distribution and it turns out that the likelihood (2.2.2) is part of that. It would be computationally convenient to have a distribution that is conjugate for the multinomial likelihood model. This is the family of Dirichlet distributions.

Let $m \in \mathbb{N}$ and assume that we are working in the $(m-1)$-simplex

$$\Theta = \left\{ (x_1, \ldots, x_m) : \ 0 \leqslant x_1, \ldots, x_m \leqslant 1, \sum_{k=1}^{m} x_i = 1 \right\} \subseteq [0,1]^m.$$

For $m = 2$, the probability density function (pdf) of a $\text{Beta}(p, q)$ distribution is

$$X \sim \text{Beta}(p, q) \quad \Leftrightarrow \quad f_X(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}.$$

The Dirichlet distribution is the multidimensional analogue for the Beta distribution. We say

that a random variable $X = (X_1, \dots, X_m)$ follows a Dirchlet distribution with parameters $c_1, \dots, c_m > 0$ and write $X \sim \text{Dir}(c_1, \dots, c_m)$ if its support is the $(m-1)$-simplex Θ, living in an m-dimensional space, and its pdf is given by

$$f_X(x_1, \dots, x_m) = \frac{\Gamma\left(\sum_{i=1}^{m} c_i\right)}{\prod_{i=1}^{m} \Gamma(c_i)} \prod_{i=1}^{m} x_i^{c_i - 1}.$$

If we view a simplex as a discrete probability distribution, then the Dirichlet distribution assigns a probability at each probability vector. In other words, the Dirichlet distribution can be viewed as a distribution over distributions and has many interesting properties. Let $X = (X_1, \dots, X_m) \sim \text{Dir}(c_1, \dots, c_m)$. The mean value of each component is

$$\mathbb{E}[X_i] = \frac{c_i}{c_1 + \cdots + c_m}, \quad i = 1, \dots, m,$$

assigning in a way the relative importance of each parameter. This result can be derived from the fact that the marginal distributions of X's components are Beta distributions. In particular, if $X = (X_1, \dots, X_m) \sim$Dir$(c_1, \dots, c_m)$, then

$$X_k \sim \text{Beta}(c_k\,,\, c_1 + \cdots + c_{k-1} + c_{k+1} + \cdots + c_m), \quad 1 \leqslant k \leqslant m.$$

In order to simulate a single observation $(x_1, \dots, x_m)$ from $\text{Dir}(c_1, \dots, c_m)$, we can first simulate m independent observations $y_i \sim \text{Gamma}(c_i, 1)$ and then set

$$(x_1, \dots, x_m) = \left(\frac{y_1}{\sum_{i=1}^{m} y_i}\,,\, \dots\,,\, \frac{y_m}{\sum_{i=1}^{m} y_i} \right).$$

2.3.3 Posterior Inference

In our problem, the unknown parameter is $\theta = P$ and the likelihood is the same as in (2.2.2), i.e.,

$$f(x \mid P) = L_n(P) = \mathbb{P}(X_1, \dots, X_n; P \mid x_0) = \prod_{i \in S} \prod_{j \in S} p_{ij}^{n_{ij}(n)} = \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{n_{ij}(n)}.$$

We need to find a suitable prior distribution for the matrix P. A reasonable idea is to study each row independently and assign a Dirichlet distribution, which is conjugate for the multinomial likelihood model. In particular, if $P_{i\cdot} = (p_{i1}, \dots, p_{im})$ denotes the i-th row of P, we can take a prior $P_{i\cdot} \sim \text{Dir}(c_{i1}, \dots, c_{im})$ assigning density

$$p(P_{i\cdot}) \propto p_{i1}^{c_{i1}-1} \cdots p_{im}^{c_{im}-1}$$

at each probability vector $P_{i\cdot}$, so the prior of the whole transition matrix P becomes

$$p(P) = \prod_{i=1}^{m} p(P_{i\cdot}) \propto \prod_{i=1}^{m} p_{i1}^{c_{i1}-1} \cdots p_{im}^{c_{im}-1}.$$

Combining the likelihood with the prior, the posterior distribution becomes

$$\begin{aligned}
\pi(P \mid x) \propto f(x \mid P)\, p(P) &\propto \prod_{i=1}^{m}\prod_{j=1}^{m} p_{ij}^{n_{ij}(n)} \prod_{k=1}^{m} p_{k1}^{c_{k1}-1} \cdots p_{km}^{c_{km}-1} \\
&= \prod_{i=1}^{m} \left(p_{i1}^{n_{i1}(n)-1} \cdot p_{i2}^{n_{i2}(n)-1} \cdots p_{im}^{n_{im}(n)-1} \right) \cdot \prod_{i=1}^{m} \left(p_{i1}^{c_{i1}-1} \cdots p_{im}^{c_{im}-1} \right) \\
&= \prod_{i=1}^{m} \left(p_{i1}^{n_{i1}(n)+c_{i1}-1} \cdots p_{i1}^{n_{im}(n)+c_{im}-1} \right).
\end{aligned}$$

This shows that the prior independence of each row leads to posterior independence of each row, i.e., $P_{i\cdot} \mid x \sim \text{Dir}(n_{i1}(n) + c_{i1}, \ldots, n_{im}(n) + c_{im})$. Now that we have the posterior distribution of each row and the transition matrix P in its whole, one can make inference using probabilistic arguments on these posterior distributions. For instance, the posterior mean of each transition probability is

$$\mathbb{E}[p_{ij} \mid x] = \frac{n_{ij}(n) + c_{ij}}{\sum_{k=1}^{m} (n_{ik}(n) + c_{ik})}.$$

The main difference of the Bayesian approach in comparison to the frequentist one, is that now we are not trying to study some characteristics (e.g. its standard error or sampling distribution) of a statistic that estimates the parameter vector of interest, but rather we find the whole posterior distribution of this parameter vector.

Chapter 3

Bootstrapping Finite Markov Chains

3.1 Frequentist Bootstrap

Consider an ergodic (positive recurrent, irreducible, aperiodic), time-homogeneous Markov chain $(X_n)_{n\geqslant 0}$, with finite state space and transition matrix P. Let $\hat{P}_n$ denote its maximum likelihood estimator based on an observed path $\mathbf{x} = (x_0, x_1, \ldots, x_n)$, as discussed in (2.2.1). Our goal is to estimate the sample distribution of the random variable

$$R(\mathbf{X}, P) := \sqrt{n}\,(\hat{P}_n - P).$$

For that, we adopt the *parametric bootstrap* approach, which consists of the following steps:

(1) Find an estimate of P. Here, we will use the maximum likelihood estimator $\hat{P}_n$.

(2) Use $\hat{P}_n$ as a transition matrix and generate a path $(X_0^*, X_1^*, \ldots, X_{N_n}^*)$, where $N_n + 1$ is the length of the generated path.

(3) Using the plug-in principle, approximate the sample distribution of $R(\mathbf{X}, P)$ by the distribution of $R^* := R(\mathbf{X}^*, \hat{P}_n)$. That is, find the maximum likelihood estimator $\tilde{P}_n$ as if the matrix $\hat{P}_n$ that generated the data was the unknown transition matrix, and study the distribution of $\sqrt{N_n}\,(\tilde{P}_n - \hat{P}_n)$.

In practice, we implement the boostrap method using Monte Carlo simulation to generate paths from the desired Markov chain whose transition matrix is to be estimated. The idea behind the procedure described above is the following. We would like to estimate the sampling distribution of the maximum likelihood estimator (MLE) of the transition probability matrix, since this is the quantity that governs the behavior of the Markov chain (assuming a fixed initial state). However, based on a given dataset, we only have one path, and hence one MLE, so we cannot know how good this estimate is, what is its variability etc. For that, we adopt the parametric bootstrap approach, which allows us to mimic the data obtaining process and obtain several MLEs. Based on them, we can then provide point estimates and

confidence intervals for the transition probabilities.

We shall see below that there is a nice asymptotic verification of the validity of the procedure described above. It turns out that $\sqrt{N_n}\,(\tilde{P}_n - \hat{P}_n)$ converges in distribution to the same distribution as $\sqrt{n}\,(\hat{P}_n - P)$ does, i.e., $N(0, \Sigma)$ (Theorem 2.2.2).

Before we state and prove this Theorem, we first state one of the most important Theorems in Probability Theory, namely the *Lindeberg-Feller Central Limit Theorem*. This theorem generalizes the well-known Lindeberg-Lévy Central Limit Theorem, which applies only in the iid case. Now we drop the "identically distributed" condition and deal with independent rows of independent random variables, possibly with different distributions each. If a technical condition (which intuitively expresses the idea that the random variables do not take large values with high probability) is satisfied, then the sum of these random variables converges in distribution to a random variable following a Normal distribution. We will state only the one (and most important) direction of the theorem. It is the case that if an additional condition holds, then the converse is also true (but this is of less interest in our case).

Theorem 3.1.1 (Lindeberg-Feller CLT). *Let $\{X_{nj} : 1 \leqslant j \leqslant k_n\}$ be a triangular array of independent random variables, that is*

$$\begin{array}{cccc} X_{11} & X_{12} & \cdots & X_{1k_1} \\ X_{21} & X_{22} & \cdots & X_{2k_2} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{nk_n} \\ \cdots & \cdots & \cdots & \cdots, \end{array}$$

where, for every n, the random variables $X_{n1}, \ldots, X_{nk_n}$ are independent and satisfy

$$\mathbb{E}[X_{nj}] = 0, \quad \sigma_{nj}^2 = \mathbb{E}[X_{nj}^2] < \infty, \quad s_n^2 = \sum_{j=1}^{k_n} \sigma_{nj}^2 > 0.$$

If the Lindeberg condition

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^{k_n} \mathbb{E}\Big[X_{nj}^2\, I(|X_{nj}| \geqslant \varepsilon s_n)\Big] = 0 \tag{3.1.1}$$

is satisfied, and we set $S_n = X_{n1} + \cdots + X_{nk_n}$, then we have that

$$\frac{S_n}{s_n} \xrightarrow{d} N(0,1). \tag{3.1.2}$$

If we take the normalized version of the random variables, we get an equivalent formulation of the Lindeberg-Feller CLT:

Theorem 3.1.2 (Lindeberg-Feller CLT). *Let $\{X_{nj} : 1 \leqslant j \leqslant k_n\}$ be a triangular array of*

independent random variables, satisfying

$$\mathbb{E}[X_{nj}] = 0, \quad \sigma_{nj}^2 = \mathbb{E}[X_{nj}^2] < \infty, \quad s_n^2 = \sum_{j=1}^{k_n} \sigma_{nj}^2 = 1.$$

Then the following are equivalent:

A. $S_n \xrightarrow{d} N(0,1)$, *where* $S_n :=$ *and* $\max_{1 \leqslant k \leqslant k_n} \mathbb{E}[X_{nk}^2] \xrightarrow[n\to\infty]{} 0$

B. The Lindeberg condition holds:

$$\forall \varepsilon > 0, \quad L_n(\varepsilon) := \sum_{j=1}^{k_n} \int_{|x|>\varepsilon} x^2 \, dF_{nj}(x) \xrightarrow[n\to\infty]{} 0,$$

where F_{nj} *is the cdf of the random variable* X_{nj}.

