Dynamics of Markov Chains for Undergraduates

Ursula Porod

February 27, 2024

Contents

Preface

This book is about the theory of Markov chains and their long-term dynamical properties. It is written for advanced undergraduates who have taken a course in calculus-based probability theory and are familiar with the classical limit theorems such as the Central Limit Theorem and the Strong Law of Large Numbers. Knowledge of linear algebra and a basic familiarity with groups are expected. Measure theory is not a prerequisite.

The book covers in depth the classical theory of discrete-time Markov chains with countable state space and introduces the reader to more contemporary areas such as Markov chain Monte Carlo methods and the study of convergence rates of Markov chains. For example, it includes a study of random walks on the symmetric group S_n as a model of card shuffling and their rates of convergence. A possible novelty for an undergraduate text is the book's approach to studying convergence rates for natural sequences of Markov chains with increasing state spaces of N elements, rather than for fixed sized chains. This approach allows for a simultaneous time T and size N asymptotics which reveals, in some cases, the so-called cut-off phenomenon, a kind of phase transition that occurs as these Markov chains converge to stationarity. The book also covers martingales and it covers random walks on graphs as electric networks. The analogy with electric networks reveals interesting connections between certain laws of Physics, discrete harmonic functions, and the study of reversible Markov chains, in particular for computations of their cover times and hitting times. Most of the currently available undergraduate textbooks do not include these topics in any detail.

The following is a brief summary of the book's chapters. Chapters 1, 2, and 3 cover the classical theory of discrete-time Markov chains. They are the foundation for the rest of the book. Chapter 4 focuses in detail on path properties for simple random walk on $\mathbb{Z}$, a topic that is also relevant for a future study of Brownian motion. The classical Galton-Watson branching process is covered in Chapter 5. Chapter 6 introduces the reader to martingales and some of their applications to Markov chains. It includes a discussion of the Optional Stopping Theorem and the Martingale Convergence Theorem. Chapter 7 collects material about reversible processes, a large class of processes that can be viewed as resistor networks. The material is a prerequisite for the following chapters. Chapter 8

treats reversible Markov chains as electrical networks, a fruitful approach that was first introduced by Kakutani in [19] and later popularized by Doyle and Snell in [11]. Markov chain Monte Carlo algorithms and some of their applications are introduced in Chapter 9. Chapter 10 introduces the reader to random walks on finite groups and, as an example, introduces card shuffling. Chapter 11 focuses on rates of convergence for Markov chains. It introduces the reader to the large N and T asymptotics and the cut-off phenomenon. Three Appendices collect necessary background material from probability, analysis, and linear algebra.

The book has more material than a standard one-semester course will cover. It is designed to lead students from the basics of Markov chains to interesting and advanced topics in the field. A one-semester introductory course might cover most of Chapters 1-4 and a selection of topics from Chapters 5 and 9-11.

Chapter 1

Markov Chains: Construction and Basic Notions

1.1 Introduction

A *stochastic process* is a mathematical model for the random evolution of a system in time. More precisely, it is a collection of random variables $X_t(\omega)$ on a probability space Ω, indexed by a time parameter $t \in I$ from some index set I, and taking values in a common *state space* $\mathcal{S}$.

In this book, the time index set I will always be $\mathbb{N}_0$. We call such a process $(X_n)_{n\geq 0}$ a *discrete time* stochastic process. The random variable X_n gives the position (or state) at time n. The state space $\mathcal{S}$ will always be a discrete set $\mathcal{S}$, either finite or countably infinite.

The fundamental assumption on the stochastic processes $(X_n)_{n\geq 0}$ is the so-called *Markov property*. The Markov property can be described informally in the following way: At each time n, the future positions of the process only depend on the position at time n and not on the positions of X_s for $s < n$.

A stochastic process $(X_n)_{n\geq 0}$ with discrete state space $\mathcal{S}$ that has the Markov property is called a *Markov chain*. We will give a precise definition of the Markov property and a mathematically rigorous construction of a Markov chain in the following section.

The first three chapters of this book cover classical material on Markov chains, such as their construction, properties, and convergence behavior. These chapters are the foundation for all later chapters.

1.2 Construction of a Markov chain

1.2.1 Finite-length trajectories

We start with a discrete (finite or countably infinite) *state space* $\mathcal{S}$. Fix a positive integer $n \geq 1$ and consider the direct product space

$$\Omega = \mathcal{S} \times \mathcal{S} \times \cdots \times \mathcal{S} \qquad \text{with } (n+1) \text{ factors } \mathcal{S}$$

which is denoted by $\Omega = \mathcal{S}^{n+1}$. Recall that this means

$$\Omega = \{(x_0, x_1, ..., x_n) \,|\, x_i \in \mathcal{S} \text{ for } 0 \leq i \leq n\}\,.$$

We will think of the index set $\{0, 1, ..., n\}$ as modeling time. Let π_0 be a probability distribution on $\mathcal{S}$, which we call the *initial distribution*. Furthermore, let $p : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be a function with the property

$$\sum_{x_j \in \mathcal{S}} p(x_i, x_j) = 1 \quad \text{for all } x_i \in \mathcal{S}\,. \tag{1.1}$$

Let $\mathcal{F} = \mathcal{P}(\Omega)$ be the power set (the set of all subsets) of Ω. The power set $\mathcal{F}$ consists of all sets of the form

$$F = F_0 \times F_1 \times \cdots \times F_n \quad \text{with } F_0, F_1, ..., F_n \subseteq \mathcal{S}\,.$$

$(\Omega, \mathcal{F})$ is a σ-algebra (see Appendix B for definitions). We construct a probability $\mathbb{P}$ on $(\Omega, \mathcal{F})$ with the use of p in the following way: For all $\omega = (x_0, x_1, ..., x_n) \in \Omega$, define

$$\mathbb{P}(\omega) = \pi_0(x_0)\, p(x_0, x_1)\, p(x_1, x_2) \cdots p(x_{n-1}, x_n)\,, \tag{1.2}$$

and from this for all $F \in \mathcal{F}$, since Ω is discrete,

$$\mathbb{P}(F) = \sum_{\omega \in F} \mathbb{P}(\omega)\,.$$

We verify that (1.1) and (1.2) imply $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$. Thus $(\Omega, \mathcal{F}, \mathbb{P})$ defines a probability space. Next, we consider the *coordinate random variables* $X_0, X_1, ..., X_n$ on $(\Omega, \mathcal{F}, \mathbb{P})$ defined by

$$X_i(\omega) = x_i \quad \text{for } \omega = (x_0, x_1, ..., x_n) \in \Omega,\ 0 \leq i \leq n\,.$$

As a consequence of this construction, conditional probabilities involving the coordinate random variables X_i have the following important property, called the **Markov property**:

Proposition 1.2.1 (Markov property)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space as defined above and $(X_k)_{0\le k\le n}$ the random vector whose components are the coordinate random variables as defined above. Then for all $0 \le i \le n-1$ and for all $x_0, x_1, ..., x_{i+1} \in \mathcal{S}$,*

$$\begin{aligned}\mathbb{P}(X_{i+1} = x_{i+1} \,|\, X_0 = x_0, ..., X_i = x_i) \\ = \mathbb{P}(X_{i+1} = x_{i+1} \,|\, X_i = x_i) = p(x_i, x_{i+1})\,.\end{aligned} \tag{1.3}$$

Proof. By (1.2),

$$\begin{aligned}\mathbb{P}(X_{i+1} = x_{i+1} \,|\, X_0 = x_0, ..., X_i = x_i) &= \frac{\mathbb{P}(X_0 = x_0, X_1 = x_1, ..., X_{i+1} = x_{i+1})}{\mathbb{P}(X_0 = x_0, X_1 = x_1, ..., X_i = x_i)} \\ &= \frac{\pi_0(x_0)\, p(x_0, x_1)\, p(x_1, x_2) \cdots p(x_i, x_{i+1})}{\pi_0(x_0)\, p(x_0, x_1)\, p(x_1, x_2) \cdots p(x_{i-1}, x_i)} \\ &= p(x_i, x_{i+1})\,.\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{P}(X_{i+1} = x_{i+1} \,|\, X_i = x_i) &= \frac{\mathbb{P}(X_i = x_i,\ X_{i+1} = x_{i+1})}{\mathbb{P}(X_i = x_i)} \\ &= \frac{\displaystyle\sum_{x_0,...,x_{i-1}\in\mathcal{S}} \pi_0(x_0)\, p(x_0, x_1) \cdots p(x_{i-1}, x_i)\, p(x_i, x_{i+1})}{\displaystyle\sum_{x_0,...,x_{i-1}\in\mathcal{S}} \pi_0(x_0)\, p(x_0, x_1) \cdots p(x_{i-1}, x_i)} \\ &= \frac{\mathbb{P}(X_i = x_i) p(x_i, x_{i+1})}{\mathbb{P}(X_i = x_i)} = p(x_i, x_{i+1})\,.\end{aligned}$$

This proves (1.3). □

Conversely, let us now assume that we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a discrete set $\mathcal{S}$, and a sequence of random variables $X_0, X_1, ..., X_n$ with $X_i : \Omega \to \mathcal{S}$ for $0 \le i \le n$. Let $X_0 \sim \pi_0$ and assume that there exists a function $p : \mathcal{S} \times \mathcal{S} \to [0, 1]$ such that for all $0 \le i \le n-1$ the Markov property (1.3) holds. Then it is straightforward to show (via sequential conditioning) that for all $\omega = (x_0, x_1, ..., x_n) \in \mathcal{S}^{n+1}$ we have

$$\mathbb{P}(X_0 = x_0, ..., X_n = x_n) = \pi_0(x_0)\, p(x_0, x_1) \cdots p(x_{n-1}, x_n)\,, \tag{1.4}$$

which describes the joint distribution of the random vector $(X_k)_{0\le k\le n}$.

Notation: From now onwards, we will write P_{xy} for the conditional probabilities $p(x,y) = \mathbb{P}(X_1 = y \,|\, X_0 = x)$ for $x, y \in \mathcal{S}$.

Definition 1.2.1. *We call $(X_k)_{0\le k\le n}$ as constructed above a* **homogeneous Markov chain of length n with state space $\mathcal{S}$, one-step transition probabilities P_{xy}, $x, y \in \mathcal{S}$, and initial distribution π_0.**

Note: The Markov property (1.3) implies
$\mathbb{P}(X_{i+1} = x_{i+1}, ..., X_m = x_m \,|\, X_0 = x_0, X_1 = x_1, ..., X_i = x_i) =$

$$\mathbb{P}(X_{i+1} = x_{i+1}, ..., X_m = x_m \,|\, X_i = x_i) \qquad \text{for } i+1 \le m \le n,$$

and more generally,
$\mathbb{P}(X_{i+k} = x_{i+k}, ..., X_m = x_m \,|\, X_0 = x_0, X_1 = x_1, ..., X_i = x_i) =$

$$\mathbb{P}(X_{i+k} = x_{i+k}, ..., X_m = x_m \,|\, X_i = x_i) \qquad \text{for } 1 \le k \text{ and } i+k \le m \le n\,. \tag{1.5}$$

As already mentioned, we think of the index k for $(X_k)_{0\le k\le n}$ as denoting time. If we consider time i as the presence, then (1.5) can loosely be summarized by saying "Given the present state of the Markov chain, future events and past events for the chain are probabilistically independent".

1.2.2 From finite to infinite-length trajectories

Usually, when constructing a Markov chain, we start with a state space $\mathcal{S}$, an initial distribution π_0 on $\mathcal{S}$, and a set of one-step transition probabilities P_{xy}, $x, y \in \mathcal{S}$, that model the inherent probabilistic properties of the process. As we have seen in the previous section, this information allows us to compute all *finite-dimensional* joint distributions of the random variables X_n, $n \ge 0$, for the process (see (1.4)). However, questions such as "Will the process ever visit state x?" or "How often, on average, does the process visit state y?" are questions about the long-run behavior of the process. Being able to answer such questions requires the existence of an underlying probabilistic structure on the space of all *infinte-length* trajectories. That is to say, it requires the existence of an underlying *common* probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which *all* random variables $X_n : \Omega \to \mathcal{S}$ for $n \ge 0$ are defined, and which is consistent with the finite-dimensional joint marginal distributions as computed in (1.4).

So how do we go from what we already have, namely knowledge of all finite-dimensional joint distributions, to proving the existence of an underlying common probability space

for the infinite-length process $(X_n)_{n\geq 0}$? The answer lies in the *Kolmogorov*[1] *Extension Theorem*. It is a deep result in measure theory, and we will only quote it here. See the classic work by Kolmogorov [20], or as another reference, [33]. We first define the necessary consistency conditions that underly Kolmogorov's theorem.

Definition 1.2.2 (Kolmogorov Consistency Conditions)**.** *Let $\mathcal{S}$ be a discrete space with σ-algebra $\mathcal{A} = \mathcal{P}(\mathcal{S})$ being the power set of $\mathcal{S}$. Assume that for all $k \geq 1$ and all $0 \leq n_1 < \cdots < n_k$ there exists a k-dimensional probability distribution $\pi_{n_1,...,n_k}$ on $\mathcal{S}^k$. We say that the joint distributions $\pi_{n_1,...,n_k}$ satisfy the Kolmogorov consistency conditions iff*

(a) for all $k, m \geq 1$ and for all $E_1, ..., E_k \subseteq \mathcal{S}$ we have

$$\pi_{n_1,...,n_{k+m}}(E_1, ..., E_k, \mathcal{S}, ..., \mathcal{S}) = \pi_{n_1,...,n_k}(E_1, ..., E_k) \,,$$

and

(b) for all $k \geq 1$ and any permutation σ of $(1, ..., k)$ we have

$$\pi_{n_1,...,n_k}(E_1, ..., E_k) = \pi_{n_{\sigma(1)},...,n_{\sigma(k)}}(E_{\sigma(1)}, ..., E_{\sigma(k)}) \,.$$

It is straightforward to show using (1.4) that, for a given $\mathcal{S}$, a given initial distribution π_0 on $\mathcal{S}$, and a given set of one-step transition probabilities P_{xy}, $x, y \in \mathcal{S}$, the Kolmogorov consistency conditions hold for the joint distributions $\pi_{n_1,...,n_k}$ of the random vectors $(X_{n_1}, ..., X_{n_k})$ for all $k \geq 1$ and all $0 \leq n_1 < \cdots < n_k$.

Theorem 1.2.2 (Kolmogorov Extension Theorem)**.** *Let $\mathcal{S}$ be a discrete space. Assume that for all $k \geq 1$ and $0 \leq n_1 < ... < n_k$ there exists a probability measure $\pi_{n_1,...,n_k}$ on $\mathcal{S}^k$ such that this family of probability measures $\{\pi_{n_1,...,n_k}\}$ satisfies the Kolmogorov consistency conditions (Definition 1.2.2). Then there exists a unique probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a collection of random variables $(X_n)_{n\geq 0}$ on this space such that for all $k \geq 0$ and $0 \leq n_1 < ... < n_k$ the probability distributions $\pi_{n_1,...,n_k}$ are the joint marginals of $X_{n_1}, ..., X_{n_k}$.*

Notes: (1) Theorem 1.2.2 guarantees the existence and uniqueness of an infinite-length stochastic process $(X_n)_{n\geq 0}$ for a given state space $\mathcal{S}$, a given initial distribution π_0 on $\mathcal{S}$,

[1] Andrey Nikolaevich Kolmogorov (1903-1987), Soviet mathematician. He laid the mathematical foundations of modern Probability theory.

a given set of one-step transition probabilities P_{xy}, $x, y \in \mathcal{S}$, and for which the Markov property holds. By Theorem 1.2.2, Markov chains as defined in Definition 1.2.3 exist.

(2) Usually, we will not know the space Ω explicitly, and in fact, won't need to know it. What we do need to know and work with are the finite-dimensional joint distributions of the random variables $(X_n)_{n\geq 0}$, from which we can compute probabilities for events defined for infinite-length trajectories by applying the *continuity property of probability* (see Lemma B.1.1).

(3) For intuition, we can always think of Ω as the so-called **canonical path space** $\mathcal{S}^{\mathbb{N}_0}$ which is the space of all infinte-length trajectories (i.e., sample paths) for the process $(X_n)_{n\geq 0}$. The random variables X_n are then projections onto the nth coordinate: If $\omega \in \mathcal{S}^{\mathbb{N}_0}$ with $\omega = (\omega_0, \omega_1, ...)$, then $X_n(\omega) = \omega_n$. Theorem 1.2.2 guarantees the existence and uniqueness of a probability measure $\mathbb{P}_{\text{path}}$ on $\mathcal{S}^{\mathbb{N}_0}$ (to be precise, on the induced σ-algebra on $\mathcal{S}^{\mathbb{N}_0}$) which is consistent with the finite-dimensional marginals computed in (1.4). More precisely, $\mathbb{P}_{\text{path}}$ is the *push-forward* measure to $\mathcal{S}^{\mathbb{N}_0}$ under $(X_n)_{n\geq 0}$ of the probability measure $\mathbb{P}$ on Ω.

(4) The canonical path space $\mathcal{S}^{\mathbb{N}_0}$ is *uncountable*.

We are now ready for the definition of a Markov chain $(X_n)_{n\geq 0}$.

Definition 1.2.3 (Markov chain)**.** *Let $\mathcal{S}$ be a finite or countably infinite set.*

- *A discrete-time stochastic process $(X_n)_{n\geq 0}$ with state space $\mathcal{S}$ is called a* **Markov chain** *if for all $n \geq 0$ and all states $x_0, ..., x_{n-1}, x, y \in \mathcal{S}$,*

$$\mathbb{P}(X_{n+1} = y \,|\, X_0 = x_0, ..., X_{n-1} = x_{n-1}, X_n = x) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) \tag{1.6}$$

  whenever both conditional probabilities are well-defined.

- *If, in addition, the conditional probabilities* (1.6) *do not depend on time n, we call the Markov chain* **time-homogeneous** *or simply homogeneous. We will then use the notation $P_{xy} = \mathbb{P}(X_1 = y \,|\, X_0 = x)$ for $x, y \in \mathcal{S}$ and call the P_{xy} the* **one-step transition probabilities** *for the homogeneous Markov chain.*

- *If $X_0 \sim \pi_0$, we call π_0 the* **initial distribution** *of the Markov chain.*

Note: Unless otherwise noted, we will always work with time-homogeneous Markov chains.

Since we will always work with a discrete state space $\mathcal{S}$, we can take $\mathcal{S} = \{0, 1, ..., N\}$, or in the infinite case, $\mathcal{S} = \mathbb{N}_0$. Using the natural ordering of the elements in $\mathcal{S}$, it will often be convenient to write the transition probabilities P_{xy} in matrix format. This results in the (finite or infinite) **one-step transition matrix**, or simply **transition matrix**,

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & \cdots & P_{0N} \\ P_{10} & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ P_{N0} & \cdots & \cdots & P_{NN} \end{pmatrix} \quad \text{or} \quad \mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & \cdots & \cdots \\ \vdots & \vdots & & \\ \vdots & \vdots & & \end{pmatrix}.$$

Definition 1.2.4. *A square matrix* $\mathbf{P}$ *is called a* **stochastic matrix**, *if*

- *all matrix entries are nonnegative, and*
- *each row sums to* 1.

The transition matrix $\mathbf{P}$ for a Markov chain is always a stochastic matrix.

Note: In working with an infinite transition matrix $\mathbf{P}$, we will use the same basic formal rules for matrix addition and matrix multiplication that hold for finite matrices. For example, if $\mathbf{P}$ and $\mathbf{P}'$ are two infinite transition matrices for infinite state space $\mathcal{S}$, then their product $\mathbf{PP}'$ is the matrix $\mathbf{P}''$ with entries $P''_{xy} = \sum_{z\in\mathcal{S}} P_{xz}P'_{zy}$. In general, matrix multiplication for infinite matrices is not associative. However, matrix multiplication for infinite *stochastic* matrices is associative (see Corollary A.4.4 in Appendix A).

Proposition 1.2.3. *Let* $(X_n)_{n\geq 0}$ *be a Markov chain on state space* $\mathcal{S}$. *Let* $i \geq 1$ *and* $x_i \in \mathcal{S}$. *Consider time* i *as representing the presence. Then conditional on the event* $\{X_i = x_i\}$, *the past and the future of the process are independent. That is, for all* $n > i$ *and for all* $x_0, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n \in \mathcal{S}$,

$$\begin{aligned} &\mathbb{P}(X_0 = x_0, ..., X_{i-1} = x_{i-1},\ X_{i+1} = x_{i+1}, ..., X_n = x_n \,|\, X_i = x_i) \\ &= \mathbb{P}(X_0 = x_0, ..., X_{i-1} = x_{i-1} \,|\, X_i = x_i)\, \mathbb{P}(X_{i+1} = x_{i+1}, ..., X_n = x_n \,|\, X_i = x_i), \end{aligned}$$

provided all conditional probabilities are defined.

Proof. The statement follows from the familiar fact

$$\mathbb{P}(E \cap F \,|\, G) = \mathbb{P}(E \,|\, F \cap G)\, \mathbb{P}(F \,|\, G)$$

for conditional probabilities for events E, F, and G, and from the Markov property. □

Definition 1.2.5. *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$. A state $x \in \mathcal{S}$ is called an* **absorbing state** *if $P_{xx} = 1$.*

Example 1.2.1 (The 2-state chain)**.** The simplest example of a Markov chain is the 2-state chain for which we can take $\mathcal{S} = \{0, 1\}$. We write the four one-step transition probabilities $P_{00}, P_{01}, P_{1,0}, P_{11}$ as entries in a (2×2)-transition matrix

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

To avoid trivial cases, we assume $a, b \in (0, 1)$. We can represent the transition mechanism, i.e., the information given by $\mathbf{P}$, as a *directed, weighted graph* whose vertex set is the state space $\mathcal{S}$ and for which a weight assigned to a directed edge is the corresponding one-step transition probability. If a one-step transition probability $P_{xy} = 0$, the graph will not have a directed edge leading from x to y. We call such a graph the **transition graph** for the Markov chain. The transition graph for the 2-state chain is shown in Figure 1.1.

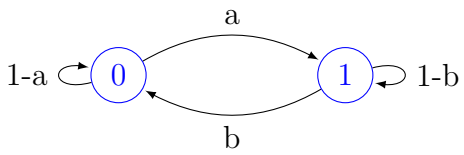


Figure 1.1: The 2-state chain

Here Ω can be identified with the set of infinite binary sequences. □

Example 1.2.2 (A non-Markov chain)**.** An box contains one red ball and one green ball. At each time step, one ball is drawn uniformly at random, its color noted and then, together with one additional ball of the *same* color, put back into the box. Define the following stochastic process $(Y_n)_{n\geq 1}$ with state space $\mathcal{S} = \{0, 1\}$: Let $Y_n = 1$ if the nth ball drawn is red, and let $Y_n = 0$ if the nth ball drawn is green. The process $(Y_n)_{n\geq 1}$ is not a Markov chain. Indeed, we see that

$$\mathbb{P}(Y_3 = 1 \,|\, Y_1 = 1, Y_2 = 1) = \frac{(1/2)(2/3)(3/4)}{(1/2)(2/3)} = \frac{3}{4}$$

and

$$P(Y_3 = 1 \,|\, Y_1 = 0, Y_2 = 1) = \frac{(1/2)(1/3)(2/4)}{(1/2)(1/3)} = \frac{1}{2}$$

are not equal, and so the Markov property (1.3) does not hold for $(Y_n)_{n\geq 1}$. This process a special case of *Pólya's Urn* which we will introduce in more detail in Section 1.5. □

Example 1.2.3 (A queuing model)**.** Consider people waiting in line at the post office waiting for service. At each time interval, the probability that somebody new joins the queue is p, and the probability that somebody finishes their service and leaves the queue is q. All people arrive and leave the queue independently of each other. We assume $p, q \in (0,1)$, and the queue starts with x_0 persons in line. Let X_n be the number of people in line at time n. We have $\mathbb{P}(X_0 = x_0) = 1$. The process $(X_n)_{n\geq 0}$ is a Markov chain with state space is $\mathcal{S} = \mathbb{N}_0$ and one-step transition probabilities $P_{00} = 1-p\,,\ P_{01} = p$, and

$$P_{xx} = (1-p)(1-q) + pq\,, \quad P_{x,x+1} = p(1-q)\,, \quad P_{x,x-1} = q(1-p)$$

for $x \geq 1$, and 0 otherwise. The transition graph is shown in Figure 1.3

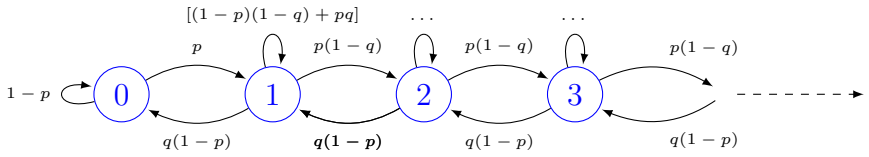


Figure 1.2

We verify that, as expected,

$$\sum_{y\in\mathcal{S}} P_{0y} = 1 \quad \text{and} \quad \sum_{y\in\mathcal{S}} P_{xy} = P_{xx} + P_{x,x+1} + P_{x,x-1} = 1 \ \text{ for } x \geq 1\,.$$

This queuing model is an example of a so-called *birth/death chain* (a Markov chain for which, at each step, the state of the system can only change by at most 1) which we will introduce in more detail in Section 1.5. □

1.3 Basic computations for Markov chains

Consider a Markov chain $(X_n)_{n\geq 0}$ on (finite or countably infinite) state space $\mathcal{S}$ with initial distribution π_0 and transition matrix $\mathbf{P}$. Here we show how various probabilities associated with $(X_n)_{n\geq 0}$ can be computed from π_0 and $\mathbf{P}$. First, we will compute the distributions of the random variables X_n, $n \geq 1$.

Note: We will denote the distribution of X_n by π_n. By this we mean $\mathbb{P}(X_n = x) = \pi_n(x)$ for $x \in \mathcal{S}$. It will be convenient to consider π_n as a row vector $\pi_n = (\pi_n(x_1), \pi_n(x_2), ...)$ with respect to some natural ordering of the elements $x_1, x_2, ...$ of $\mathcal{S}$. We will denote the transpose of π_n by π_n^{t}.

Computing π_1: By the law of total probability, we have

$$\pi_1(y) = \mathbb{P}(X_1 = y) = \sum_{x \in \mathcal{S}} \pi_0(x) P_{xy}$$

which is the row vector π_0 multiplied by the yth column vector of the matrix $\mathbf{P}$. Hence we get

$$\pi_1 = \pi_0 \, \mathbf{P} \, .$$

Computing π_2: Applying again the law of total probability, we get

$$\pi_2(y) = \mathbb{P}(X_2 = y) = \sum_{x \in \mathcal{S}} \pi_0(x) \mathbb{P}(X_2 = y \,|\, X_0 = x) \, , \tag{1.7}$$

or, alternatively,

$$\pi_2(y) = \sum_{z \in \mathcal{S}} \pi_1(z) P_{zy}$$

which we can rewrite as

$$\pi_2(y) = \sum_{z \in \mathcal{S}} \left(\sum_{x \in \mathcal{S}} \pi_0(x) P_{xz} \right) P_{zy} = \sum_{x \in \mathcal{S}} \pi_0(x) \left(\sum_{z \in \mathcal{S}} P_{xz} \, P_{zy} \right) \, . \tag{1.8}$$

The change of summation in the second equality of (1.8) is justified for infinite state space $\mathcal{S}$ because of the absolute convergence of the double sum. Comparing (1.7) and (1.8) (and by taking π_0 to be unit mass at x), we get for the conditional probability $\mathbb{P}(X_2 = y \,|\, X_0 = x)$ which we call the 2-*step transition probability from* x *to* y,

$$\mathbb{P}(X_2 = y \,|\, X_0 = x) = \sum_{z \in \mathcal{S}} P_{xz} \, P_{zy} = [\mathbf{P}^2]_{x,y}$$

which is the (x, y)-entry in the square $\mathbf{P}^2$ of the one-step transition matrix $\mathbf{P}$. We have

$$\pi_2 = \pi_0 \, \mathbf{P}^2 \, ,$$

and by induction,

$$\pi_n = \pi_0 \, \mathbf{P}^n \qquad \text{for all } n \geq 1 \, . \tag{1.9}$$

From (1.9) (and by taking π_0 to be unit mass at x), we get for all $n \geq 1$,

$$\mathbb{P}(X_n = y \,|\, X_0 = x) = [\mathbf{P}^n]_{x,y}$$

and by homogeneity,

$$\mathbb{P}(X_{m+n} = y \,|\, X_m = x) = [\mathbf{P}^n]_{x,y} \qquad \text{for all } m \geq 0 \, .$$

Definition 1.3.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix $\mathbf{P}$. For $n \geq 1$, we call the matrix $\mathbf{P}^n$ the* **n-step transition matrix** *for the chain. The entries of $\mathbf{P}^n$ are called the* **n-step transition probabilities**. *For all $x, y \in \mathcal{S}$, we denote the n-step transition probability from x to y by P_{xy}^n.*

Attention: $P_{xy}^n \neq (P_{xy})^n$. The n-step transition probability P_{xy}^n is the (x, y)-entry in the matrix power $\mathbf{P}^n$.