Before we state the theorem about the asymptotic normality of the bootstrap estimator, we state another very useful Proposition that helps us in proving things in a multidimensional framework by simply examining 1-dimensional projections. This technique is generally referred to as the *Cramér-Wold device*. A short proof of this result can be obtained with the use of characteristic functions (Billingsley [4], page 383).

Proposition 3.1.1 (Cramér-Wold). *Let* $X_n = (X_{n1}, \ldots, X_{nk})$, $n \in \mathbb{N}$, *be a sequence of random vectors in* $\mathbb{R}^k$, *and* $X = (X_1, \ldots, X_k) \in \mathbb{R}^k$. *Then,* $X_n \xrightarrow{d} X$ *if, and only if,* $a_1 X_{n1} + \cdots + a_k X_{nk} \xrightarrow{d} a_1 X_1 + \cdots + a_k X_k$ *for all* $a_1, \ldots, a_k \in \mathbb{R}$.

Now we are ready to state the main theorem regarding the asymptotic behavior of the frequentist bootstrap procedure that we use in order to tackle the initial problem of interest. The proof of this result is highly technical and is presented in detail in Appendix A.

Theorem 3.1.3. *Let* $(X_n)_{n \geqslant 0}$ *be an ergodic, time-homogeneous Markov chain with transition matrix* P *and finite state-space* $S = \{1, \ldots, s\}$. *Let* $\hat{P}_n$ *be its maximum likelihood estimator based on an observed path* $\mathbf{x} = (x_0, x_1, \ldots, x_n)$. *For* $\tilde{P}_n$ *as described above and matrix* Σ *as in (2.2.2), we have that*

$$\sqrt{N_n}\,(\tilde{P}_n - \hat{P}_n) \xrightarrow{d} \mathscr{N}_{s^2}(0, \Sigma), \tag{3.1.3}$$

as $n \to \infty$ *and* $N_n \to \infty$.

3.2 Bayesian Bootstrap

A simple way of illustrating the general idea of Bayesian bootstrap is the following: suppose the statistic of interest is the sample mean.

- The frequentist bootstrap procedure yields the bootstrap value $\sum_{i=1}^{n} w_i x_i$ at each step, where $w_i \sim$Multinomial$(1, 1/n, \ldots, 1/n), i = 1, \ldots, n$.

- Problem: the discrete nature of the Multinomial distribution leads to non-smooth estimators.

The Bayesian Bootstrap solves the above problem by using a continuous version of the weights' distribution, the Dirichlet distribution. We have the following:

(i) Take a sample $\{x_1, \ldots, x_n\}$ and calculate the sample mean.

(ii) Generate the weights $w_i \sim \text{Dir}(1, \ldots, 1)$, $i = 1, \ldots, n$.

(iii) Take the bootstrap value $\sum_{i=1}^{n} w_i x_i$.

(iv) Iterate the previous steps and take the mean of all the bootstrap values.

The above procedure leads to smoother estimators. Note that now the weights are random variables.

3.2.1 Bayesian Bootstrap for Markov Chains

Let B_{ij} denote the set $B_{ij} = \left\{ \sum_{\ell=1}^{j} n_{i(\ell-1)} + 1, \ldots, \sum_{\ell=1}^{j} n_{i\ell} \right\}$, $i, j \in S$, where we define $n_{i0} = 0$ for all $i \in S$. Note that each B_{ij} can be written as

$$B_{ij} = \{n_{i1} + \cdots + n_{i(j-1)} + 1, \ldots, n_{i1} + \cdots + n_{i(j-1)} + n_{ij}\},$$

from which we conclude that each B_{ij} has n_{ij} elements. Hence, the maximum likelihood estimator (2.2.1) can be written as

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i} = \frac{\sum_{t \in B_{ij}} 1}{\sum_{t=1}^{n_i} 1}, \tag{3.2.1}$$

which works as a useful connection between the asymptotic and the Bayesian bootstrap approach. We will use this connection to define the Bayesian bootstrap estimators for the transition probabilities.

Example 3.2.1. We will give an example in order to understand the notation better. Let $N = 20$ and suppose we have a Markov chain path such that $n_{11} = 2, n_{12} = 2, n_{13} = 2, n_{21} = 1, n_{22} = 1, n_{23} = 4, n_{31} = 2, n_{32} = 4, n_{33} = 1$. Then we have that $n_1 = n_{11} + n_{12} + n_{13} = 6$, $n_2 = n_{21} + n_{22} + n_{23} = 6$ and $n_3 = n_{31} + n_{32} + n_{33} = 7$. The sets B_{ij} are defined as:

$$B_{11} = \{n_{10} + 1, \ldots, n_{11}\} = \{1, 2\} \rightsquigarrow |B_{11}| = n_{11} = 2$$

$$B_{12} = \{n_{10} + n_{11} + 1, \ldots, n_{11} + n_{12}\} = \{3, 4\} \rightsquigarrow |B_{12}| = n_{12} = 2$$

$$\vdots$$

$$B_{32} = \{n_{30} + n_{31} + 1, \ldots, n_{31} + n_{32}\} = \{3, 4, 5, 6\} \rightsquigarrow |B_{32}| = n_{32} = 4$$

$$B_{33} = \{n_{30} + n_{31} + n_{32} + 1, \ldots, n_{31} + n_{32} + n_{33}\} = \{7\} \rightsquigarrow |B_{33}| = n_{33} = 1$$

Notice that the sets $B_{i1}, B_{i2}, \ldots, B_{im}$ *partition* the set $\{1, 2, \ldots, n_i\}$, for all $i \in [m]$. □

We will use an estimator similar to the M.L.E. (3.2.1), but now we are going to include simulation.

Definition 3.2.1 (Bayesian Bootstrap Estimator). For every $i, j \in S = [m]$, the Bayesian bootstrap estimator of p_{ij} is defined to be

$$\hat{p}_n^*(i,j) := \frac{\sum_{t \in B_{ij}} Z_{it}}{\sum_{t=1}^{n_i} Z_{it}}, \tag{3.2.2}$$

where $Z_{it} \sim \text{Exp}(1)$ are iid random variables, for $i = 1, \ldots, m$, $t = 1, \ldots, n_i$.

Algorithm 1 Bayesian Bootstrap for Finite Markov Chains

1: Simulate iid $Z_{it} \sim \text{Exp}(1)$, $i = 1, \ldots, m$, $t = 1, \ldots, n_i$.

2: Calculate $\hat{p}_n^*(i,j) := \dfrac{\sum_{t \in B_{ij}} Z_{it}}{\sum_{t=1}^{n_i} Z_{it}}$ and obtain the matrix estimator $\hat{P}^*$.

3: Repeat the previous steps B times to obtain the estimators $\hat{P}_{n1}^*, \ldots, \hat{P}_{nB}^*$ and approximate the posterior distribution $\pi(P \mid \mathbf{x})$ using them.

Let $P = (P_{1\cdot}\ P_{2\cdot} \cdots P_{m\cdot})^T$ denote the unknown transition matrix P and $\hat{P}_n^* = (\hat{p}_n^*(i,j))$ the Bayesian bootstrap estimator. It can be shown that the joint distribution of each row of $\hat{P}_n^*$ is Dirichlet with parameters involving the quantities n_{ij}. In particular,

$$\hat{P}_{i\cdot}^* \sim \text{Dir}(n_{i1}, \ldots, n_{im}).$$

Thus the joint distribution of the Bayesian bootstrap estimator $\hat{P}_n^*$ is a matrix-beta distribution (product of independent Dirichlet distributions).

We now state the Central Limit Theorem for the Bayesian bootstrap procedure that we described. The proof is presented in detail in Appendix A.

Theorem 3.2.1 (CLT for the Bayesian Bootstrap). *Let $i, j \in S$ and $\hat{p}_{ij}$ denote the maximum likelihood estimator of the transition probability p_{ij}. Then, for almost all sample sequences $\mathbf{x} = (x_0, x_1, \ldots, x_n)$ we have*

$$\frac{\sqrt{n_i}(\hat{p}_n^*(i,j) - \hat{p}_{ij})}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})}} \,\Big|\, \mathbf{x} \xrightarrow{d} N(0,1), \tag{3.2.3}$$

as $n \to \infty$. That is, the Bayesian bootstrap estimators are asymptotically normal.

Chapter 4

Applications

We wish to illustrate the validity of Theorem 3.1.3 empirically, with the use of simulated and real data. For simplicity, we do it for a Markov chain with a state-space that only has two elements. The functions used to simulate and estimate a discrete-time Markov chain, along with the bootstrapping procedure that is followed, are presented below.

A. Define a function that has the initial state, the transition probability matrix and the length n as inputs, and outputs n observations simulated from this Markov chain. The pseudocode for this function is shown in Algorithm 2.

Algorithm 2 Simulate Markov Chain path

```
function SIMULATE_DTMC(init_state, P, n)
    states ← {1, 2}
    Initialize: path ← rep(0, n)
    path[1] ← init_state
    for i from 2 to n do
        path[i] ← sample(states, 1, P[path[i − 1]])
    end for
    return path
end function
```

B. Define a function that takes as input the path of a Markov chain and outputs the MLE of its transition probability matrix. The pseudocode for this function is shown in Algorithm 3.

C. Define a function that takes as input a transition matrix (in our case, the estimated transition matrix), the original path length n, the bootstrap path length N and the number of bootstrap iterations B, and outputs a vector of the bootstrap results, which are the values of the desired CLT type of quantity stated in Theorem 3.1.3. The pseudocode for this function is shown in Algorithm 4.

Algorithm 3 Estimate Transition Probability Matrix

function ESTIMATE_MATRIX($path$)
    Initialize: $est\_P \leftarrow$ matrix$(c(0,0,0,0),2,2)$
    Initialize: $P\_hat \leftarrow$ matrix$(c(0,0,0,0),2,2)$
    for i from 1 to (length$(path)-1$) **do**
        $prev\_state \leftarrow path[i]$
        $next\_state \leftarrow path[i+1]$
        $est\_P[prev\_state, next\_state] \leftarrow est\_P[prev\_state, next\_state] + 1$
    end for
    for each row i in est_P **do**
        $s \leftarrow$ sum(elements in row i)
        for each element j in row i **do**
            $P\_hat[i,j] \leftarrow \frac{est\_P[i,j]}{s}$
        end for
    end for
    return P_hat
end function

Algorithm 4 Bootstrap Transition Probability Matrix

function BOOTSTRAP(P_hat, N, B)
    Initialize: $clt \leftarrow rep(0, B)$
    for i in 1 to B **do**
        $init\_state \leftarrow$ sample($\{1,2\}$, 1, prob=$\{0.5, 0.5\}$)
        $boot\_path \leftarrow$ SIMULATE_DTMC($init\_state$, P_hat, N)
        $P\_tilde \leftarrow$ ESTIMATE_MATRIX($boot\_path$)
        $clt[i] \leftarrow \sqrt{N} * (P\_tilde[1,1] - P\_hat[1,1])$
    end for
    return clt
end function

4.1 Simulated Data

In this section, we use 1000 data points simulated by a Markov chain $\{X_n\}_{n\geq 0}$ with state-space $S = \{1, 2\}$, initial state $X_0 = 1$ and transition probability matrix

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

We run 1000 bootstrap replications and are interested in the quantity $\sqrt{N_n}(\tilde{P}_n - \hat{P}_n)$ of the main theorem 3.1.3. For simplicity, we study $\sqrt{N_n}(\tilde{P}_n[1,1] - \hat{P}_n[1,1])$, the one-dimensional counterpart of the quantity of interest, and plot its histogram along with the normal density predicted by Theorem 3.1.3 in Figure 4.1. The histogram shows a strong concentration to the normal distribution, as the main theorem suggests.

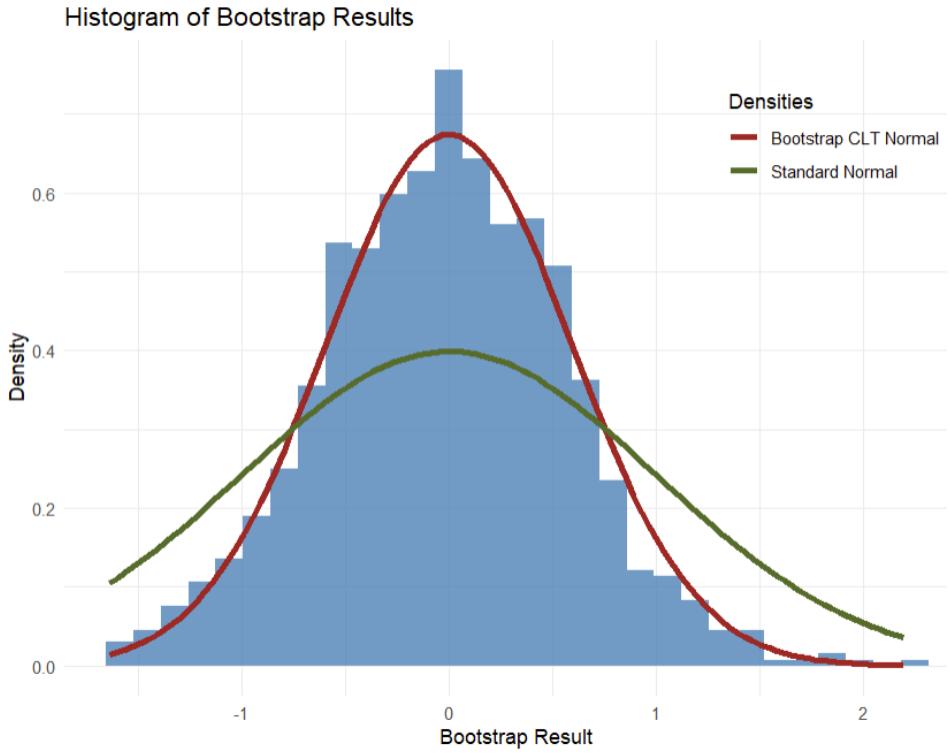

FIGURE 4.1: Histogram of $\sqrt{N_n}(\tilde{P}_n[1,1] - \hat{P}_n[1,1])$.

4.2 Real Data

We now apply the bootstrap method we introduced in Chapter 3 to a real dataset embedded in R. In particular, we use the `weather` dataset, a data frame containing information on the weather in Canberra, Australia, in the span of one year. For the full documentation of the `rattle` package and the `weather` dataset included in it, the reader is referred to rattle.data.

Even though the data frame has dimension 366×24, we will only look at the trajectory of

one variable, namely the `RainTomorrow` variable. This variable can theoretically take three values: 'Yes', 'No', 'NA', although in our variable of interest there were no 'NA' values. We convert 'No' to 1 and 'Yes' to 2. We assume that the modified `RainTomorrow` vector constitutes a path of 366 observations from a discrete-time Markov chain with state-space $S = \{1, 2\}$. To empirically verify the validity of this assumption, we look at the ACF plot of our time-series data, as shown in Figure 4.2. We observe only one statistically significant spike at lag 1, which shows that fitting a first order discrete-time Markov chain on the data is a plausible assumption.

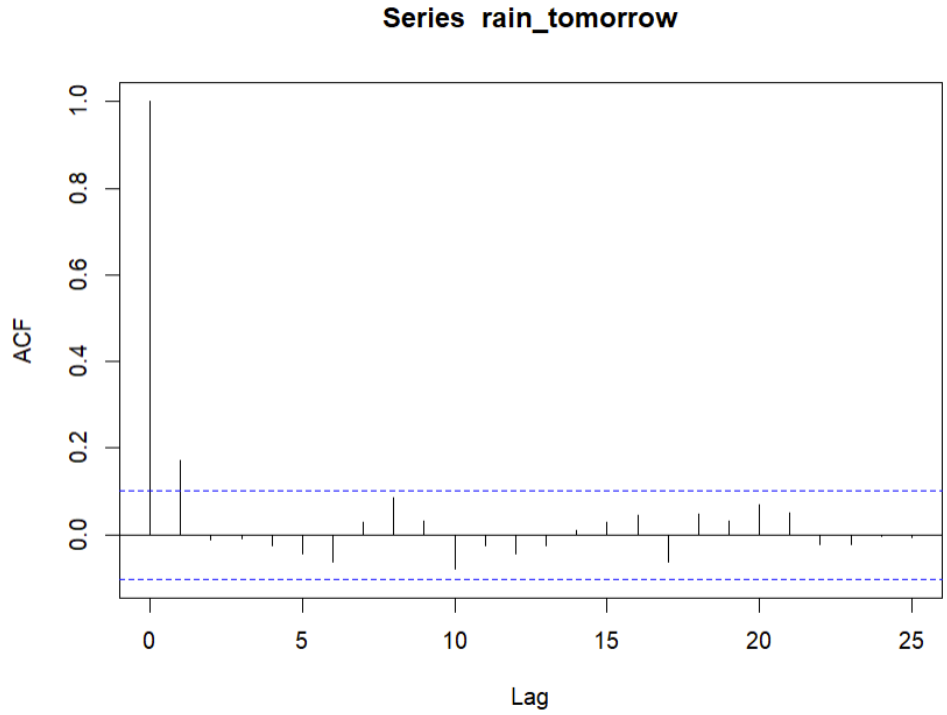

FIGURE 4.2: ACF plot of `RainTomorrow` data.

To give stronger evidence for the plausibility of fitting a first-order Markov chain to our data, we compare it to fitting a second-order Markov chain, in which the dynamics are governed by the following property:

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0) \\ = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}).\end{aligned}$$

In our case, the possible states are four: $(1,1), (1,2), (2,1), (2,2)$. Note that there are fewer than 16 parameters to estimate, since some transitions, e.g. $(1,1) \to (2,2)$, are not possible. In particular, there are only 4 free parameters to be estimated, since at each state there are only two possible transitions (e.g. from the state $(1,1)$ the chain can only visit either $(1,1)$ or $(1,2)$) and the matrix is stochastic, so we lose one degree of freedom for each row. Based on the same observed path we used for the first-order Markov chain, we estimate the non-

zero transition probabilities of the fitted second-order Markov chain. Now that we have the MLEs of both models, we plug them into their log-likelihoods and compute the AIC and BIC for both models. The results are shown in table 4.1. Based on these results, we conclude that among the two candidate models, the first-order Markov chain is the preferable one.

	1st-order DTMC	2nd-order DTMC
AIC	336.38	337.25
BIC	344.18	352.86

TABLE 4.1: AIC and BIC for the two models

The goal is to use the parametric bootstrap algorithm described previously in order to estimate the MLE of the probability transition matrix of the DTMC, i.e., we want the MLE of the probability that tomorrow it will either rain or not, given that today it rained or not. The summary statistics of the estimated transition probabilities based on the bootstrap method are given in the table 4.2, and the histogram of the quantity used in the main theorem is given in Figure 4.3.

	Bootstrap Mean	Bootstrap SD	95% CI
$\tilde{p}_{11}$	0.8523	0.0198	$(0.8114, 0.89)$
$\tilde{p}_{12}$	0.1477	0.0198	$(0.11, 0.1886)$
$\tilde{p}_{21}$	0.6893	0.0583	$(0.5789, 0.8039)$
$\tilde{p}_{22}$	0.3107	0.0583	$(0.1961, 0.4211)$

TABLE 4.2: Summary statistics for the entries of the 2×2 bootstrap transition probability matrix $\tilde{P}$

As a quick sanity check, we notice that $\tilde{p}_{11} + \tilde{p}_{12} = 1$ and $\tilde{p}_{21} + \tilde{p}_{22} = 1$, as expected since $\tilde{P}$ is a stochastic matrix. To see how robust the bootstrap mean estimator is, we plot the trajectory of the bootstrap mean of $\tilde{p}_{11}$ across the 1000 bootstrap replicates. The result is shown in Figure 4.4. We observe that after a very short burn-in period, the estimate of $\tilde{p}_{11}$ reaches the mean value of 0.85 in the first few bootstrap replications and fluctuates very little until it reaches stationarity.

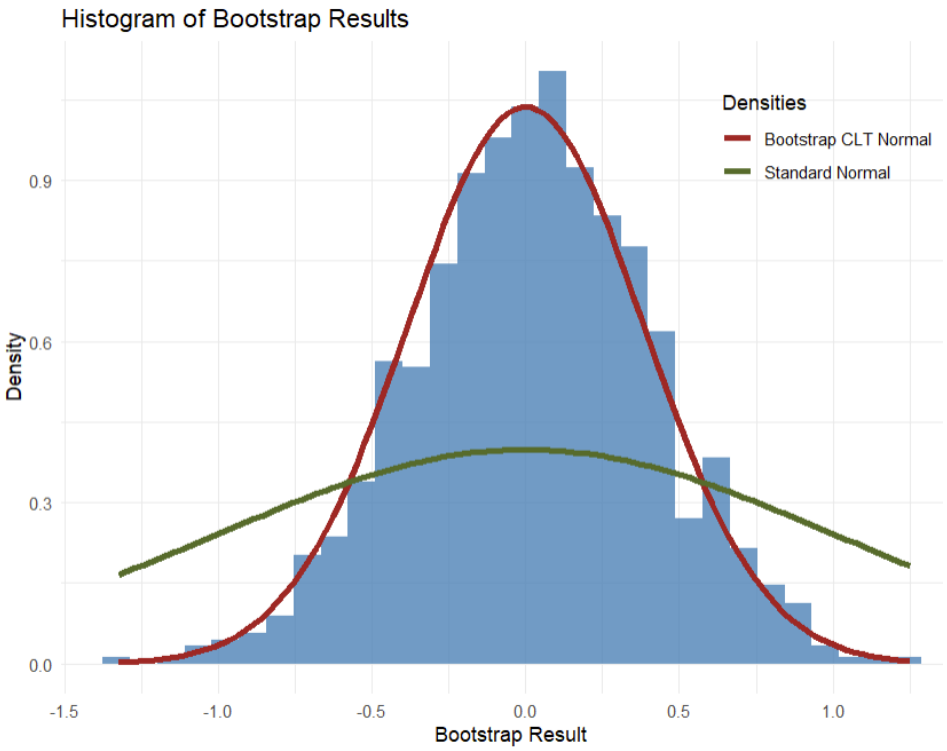

FIGURE 4.3: Histogram of $\sqrt{N_n}(\tilde{P}_n[1,1] - \hat{P}_n[1,1])$ based on the `weather` data