Notes: (1) For any $m, n \geq 1$ we have

$$\mathbf{P}^{m+n} = \mathbf{P}^m \, \mathbf{P}^n \,,$$

and hence

$$P_{xy}^{m+n} = \sum_{z \in \mathcal{S}} P_{xz}^m \, P_{zy}^n \,.$$

(2) If x is an absorbing state, then $P_{xx}^n = 1$ for all $n \geq 1$.

Example 1.3.1. Consider a Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{S} = \{1, 2, 3\}$. Let

$$\mathbf{P} = \begin{pmatrix} 0 & 0.7 & 0.3 \\ 0.4 & 0.2 & 0.4 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

and assume the Markov chain starts in state 2. Compute

(a) $\mathbb{P}(X_1 = 3, X_3 = 2, X_6 = 3)$
(b) $\mathbb{P}(X_5 = 2 \,|\, X_2 = 1, X_3 = 3)$
(c) $\mathbb{P}(X_3 = 2 \,|\, X_2 = 3, X_4 = 1)$
(d) $\mathbb{P}(X_2 = 3, X_4 = 1 \,|\, X_3 = 2)$.

The transition graph for this chain is shown in Figure **??**.
The 2-step and 3-step transition matrices are

$$\mathbf{P}^2 = \begin{pmatrix} 0.43 & 0.14 & 0.43 \\ 0.28 & 0.32 & 0.4 \\ 0.25 & 0.35 & 0.4 \end{pmatrix} \qquad \mathbf{P}^3 = \begin{pmatrix} 0.271 & 0.329 & 0.4 \\ 0.328 & 0.26 & 0.412 \\ 0.34 & 0.245 & 0.415 \end{pmatrix}$$

(use a matrix calculator).
Answers:

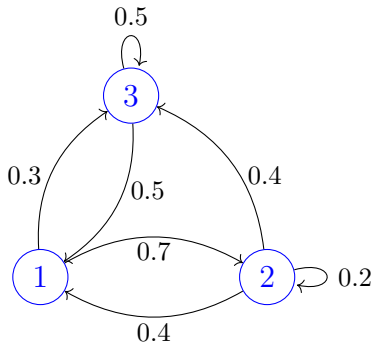


Figure 1.3

(a) $\mathbb{P}(X_1 = 3, X_3 = 2, X_6 = 3) = \left(\sum_{x=1}^{3} \pi_0(x)\, P_{x3}\right) P^2_{3,2}\, P^3_{2,3} = (1{\cdot}0.4)\,(0.35)\,(0.412) \approx 0.06$

(b) $\mathbb{P}(X_5 = 2 \,|\, X_2 = 1, X_3 = 3) = \mathbb{P}(X_5 = 2 \,|\, X_3 = 3) = P^2_{3,2}$ by the Markov property. Indeed,

$$\mathbb{P}(X_5 = 2 \,|\, X_2 = 1, X_3 = 3) = \frac{\mathbb{P}(X_5 = 2, X_2 = 1, X_3 = 3)}{\mathbb{P}(X_2 = 1, X_3 = 3)} = \frac{\pi_2(1)\, P_{1,3}\, P^2_{3,2}}{\pi_2(1)\, P_{1,3}} = P^2_{3,2}\,.$$

Hence $\mathbb{P}(X_5 = 2 \,|\, X_2 = 1, X_3 = 3) = 0.35$.

(c)

$$\mathbb{P}(X_3 = 2 \,|\, X_2 = 3, X_4 = 1) = \frac{\mathbb{P}(X_2 = 3, X_3 = 2, X_4 = 1)}{\mathbb{P}(X_2 = 3, X_4 = 1)} = \frac{\pi_2(3)\, P_{3,2}\, P_{2,1}}{\pi_2(3)\, P^2_{3,1}} = \frac{0 \cdot 0.4}{0.25} = 0\,.$$

(d) By Proposition 1.2.3,

$$\mathbb{P}(X_2 = 3, X_4 = 1 \,|\, X_3 = 2) = \mathbb{P}(X_4 = 1 \,|\, X_3 = 2)\, \mathbb{P}(X_2 = 3 \,|\, X_3 = 2)\,.$$

So

$$\mathbb{P}(X_2 = 3, X_4 = 1 \,|\, X_3 = 2) = P_{2,1}\, \frac{\mathbb{P}(X_2 = 3, X_3 = 2)}{\pi_3(2)} = P_{2,1}\, \frac{\pi_2(3)\, P_{3,2}}{\pi_3(2)}\,.$$

We compute $\pi_2(3) = 0.4$ and $\pi_3(2) = 0.26$. Hence we get

$$P(X_2 = 3, X_4 = 1 \,|\, X_3 = 2) = \frac{(0.4)\,(0.4) \cdot 0}{0.26} = 0\,.$$

□

In most cases it is too difficult to compute an explicit formula for the n-step transition matrix $\mathbf{P}^n$ for a given Markov chain (unless $\mathbf{P}$ is diagonalizable). But for the simplest case, the 2-state chain, we can find explicit formulas for the n-step transition probabilities without too much work. We will show this in the next example.

Example 1.3.2. Recall the 2-state chain from Example 1.2.1. We take $\mathcal{S} = \{0, 1\}$. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

for which we assume $a, b \in (0, 1)$. Perhaps the easiest approach to compute higher powers $\mathbf{P}^n$ of the transition matrix is via diagonalization (if $\mathbf{P}$ is in fact diagonalizable). Since $\mathbf{P}$ is a stochastic matrix, the column vector $(1, 1)^{\mathrm{t}}$ is a right eigenvector corresponding to eigenvalue $\lambda_1 = 1$. Since $\mathrm{trace}(\mathbf{P}) = 2 - a - b$, we know that $\lambda_2 = 1 - a - b$ is also an eigenvalue. Per our assumptions on a and b, we have $-1 < \lambda_2 < 1$. So $\mathbf{P}$ is diagonalizable. There exists an invertible 2×2-matrix $\mathbf{U}$ such that

$$\mathbf{P} = \mathbf{U} \begin{pmatrix} 1 & 0 \\ 0 & (1-a-b) \end{pmatrix} \mathbf{U}^{-1}.$$

This implies

$$\mathbf{P}^n = \mathbf{U} \begin{pmatrix} 1 & 0 \\ 0 & (1-a-b)^n \end{pmatrix} \mathbf{U}^{-1}$$

for all $n \geq 1$. Hence each entry P_{ij}^n, $i, j \in \{0, 1\}$, of the matrix $\mathbf{P}^n$ is of the form $\alpha_{ij} + \beta_{ij}(1-a-b)^n$ for some constants α_{ij}, β_{ij} that do not depend on n.
Let us first compute α_{00} and β_{00}. Since $P_{00}^0 = 1$ and $P_{00}^1 = P_{00} = 1 - a$, we get the following system of equations for computing α_{00} and β_{00}:

$$\begin{aligned} 1 &= \alpha_{00} + \beta_{00} \\ 1-a &= \alpha_{00} + \beta_{00}(1-a-b) \end{aligned}$$

which yields

$$\alpha_{00} = \frac{b}{a+b} \quad \text{and} \quad \beta_{00} = \frac{a}{a+b}.$$

Thus $P_{00}^n = \frac{b}{a+b} + \frac{a}{a+b}(1-a-b)^n$. Since $P_{01}^n = 1 - P_{00}^n$, we also get $P_{01}^n = \frac{a}{a+b} - \frac{a}{a+b}(1-a-b)^n$.

Similarly, we set $P_{10}^n = \alpha_{10} + \beta_{10}(1-a-b)^n$, and for $n = 0$ and $n = 1$, get the system of equations

$$\begin{aligned} 0 &= \alpha_{10} + \beta_{10} \\ b &= \alpha_{10} + \beta_{10}(1-a-b) \end{aligned}$$

which yields

$$\alpha_{10} = \frac{b}{a+b} \quad \text{and} \quad \beta_{10} = \frac{-b}{a+b}.$$

Thus $P_{10}^n = \frac{b}{a+b} - \frac{b}{a+b}(1-a-b)^n$ and $P_{11}^n = \frac{a}{a+b} + \frac{b}{a+b}(1-a-b)^n$.

Altogether, we have computed the n-step transition matrix

$$\mathbf{P}^n = \begin{pmatrix} \frac{b}{a+b} + \frac{a}{a+b}(1-a-b)^n & \frac{a}{a+b} - \frac{a}{a+b}(1-a-b)^n \\ \frac{b}{a+b} - \frac{b}{a+b}(1-a-b)^n & \frac{a}{a+b} + \frac{b}{a+b}(1-a-b)^n \end{pmatrix}.$$

An interesting situation arises: Since $|1-a-b| < 1$, we have

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}.$$

Hence for *any* initial distribution π_0 on $\mathcal{S}$,

$$\pi_0 \mathbf{P}^n = \pi_n \xrightarrow{n\to\infty} (\frac{b}{a+b}, \frac{a}{a+b}), \tag{1.10}$$

and so in the long run, the process "settles down" in the sense that its distribution approaches a unique *limiting distribution*, here $(\frac{b}{a+b}, \frac{a}{a+b})$. We will discuss conditions under which, more generally, such *convergence* occurs in Section 3.2. □

Definition 1.3.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$. A probability distribution π on $\mathcal{S}$ is called a* **stationary** *or* **invariant distribution** *if*

$$\pi(y) = \sum_{x\in\mathcal{S}} \pi(x) P_{xy} \quad \text{for all } y \in \mathcal{S}, \tag{1.11}$$

or equivalently (with π as a row vector), if

$$\pi\mathbf{P} = \pi.$$

If π is a stationary distribution, It follows by induction that $\pi\mathbf{P}^n = \pi$ for all $n \geq 1$. Thus if π is the initial distribution for a Markov chain, then at each future time n, the Markov chain will be in distribution π. We then say the Markov chain is *in stationarity*. A stationary distribution π represents a sort of equilibrium situation for the Markov chain.

Example 1.3.3. We return to the 2-state chain from Example 1.3.2. Let

$$\pi = (\frac{b}{a+b}, \frac{a}{a+b}).$$

We directly verify that

$$\pi\mathbf{P} = \pi,$$

and so π is a stationary distribution. Notice that by (1.10), we have $\lim_{n\to\infty} \pi_n = \pi$ for any initial distribution π_0. In the long run, the Markov chain *converges to stationarity*. We will discuss this phenomenon in detail in Section 3.2. □

1.4 Strong Markov Property

Consider a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S}$ and with transition probabilities P_{xy} for $x, y \in \mathcal{S}$. Usually, we observe the Markov chain from time $n = 0$ onwards. However, the Markov property guarantees that for *any fixed time* $n_0 > 0$, the process $(X_{n_0+n})_{n\geq 0}$ is also a Markov chain and has the same transition probabilities P_{xy} as $(X_n)_{n\geq 0}$. Moreover, the process $(X_{n_0+n})_{n\geq 0}$ is independent of the past history $X_0, X_1, ..., X_{n_0-1}$. Loosely speaking, for any fixed time $n_0 > 0$, and conditional on the event $\{X_{n_0} = x\}$, the Markov chain probabilistically "starts afresh" in state x.

What if, instead of conditioning on the event $\{X_{n_0} = x\}$, we waited until the Markov chain hits x at some *random time* T, and then observed the process from that time onwards? It turns out that under certain conditions on the random time T, the process $(X_{T+n})_{n\geq 0}$ is again a Markov chain and has the same transition probabilities as the original chain. This property is called the **strong Markov property**. In the following we will make this precise.

Definition 1.4.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain. A* **stopping time** *T is a random variable taking values in $\mathbb{N}_0 \cup \{\infty\}$ such that for all $m \in \mathbb{N}_0$, the event $\{T = m\}$ can be determined from the values of $X_0, X_1, ..., X_m$.*

By watching the process from time 0 onwards, we know at any time whether or not a stopping time T has occurred. At time m, we do not need information about the future of the process $X_{m+1}, X_{m+2}, ...$ to determine whether or not T has occurred at time m. Note that this is equivalent to saying that for every $m \in \mathbb{N}_0$, the indicator random variable $\mathbb{1}_{\{T=m\}}$ is a deterministic function of the random variables $X_0, X_1, ...X_m$.

Example 1.4.1. (a) The *first hitting time* $T^x = \min\{n \geq 1 : X_n = x\}$ of state x is a stopping time because

$$\{T^x = n\} = \{X_1 \neq x, X_2 \neq x, ..., X_{n-1} \neq x, X_n = x\}.$$

Similarly, the *second hitting time* $T^x_{(2)} = \min\{n > T^x : X_n = x\}$, and defined analogously, the *third hitting time* $T^x_{(3)}$ and so on, are stopping times as well.