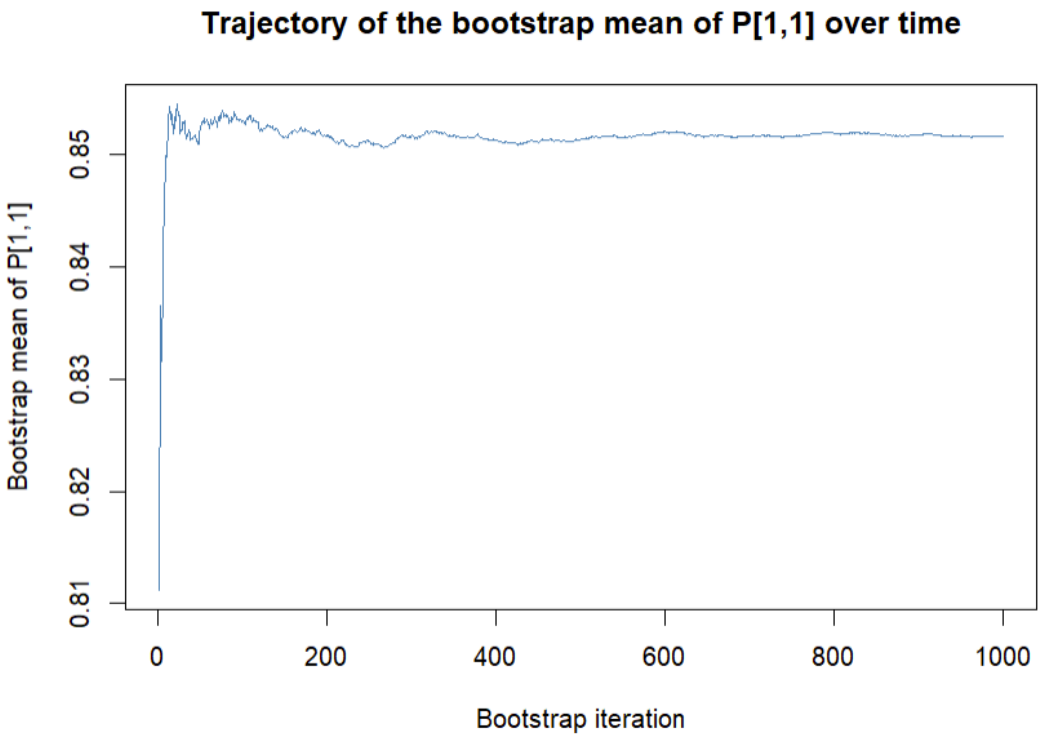

FIGURE 4.4: Trajectory of the bootstrap mean of $\tilde{p}_{11}$

Appendix A

Technical Proofs

In order to prove the pointwise ergodic theorem, we will first prove a useful lemma, often referred to as the *Maximal Ergodic Lemma.*

Lemma A.0.1. *Suppose Z is a random variable satisfying $\mathbb{E}|Z| < \infty$ and $\mathbb{E}[Z \mid \mathscr{G}] > 0$ a.s. Then,*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Z \circ T^k \geq 0 \quad \textit{a.s.}, \tag{A.0.1}$$

where T is a measure preserving transformation.

Proof. For notational convenience, denote $S_n := \sum_{k=0}^{n-1} Z \circ T^k$, $n \geq 1$. Note that

$$\mathbb{E}|S_n| \leq \sum_{k=0}^{n-1} \mathbb{E}|Z \circ T^k| \leq \sum_{k=0}^{n-1} \mathbb{E}|Z| = n\,\mathbb{E}|Z|.$$

Define the quantities

$$L_n = \inf_{1\leq k\leq n} S_k \quad \text{and} \quad A = \Big\{\omega \in \Omega : \inf_{n\geq 1} L_n(\omega) = -\infty\Big\}.$$

Observe that the statement of the lemma will follow immediately if we show that $\mathbb{P}(A) = 0$. For that, we note the following. L_n is a random variable for $n \geq 1$, and A is a measurable set. We have $|Z| < \infty$ a.s., since $\mathbb{E}|Z| < \infty$ by assumption, and

$$\Big\{ \inf_{n\geq 1} S_n = -\infty \Big\} = \Big\{ \inf_{n\geq 1} S_n \circ T = -\infty \Big\} \quad \text{a.s.},$$

so A is T-invariant, i.e., $A = T^{-1}(A)$. Indeed,

$$T^{-1}(A) = \{\omega \in \Omega : T(\omega) \in A\} = \{\omega \in \Omega : \inf L_n(T(\omega)) = -\infty\},$$

L_n is given through the S_k's, and

$$\{\inf S_n = -\infty\} = \{\inf S_n \circ T = -\infty\} \quad \text{a.s.}$$

Note also that the sequence $\{L_n\}_{n=1}^{\infty}$ is decreasing, so we have

$$\begin{aligned} L_n &= \inf_{1\le k\le n} S_k = Z + \inf_{1\le k\le n}\{S_k - Z\} \\ &= Z + \inf_{1\le k\le n}\left\{\sum_{m=0}^{k-1} Z\circ T^m - Z\right\} \\ &= Z + \min\left\{0, \inf_{1\le k\le n-1}\sum_{m=0}^{k-1} Z\circ T^{m+1}\right\} \\ &= Z + \min\{0, L_{n-1}\circ T\} \\ &\ge Z + \min\{0, L_n\circ T\}. \end{aligned}$$

We thus get $Z \le L_n - \min\{0, L_n\circ T\} = L_n + (L_n\circ T)^- = L_n + L_n^-\circ T$ a.s., where $(x)^+ := \max\{x, 0\}$ and $(x)^- := \max\{-x, 0\}$ denote the positive and the negative part of a quantity x, respectively. Recall that A is invariant, so $1_A = 1_A\circ T$, thus we get the useful inequality

$$\begin{aligned} \mathbb{E}[1_A Z] &\le \mathbb{E}[1_A\cdot(L_n + L_n^-\circ T)] = \mathbb{E}[1_A L_n] + \mathbb{E}[1_A L_n^-\circ T] \\ &= \mathbb{E}[1_A L_n] + \mathbb{E}[1_A\circ T L_n^-\circ T] \le \mathbb{E}[1_A L_n] + \mathbb{E}[1_A L_n^-] \\ &= \mathbb{E}[1_A L_n + 1_A L_n^-] = \mathbb{E}[1_A(L_n + L_n^-)] = \mathbb{E}[1_A L_n^+]. \end{aligned}$$

We want to show that $\lim_{n\to\infty}\mathbb{E}[1_A L_n^+] = \mathbb{E}\left[\lim_{n\to\infty} 1_A L_n^+\right] = 0$, to get $\mathbb{E}[1_A Z] = 0$ from the preceding inequality. This follows from the Dominating Convergence Theorem, upon noting that $1_A L_n^+ \overset{\text{a.s.}}{\to} 0$, $L_n^+ \le Z^+$ and $\mathbb{E}Z^+ \le \mathbb{E}|Z| < \infty$. Thus, $\mathbb{E}[1_A Z] = 0$ and we get

$$\mathbb{E}[1_A\mathbb{E}[Z\mid\mathscr{G}]] = \mathbb{E}[1_A Z] = 0,$$

where the first equality comes from the definition of conditional expectation. Now recall that $\mathbb{E}[Z\mid\mathscr{G}] > 0$ a.s. by assumption, so the previous equality yields $\mathbb{P}(A) = 0$ or, equivalently,

$$\mathbb{P}\left(\inf_{n\ge1} L_n = -\infty\right) = \mathbb{P}\left(\inf_{n\ge1}\inf_{1\le k\le n} S_k = -\infty\right) = 0,$$

showing that $\liminf_{n\to\infty}\frac{S_n}{n} \ge 0$ a.s., which is what we wanted to prove. □

We are now ready to prove Birkhoff's Ergodic Theorem.

Proof of Theorem 1.2.3. Let $\varepsilon > 0$. The idea is to utilize the Maximal Ergodic Lemma for a suitably chosen random variable Z. For that, define $Z = Y - \mathbb{E}[Y \mid \mathscr{G}] + \varepsilon$. Note that

$$\mathbb{E}|Z| = \mathbb{E}|Y - \mathbb{E}[Y \mid \mathscr{G}] + \varepsilon| \leq \mathbb{E}|Y| + \mathbb{E}[\mathbb{E}[|Y|| \mathscr{G}]] + \varepsilon = \mathbb{E}|Y| + \mathbb{E}|Y| + \varepsilon = 2\mathbb{E}|Y| + \varepsilon$$

and

$$\mathbb{E}[Z \mid \mathscr{G}] = \mathbb{E}[Y \mid \mathscr{G}] - \mathbb{E}[\mathbb{E}[Y \mid \mathscr{G}]] + \varepsilon = \mathbb{E}[Y \mid \mathscr{G}] - \mathbb{E}[Y \mid \mathscr{G}] + \varepsilon = \varepsilon > 0.$$

Since $\mathbb{E}[Y \mid \mathscr{G}]$ is $\mathscr{G}$-measurable, it is also T-invariant. Thus, from Lemma (A.0.1) we get

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Z \circ T^k \geq 0 \quad \text{a.s.},$$

which yields

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \geq \mathbb{E}[Y \mid \mathscr{G}] - \varepsilon \quad \text{a.s.}$$

Using $-Y$ instead of Y in the last inequality, we get

$$-\limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \geq -\mathbb{E}[Y \mid \mathscr{G}] - \varepsilon,$$

which yields

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \leq \mathbb{E}[Y \mid \mathscr{G}] + \varepsilon \quad \text{a.s.}$$

Combining these two inequalities, we get

$$-\varepsilon + \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \leq \mathbb{E}[Y \mid \mathscr{G}] \leq \liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k + \varepsilon \quad \text{a.s.}$$

Since $\varepsilon > 0$ was arbitrary, this proves that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \leq \mathbb{E}[Y \mid \mathscr{G}] \leq \liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \quad \text{a.s.}$$

Since it is always true that $\liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k$, the above yields

$$\frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k \overset{\text{a.s.}}{\to} \mathbb{E}[Y \mid \mathscr{G}].$$

□

Theorem A.0.1. *Let $(X_n)_{n\geqslant 0}$ be an ergodic, time-homogeneous Markov chain with transition matrix P and finite state-space $S = \{1, \dots, s\}$. Let $\hat{P}_n$ be its maximum likelihood estimator based on an observed path $\mathbf{X} = (x_0, x_1, \dots, x_n)$. For $\tilde{P}_n$ as described above and matrix Σ as in (2.2.2), we have that*

$$\sqrt{N_n}\,(\tilde{P}_n - \hat{P}_n) \xrightarrow{d} \mathscr{N}_{s^2}(0, \Sigma), \tag{A.0.2}$$

as $n \to \infty$ and $N_n \to \infty$.