(b) More generally, the *first hitting time* T^A of a set $A \subset \mathcal{S}$ is a stopping time because

$$\{T^A = n\} = \{X_1 \notin A, X_2 \notin A, ..., X_{n-1} \notin A, X_n \in A\}.$$

(c) The *last exit time* L^A defined by

$$L^A = \max\{n \geq 0 : X_n \in A\}$$

is **not** a stopping time because the event $\{L^A = n\}$ depends on whether or not any of the random variables X_{n+m} for $m \geq 1$ take values in A. □

Theorem 1.4.1 (Strong Markov Property)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with transition probabilities P_{xy} for $x, y \in \mathcal{S}$ and let T be a stopping time. Then conditional on the event $\{T < \infty\}$, the process $(X_{T+n})_{n\geq 0}$ is a Markov chain with the same transition probabilities P_{xy}, and it is independent of the random variables $X_0, X_1, ..., X_{T-1}$.*

Proof. We will show that, conditional on $\{T < \infty\}$ and $\{X_T = x\}$, the sequence of random variables $(X_{T+n})_{n\geq 0}$ forms a Markov chain that is independent of $X_0, X_1, ..., X_{T-1}$ and that proceeds according to the same original transition probabilities, by proving

$$\begin{aligned}\mathbb{P}(X_{T+1} = y_1, ..., X_{T+n} = y_n \,|\, X_k = u_k \text{ for } 0 \leq k < T,\ X_T = x,\ T < \infty) \\ = \ P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n}\,. \end{aligned} \tag{1.12}$$

First we consider the joint probability

$$\mathbb{P}(X_{T+1} = y_1, ..., X_{T+n} = y_n;\ X_k = u_k \text{ for } 0 \leq k < T,\ X_T = x,\ T < \infty)$$

and, using the Law of total probability, rewrite it as the sum

$$\sum_{s=0}^{\infty} \mathbb{P}(X_{T+1} = y_1, ..., X_{T+n} = y_n;\ X_k = u_k \text{ for } 0 \leq k < T,\ X_T = x,\ T = s)$$

$$= \sum_{s=0}^{\infty} \mathbb{P}(X_{s+1} = y_1, ..., X_{s+n} = y_n;\ X_k = u_k \text{ for } 0 \leq k < s,\ X_s = x,\ T = s)\,.$$

Using conditioning, we rewrite the last sum as

$$\sum_{s=0}^{\infty} [\mathbb{P}(T = s \,|\, X_{s+1} = y_1, ..., X_{s+n} = y_n;\ X_k = u_k \text{ for } 0 \leq k < s,\ X_s = x)$$

$$\times \quad \mathbb{P}(X_{s+1} = y_1, ..., X_{s+n} = y_n;\ X_k = u_k \text{ for } 0 \leq k < s,\ X_s = x)]\ ,$$

and, after further conditioning, as

$$\sum_{s=0}^{\infty} [\mathbb{P}(T = s \,|\, X_{s+1} = y_1, ..., X_{s+n} = y_n;\ X_k = u_k \text{ for } 0 \leq k < s,\ X_s = x)$$

$$\times \quad \mathbb{P}(X_{s+1} = y_1, ..., X_{s+n} = y_n \,|\, X_k = u_k \text{ for } 0 \leq k < s,\ X_s = x)$$

$$\times \quad \mathbb{P}(X_k = u_k \text{ for } 0 \le k < s,\ X_s = x)] \ . \tag{1.13}$$

We now take a closer look at the first two factors in each summand in (1.13). Since T is a *stopping time*, the event $\{T = s\}$ is independent of any event defined by the random variables X_{s+k} for $k \ge 1$. As a consequence, the first factor in the sum (1.13) is equal to

$$\mathbb{P}(T = s \,|\, X_k = u_k \text{ for } 0 \le k < s,\ X_s = x) \, .$$

By the Markov property, the second factor in the sum (1.13) is equal to

$$P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n} \, .$$

Thus the sum (1.13) becomes

$$P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n} \sum_{s=0}^{\infty} \mathbb{P}(T = s \,|\, X_k = u_k \text{ for } 0 \le k < s,\ X_s = x)\, \mathbb{P}(X_k = u_k \text{ for } 0 \le k < s,\ X_s = x)$$

$$= P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n} \sum_{s=0}^{\infty} \mathbb{P}(T = s, X_k = u_k \text{ for } 0 \le k < s,\ X_s = x)$$

$$= P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n} \mathbb{P}(T < \infty, X_k = u_k \text{ for } 0 \le k < T,\ X_T = x) \, .$$

Altogether, we get

$$\mathbb{P}(X_{T+1} = y_1, ..., X_{T+n} = y_n \,|\, X_k = u_k \text{ for } 0 \le k < T,\ X_T = x,\ T < \infty)$$

$$= \frac{P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n} \mathbb{P}(T < \infty, X_k = u_k \text{ for } 0 \le k < T,\ X_T = x)}{\mathbb{P}(T < \infty, X_k = u_k \text{ for } 0 \le k < T,\ X_T = x)}$$

$$= P_{xy_1} P_{y_1 y_2} \cdots P_{y_{n-1} y_n}$$

which proves (1.12). This completes the proof of Theorem 1.4.1. □

1.5 Examples of Markov Chains

(1) Markov chains derived from a sequence of i.i.d. random variables

Example 1.5.1. Let X be a discrete random variable taking values in $\mathbb{N}_0$ and $\pi = (p_0, p_1, p_2, ...)$ its distribution. Consider a sequence $(X_n)_{n \ge 0}$ of i.i.d. random variables

with $X_n \sim X$. The sequence $(X_n)_{n\geq 0}$ is a Markov chain. Its transition matrix $\mathbf{P}$ has equal rows and is of the form

$$\mathbf{P} = \begin{pmatrix} p_0 & p_1 & p_2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \\ \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

Conversely, if $(Y_n)_{n\geq 0}$ is a Markov chain whose transition matrix has *equal rows*, then the random variables Y_n, $n \geq 0$, are i.i.d. and their distribution is given by any one of the (equal) rows of the transition matrix. □

Example 1.5.2 (Successive maxima)**.** Consider the above defined sequence $(X_n)_{n\geq 0}$ of i.i.d. random variables. We define the new Markov chain $(M_n)_{n\geq 0}$ of successive maxima by

$$M_n = \max\{X_0, X_1, ..., X_n\} \quad \text{for } n \geq 0\,.$$

Note that $M_{n+1} = \max\{M_n, X_{n+1}\}$. From this we get the transition probabilities

$$P_{xy} = \begin{cases} p_y & \text{if } y > x \\ \displaystyle\sum_{0\leq i\leq x} p_i & \text{if } y = x \\ 0 & \text{if } y < x\,. \end{cases}$$

□

Example 1.5.3 (Random walk on $\mathbb{Z}$)**.** Let Y be a discrete random variable taking values in $\mathbb{Z}$ and $\mu = (..., \mu(-1), \mu(0), \mu(1), ...)$ its distribution. Let $(Y_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with $Y_k \sim Y$. We define the process $(S_n)_{n\geq 0}$ of *successive partial sums* by

$$S_n = \sum_{k=1}^{n} Y_k \qquad \text{for } n \geq 1$$

and $S_0 = 0$. The process $(S_n)_{n\geq 0}$ is called a **random walk** on $\mathbb{Z}$ with **step distribution** μ. At each time interval, the walk takes a step according to μ, and all steps are chosen independently. The transition probabilities are

$$\mathbb{P}(S_{n+1} = y \,|\, S_n = x) = P_{xy} = \mu(y - x) \quad \text{for all } x, y \in \mathbb{Z}\,.$$

We have $S_1 \sim \mu$. The distribution of $S_2 = Y_1 + Y_2$ is the **convolution** $\mu \star \mu$ defined by

$$\mu \star \mu(y) = \sum_{x\in\mathbb{Z}} \mu(x)\mu(y - x) \quad \text{for all } y \in \mathbb{Z}\,.$$

We write $\mu^{\star 2} = \mu \star \mu$. From this we derive the distribution of $S_3 = S_2 + Y_3$. It is $\mu^{\star 3}$ and computed by

$$\mu^{\star 3}(y) = \mu^{\star 2} \star \mu(y) = \sum_{x \in \mathbb{Z}} \mu^{\star 2}(x)\mu(y-x) \quad \text{for all } y \in \mathbb{Z}\,.$$

By induction we get $S_n \sim \mu^{\star n}$ where

$$\mu^{\star n}(y) = \sum_{x \in \mathbb{Z}} \mu^{\star (n-1)}(x)\mu(y-x) \quad \text{for all } y \in \mathbb{Z}\,.$$

Thus the n-step transition probabilities for a random walk on $\mathbb{Z}$ are

$$P^n_{xy} = \mu^{\star n}(y-x) \quad \text{for all } x, y \in \mathbb{Z}\,.$$

□

The following example is a special case of Example 1.5.3.

Example 1.5.4 (Simple random walk on $\mathbb{Z}$)**.** Let $p \in (0,1)$. Random walk $(S_n)_{n \geq 0}$ with step distribution μ defined by $\mu(1) = p$ and $\mu(-1) = 1 - p$ is called **simple random walk on** $\mathbb{Z}$. We can best visualize trajectories of simple random walk on $\mathbb{Z}$ by plotting the location of the walk against time (adding connecting line segments). Figure 1.4 shows a sample trajectory of finite length.

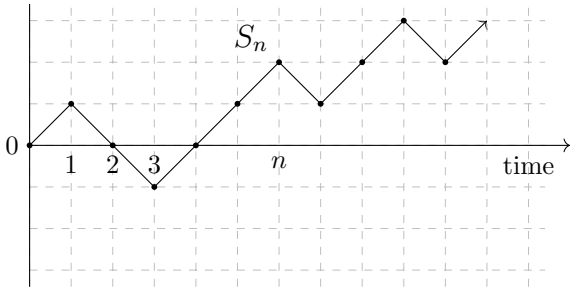


Figure 1.4: Simple random walk on $\mathbb{Z}$

□

We discuss simple random walk on $\mathbb{Z}$ in detail in Chapter 4.

(2) Birth/death chains

A **birth/death chain** is a Markov chain whose state space is a set of consecutive integers and which can only possibly move from x to $x+1$ (a *birth*), or from x to $x-1$ (a *death*),

or stay in place in one step. Usually, the state space $\mathcal{S}$ will be either $\mathbb{N}_0$ or $\mathbb{Z}$, or for the finite-state case, $\{0, 1, ..., N\}$. The transition probabilities are

$$P_{xy} = \begin{cases} q_x & \text{if } y = x - 1 \\ p_x & \text{if } y = x + 1 \\ r_x & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

with $p_x + q_x + r_x = 1$ for all $x \in \mathcal{S}$. A transition graph is shown in Figure 1.5.

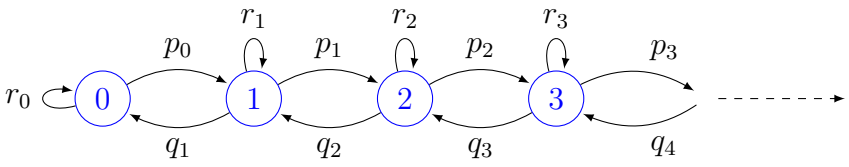


Figure 1.5: Birth/death chain

Birth/death chains frequently arise as models for real-life processes. Due to their relatively simple structure, they can often be analyzed in detail.

(3) Random walks on graphs

A **graph** $G(V, E)$ consists of a finite or countably infinite **vertex set** V and an **edge set** E. The edge set E consists of unordered pairs $\{v, w\}$ of vertices $v, w \in V$ with $v \neq w$. We say two vertices v and w are **neighbors** if $\{v, w\} \in E$. Graphically, this means that the vertices v and w are joined by a line segment. If $\{v, w\} \in E$, we write $v \sim w$. The **degree** $\deg(v)$ of a vertex v is defined as the number of neighbors of v.

Simple random walk on $G(V, E)$ is a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S} = V$ and transition probabilities

$$P_{vw} = \begin{cases} 1/\deg(v) & \text{if } v \sim w \\ 0 & \text{otherwise}\,. \end{cases}$$

At each step, the Markov chain chooses its next location (vertex) uniformly at random from the neighbors of its current location.

Example 1.5.5. Consider simple random walk on the following graph $G(V, E)$ in Figure 1.6. The state space is $V = \{1, 2, 3, 4, 5\}$ and the transition matrix is

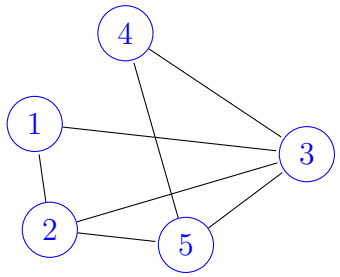


Figure 1.6

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix} .$$

□

Example 1.5.6 (Simple random walk on the hypercube $\mathbb{Z}_2^k$)**.** Here the vertex set V consists of the vertices of a unit cube in $\mathbb{R}^k$. Thus V can be identified with the set of binary k-tuples

$$V = \{(x_1, ..., x_k) : x_i \in \{0, 1\}, 1 \leq i \leq k\} .$$

Figure 1.7 shows a picture of $\mathbb{Z}_2^3$.

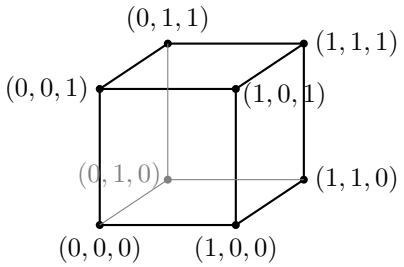


Figure 1.7: The hypercube $\mathbb{Z}_2^k$ for $k = 3$

There are k edges emanating from each vertex. Assume the walk is currently at vertex $v = (x_1, ..., x_k)$. For the next step, choose uniformly at random an index j from $\{1, ..., k\}$. If $x_j = 0$, switch it to 1. If $x_j = 1$, switch it to 0. Thus with each step, exactly one of the entries in the current state, the binary k-tuple v, changes. This is nearest neighbor random walk on the k-hypercube. □

Random walk on a weighted graph. A graph $G(V, E)$ may have a positive **edge weight** $C(v, w)$ associated with each edge $\{v, w\} \in E$. If this is the case, the transition probabilities are proportional to the given edge weights in the following way. Define $C(v) = \sum_{w:w\sim v} C(v, w)$. The transition probabilities for random walk on a **weighted graph** $G(V, E)$ with weight function $C : E \to \mathbb{R}^+$ are defined by

$$P_{vw} = \begin{cases} C(v,w)/C(v) & \text{if } w \sim v \\ 0 & \text{otherwise}\,. \end{cases}$$

Notice that *simple* random walk is a special case with $C(v, w) = 1$ for all edges $\{v, w\} \in E$.

Example 1.5.7. Consider random walk on the weighted graph shown in Figure 1.8.

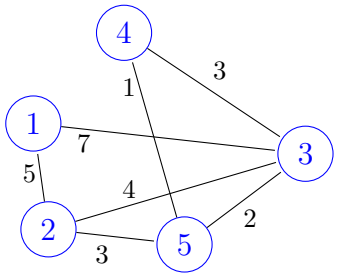


Figure 1.8

The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{5}{12} & \frac{7}{12} & 0 & 0 \\ \frac{5}{12} & 0 & \frac{1}{3} & 0 & \frac{1}{4} \\ \frac{7}{16} & \frac{1}{4} & 0 & \frac{3}{16} & \frac{1}{8} \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{6} & 0 \end{pmatrix}.$$

□

Note: In earlier examples, we have introduced the *transition graph* for a Markov chain. It is a *directed*, weighted graph, and every Markov chain has an associated transition graph. Random walks on weighted graphs, that is, random walks on graphs with *undirected edges*, constitute a more special category of Markov chains. Many (but by no means all) Markov chains can be interpreted as a random walk on an undirected weighted graph which, if applicable, is often a useful viewpoint for computations. We will discuss random walks on weighted graphs in detail in Chapter 8.

(4) Urn models

Urn models have a long history as models for real-world processes such as they arise in theoretical physics, statistics, population genetics (see [16]). Their set-up always involves a number of urns (boxes) and balls (often of different or changing color) and a prescribed scheme by which balls are drawn and returned into urns. When a ball is drawn, it is assumed to be drawn uniformly at random from all balls in the urn, regardless of color. The following questions often arise in connection with urn models:

1. What is the distribution of balls of a certain color in a specified urn (or urns) after n steps?
2. What is the distribution of the waiting time until a specified condition of the system occurs?
3. What are the asymptotic properties of the system? Will the system settle down in the long-run to a predictable state?

We introduce several classical urn models.

Ehrenfest chain:

This is a model of gas diffusion through a porous membrane and was suggested by physicists Paul and Tatiana Ehrenfest in 1907. Consider two boxes and N indistinguishable balls that are distributed between Box 1 and Box 2. At each step, one of the balls is chosen uniformly at random from the N balls and moved to the opposite box. See Figure 1.9. The number of balls in Box 1 describes the state of the system, and so the state space is $\mathcal{S} = \{0, 1, ..., N\}$. The transition probabilities are

$$P_{x,x-1} = \frac{x}{N} \text{ for } 1 \le x \le N, \quad P_{x,x+1} = \frac{N-x}{N} \text{ for } 0 \le x \le N-1, \quad \text{and 0 otherwise}.$$

The Ehrenfest chain is a birth/death chain with "pull towards the center", as the transition probabilities that lead towards the center state $N/2$ (or the nearest integer to $N/2$) grow with increasing distance from the center state. We can verify that the following equations hold:

$$\binom{N}{x} P_{x,x-1} = \binom{N}{x-1} P_{x-1,x}\,,$$

that is,

$$\binom{N}{x} \frac{x}{N} = \binom{N}{x-1} \frac{N-x+1}{N}. \tag{1.14}$$

From this, and since $\sum_{x=0}^{N} \binom{N}{x} = 2^N$, we conclude that $\pi \sim \text{Bin}(N, \frac{1}{2})$, so

$$\pi(x) = \binom{N}{x} \frac{1}{2^N} \quad \text{for } x = 0, 1, ..., N,$$

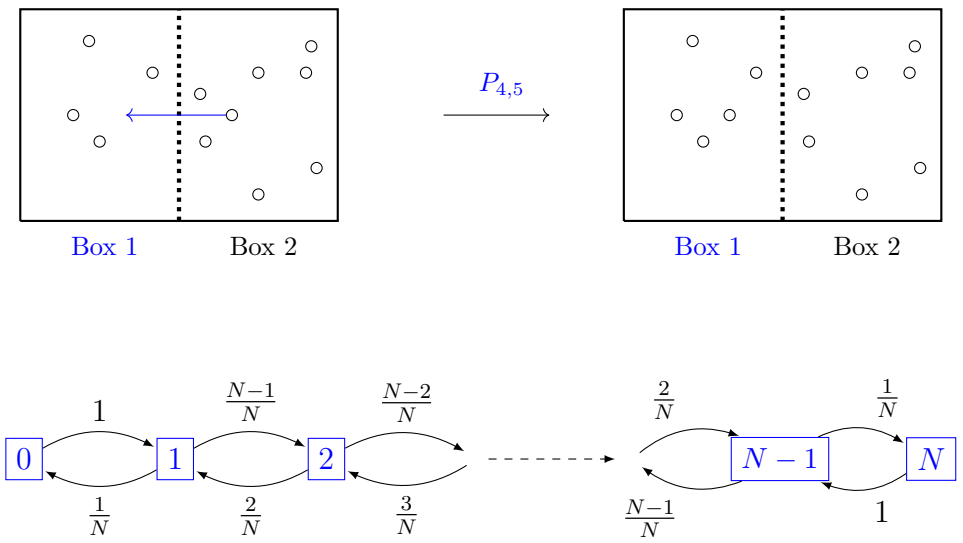


Figure 1.9: Ehrenfest chain

is a stationary distribution for the Ehrenfest chain (See Section 7.2 for more on this).

Bernoulli-Laplace model of diffusion:

This model was originally introduced by D. Bernoulli in 1769 and later analyzed by Laplace in 1812. It is a model for the diffusion of two incompressible liquids. There are two boxes, and initially there are N blue balls in Box 1 and N green balls in Box 2. At each step, one ball is chosen uniformly at random from each box (the two choices are made independently). Then the two chosen balls switch boxes. Note that the number of balls (molecules) remains the same N in each box throughout the process. See Figure 1.11. The number of blue balls in Box 1 describes the system, and so the state space is $\mathcal{S} = \{0, 1, ..., N\}$. The transition probabilities are

$$\begin{aligned} P_{x,x-1} &= (\tfrac{x}{N})^2 \qquad \text{for } 1 \le x \le N \\ P_{x,x+1} &= (\tfrac{N-x}{N})^2 \qquad \text{for } 0 \le x \le N-1 \\ P_{x,x} &= 2\tfrac{x(N-x)}{N^2} \qquad \text{for } 1 \le x \le N-1 \end{aligned}$$

and 0 otherwise. This is a birth/death chain. Since (1.14) holds, we can verify that for this Markov chain the following equations hold:

$$\binom{N}{x}^2 P_{x,x-1} = \binom{N}{x-1}^2 P_{x-1,x}\,,$$

that is,

$$\binom{N}{x}^2 \left(\frac{x}{N}\right)^2 = \binom{N}{x-1}^2 \left(\frac{N-x+1}{N}\right)^2.$$

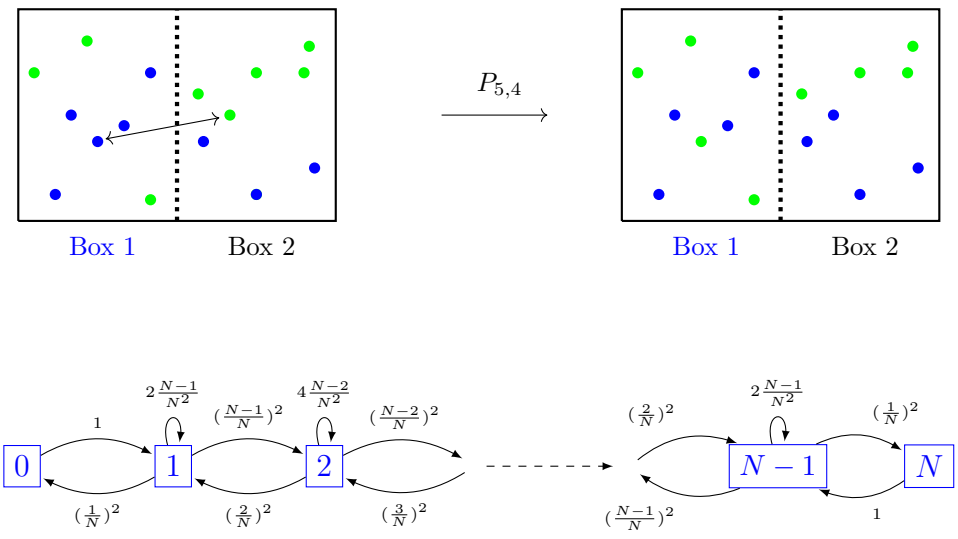


Figure 1.10: Bernoulli-Laplace model of diffusion

From this, and by using the binomial identity

$$\sum_{x=0}^{N} \binom{N}{x}^2 = \binom{2N}{N},$$

we conclude that π with

$$\pi(x) = \binom{N}{x}^2 / \binom{2N}{N} \quad \text{for } x = 0, 1, ..., N$$

is a stationary distribution (See Section 7.2). Note that π is a hypergeometric distribution. It can be interpreted as the distribution of the number of red balls among N balls that have been drawn uniformly at random and without replacement from a collection of N red and N green balls.

Wright-Fisher[2] model for genetic drift:

The Wright-Fisher model was introduced in 1931 as a model for *genetic drift* in a fixed size population. Genetic drift describes the change in relative allele frequencies in a population over time (the change that is caused by the inherent randomness of the process, but not by any outside factors). The Wright-Fisher model does not take mutation, selection, or environmental factors into account. It starts with a population of fixed size N. Each individual possesses in a certain locus of the chromosome an allele of one of two types,

[2]Sewall Wright (1889-1988), American geneticist, and Sir Ronald A. Fisher (1890-1962), British statistician and geneticist.

either type a or type A. We assume generations for this population do not overlap (such as for annual plants).
The Wright-Fisher chain $(X_n)_{n\geq 0}$ describes the evolving count of a specific type of allele, say type A, over time. X_n is the number of alleles of type A that are present in Generation n. Assuming the population is haploid (with a single chromosome), the state space is $\mathcal{S} = \{0, 1, ..., N\}$. $(X_n)_{n\geq 0}$ is a Markov chain. Its transition probabilities are defined by

$$P_{xy} = \binom{N}{y} \left(\frac{x}{N}\right)^y \left(\frac{N-x}{N}\right)^{N-y} \quad \text{for } x, y \in \mathcal{S}\,. \tag{1.15}$$

We can interpret the transition probabilities (1.15) in the following way: The $(n+1)^{th}$ generation is created by sampling with replacement from the n^{th} generation with parameter $p = X_n/N$. Or each of the N individuals in the $(n+1)^{th}$ generation inherits its allele type from a uniformly at random chosen parent in the n^{th} generation, and all choices are independent. Hence the distribution of X_{n+1} is binomial with parameters N and $p = X_n/N$.
Note that states 0 and N are absorbing states for the Wright-Fisher chain. Eventually, either allele a or allele A becomes *fixated*, while the other allele dies out. Of natural interest is the probability that a particular allele gets fixated. We will answer this question in Chapter 5.

Moran[3] model for genetic drift:

This model was proposed in 1958. As with the Wright-Fisher model, the Moran model describes the evolution over time of an allele frequency in a fixed size population. Here generations are modeled as overlapping. At each time step, only one uniformly randomly chosen individual reproduces and passes on its allele type, and another uniformly randomly (and independently from the first) chosen individual dies. We can model the process as an urn process: The urn contains N balls of two colors. At each time step, one ball is chosen uniformly at random, its color noted, and then returned to the urn. A second ball is chosen uniformly at random and removed from the urn, and in its place a ball of the color of the first ball is put into the urn. The Moran process $(X_n)_{n\geq 0}$ is a Markov chain on state space $\mathcal{S} = \{0, 1, ..., N\}$ where X_n is the number of alleles of type A present in the population at time n. The transition probabilities are

$$P_{x,x+1} = P_{x,x-1} = \frac{(N-x)x}{N^2} \qquad P_{xx} = 1 - 2\frac{(N-x)x}{N^2}\,.$$

[3]Patrick Moran (1917-1988), Australian statistician and geneticist.

States 0 and N are absorbing states. Note that this is a birth/death chain. We will compute the probability of eventual fixation in state N (i.e., the state in which the entire population has allele A) in Chapter 5.

Pólya's[4] Urn:

The following model was introduced by Eggenberger and Pólya ([13]) in 1923: An urn contains b blue balls and g green balls. At each time step, a ball is drawn uniformly at random from the urn, its color noted, and then, together with c additional balls of the *same* color, put back into the urn. Thus the number of balls in the urn is increasing by c with each step. Among many other applications, the process can be viewed as a model for the spread of a contagious disease. It is an example of a so-called *reinforced process*. In the following we mention several stochastic processes related to Pólya's urn model (not all of which are Markov chains). We will return to Pólya's urn using martingale theory in Section 6.6.4.

• *Process* $(B_n)_{n\geq 0}$. Let B_n denote the number of blue balls in the urn at time n. The process $(B_n)_{n\geq 0}$ is a time-inhomogeneous Markov chain on state space $\mathcal{S} = \{b, b+c, b+2c, ...\}$. The time dependent transition probabilities are

$$\mathbb{P}(B_{n+1} = k + c \,|\, B_n = k) = \frac{k}{b+g+cn}, \quad \mathbb{P}(B_{n+1} = k \,|\, B_n = k) = 1 - \frac{k}{b+g+cn}\,.$$

• *Process* $(B_n, G_n)_{n\geq 0}$. Let G_n denote the number of green balls in the urn at time n. Then $B_n + G_n = b + g + cn$. The process $(B_n, G_n)_{n\geq 0}$ *is* a time-homogeneous Markov chain. See Figure 1.11 for an illustration where $c = 2$. The process has state space $\mathcal{S} = \{b, b+c, b+2c, ...\} \times \{g, g+c, g+2c, ...\}$.

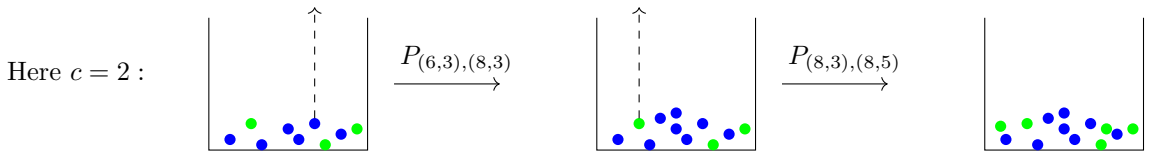


Figure 1.11: The process $(B_n, G_n)_{n\geq 0}$ for Pólya's urn

The transition probabilities for the process $(B_n, G_n)_{n\geq 0}$ are

$$P_{(x,y),(x+c,y)} = \frac{x}{x+y}, \quad P_{(x,y),(x,y+c)} = \frac{y}{x+y} \quad \text{for } (x,y) \in \mathcal{S}\,.$$

[4]George Pólya (1887–1985), Hungarian mathematician.