Proof. Let $\{X_{nr} : 1 \leqslant r \leqslant N_n\}$ denote the bootstrap sample that occurs as a Markov path generated by the transition matrix $\hat{P}_n$. We define the quantities

$$m_i^{(n)} := \sum_{r=1}^{N_n} 1(X_{nr} = i) \quad \text{and} \quad m_{ij}^{(n)} := \sum_{r=0}^{N_n - 1} 1(X_{nr} = i, X_{n,r+1} = j).$$

Then, the desired matrix statistic takes the form

$$\begin{aligned}\sqrt{N_n}\,(\tilde{P}_n - \hat{P}_n) &= \left(\sqrt{N_n} \cdot \left(\frac{m_{ij}^{(n)}}{m_i^{(n)}} - \frac{n_{ij}}{n_i}\right)\right)_{i,j=1,\dots,s} \\ &= \left(\sqrt{N_n} \cdot \left(\frac{m_{ij}^{(n)}}{m_i^{(n)}} - p_{nij}\right)\right)_{i,j=1,\dots,s},\end{aligned}$$

where $p_{nij} = (\hat{P}_n)_{ij}$ is the (i,j)-element of the MLE $\hat{P}_n$. By the Cramér-Wold device 3.1.1, it suffices to prove that for all $\ell_{ij} \in \mathbb{R}$ the sequence

$$\sum_{i=1}^{k}\sum_{j=1}^{k} \ell_{ij}\sqrt{N_n}\left(\frac{m_{ij}^{(n)}}{m_i^{(n)}} - p_{nij}\right)$$

converges in distribution to a Normal distribution. Since a linear combination of Normal distributions is a Normal distribution, it suffices to prove that the sequence

$$\sqrt{N_n}\left(\frac{m_{ij}^{(n)}}{m_i^{(n)}} - p_{nij}\right)$$

converges in distribution to a Normal distribution. The rest of the proof deals with this goal.

Let $\{W_{it}^{(n)}\}$, $i = 1, \dots, s$, $t \in \mathbb{N}$, be a sequence of independent random variables such that $\mathbb{P}(W_{it}^{(n)} = j) = p_{nij}$, for all $j = 1, \dots, s$, and recursively define

$$X_0^{(n)} = 1 \quad \text{and} \quad X_{i+1}^{(n)} = W^{(n)}_{X_i^{(n)} m}, i \geqslant 0,$$

where $m = 1 + \#\{\ell : 1 \leqslant \ell \leqslant N_n, X_\ell^{(n)} = X_i^{(n)}\}$. This is a way to describe how the bootstrap sample $\{X_{nr} : 1 \leqslant r \leqslant N_n\}$ is generated.

Fix $i, j \in S$ and define the random variables

$$C_t^{(n)}(j) = \begin{cases} 1, & \text{if } W_{it}^{(n)} = j,\ t = 1, \dots, m_i^{(n)} \\ 0, & \text{else} \end{cases}.$$

By definition, we have that $C_t^{(n)}(j) \sim \text{Bernoulli}(p_{nij})$, since

$$\mathbb{P}(C_t^{(n)}(j) = 1) = \mathbb{P}(W_{it}^{(n)} = j) = p_{nij}.$$

Let $\hat{\pi}_n = (\pi_{n1}, \dots, \pi_{ns})$ be the limit distribution of $\hat{P}_n$. We denote by $\lfloor x \rfloor$ the largest integer which is smaller than x. For a fixed $i \in S$ with $\pi_{ni} > 0$, we define

$$d_t^{(n)}(j) := \frac{C_t^{(n)}(j) - p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1 - p_{nij})}}, \quad n \in \mathbb{N},\ t = 1, \dots, \lfloor N_n \pi_{ni} \rfloor.$$

Using the fact that $C_t^{(n)}(j) \sim \text{Bernoulli}(p_{nij})$, we get that

$$\mathbb{E}\left[d_t^{(n)}(j)\right] = \frac{\mathbb{E}\left[C_t^{(n)}(j)\right] - p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1 - p_{nij})}} = \frac{p_{nij} - p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1 - p_{nij})}} = 0$$

and

$$\text{Var}(d_t^{(n)}(j)) = \frac{\text{Var}\left(C_t^{(n)}(j)\right)}{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1 - p_{nij})} = \frac{p_{nij}(1 - p_{nij})}{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1 - p_{nij})} = \frac{1}{\lfloor N_n \pi_{ni} \rfloor}.$$

First, we will show that

$$S_{\lfloor N_n \pi_{ni} \rfloor}^{(n)}(j) := \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} d_t^{(n)}(j) = \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \frac{C_t^{(n)}(j) - p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1 - p_{nij})}} \xrightarrow{d} N(0, 1).$$

For that, we will use the Lindeberg-Feller Central Limit Theorem. Notice that

$$\text{Var}\left(d_t^{(n)}(j)\right) = \frac{1}{\lfloor N_n \pi_{ni} \rfloor} \Rightarrow s_n^2 = \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \text{Var}\left(d_t^{(n)}(j)\right) = \lfloor N_n \pi_{ni} \rfloor \cdot \frac{1}{\lfloor N_n \pi_{ni} \rfloor} = 1,$$

hence we will use the 2nd version of Lindeberg-Feller (Theorem 3.1.2). We thus have to check the Lindeberg condition

$$\forall \varepsilon > 0 \quad L_n(\varepsilon) := \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \int_{|x| > \varepsilon} x^2 \, dF_{nt}(x) \xrightarrow[n \to \infty]{} 0,$$

where F_{nt} is the cdf of the random variable $d_t^{(n)}(j)$. We have the following bound:

$$\left| d_t^{(n)}(j) \right| = \frac{\left| C_t^{(n)}(j) - p_{nij} \right|}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}} \leqslant \frac{\overbrace{\left| C_t^{(n)}(j) \right|}^{\leqslant 1} + \overbrace{\left| p_{nij} \right|}^{\leqslant 1}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}}$$
$$\leqslant \frac{2}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}}.$$

Since $\hat{P}_n \xrightarrow{a.s.} P$ and $\hat{\pi}_n \xrightarrow{a.s.} \pi$, where $\pi = \pi P$, we have that

$$\pi_{nj} p_{nij}(1-p_{nij}) \xrightarrow{a.s.} \pi_j p_{ij}(1-p_{ij}) > 0.$$

Let $\varepsilon > 0$. Since $N_n \xrightarrow[n\to\infty]{} \infty$ and N_n appears in the denominator of $d_t^{(n)}(j)$, there exists an $n_0 \in \mathbb{N}$ such that for every $n > n_0$ we have $\sup_t d_t^{(n)}(j) < \varepsilon$. Hence,

$$\int_{|x|>\varepsilon} x^2 \, dF_{nt}(x) = 0, \quad \forall n > n_0, \ t = 1, \ldots, \lfloor N_n \pi_{ni} \rfloor,$$

which yields

$$L_n(\varepsilon) := \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \int_{|x|>\varepsilon} x^2 \, dF_{nt}(x) = 0, \quad \forall n > n_0$$

and thus Lindeberg's condition is satisfied. From the Lindeberg-Feller CLT, we obtain

$$S^{(n)}_{\lfloor N_n \pi_{ni} \rfloor}(j) := \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} d_t^{(n)}(j) = \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \frac{C_t^{(n)}(j) - p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}} \xrightarrow{d} N(0,1). \quad \text{(A.0.3)}$$

We now observe that

$$p_{nj} \xrightarrow{a.s.} p_{ij} \Rightarrow p_{nj} \xrightarrow{p} p_{ij} \Rightarrow \frac{1}{\sqrt{p_{nij}(1-p_{nij})}} \xrightarrow{p} \frac{1}{\sqrt{p_{ij}(1-p_{ij})}}$$

and

$$\frac{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{ij}(1-p_{ij})}} \xrightarrow{p} 1.$$

Using (A.0.3) and Slutsky's Lemma, we obtain

$$\frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \left(C_t^{(n)}(j) - p_{nij} \right)}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{ij}(1-p_{ij})}} = \frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \left(C_t^{(n)}(j) - p_{nij} \right)}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}} \cdot \frac{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{nij}(1-p_{nij})}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{ij}(1-p_{ij})}} \xrightarrow{d} N(0,1).$$

Now we define the quantity

$$\dot{m}_{ij}^{(n)} := \sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} 1(X_{t-1} = i, X_t = j)$$

and write

$$\begin{aligned}\frac{m_{ij}^{(n)} - m_i^{(n)} p_{nij}}{\sqrt{m_i^{(n)}}} &= \frac{m_{ij}^{(n)} - \dot{m}_{ij}^{(n)} + \dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij} + \lfloor N_n \pi_{ni} \rfloor p_{nij} - m_i^{(n)} p_{nij}}{\sqrt{m_i^{(n)}}} \\ &= \frac{\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{m_i^{(n)}}} + \frac{\left(m_{ij}^{(n)} - m_i^{(n)} p_{nij}\right) - \left(\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}\right)}{\sqrt{m_i^{(n)}}} \\ &= \frac{\sqrt{\lfloor N_n \pi_{ni} \rfloor}}{\sqrt{m_i^{(n)}}} \cdot \left(\frac{\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}} \right) + \\ &\quad + \frac{\sqrt{N_n}}{\sqrt{m_i^{(n)}}} \cdot \frac{\left(m_{ij}^{(n)} - m_i^{(n)} p_{nij}\right) - \left(\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}\right)}{\sqrt{N_n}}. \end{aligned} \tag{A.0.4}$$

Recall that

$$\frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \left(C_t^{(n)}(j) - p_{nij} \right)}{\sqrt{\lfloor N_n \pi_{ni} \rfloor p_{ij}(1 - p_{ij})}} \xrightarrow{d} N(0,1),$$

from which we get

$$\frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \left(C_t^{(n)}(j) - p_{nij} \right)}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}} \xrightarrow{d} N(0, p_{ij}(1 - p_{ij}))$$

and we observe that

$$\frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \left(C_t^{(n)}(j) - p_{nij} \right)}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}} = \frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} C_t^{(n)}(j) - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}}.$$

From the definition of $\dot{m}_{ij}^{(n)}$ we see that $\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} C_t^{(n)}(j) \overset{d}{=} \dot{m}_{ij}^{(n)}$, i.e., they have the same distribution. From this, we get that

$$\frac{\sum_{t=1}^{\lfloor N_n \pi_{ni} \rfloor} \left(C_t^{(n)}(j) - p_{nij} \right)}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}} \overset{d}{=} \frac{\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}},$$

which yields

$$\frac{\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}} \xrightarrow{d} N(0, p_{ij}(1 - p_{ij})). \tag{A.0.5}$$

Our goal now is to show the following two convergences:

(i)

$$\frac{m_i^{(n)}}{N_n} - \pi_{ni} \xrightarrow{p} 0$$

(ii)

$$\eta_{nij} := \frac{m_{ij}^{(n)} - m_i^{(n)} p_{nij}}{\sqrt{N_n}} - \frac{\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{N_n}} \xrightarrow{p} 0.$$