• *Process* $(Y_n)_{n\geq 0}$. We now consider the process $(Y_n)_{n\geq 1}$ where $Y_n \in \{0,1\}$ with $Y_n = 1$ if the nth ball drawn is blue and $Y_n = 0$ if the nth ball drawn is green. Note that $(Y_n)_{n\geq 1}$ is *not* a Markov chain. Indeed,

$$\mathbb{P}(Y_3 = 1 \,|\, Y_2 = 1,\, Y_1 = 1) = \frac{b+2c}{b+g+2c} \neq \frac{b+c}{b+g+2c} = \mathbb{P}(Y_3 = 1 \,|\, Y_2 = 1,\, Y_1 = 0)\,.$$

The stochastic process $(Y_n)_{n\geq 1}$ does however have other interesting properties which we will discuss in the following. First, we introduce the notion of *exchangeability* for a sequence of random variables.

Definition 1.5.1. *Let $(Z_n)_{n\geq 1}$ be a sequence of random variables taking values in a discrete state space $\mathcal{S}$. We say the stochastic process $(Z_n)_{n\geq 1}$ is* **exchangeable** *if for all $n \geq 2$ and any permutation π of $\{1, 2, ..., n\}$ the distribution of $(Z_1, ..., Z_n)$ is the same as the distribution of $(Z_{\pi(1)}, ..., Z_{\pi(n)})$.*

Note that an exchangeable stochastic process $(Z_n)_{n\geq 1}$ is a **stationary process**, that is, its distribution is invariant under time shifts. More precisely, , for all $n, k \geq 1$ and $x_1, ..., x_n \in \mathcal{S}$,

$$\mathbb{P}(Z_1 = x_1, ..., Z_n = x_n) = \mathbb{P}(Z_{1+k} = x_1, ..., Z_{n+k} = x_n)\,. \tag{1.16}$$

Indeed, Property (1.16) follows from exchangeability, since
$\mathbb{P}(Z_1 = x_1, ..., Z_n = x_n, Z_{n+1} \in \mathcal{S}, ..., Z_{n+k} \in \mathcal{S})$
$$= \mathbb{P}(Z_1 \in \mathcal{S}, ..., Z_k \in \mathcal{S}, Z_{k+1} = x_1, ..., Z_{k+n} = x_n)\,.$$
In particular, the random variables Z_n, $n \geq 1$, have *identical* distribution.

We now show that the process $(Y_n)_{n\geq 1}$ connected with Pólya's urn is exchangeable. Let $n \geq 1$, $0 \leq k \leq n$ and consider the n-tupel $\omega_n = (1, ..., 1, 0, ..., 0)$ with exactly k 1's. We write $\mathbb{P}(\omega_n)$ as a product of successive conditional probabilities:

$$\mathbb{P}(\omega_n) = \; P(Y_1 = 1, ..., Y_k = 1, Y_{k+1} = 0, ..., Y_n = 0) =$$

$$\frac{b}{(b+g)}\frac{(b+c)}{(b+g+c)}\cdots\frac{(b+(k-1)c)}{(b+g+(k-1)c)}\frac{g}{(b+g+kc)}\frac{(g+c)}{(b+g+(k+1)c)}\cdots\frac{(g+(n-k-1)c)}{(b+g+(n-1)c)}\,.$$

Similarly, for any reordering of the k 1's and the $(n-k)$ 0's resulting in a binary n-tuple $\tilde{\omega}_n$, we can write $\mathbb{P}(\tilde{\omega}_n)$ as a product of n successive conditional probabilities. In this product, the denominator will remain the same $(r+g)(r+g+c)\cdots(r+g+(n-1)c)$. The product in the numerator will also remain the same, but the factors will appear reordered

according to the reordering of the 1's and 0's in $\tilde{\omega}_n$. From this we see that $\mathbb{P}(\tilde{\omega}_n) = \mathbb{P}(\omega_n)$, and so $(Y_n)_{n\geq 1}$ is exchangeable.

Proposition 1.5.1. *Consider Pólya's urn with parameters b, g, c. Let $X_n = \sum_{i=1}^n Y_i$ be the number of blue balls drawn up to time n. Then*

$$\mathbb{P}(X_n = k) = \binom{n}{k} \frac{b(b+c)\cdots(b+(k-1)c)g(g+1)\cdots(g+(n-k-1)c)}{(b+g)(b+g+c)\cdots(b+g+(n-1)c)}.$$

Proof. The formula follows from exchangeability of the process $(Y_n)_{n\geq 1}$ and the above formula for $\mathbb{P}(\omega_n)$. □

Recall that $B_n = b + cX_n$. The distribution of X_n is called the **Pólya-Eggenberger distribution**. We point out a few special cases:

(a) For $c = 0$, we have sampling with replacement. In this case the distribution of X_n is the binomial distribution $\text{Bin}(n, \frac{b}{b+g})$.

(b) For $c = -1$, we have sampling without replacement. The distribution of X_n is the hypergeometric distribution with parameters $n, (b+g), b$.

(c) And for the **special case $b = g = c = 1$, the distribution of B_n, the number of blue balls in the urn at time n, is uniform distribution on** $\{1, 2, ..., n+1\}$ for $n \geq 1$:

$$\mathbb{P}(B_n = k) = \mathbb{P}(X_n = k-1) = \binom{n}{k-1} \frac{(k-1)!(n-(k-1))!}{(n+1)!} = \frac{1}{n+1}.$$

The result of Proposition 1.5.1 can easily be generalized to a **k-color Pólya's urn**. For this process, we consider k distinct colors $\mathcal{C}_1, ..., \mathcal{C}_k$. The process starts with c_i balls of color $\mathcal{C}_i$, $i = 1, ..., k$, in the urn. At each step, a ball is drawn uniformly at random, its color noted and then, together with c additional balls of the same color, put back into the urn. The following result will be useful for a later chapter.

Proposition 1.5.2. *Consider a k-color Pólya's urn that starts with 1 ball of each color and for which $c = 1$. Let X_n^i, $i = 1, .., k$, denote the number of balls of color $\mathcal{C}_i$ that are in the urn after n time steps. For any time $n \geq 1$, the random vector $(X_n^1, ..., X_n^k)$ is uniformly distributed over the set*

$$V_n = \{(x_1, ..., x_k) \in (\mathbb{Z}^+)^k : x_1 + \cdots + x_k = n + k\}.$$

Proof. We first note that

$$|V_n| = \binom{n+k-1}{k-1}$$

since $|V_n|$ is equal to the number of ways in which n indistinguishable balls can be distributed over k distinguishable boxes (empty boxes are allowed). Fix $n \geq 1$, a state $(x_1, ..., x_k) \in V_n$, and set $x_i = 1 + a_i$ for $i = 1, ..., k$. For Pólya's urn to be in state $\vec{x} = (x_1, ..., x_k)$ at time n, color $\mathcal{C}_i$ was drawn exactly a_i times for $i = 1, ..., k$. This can happen in $\binom{n}{a_1\, a_2 \ldots a_k}$ ways, according to the ordering in which the colors were drawn. By exchangeability, all orderings are equally likely to occur (we also see this by directly computing the probability of occurrence of a specific ordering of colors drawn). Thus

$$\begin{aligned}\mathbb{P}((X_n^1, ..., X_n^k) = \vec{x}) &= \binom{n}{a_1\, a_2 \ldots a_k} \frac{a_1! a_2! \cdots a_k!}{k(k+1)\cdots(n+k-1)} \\ &= \frac{n!(k-1)!}{(n+k-1)!} = \binom{n+k-1}{k-1}^{-1}.\end{aligned}$$

□

We return to the 2-color Pólya's urn. The exchangeability property of $(Y_n)_{n\geq 1}$ implies that $(Y_n)_{n\geq 1}$ is a stationary process and, in particular, that the probability of drawing a blue ball is the same at each step n and equal to $\mathbb{P}(Y_1 = 1) = \frac{b}{b+g}$. So the process $(Y_n)_{n\geq 1}$ is an infinite sequence of identically distributed (but not independent!) Bernoulli random variables. A theorem due to de Finetti[5] (which we more precisely quote below) states that the distribution of a sequence of exchangeable and identically distributed Bernoulli random variables is a *weighted average of the distributions of i.i.d. Bernoulli random variables*. For a reference, see [16].

Theorem 1.5.3 (de Finetti's Theorem)**.** *Let $p_0 \in (0,1)$ and $(Y_n)_{n\geq 1}$ be an infinite sequence of identically distributed Bernoulli random variables with $\mathbb{P}(Y_1 = 1) = p_0$ and $\mathbb{P}(Y_1 = 0) = 1 - p_0$. Then there exists a probability distribution dF on the interval $[0,1]$ such that for all $n \geq 1$ and for all $x_1, ..., x_n \in \{0,1\}$,*

$$\mathbb{P}(Y_1 = x_1, ..., Y_n = x_n) = \int_0^1 p^k\,(1-p)^{n-k}\, dF(p)$$

when $x_1 + \cdots + x_n = k$.

The theorem implies that, conditional on p which is chosen according to the distribution dF on $[0,1]$, the random variables $(Y_n)_{n\geq 1}$ are i.i.d. Bernoulli(p). In other words, to

[5]Bruno de Finetti (1906-1985), Italian mathematician and statistician.

generate the distribution of $(Y_n)_{n\geq 1}$, first choose the success probability p from $[0,1]$ according to dF, and then generate a sequence of i.i.d Bernoulli random variables with success probability p.

It turns out that the distribution dF is determined by its moments. This fact (which we do not prove) allows us to compute dF. Let m_k denote the kth moment of the distribution dF for $k \geq 1$. Since

$$\mathbb{P}(Y_1 = 1, Y_2 = 1, ..., Y_k = 1) = \int_0^1 p^k \, dF(p) \,,$$

we have

$$m_k = \mathbb{P}(Y_1 = 1, Y_2 = 1, ..., Y_k = 1) \,.$$

A straightforward computation (which is left as an exercise) yields

$$m_k = \frac{\Gamma(\frac{b}{c} + k)\Gamma(\frac{b+g}{c})}{\Gamma(\frac{b+g}{c} + k)\Gamma(\frac{b}{c})}$$

which we recognize as the kth moment of the beta distribution Beta$(\frac{b}{c}, \frac{g}{c})$. It follows that the distribution dF in de Finetti's theorem, as applied to the process $(Y_n)_{n\geq 0}$ for the 2-color Pólya's urn, is Beta$(\frac{b}{c}, \frac{g}{c})$. For the special case $b = g = c$, this distribution is Beta$(1,1) \sim$ Unif$([0,1])$. We will say more about Pólya's urn model in Section 6.6.4.

1.6 Irreducibility and class structure of the state space

The notion of irreducibility generalizes the notion of connectivity of graphs to Markov chains.

Definition 1.6.1 (Irreducibility)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $x, y \in \mathcal{S}$.*

(a) We say that x **leads to** *y, denoted by $x \longrightarrow y$, if there exists $n \geq 1$ such that $P^n_{xy} > 0$.*

(b) We say that x and y **communicate with each other**, *denoted by $x \longleftrightarrow y$, if $x \longrightarrow y$ and $y \longrightarrow x$.*

(c) We say that the Markov chain is **irreducible** *if for all $x, y \in \mathcal{S}$, we have $x \longrightarrow y$. Otherwise, we say the Markov chain is* **reducible**.

Notes: (1) It follows that $x \longrightarrow y$ iff there exists a finite sequence $(x, x_1, x_2, ..., x_{n-1}, y)$ of elements in $\mathcal{S}$ such that $P_{xx_1} > 0, P_{x_1x_2} > 0, ..., P_{x_{n-1}y} > 0$. Such a sequence is called a **path** from x to y.

(2) The relation $x \longleftrightarrow y$ is symmetric and transitive. Symmetry is obvious from the definition. Transitivity follows from the fact

$$P_{xz}^{n+m} \geq P_{xy}^{n} P_{yz}^{m} .$$

Example 1.6.1. Random walk on a connected graph $G(V, E)$ is irreducible. Indeed, since G is connected, for any vertices $x, y \in V$ there exists a sequence of edges

$$(\{x, x_1\}, \{x_1, x_2\}, ..., \{x_{n-1}, y\})$$

consisting of elements of E. For each edge $\{u, v\} \in E$, the transition probability P_{uv} is positive. Hence $x \longrightarrow y$. □

Example 1.6.2. The Wright-Fisher model (introduced in Section 1.5) is reducible: States $1, 2, ..., (N-1)$ lead to 0 and to N which are absorbing states. An absorbing state does not lead to any state other than itself. □

When studying reducible Markov chains, we often decompose the state space $\mathcal{S}$ into smaller, more elementary building blocks. We then study properties of the Markov chain restricted to these smaller building blocks and later reassemble the state space to deduce properties of the original chain. The main notion that is relevant for such a decomposition is the notion of an *irreducible closed class*, also called a *communication class*.

Definition 1.6.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $E \subseteq \mathcal{S}$. We say that E is an* **irreducible closed class** *or a* **communication class** *if*

(a) for all $x, y \in E$, $x \longleftrightarrow y$, and

(b) $P_{yz} = 0$ for all $y \in E$, $z \in E^c$.

Notes: (1) A Markov chain is irreducible if and only if the entire state space $\mathcal{S}$ is the only irreducible closed class.

(2) An irreducible closed class E for a Markov chain is *maximal* in the sense that if $E \subseteq F$ for an irreducible closed class F, then $E = F$. And furthermore, if E_1, E_2 are two irreducible closed classes and $E_1 \cap E_2 \neq \emptyset$, then $E_1 = E_2$. (See Exercise 1.17)

Example 1.6.3. Consider the Markov chain on state space $\mathcal{S} = \{1, 2, ..., 6\}$ with transi-

tion matrix

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0.1 & 0 & 0.9 \\ 0.1 & 0.3 & 0.2 & 0 & 0.4 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.1 & 0 & 0.4 \\ 0 & 0.6 & 0.3 & 0.1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Its transition graph is shown in Figure 1.12.