If we show these two, then from A.0.4 and A.0.5, the proof of the theorem is complete. Indeed,

$$\frac{\dot{m}_{ij}^{(n)} - \lfloor N_n \pi_{ni} \rfloor p_{nij}}{\sqrt{\lfloor N_n \pi_{ni} \rfloor}} \xrightarrow{d} N(0, p_{ij}(1 - p_{ij}))$$

and

$$\frac{m_i^{(n)}}{N_n} - \pi_{ni} \xrightarrow{p} 0 \overset{\div \pi_{ni} > 0}{\Rightarrow} \frac{m_i^{(n)}}{N_n \pi_{ni}} - 1 \xrightarrow{p} 0 \Rightarrow \frac{m_i^{(n)}}{N_n \pi_{ni}} \xrightarrow{p} 1 \Rightarrow \frac{N_n \pi_{ni}}{m_i^{(n)}} \xrightarrow{p} 1 \Rightarrow$$

$$\frac{\lfloor N_n \pi_{ni} \rfloor}{m_i^{(n)}} = \underbrace{\frac{N_n \pi_{ni}}{m_i^{(n)}}}_{\xrightarrow{p} 1} \cdot \underbrace{\frac{\lfloor N_n \pi_{ni} \rfloor}{N_n \pi_{ni}}}_{\xrightarrow{p} 1} \xrightarrow{p} 1 \Rightarrow \sqrt{\frac{\lfloor N_n \pi_{ni} \rfloor}{m_i^{(n)}}} \xrightarrow{p} 1,$$

so A.0.4 gives

$$\frac{m_{ij}^{(n)} - m_i^{(n)} p_{nij}}{\sqrt{m_i^{(n)}}} \xrightarrow{d} N(0, p_{ij}(1 - p_{ij})).$$

From the Weak Law of Large Numbers, we get $\frac{m_i^{(n)}}{N_n} \xrightarrow{p} \pi_i$, so $\frac{\sqrt{N_n}}{\sqrt{m_i^{(n)}}} \xrightarrow{p} \frac{1}{\sqrt{\pi_i}}$, and finally from Slutsky's lemma we have

$$\begin{aligned}
\sqrt{N_n}\,(\tilde{p}_{nij} - p_{nij}) &= \sqrt{N_n} \left(\frac{m_{ij}^{(n)}}{m_i^{(n)}} - p_{nij} \right) \\
&= \underbrace{\frac{\sqrt{N_n}}{\sqrt{m_i^{(n)}}}}_{\xrightarrow{p} \frac{1}{\sqrt{\pi_i}}} \cdot \underbrace{\frac{m_{ij}^{(n)} - m_i^{(n)} p_{nij}}{\sqrt{m_i^{(n)}}}}_{\xrightarrow{d} N(0, p_{ij}(1-p_{ij}))} \xrightarrow{d} \mathcal{N}\left(0, \frac{p_{ij}(1 - p_{ij})}{\pi_i}\right),
\end{aligned}$$

which is equivalent to what we wanted to show. Thus, it suffices to show (i) and (ii). We begin with (i). We will show that $\frac{m_i^{(n)}}{N_n} - \pi_{ni} \xrightarrow{p} 0$, i.e., we will show that for every $\varepsilon > 0$, $\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{m_i^{(n)}}{N_n} - \pi_{ni} \right| \right) = 0$. First, we use a result on the mixing time of Markov chains. Let $n \in \mathbb{N}$. For every $k > 0$, there exists a constant $C > 0$ such that

$$\left| p_{nij}^{(\alpha)} - \pi_{nj} \right| \leqslant C \rho_n^{\alpha},$$

where $\rho_n = 1 - \min_{i,j} p_{nij} < 1$. Since the Markov chain in our case is finite and recursive, we have $p_{nij} \to p_{ij}$, so $\rho_n \to 1 - \min_{i,j} p_{ij} < 1$ as $n \to \infty$, so there exist $\rho < 1$ and $n_0 \in \mathbb{N}$

such that

$$\forall n \geqslant n_0 \quad \left|p_{nij}^{(\alpha)} - \pi_{nj}\right| \leqslant C\rho^{\alpha}, \quad \forall \alpha \in \mathbb{N}. \tag{A.0.6}$$

Now we are ready to prove (i). Let $\varepsilon > 0$. From Markov's inequality, we have

$$\mathbb{P}\left(\left|\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right| > \varepsilon\right) \leqslant \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right)^2\right],$$

where

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right)^2\right] &= \mathbb{E}\left[\frac{\left(m_i^{(n)} - N_n\pi_{ni}\right)^2}{N_n^2}\right] \\
&= \frac{1}{N_n^2}\mathbb{E}\left[\left(m_i^{(n)} - N_n\pi_{ni}\right)^2\right] \\
&= \frac{1}{N_n^2}\mathbb{E}\left[\left(\sum_{k=1}^{N_n} I(X_{nk} = i) - N_n\pi_{ni}\right)^2\right] \\
&= \frac{1}{N_n^2}\mathbb{E}\left[\left(\sum_{k=1}^{N_n} (I(X_{nk} = i) - \pi_{ni})\right)^2\right] \\
&= \frac{1}{N_n^2}\mathbb{E}\left[\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} m_{ni}^{(k,\ell)}\right],
\end{aligned} \tag{A.0.7}$$

where $m_{ni}^{(k,\ell)} := 1_i(X_{nk})1_i(X_{n\ell}) - \pi_{ni}1_i(X_{n\ell}) - \pi_{ni}1_i(X_{nk}) + \pi_{ni}^2$. We verify the last equality in A.0.7. Using the definition and expanding, we see that

$$\begin{aligned}
\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} m_{ni}^{(k,\ell)} &= \sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} 1_i(X_{nk})1_i(X_{n\ell}) - \sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} \pi_{ni}1_i(X_{n\ell}) - \sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} \pi_{ni}1_i(X_{nk}) + \sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} \pi_{ni}^2 \\
&= \sum_{k=1}^{N_n} 1_i(X_{nk}) \sum_{\ell=1}^{N_n} 1_i(X_{n\ell}) - \pi_{ni}N_n \sum_{\ell=1}^{N_n} 1_i(X_{n\ell}) - \pi_{ni}N_n \sum_{k=1}^{N_n} 1_i(X_{nk}) + N_n^2\pi_{ni}^2 \\
&= \left(\sum_{k=1}^{N_n} 1_i(X_{nk})\right)^2 - 2\pi_{ni}N_n \sum_{k=1}^{N_n} 1_i(X_{nk}) + N_n^2\pi_{ni}^2 \\
&= \left(\sum_{k=1}^{N_n} 1_i(X_{nk}) - N_n\pi_{ni}\right)^2 = \left(\sum_{k=1}^{N_n} (1_i(X_{nk}) - \pi_{ni})\right)^2,
\end{aligned}$$

so A.0.7 holds and we get

$$\mathbb{E}\left[\left(\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right)^2\right] = \frac{1}{N_n^2}\mathbb{E}\left[\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} m_{ni}^{(k,\ell)}\right]. \tag{A.0.8}$$

Recall that throughout the proof we are using the notation

$$1_i(X_{nk}) := 1(X_{nk} = i) = \begin{cases} 1, & \text{if } X_{nk} = i \\ 0, & \text{if } X_{nk} \neq i \end{cases},$$

so we have $\mathbb{E}[1_i(X_{nk})] = \mathbb{E}[1(X_{nk} = i)] = \mathbb{P}(X_{nk} = i) = p_{n1i}^{(k)}$, since $X_0^{(n)} = 1$ by assumption. Similarly, $\mathbb{E}[1_i(X_{n\ell})] = p_{n1i}^{(\ell)}$ and $\mathbb{E}[1_i(X_{nk})1_i(X_{n\ell})] = p_{n1i}^{(k\wedge\ell)} p_{nii}^{(|k-\ell|)}$, since the Markov chain goes from state 1 to state i in $k \wedge \ell$ steps and from i to i in the remaining $|k - \ell|$ steps. Thus,

$$\mathbb{E}\, m_{ni}^{(k,\ell)} = p_{n1i}^{(k\wedge\ell)} p_{nii}^{(|k-\ell|)} - \pi_{ni} p_{n1i}^{(\ell)} - \pi_{ni} p_{n1i}^{(k)} + \pi_{ni}^2.$$

Using A.0.6, we can write $p_{nij}^{(\alpha)} = \pi_{nj} + \varepsilon_{nj}^{(\alpha)}$, where $|\varepsilon_{nj}^{(\alpha)}| \leqslant C\rho^\alpha$. For notational convenience, let $s = k \wedge \ell$ and $t = |k - \ell|$, so we get

$$\begin{aligned}
\mathbb{E} m_{ni}^{(k,\ell)} &= p_{n1i}^{(s)} p_{nii}^{(t)} - \pi_{ni} p_{n1i}^{(\ell)} - \pi_{ni} p_{n1i}^{(k)} + \pi_{ni}^2 \\
&= (\pi_{ni} + \varepsilon_{ni}^{(s)})(\pi_{ni} + \varepsilon_{ni}^{(t)}) - \pi_{ni} p_{n1i}^{(\ell)} - \pi_{ni} p_{n1i}^{(k)} + \pi_{ni}^2 \\
&= \cancel{\pi_{ni}^2} + \pi_{ni}\varepsilon_{ni}^{(t)} + \pi_{ni}\varepsilon_{ni}^{(s)} + \varepsilon_{ni}^{(s)}\varepsilon_{ni}^{(t)} - \cancel{\pi_{ni}^2} - \pi_{ni}\varepsilon_{ni}^{(\ell)} - \cancel{\pi_{ni}^2} - \pi_{ni}\varepsilon_{ni}^{(k)} + \cancel{\pi_{ni}^2} \\
&= \pi_{ni}\varepsilon_{ni}^{(t)} + \pi_{ni}\varepsilon_{ni}^{(s)} + \varepsilon_{ni}^{(s)}\varepsilon_{ni}^{(t)} - \pi_{ni}\varepsilon_{ni}^{(\ell)} - \pi_{ni}\varepsilon_{ni}^{(k)},
\end{aligned}$$

so A.0.6 yields that for every $n \geqslant n_0$

$$\begin{aligned}
\left|\mathbb{E} m_{ni}^{(k,\ell)}\right| &\leqslant \pi_{ni}|\varepsilon_{ni}^{(t)}| + \pi_{ni}|\varepsilon_{ni}^{(s)}| + |\varepsilon_{ni}^{(s)}\varepsilon_{ni}^{(t)}| + \pi_{ni}|\varepsilon_{ni}^{(\ell)}| + \pi_{ni}|\varepsilon_{ni}^{(k)}| \\
&= \pi_{ni}\left(|\varepsilon_{ni}^{(t)}| + |\varepsilon_{ni}^{(s)}| + \frac{1}{\pi_{ni}}|\varepsilon_{ni}^{(s)}\varepsilon_{ni}^{(t)}| + |\varepsilon_{ni}^{(\ell)}| + |\varepsilon_{ni}^{(k)}|\right) \\
&\leqslant \pi_{ni}\left(C\rho^s + C\rho^t + \frac{1}{\pi_{ni}} C\rho^s\rho^t + \rho^\ell + \rho^k\right) \leqslant C'(\rho^s + \rho^t + \rho^k + \rho^\ell),
\end{aligned}$$

for some constant $C' > 0$. So now the bound in A.0.8 becomes

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right)^2\right] &= \frac{1}{N_n^2}\mathbb{E}\left[\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} m_{ni}^{(k,\ell)}\right] = \frac{1}{N_n^2}\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n}\mathbb{E}\, m_{ni}^{(k,\ell)} \\
&\leqslant \frac{1}{N_n^2}\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n} C'(\rho^s + \rho^t + \rho^k + \rho^\ell) \\
&\leqslant \frac{C'}{N_n^2}\sum_{k=1}^{N_n}\sum_{\ell=1}^{N_n}(2\rho^k + \rho^{\ell-k} + \rho^\ell)
\end{aligned}$$

$$\leqslant \frac{C'}{N_n^2}\left(3N_n\sum_{k=1}^{\infty}\rho^k + \sum_{\ell=1}^{N_n}\sum_{k=1}^{\ell}\rho^{\ell-k}\right)$$
$$\leqslant \frac{C'}{N_n^2}\left(3N_n\cdot\frac{1}{1-\rho} + N_n\cdot\frac{1}{1-\rho}\right)$$
$$= \frac{C'}{N_n^2}\cdot\frac{4N_n}{1-\rho} = \frac{4C'}{N_n(1-\rho)} \to 0 \quad \text{as } n\to\infty,$$

since $N_n \to \infty$ by assumption. Hence, Markov's inequality yields

$$\mathbb{P}\left(\left|\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right| > \varepsilon\right) \leqslant \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\frac{m_i^{(n)}}{N_n} - \pi_{ni}\right)^2\right] \leqslant \frac{4C'}{\varepsilon^2 N_n(1-\rho)} \to 0 \quad \text{as } n\to\infty,$$

which proves (i). Now we will prove (ii), i.e., $\eta_{nij} \xrightarrow{p} 0$. Define $S_m := \sum_{t=1}^{m}\left(C_t^{(n)}(j) - p_{nij}\right)$, where

$$C_t^{(n)}(j) = \begin{cases} 1, & \text{if } W_{it}^{(n)} = j,\ t = 1,\dots,m_i^{(n)} \\ 0, & \text{else} \end{cases}.$$