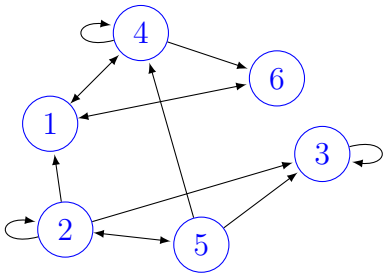


Figure 1.12

The irreducible closed classes are circled (dashed) in Figure 1.14. The Markov chain is reducible. Its irreducible closed classes are $R_1 = \{3\}$ and $R_2 = \{1, 4, 6\}$. Note that state 3 is an absorbing state. Any absorbing state forms its own singleton irreducible closed class. □

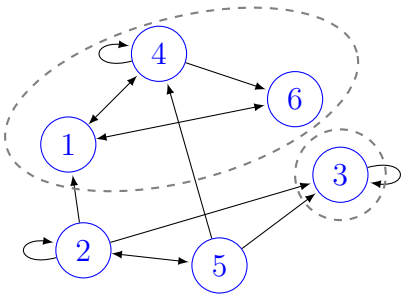


Figure 1.13

1.7 Functions of Markov chains

Often, when we consider a Markov chain, we are only interested in a particular feature of the elements in the state space and don't care to distinguish between elements that have the same feature. We would like to *lump* together all elements with the same "microscopic feature" and consider a state space of "macroscopic states". The question becomes, is the new process with the smaller and coarser state space still a Markov chain? More precisely, for a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S}$ and a given non-injective function f on $\mathcal{S}$, when is it true that the process $(f(X_n))_{n\geq 0}$ is again a Markov chain?

Example 1.7.1. Let $(X_n)_{n\geq 0}$ be the Markov chain on state space $\mathcal{S} = \{x, y, z\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

and initial distribution $\mu_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Consider the function $f : \mathcal{S} \to \mathbb{R}$ defined by $f(x) = f(y) = 0$ and $f(z) = 1$. Define the process $(Y_n)_{n\geq 0}$ by $Y_n = f(X_n)$. Is $(Y_n)_{n\geq 0}$ a Markov chain? Note that the state space of $(Y_n)_{n\geq 0}$ is $\{0, 1\}$ (which we could identify with the partition $\{\{x, y\}, \{z\}\}$ of $\mathcal{S}$). We compute

$$\mathbb{P}(Y_2 = 0 \,|\, Y_0 = 1, Y_1 = 0) = \frac{\mathbb{P}(X_0 = z,\, X_1 \in \{x, y\},\, X_2 \in \{x, y\})}{\mathbb{P}(X_0 = z,\, X_1 \in \{x, y\})} = \frac{1}{2}$$

and

$$\mathbb{P}(Y_2 = 0 \,|\, Y_0 = 0, Y_1 = 0) = \frac{\mathbb{P}(X_0 \in \{x, y\},\, X_1 \in \{x, y\},\, X_2 \in \{x, y\})}{\mathbb{P}(X_0 \in \{x, y\},\, X_1 \in \{x, y\})} = 0\,.$$

Here the Markov property does not hold, and so $(Y_n)_{n\geq 0}$ is not a Markov chain. □

Let $(X_n)_{n\geq 0}$ be a Markov chain on a discrete state space $\mathcal{S}$ and $f : \mathcal{S} \to \mathbb{R}$ a function. For any $x_i \in \mathrm{Im}(f)$, we set $A_i = f^{-1}(x_i)$. Then $\mathcal{A} = \{A_1, A_2, ...\}$ forms a partition of $\mathcal{S}$. Any partition $\mathcal{A} = \{A_1, A_2, ...\}$ of $\mathcal{S}$ induces an equivalence relation $x \sim y$ on $\mathcal{S}$ defined by $x \sim y$ if x and y belong to the same subset $A_i \in \mathcal{A}$. For $x \in \mathcal{S}$, we denote the equivalence class of x by $\langle x \rangle$, that is,

$$\langle x \rangle = \{y \in \mathcal{S} \,:\, y \sim x\}\,.$$

Definition 1.7.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with initial distribution μ_0 and let $\mathcal{A} = \{A_1, A_2, ...\}$ be a partition of its state space $\mathcal{S}$. We say that $(X_n)_{n\geq 0}$ is* **lumpable** *for $\mathcal{A}$, if $(\widehat{X}_n)_{n\geq 0}$ with $\widehat{X}_n = \langle X_n \rangle$ is a Markov chain on state space $\mathcal{A}$ with initial distribution $\widehat{\mu}_0$ given by*

$$\widehat{\mu}_0(A_i) = \sum_{x\in A_i} \mu_0(x) \qquad \text{for } A_i \in \mathcal{A}\,.$$

Proposition 1.7.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with initial distribution μ_0 and let $\mathcal{A} = \{A_1, A_2, ...\}$ be a partition of its state space $\mathcal{S}$.*

(a) If we have

$$P_{x,A_j} = P_{y,A_j} \quad \text{for all } A_j \in \mathcal{A} \quad \text{and whenever } x \sim y, \tag{1.17}$$

then the Markov chain is lumpable for the partition $\mathcal{A}$. Assuming $x \in A_i$ in (1.17), the transition probabilities for the lumped chain $(\widehat{X}_n)_{n\geq 0}$ are given by

$$\widehat{P}_{A_i,A_j} = P_{x,A_j}\,. \tag{1.18}$$

(b) If $(X_n)_{n\geq 0}$ is lumpable for the partition $\mathcal{A}$ and if π is a stationary distribution for $(X_n)_{n\geq 0}$, then $\widehat{\pi}$ defined by

$$\widehat{\pi}(A_i) = \sum_{y:y\in A_i} \pi(y) \quad \text{for } A_i \in \mathcal{A}$$

is a stationary distribution for the lumped chain $(\widehat{X}_n)_{n\geq 0}$.

Proof. (a) Assume (1.17) holds. We compute

$$\begin{aligned}
\mathbb{P}(X_0 \in A_{i_0}, ..., X_n \in A_{i_n}) &= \sum_{x\in A_{i_{n-1}}} \mathbb{P}(X_0 \in A_{i_0}, ..., X_{n-2} \in A_{i_{n-2}}, X_{n-1} = x, X_n \in A_{i_n}) \\
&= \sum_{x\in A_{i_{n-1}}} \mathbb{P}(X_n \in A_{i_n} \,|\, X_{n-1} = x)\mathbb{P}(X_0 \in A_{i_0}, ..., X_{n-2} \in A_{i_{n-2}}, X_{n-1} = x) \\
&= P_{x,A_{i_n}} \mathbb{P}(X_0 \in A_{i_0}, ..., X_{n-1} \in A_{i_{n-1}}) \quad \text{for any } x \in A_{i_{n-1}}.
\end{aligned}$$

Thus

$$\mathbb{P}(X_n \in A_{i_n} \,|\, X_0 \in A_{i_0}, ..., X_{n-1} \in A_{i_{n-1}}) = P_{x,A_{i_n}} \quad \text{for any } x \in A_{i_{n-1}}.$$

This shows that, under lumping of the state space according to the partition $\mathcal{A}$, the Markov property still holds, and that the transition probabilities for the lumped chain are given by (1.18).
(b) Assume π is a stationary distribution for $(X_n)_{n\geq 0}$ which is lumpable for $\mathcal{A} = \{A_1, A_2, ...\}$. Then distribution $\widehat{\pi}$ defined by

$$\widehat{\pi}(A_i) = \sum_{y:y\in A_i} \pi(y) \quad \text{for } A_i \in \mathcal{A}$$

is a probability distribution on $\mathcal{A}$. We have

$$\begin{aligned}\sum_{A_i\in\mathcal{A}} \widehat{\pi}(A_i)\widehat{P}_{A_i,A_j} &= \sum_{A_i\in\mathcal{A}} \left(\sum_{z\in A_i} \pi(z)P_{z,A_j}\right) \\ &= \sum_{x\in\mathcal{S}} \pi(x)P_{x,A_j} = \sum_{y\in A_j}\sum_{x\in\mathcal{S}} \pi(x)P_{x,y} \\ &= \sum_{y\in A_j} \pi(y) = \widehat{\pi}(A_j)\,.\end{aligned}$$

□

Example 1.7.2. Consider a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{x, y, z, w, v\}$ with transition matrix

$$\mathbf{P} = \begin{array}{c} \\ x \\ y \\ z \\ w \\ v \end{array} \begin{array}{c} \begin{array}{ccccc} x & y & z & w & v \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{2}{3} & 0 & \frac{1}{6} & \frac{1}{6} & 0 \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{6} & \frac{1}{12} & \frac{1}{6} \\ \frac{7}{12} & 0 & \frac{1}{6} & 0 & \frac{1}{4} \\ \frac{1}{8} & \frac{11}{24} & 0 & \frac{5}{12} & 0 \end{pmatrix} \end{array}.$$

We can verify that condition (1.17) is fulfilled for the partition $\mathcal{A} = \{A_1, A_2\}$ with $A_1 = \{x, y\}$ and $A_2 = \{z, w, z\}$, so $(X_n)_{n\geq 0}$ is lumpable with respect to $\mathcal{A}$. The transition matrix $\widehat{\mathbf{P}}$ for the lumped chain is

$$\widehat{\mathbf{P}} = \begin{array}{c} \\ A_1 \\ A_2 \end{array} \begin{array}{c} \begin{array}{cc} A_1 & A_2 \end{array} \\ \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{7}{12} & \frac{5}{12} \end{pmatrix} \end{array}.$$

□

Computation of $\widehat{\mathbf{P}}^k$:

Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ with $|\mathcal{S}| = n$. Assume that $(X_n)_{n\geq 0}$ is lumpable with respect to the partition $\mathcal{A} = \{A_1, ..., A_m\}$ of $\mathcal{S}$. Let $\mathbf{B}$ be the $n \times m$ matrix whose jth column has 1's in its coordinates corresponding to the elements in A_j and 0s everywhere else. Moreover, let $\mathbf{A}$ be the $m \times n$ matrix whose ith row has the entry $1/|A_i|$ in its coordinates corresponding to the elements in A_i and 0s everywhere else. Then we have

$$\widehat{\mathbf{P}} = \mathbf{APB}\,. \tag{1.19}$$

We illustrate (1.19) with an example.

Example 1.7.3. Consider the transition matrix $\mathbf{P}$ and the partition $\mathcal{A} = \{A_1, A_2\}$ from Example 1.7.2. Here

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{A} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

We have

$$\mathbf{APB} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{2}{3} & 0 & \frac{1}{6} & \frac{1}{6} & 0 \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{6} & \frac{1}{12} & \frac{1}{6} \\ \frac{7}{12} & 0 & \frac{1}{6} & 0 & \frac{1}{4} \\ \frac{1}{8} & \frac{11}{24} & 0 & \frac{5}{12} & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \\ \frac{7}{12} & \frac{5}{12} \\ \frac{7}{12} & \frac{5}{12} \\ \frac{7}{12} & \frac{5}{12} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{7}{12} & \frac{5}{12} \end{pmatrix} = \widehat{\mathbf{P}}\,.$$

□

Note that, as can be seen in the above example, all rows of $\mathbf{PB}$ that correspond to elements in the *same* subset A_i of the partition are the same. This is the equivalent of condition (1.17). Left multiplying $\mathbf{PB}$ by $\mathbf{A}$ has the effect of collapsing all rows corresponding to elements from the same subset (hence equal rows) into one row. Further left multiplying

the result by matrix $\mathbf{B}$ has the effect of undoing this collapsing of equal rows, thus resulting back in $\mathbf{PB}$. All this is summarized in the following lemma.

Lemma 1.7.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. Assume $(X_n)_{n\geq 0}$ is lumpable with respect to the partition $\mathcal{A}$, and let $\mathbf{A}$ and $\mathbf{B}$ be the corresponding matrices as defined above. Then*

$$\mathbf{BAPB} = \mathbf{PB}\,. \tag{1.20}$$

Corollary 1.7.3. *With the notation and assumptions from Lemma 1.7.2, we have*

$$\widehat{\mathbf{P}}^k = \mathbf{AP}^k\mathbf{B} \qquad \text{for all } k \geq 1\,.$$

Proof. Use (1.20) and induction on k. □

Example 1.7.4 (Random walk on the hypercube and the Ehrenfest chain)**.** We have introduced the Ehrenfest chain (a model of gas diffusion of n particles between two containers) in Section 1.5. The Ehrenfest chain $(Y_n)_{n\geq 0}$ is a birth/death chain on $\mathcal{S} = \{0, 1, ..., n\}$ with transition probabilities

$$P_{x,x-1} = \frac{x}{n} \text{ for } 1 \leq x \leq n, \quad P_{x,x+1} = \frac{n-x}{n} \text{ for } 0 \leq x \leq n-1\,.$$

It can be derived from simple random walk on the hypercube $\mathbb{Z}_2^n$ (see Section 1.5) via lumping. Recall that $\mathbb{Z}_2^n = \{(x_1, ..., x_n) \,|\, x_i \in \{0, 1\} \text{ for } 1 \leq i \leq n\}$. Simple random walk on $\mathbb{Z}_2^n$ proceeds by choosing uniformly at random an index k from $\{1, ..., n\}$ and switching the kth entry of the current state $\mathbf{x} \in \mathbb{Z}_2^n$ from x_k to $(x_k + 1) \mod 2$. If we identify the index set $\{1, ..., n\}$ with the set of distinct balls for the Ehrenfest model, we have a one-to-one correspondence between the states $\mathbf{x} \in \mathbb{Z}_2^n$ and the possible assignments of the n balls to the two boxes: Ball k is in Box 1 iff $x_k = 1$. The **Hamming weight** $h(\mathbf{x})$ which is defined by

$$h(\mathbf{x}) = \sum_{k=1}^{n} x_k$$

counts the number of balls in Box 1. The lumpability condition (1.18) holds for simple random walk $(X_n)_{n\geq 0}$ on the hypercube $\mathbb{Z}_2^n$ and Hamming weight. Indeed, first note that $P_{\mathbf{x},\langle\mathbf{y}\rangle} \neq 0$ iff $|h(\mathbf{x}) - h(\mathbf{y})| = 1$. We have

$$P_{\mathbf{x},\langle\mathbf{y}\rangle} = \begin{cases} \frac{h(\mathbf{x})}{n} & \text{if } h(\mathbf{y}) = h(\mathbf{x}) - 1 \\ \frac{n-h(\mathbf{x})}{n} & \text{if } h(\mathbf{y}) = h(\mathbf{x}) + 1\,. \end{cases}$$