Let $\varepsilon > 0$. From (i), there exists $n_0 \in \mathbb{N}$ such that $\mathbb{P}\left(|m_i^{(n)} - \lfloor N_n\pi_{ni}\rfloor| > \varepsilon^3 N_n\right) < \varepsilon$ for every $n > n_0$. We will show that $\mathbb{P}(|\eta_{nij}| > \varepsilon) \to 0$ as $n\to\infty$. By the definition of η_{nij}, S_m and $C_t^{(n)}(j)$, we have

$$\mathbb{P}(|\eta_{nij}| > \varepsilon) = \mathbb{P}\left(\left|S_{m_i^{(n)}} - S_{\lfloor N_n\pi_{ni}\rfloor}\right| > \sqrt{N_n}\varepsilon\right)$$
$$= \mathbb{P}\left(|m_i^{(n)} - \lfloor N_n\pi_{ni}\rfloor| > \varepsilon^3 N_n\ ,\ \left|S_{m_i^{(n)}} - S_{\lfloor N_n\pi_{ni}\rfloor}\right| > \sqrt{N_n}\varepsilon\right) +$$
$$+ \mathbb{P}\left(|m_i^{(n)} - \lfloor N_n\pi_{ni}\rfloor| \leqslant \varepsilon^3 N_n\ ,\ \left|S_{m_i^{(n)}} - S_{\lfloor N_n\pi_{ni}\rfloor}\right| > \sqrt{N_n}\varepsilon\right)$$
$$\leqslant \mathbb{P}\left(|m_i^{(n)} - \lfloor N_n\pi_{ni}\rfloor| > \varepsilon^3 N_n\right) + \mathbb{P}\left(\max_{|m-\lfloor N_n\pi_{ni}\rfloor|\leqslant\varepsilon^3 N_n}|S_m - S_{\lfloor N_n\pi_{ni}\rfloor}| > \sqrt{N_n}\varepsilon\right)$$
$$\leqslant \varepsilon + 2\,\mathbb{P}\left(\max_{1\leqslant m\leqslant\varepsilon^3 N_n}|S_m| > \sqrt{N_n}\varepsilon\right) \overset{\text{Kolmogorov's maximal inequality}}{\leqslant}$$
$$\leqslant \varepsilon + 2\cdot\frac{2}{\varepsilon^2 N_n}\mathrm{Var}\left(S_{\lfloor N_n\varepsilon^3\rfloor+1}\right) \overset{\text{independence}}{=}$$
$$= \varepsilon + \frac{4}{\varepsilon^2 N_n}\left(\lfloor N_n\varepsilon^3\rfloor + 1\right)\sigma^2 \leqslant C\varepsilon,$$

for some constant $C > 0$. Thus, $\eta_{nij} \xrightarrow{p} 0$, which proves (ii). From the observation we made above, the proof of the theorem is now complete. □

Theorem A.0.2 (CLT for the Bayesian Bootstrap). *Let $i, j \in S$ and $\hat{p}_{ij}$ denote the maximum likelihood estimator of the transition probability p_{ij}. Then, for almost all sample sequences $\mathbf{x} = (x_0, x_1, \ldots, x_n)$ we have*

$$\frac{\sqrt{n_i}(\hat{p}^*_n(i,j) - \hat{p}_{ij})}{\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}} \,\Big|\, \mathbf{x} \xrightarrow{d} N(0,1), \tag{A.0.9}$$

as $n \to \infty$. That is, the Bayesian bootstrap estimators are asymptotically normal.

Proof. Let $i, j \in S$. The idea is to decompose the quantity of (3.2.3) in pieces for which the classical Lindeberg-Lévy Central Limit Theorem can be applied directly. For that, we first introduce the following convenient notation:

$$S_{n_{ij}} := \sum_{t \in B_{ij}} (Z_{it} - 1) \quad \text{and} \quad S_{n_i, n_i - n_{ij}} := \sum_{t=1}^{n_i} (Z_{it} - 1) - \sum_{t \in B_{ij}} (Z_{it} - 1).$$

Notice that, since $|B_{ij}| = n_{ij}$, the above quantities can be rewritten as

$$S_{n_{ij}} = \sum_{t \in B_{ij}} Z_{it} - n_{ij} \quad \text{and} \quad S_{n_i, n_i - n_{ij}} = \sum_{t=1}^{n_i} Z_{it} - \sum_{t \in B_{ij}} Z_{it} - n_i + n_{ij}.$$

We will exploit the fact that these quantities include exponential random variables and use Slutsky's Lemma and CLT. We proceed with the following calculations:

$$\begin{aligned}
\sqrt{n_i}(\hat{p}^*_n(i,j) - \hat{p}_{ij}) &= \sqrt{n_i}\left(\frac{\sum\limits_{t \in B_{ij}} Z_{it}}{\sum\limits_{t=1}^{n_i} Z_{it}} - \frac{n_{ij}}{n_i} \right) = \sqrt{n_i} \cdot \frac{\sum\limits_{t \in B_{ij}} Z_{it}}{\sum\limits_{t=1}^{n_i} Z_{it}} - \sqrt{n_i} \cdot \frac{n_{ij}}{n_i} \\
&= \sqrt{n_i} \cdot \frac{\sum\limits_{t \in B_{ij}} Z_{it}}{\sum\limits_{t=1}^{n_i} Z_{it}} - \frac{n_{ij}}{\sqrt{n_i}} = \frac{n_i}{\sum\limits_{t=1}^{n_i} Z_{it}} \left(\frac{\sum\limits_{t \in B_{ij}} Z_{it}}{\sqrt{n_i}} - \frac{n_{ij}}{n_i} \cdot \frac{\sum\limits_{t=1}^{n_i} Z_{it}}{\sqrt{n_i}} \right).
\end{aligned} \tag{A.0.10}$$

Notice that

$$\begin{aligned}
&\left(1 - \frac{n_{ij}}{n_i}\right) \sqrt{\frac{n_{ij}}{n_i}} \cdot \frac{S_{n_{ij}}}{\sqrt{n_{ij}}} - \frac{n_{ij}}{n_i} \sqrt{1 - \frac{n_{ij}}{n_i}} \cdot \frac{S_{n_i, n_i - n_{ij}}}{\sqrt{n_i - n_{ij}}} \\
&= \left(1 - \frac{n_{ij}}{n_i}\right) \sqrt{\frac{n_{ij}}{n_i}} \cdot \frac{\sum\limits_{t \in B_{ij}} Z_{it} - n_{ij}}{\sqrt{n_{ij}}} - \frac{n_{ij}}{n_i} \sqrt{1 - \frac{n_{ij}}{n_i}} \cdot \frac{\sum\limits_{t=1}^{n_i} Z_{it} - \sum\limits_{t \in B_{ij}} Z_{it} - n_i + n_{ij}}{\sqrt{n_i - n_{ij}}} \\
&= \left(1 - \frac{n_{ij}}{n_i}\right) \frac{\sum\limits_{t \in B_{ij}} Z_{it} - n_{ij}}{\sqrt{n_{ij}}} - \frac{n_{ij}}{n_i} \cdot \frac{n_{ij} - \sum\limits_{t \in B_{ij}} Z_{it}}{\sqrt{n_i}} - \frac{n_{ij}}{n_i} \cdot \frac{\sum\limits_{t=1}^{n_i} Z_{it} - n_i}{\sqrt{n_i}} \\
&= \frac{\sum\limits_{t \in B_{ij}} Z_{it} - n_{ij}}{\sqrt{n_{ij}}} - \cancel{\frac{n_{ij}}{n_i} \cdot \frac{\sum\limits_{t \in B_{ij}} Z_{it} - n_{ij}}{\sqrt{n_{ij}}}} + \cancel{\frac{n_{ij}}{n_i} \cdot \frac{\sum\limits_{t \in B_{ij}} Z_{it} - n_{ij}}{\sqrt{n_{ij}}}} - \frac{n_{ij}}{n_i} \cdot \frac{\sum\limits_{t=1}^{n_i} Z_{it}}{\sqrt{n_i}} + \frac{n_{ij}}{\sqrt{n_i}}
\end{aligned}$$

$$= \frac{\sum\limits_{t\in B_{ij}} Z_{it}}{\sqrt{n_i}} - \cancel{\frac{n_{ij}}{\sqrt{n_i}}} - \frac{n_{ij}}{n_i}\cdot\frac{\sum\limits_{t=1}^{n_i} Z_{it}}{\sqrt{n_i}} + \cancel{\frac{n_{ij}}{\sqrt{n_i}}} = \frac{\sum\limits_{t\in B_{ij}} Z_{it}}{\sqrt{n_i}} - \frac{n_{ij}}{n_i}\cdot\frac{\sum\limits_{t=1}^{n_i} Z_{it}}{\sqrt{n_i}},$$

thus (A.0.10) becomes

$$\begin{aligned}\sqrt{n_i}(\hat{p}_n^*(i,j) - \hat{p}_{ij}) &= \frac{n_i}{\sum\limits_{t=1}^{n_i} Z_{it}}\left[\left(1-\frac{n_{ij}}{n_i}\right)\sqrt{\frac{n_{ij}}{n_i}}\cdot\frac{S_{n_{ij}}}{\sqrt{n_{ij}}} - \frac{n_{ij}}{n_i}\sqrt{1-\frac{n_{ij}}{n_i}}\cdot\frac{S_{n_i,n_i-n_{ij}}}{\sqrt{n_i-n_{ij}}}\right] \\ &= \frac{n_i}{\sum\limits_{t=1}^{n_i} Z_{it}}\left[\underbrace{(1-\hat{p}_{ij})\sqrt{\frac{n_{ij}}{n_i}}\cdot\frac{S_{n_{ij}}}{\sqrt{n_{ij}}}}_{(I)} - \underbrace{\hat{p}_{ij}\sqrt{1-\hat{p}_{ij}}\cdot\frac{S_{n_i,n_i-n_{ij}}}{\sqrt{n_i-n_{ij}}}}_{(II)}\right].\end{aligned} \tag{A.0.11}$$

Since $Z_{it} \sim \text{Exp}(1)$, the Strong Law of Large Numbers yields

$$\frac{1}{n_i}\sum_{t=1}^{n_i} Z_{it} \xrightarrow{a.s.} \mathbb{E}[Z_{it}] = 1 \Leftrightarrow \frac{n_i}{\sum\limits_{t=1}^{n_i} Z_{it}} \xrightarrow{a.s.} 1, \tag{A.0.12}$$

thus it suffices to prove that the quantity $(I)-(II)$ converges in distribution to the standardized normal distribution. We have that

$$\begin{aligned}\frac{(I)}{(1-\hat{p}_{ij})\sqrt{\hat{p}_{ij}}} &= \frac{(1-\hat{p}_{ij})\sqrt{\frac{n_{ij}}{n_i}}\cdot\frac{S_{n_{ij}}}{\sqrt{n_{ij}}}}{(1-\hat{p}_{ij})\sqrt{\hat{p}_{ij}}} = \frac{S_{n_{ij}}}{\sqrt{n_i}\sqrt{\hat{p}_{ij}}} = \frac{S_{n_{ij}}}{\sqrt{n_i}\sqrt{\frac{n_{ij}}{n_i}}} = \frac{S_{n_{ij}}}{\sqrt{n_{ij}}} \\ &= \frac{\sum\limits_{t\in B_{ij}}(Z_{it}-1)}{\sqrt{n_{ij}}} = \frac{\sum\limits_{t\in B_{ij}} Z_{it} - n_{ij}}{\sqrt{n_{ij}}} \overset{|B_{ij}|=n_{ij}}{=} \frac{\sum\limits_{t\in B_{ij}} Z_{it} - \sum\limits_{t\in B_{ij}}\mathbb{E}[Z_{it}]}{\sqrt{n_{ij}\,\text{Var}(Z_{it})}}.\end{aligned}$$

Applying the Lindeberg-Lévy Central Limit Theorem, we get

$$\frac{(I)}{(1-\hat{p}_{ij})\sqrt{\hat{p}_{ij}}} \xrightarrow{d} N(0,1). \tag{A.0.13}$$

In the same spirit, we have that

$$\begin{aligned}\frac{(II)}{\hat{p}_{ij}\sqrt{1-\hat{p}_{ij}}} &= \frac{\hat{p}_{ij}\sqrt{1-\hat{p}_{ij}}\cdot\frac{S_{n_i,n_i-n_{ij}}}{\sqrt{n_i-n_{ij}}}}{\hat{p}_{ij}\sqrt{1-\hat{p}_{ij}}} = \frac{S_{n_i,n_i-n_{ij}}}{\sqrt{n_i-n_{ij}}} = \frac{\sum\limits_{t=1}^{n_i}(Z_{it}-1) - \sum\limits_{t\in B_{ij}}(Z_{it}-1)}{\sqrt{n_i-n_{ij}}} \\ &= \frac{\left(\sum\limits_{t=1}^{n_i} Z_{it} - \sum\limits_{t\in B_{ij}} Z_{it}\right) - (n_i-n_{ij})}{\sqrt{n_i-n_{ij}}} = \frac{\left(\sum\limits_{t=1}^{n_i} Z_{it} - \sum\limits_{t\in B_{ij}} Z_{it}\right) - \mathbb{E}\left[\sum\limits_{t=1}^{n_i} Z_{it} - \sum\limits_{t\in B_{ij}} Z_{it}\right]}{\sqrt{n_i-n_{ij}}}.\end{aligned}$$

Applying the Lindeberg-Lévy Central Limit Theorem, we get

$$\frac{(II)}{\hat{p}_{ij}\sqrt{1-\hat{p}_{ij}}} \xrightarrow{d} N(0,1). \tag{A.0.14}$$

From (A.0.13) and (A.0.14), we receive that

$$\frac{(I)}{\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}} = \frac{(I)}{(1-\hat{p}_{ij})\sqrt{\hat{p}_{ij}}} \cdot \sqrt{1-\hat{p}_{ij}} \xrightarrow{d} N(0, 1-\hat{p}_{ij})$$

and

$$\frac{(II)}{\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}} = \frac{(II)}{\hat{p}_{ij}\sqrt{1-\hat{p}_{ij}}} \cdot \sqrt{\hat{p}_{ij}} \xrightarrow{d} N(0, \hat{p}_{ij}).$$

Finally, from Slutsky's Lemma we get

$$\frac{\sqrt{n_i}(\hat{p}^*_n(i,j) - \hat{p}_{ij})}{\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}} \,\Big|\, \mathbf{x} = \frac{n_i}{\sum_{t=1}^{n_i} Z_{it}} \cdot \left[\frac{(I)}{\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}} - \frac{(II)}{\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})}} \right] \xrightarrow{d} N(0,1)$$

and the proof is complete. □

Bibliography

[1] T.W. Anderson, L.A. Goodman. *Statistical Inference about Markov Chains*. The Annals of Mathematical Statistics, 1957.

[2] A. Beskos, et al. *MCMC methods for diffusion bridges*. Stochastics and Dynamics 8.03 (2008): 319-350.

[3] M. Betancourt. *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv preprint arXiv:1701.02434 (2017).

[4] P. Billingsley. *Probability and Measure*. John Wiley and Sons, 1995. Link

[5] P. Billingsley. *Statistical Inference for Markov Processes*. The University of Chicago Press, 1961.

[6] P. Billingsley. *Statistical Methods in Markov Chains*. The Annals of Mathematical Statistics, 1961.

[7] P. Billingsley. *Ergodic Theory and Information*. John Wiley & Sons, Inc., New York, 1965.

[8] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation and queues*. Texts in applied mathematics. Springer, New York, 1999.

[9] C. Casarotto. *Markov Chains and the Ergodic Theorem*. The University of Chicago, 2007.

[10] D. Cheliotis. *Intorduction to Stochastic Calculus* (in Greek). National and Kapodistrian University of Athens, Kallipos, 2020.

[11] T. Chen, E. Fox, and C. Guestrin. *Stochastic gradient hamiltonian monte carlo*. International conference on machine learning. PMLR, 2014.

[12] R. Douc, E. Moulines, P. Priouret, P. Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering, 2018.

[13] R. Durrett. *Probability: Theory and Examples (4th edition)*. Cambridge Series in Statistical and Probabilistic Mathematics, 2010.

[14] B. Efron. *Bootstrap Methods: Another Look at the Jackknife*. The Annals of Statistics, 1979.

[15] B. Efron, R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

[16] M. Einsiedler, T. Ward. *Ergodic Theory with a view towards Number Theory*. Springer Verlag, London, 2011.

[17] B. Fristedt, L. Gray. *A Modern Approach to Probability Theory*. Birkhauser, Boston, 1997.

[18] C.D. Fuh. *The bootstrap method for Markov chains*. Retrospective Theses and Dissertations, Iowa State University, 1989.

[19] C.D. Fuh, T.S. Fan. *A Bayesian Bootstrap for Finite Markov-Chains*. Institute of Statistical Science, Academia Sinica, 1997.

[20] A. Gelman et al. *Bayesian Data Analysis*. Texts in Statistical Science, Chapman and Hall/CRC, 2013.

[21] M. Girolami, B. Calderhead. *Riemann manifold langevin and hamiltonian monte carlo methods*. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.2 (2011): 123-214.

[22] G.H. Givens, J.A. Hoeting. *Computational Statistics, 2nd edition*. John Wiley & Sons, Inc., 2012. Link

[23] F. Hollander. *Probability Theory: The Coupling Method*. Mathematical Institute, Leiden University, 2012.

[24] O. Häggström. *Finite Markov Chains and Algorithmic Applications*. London Mathematical Society Student Texts 52, Cambridge University Press, 2002.

[25] R. van der Hofstad et al. *Local Weak Convergence for PageRank*. Ann. Appl. Probab. 30(1): 40-79 (February 2020). DOI: 10.1214/19-AAP1494.

[26] M. Johannes, N. Polson. *MCMC methods for continuous-time financial econometrics*. Handbook of Financial Econometrics: Applications. Elsevier, 2010. 1-72.

[27] V. Katsianos. *Likelihood-Based Inference and Model Selection for Discrete-Time Finite State-Space Hidden Markov Models*. MSc Thesis, National and Kapodistrian University of Athens, 2018. Link

[28] V.G. Kulkarni. *Modeling and Analysis of Stochastic Systems (2nd edition)*. Chapman & Hall/CRC Texts in Statistical Science, 2010.

[29] J. Lin. *On the Dirichlet Distribution*. Queen's University, Kingston, Ontario, Canada, 2016.

[30] D. Logothetis. *Algorithmic techniques of Bayesian and Classical approach in plant growth models and convergence issues in the boundary of the parameter space* (in Greek). MSc Thesis, National and Kapodistrian University of Athens, 2016.

[31] M. Loulakis. *Stochastic Processes* (in Greek). National Technical University of Athens, Kallipos, 2015.

[32] A. Menegaki. *Ergodic Theory notes* (in Greek). National and Kapodistrian University of Athens, 2017.

[33] S.P. Meyn, R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[34] J.R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[35] M. Olvera-Cravioto et al. *PageRank's behavior under degree correlations*. Annals of Applied Probability, Vol. 1, No. 3, pp. 1403-1442, 2021.

[36] M. Olvera-Cravioto et al. *PageRank asymptotics on directed preferential attachment networks*. Annals of Applied Probability, Vol. 32, No. 4, pp. 3060-3084, 2021.

[37] M. Olvera-Cravioto et al. *PageRank Nibble on the sparse directed stochastic block model*. Proceedings of the 18th Workshop on Algorithms and Models for the Web Graph, Toronto, Canada, March 2023.

[38] N. Papadatos. *Probability Theory* (in Greek). National and Kapodistrian University of Athens, 2006.

[39] L.P. Peralta. *Finite Markov Chains*. Universitat de Barcelona, 2015.

[40] D.B. Rubin. *The Bayesian Bootstrap*. The Annals of Statistics, 1981. Link

[41] R. Tibshirani, T. Hastie, et al. *The elements of statistical learning: data mining, inference and prediction*. The Mathematical Intelligencer, 2005.

[42] R. Tibshirani, T. Hastie, et al. *An Introduction to Statistical Learning: with Applications in R* . Springer Texts in Statistics, 2013.

[43] R. Tibshirani. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, 1996.

[44] S. Trevezas. *Introduction to the statistical analysis of finite Markov chains*. Laboratoire MAS-ECP, 2013.

[45] S. Trevezas. *Nonparametric Statistics* (in Greek). National and Kapodistrian University of Athens, 2020.

[46] S. Trevezas. *Statistics for Stochastic Processes* (in Greek). National and Kapodistrian University of Athens, 2021.

[47] S.R.S. Varadhan. *Probability Theory*. Courant Lecture Notes, 2000.

[48] Kernels and Operators

[49] Ergodic Theorems History

[50] Inference on Markov Chains (University of Washington)

[51] Maximum Likelihood Estimation for Markov Chains (Carnegie Mellon)

[52] Probability Theory (Berkeley)