Since $h(\mathbf{x})$ is constant on $\langle \mathbf{x} \rangle$, the above transition probabilities are the same for all $\mathbf{x} \in \langle \mathbf{x} \rangle$. This satisfies (1.18) and shows that the Ehrenfest chain is $(Y_n)_{n\geq 0} = (h(X_n))_{n\geq 0}$ is a lumped version of random walk on the hypercube. We can verify that uniform distribution on $\mathbb{Z}_2^n$, that is π with

$$\pi(\mathbf{x}) = 1/2^n \qquad \text{for all } \mathbf{x} \in \mathbb{Z}_2^n \,,$$

is a stationary distribution (in fact, the unique stationary distribution) for random walk on the hypercube. It follows that for its lumped version that is the Ehrenfest chain, the distribution $\widehat{\pi}$ given by

$$\widehat{\pi}(y) = |\{\mathbf{x} \in \mathbb{Z}_2^n \,|\, h(\mathbf{x}) = y\}|/2^n = \binom{n}{y}/2^n \qquad \text{for } \; 0 \leq y \leq n$$

is a stationary distribution (again, the unique stationary distribution). Notice that $\widehat{\pi} \sim \text{Bin}(n, \frac{1}{2})$. □

Example 1.7.5 (Simple random walk on the discrete circle and random walk on a chain graph)**.** Consider simple symmetric random walk on the discrete (unit) cycle $\mathbb{Z}_n$ for n even. For each $1 \leq k \leq \frac{n-2}{2}$, we can lump the points $e^{2\pi ik/n}$ and $e^{-2\pi ik/n}$ and thus get the partition $\{\{1\}, \{e^{2\pi i/n}, e^{-2\pi i/n}\}, ..., \{e^{\pi i(n-1)/n}, e^{-\pi i(n-2)/n}\}, \{-1\}\}$ of $\mathbb{Z}_n$. Because of symmetry of the random walk on $\mathbb{Z}_n$, we easily verify that the lumpability condition (1.18) holds. The lumped chain can be identified with simple symmetric random walk on the integers $\{1, 2, ..., \frac{n}{2} + 1\}$ *with reflecting boundary* at the two endpoints 1 and $\frac{n}{2} + 1$ (that is, $P_{1,2} = 1$ and $P_{\frac{n}{2}+1, \frac{n}{2}-1} = 1$). See Figure 1.14.

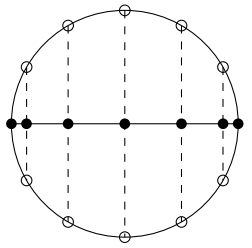


Figure 1.14

We can directly verify that for simple random walk on the discrete cycle, $\pi \sim \text{Unif}(\mathbb{Z}_n)$ is a stationary distribution (in fact, it is the unique stationary distribution). Thus for the lumped walk on the integers $\{1, 2, ..., \frac{n}{2} + 1\}$, the distribution $\widehat{\pi}$ defined by

$$\begin{aligned} \widehat{\pi}(y) &= \frac{2}{n} \quad \text{for } 2 \leq y \leq \frac{n}{2} - 1, \text{ and} \\ \widehat{\pi}(1) &= \widehat{\pi}(\frac{n}{2} + 1) = \frac{1}{n} \end{aligned}$$

is a stationary distribution (again, the unique stationary distribution). □

As we will discuss in later chapters, the eigenvalues of the transition matrix $\mathbf{P}$ play an important role in the long-term evolution of the Markov chain. The following proposition describes the connection between the eigenvalues and eigenvectors of a lumped chain and those of the original chain.

Proposition 1.7.4. *Let $(X_n)_{n\geq 0}$ be a Markov chain with tranisition matrix $\mathbf{P}$ on state space $\mathcal{S}$ and $\mathcal{A} = \{A_1, A_2, ...\}$ a partition of $\mathcal{S}$. Assume $(X_n)_{n\geq 0}$ is lumpable for $\mathcal{A}$. Denote the lumped chain by $(\widehat{X}_n)_{n\geq 0}$ and its transition matrix by $\widehat{\mathbf{P}}$. Then we have the following.*

(a) Let $\mathbf{s}$ be a right eigenvector of $\mathbf{P}$ corresponding to eigenvalue λ. We view $\mathbf{s}$ as a function $\mathbf{s} : \mathcal{S} \to \mathbb{R}$. If for each $A_i \in \mathcal{A}$, the right eigenfunction $\mathbf{s}$ is constant on A_i, then the projection $\widehat{\mathbf{s}} : \mathcal{A} \to \mathbb{R}$ defined by

$$\widehat{\mathbf{s}}(A_i) = \mathbf{s}(x) \quad \text{if } x \in A_i \text{ and for all } A_i \in \mathcal{A}$$

is a right eigenfunction (right eigenvector) of $\widehat{\mathbf{P}}$ corresponding to eigenvalue λ.

(b) Conversely, if $\widehat{\mathbf{s}}$ is a right eigenfunction of $\widehat{\mathbf{P}}$ corresponding to eigenvalue λ, then its lift $\mathbf{s} : \mathcal{S} \to \mathbb{R}$ defined by $\mathbf{s}(x) = \widehat{\mathbf{s}}(A_i)$ if $x \in A_i$ is a right eigenfunction of $\mathbf{P}$ corresponding to eigenvalue λ.

Proof. (a) Assume $\mathbf{s}$ is an eigenfunction of $\mathbf{P}$ and $x \in A_i$ and $y \in A_j$. We have

$$\begin{aligned}(\widehat{\mathbf{P}}\widehat{\mathbf{s}})(A_i) &= \sum_{A_j \in \mathcal{A}} \widehat{P}_{A_i,A_j} \widehat{\mathbf{s}}(A_j) = \sum_{A_j \in \mathcal{A}} P_{x,A_j} \mathbf{s}(y) = \sum_{A_j \in \mathcal{A}} \sum_{z \in A_j} P_{x,z} \mathbf{s}(z) \\ &= \sum_{z \in \mathcal{S}} P_{x,z} \mathbf{s}(z) = (\mathbf{P}\mathbf{s})(x) = \lambda \mathbf{s}(x) = \lambda \widehat{\mathbf{s}}(A_i)\,.\end{aligned}$$

(b) Assume $\widehat{\mathbf{s}}$ is an eigenfunction of $\widehat{\mathbf{P}}$ and $x \in A_i$ and $y \in A_j$. Let $\mathbf{s}$ be the lift (to $\mathcal{S}$) of $\widehat{\mathbf{s}}$. We have

$$\begin{aligned}(\mathbf{P}\mathbf{s})(x) &= \sum_{z \in \mathcal{S}} P_{x,z} \mathbf{s}(z) = \sum_{A_j \in \mathcal{A}} \sum_{z \in A_j} P_{x,z} \mathbf{s}(z) = \sum_{A_j \in \mathcal{A}} P_{x,A_j} \mathbf{s}(y) \\ &= \sum_{A_j \in \mathcal{A}} \widehat{P}_{A_i,A_j} \widehat{\mathbf{s}}(A_j) = (\widehat{\mathbf{P}}\widehat{\mathbf{s}})(A_i) = \lambda \widehat{\mathbf{s}}(A_i) = \lambda \mathbf{s}(x)\,.\end{aligned}$$

□

Corollary 1.7.5. *Let $(X_n)_{n\geq 0}$ be a Markov chain with tranisition matrix $\mathbf{P}$ on state space $\mathcal{S}$ and $\mathcal{A} = \{A_1, A_2, ...\}$ a partition of $\mathcal{S}$. Assume $(X_n)_{n\geq 0}$ is lumpable for $\mathcal{A}$, and denote the transition matrix for the lumped chain by $\widehat{\mathbf{P}}$. Then the set of eigenvalues of $\widehat{\mathbf{P}}$ is a subset of the set of eigenvalues of $\mathbf{P}$.*

Exercises

Exercise 1.1. A deck of cards initially consists of 3 cards of which one is black, one is white, and one is green. At each time interval, a card is selected uniformly at random from the deck. If it is black, it is removed from the deck. If it is white, it is replaced into the deck. If the card is green, we replace it by a new black card. For each of the following processes, determine whether or not the process is a Markov chain. If it is a Markov chain, give the state space and the transition matrix. Otherwise, give a reason for why it is not a Markov chain.

(a) $(X_n)_{n\geq 0}$ where X_n is the number of black cards in the deck at time n.

(b) $(Y_n)_{n\geq 0}$ Where Y_n) is the number of green cards in the deck at time n.

(c) $(Z_n)_{n\geq 0}$ where $Z_n = (X_n, Y_n)$ is the vector that gives the number of black and green cards in the deck at time n.

Exercise 1.2. *Time shift.* Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$. Show that for any fixed time $n_0 > 0$, the process $(Y_n)_{n\geq 0}$ defined by $Y_n = X_{n_0+n}$ for $n \geq 0$ is a Markov chain that has the same transition probabilities as $(X_n)_{n\geq 0}$ and whose initial distribution is the distribution of X_{n_0}.

Exercise 1.3. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. Fix $c \in \mathbb{N}$. Show that the process $(Y_n)_{n\geq 0}$ defined by $Y_n = X_{cn}$ is a Markov chain and determine its transition matrix.

Exercise 1.4. Consider the following elementary urn model for a chemical reaction: An urn contains 8 balls of which four are black and four are white. Two balls are randomly drawn from the urn. If one ball is black and the other ball is white, then the selected balls are discarded and two green balls are returned to the urn. Otherwise, the selected balls are returned to the urn. This process continues until the urn contains only green balls. Let X_n be the random variable "number of black balls in the urn after the n'th draw". Give the one-step transition matrix $\mathbf{P}$ for this Markov chain.

Exercise 1.5. Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$. Prove that for all $n \geq 1$, for all states $x, y \in \mathcal{S}$, and for all subsets $A_0, ..., A_{n-1} \subseteq \mathcal{S}$,

$$\mathbb{P}(X_{n+1} = y \,|\, X_0 \in A_0, ..., X_{n-1} \in A_{n-1}, X_n = x) = P_{xy}$$

whenever both sides are well-defined. (Note however that, in general,

$$\mathbb{P}(X_{n+1} = y \,|\, X_0 \in A_0, ..., X_{n-1} \in A_{n-1}, X_n \in A_n) \neq \mathbb{P}(X_{n+1} = y \,|\, X_n \in A_n)$$

if the set A_n is **not a singleton set**. See Exercise 1.6 for an illustration.)

Exercise 1.6. Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S} = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

Show that

$$\mathbb{P}(X_2 = 1 \,|\, X_0 = 1, X_1 \in \{0, 2\}) \neq \mathbb{P}(X_2 = 1 \,|\, X_0 = 2, X_1 \in \{0, 2\}).$$

Exercise 1.7. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S} = \{1, 2, 3\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \\ \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{pmatrix}.$$

Consider the process $(Y_n)_{n\geq 0}$ that tracks $(X_n)_{n\geq 0}$ when it moves a to a new state while ignoring any holding periods. More precisely, define $T_0 = 0$ and

$$T_n = \min\{m > T_{n-1} : X_m \neq X_{T_{n-1}}\}$$

for $n \geq 1$, and set

$$Y_n = X_{T_n}.$$

Is the process $(Y_n)_{n\geq 0}$ a Markov chain? If so, determine its transition matrix.

Exercise 1.8. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$.

(a) Define the (trivariate) *moving window* process $(Y_n)_{n\geq 1}$ by

$$Y_n = (X_{n-1}, X_n, X_{n+1})$$

for $n \geq 1$. Is $(Y_n)_{n\geq 1}$ a Markov chain? If so, what are the one-step transition probabilities?

(b) Define the (bivariate) *moving open window* process $(Z_n)_{n\geq 1}$ by

$$Z_n = (X_{n-1}, X_{n+1})$$

for $n \geq 1$. Is $(Z_n)_{n\geq 1}$ a Markov chain? If so, what are the one-step transition probabilities?

Exercise 1.9. Let $(X_n)_{n\geq 0}$ be a two-state Markov chain with state space $\mathcal{S} = \{-1, 1\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Show that the *moving average process* $(Z_n)_{n\geq 1}$ defined by

$$Z_n = \frac{1}{2}(X_{n-1} + X_n)$$

for $n \geq 1$ is not a Markov chain.

Exercise 1.10. Let $(Y_n)_{n\geq 1}$ be an i.i.d. sequence of random variables taking values in a space $\mathcal{Y}$. Let X_0 be random variable taking values in a discrete state space $\mathcal{S}$. We assume that X_0 is independent of the $Y_1, Y_2, Y_3, \ldots$. Let $f : \mathcal{S} \times \mathcal{Y} \to \mathcal{S}$ be a fixed function. Prove that the process $(X_n)_{n\geq 0}$ defined recursively by

$$X_{n+1} = f(X_n, Y_{n+1}) \qquad \text{for } n \geq 0$$

is a Markov chain with state space $\mathcal{S}$. Describe its transition probabilities P_{xy} for $x, y \in \mathcal{S}$.

Exercise 1.11. Let T' and T'' be stopping times for a Markov chain $(X_n)_{n\geq 0}$. Show that the following are also stopping times for $(X_n)_{n\geq 0}$:
(a) $T = T' + T''$
(b) $T = T' \wedge T'' := \min\{T', T''\}$
(c) $T = T' \vee T'' := \max\{T', T''\}$

Exercise 1.12. Consider a Markov chain $(X_n)_{\geq 0}$, a stopping time T for the process, and the sequence of stopping times $T \wedge n$, $n \geq 1$. Show that if $\mathbb{P}(T < \infty) = 1$, then

$$\lim_{n\to\infty} T \wedge n = T \qquad \text{with probability 1,}$$

and

$$\lim_{n\to\infty} \mathbb{E}(T \wedge n) = \mathbb{E}(T).$$

Exercise 1.13. Let x and y be distinct states of a finite-state Markov chain with $|\mathcal{S}| = N$, and suppose x leads to y. Let n_0 be the smallest positive integer such that $P_{xy}^{n_0} > 0$. Prove that $n_0 \leq N - 1$.

Exercise 1.14. Let $(X_n)_{n\geq 0}$ be a finite-state Markov chain with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. Assume $|\mathcal{S}| = N < \infty$. Prove that $(X_n)_{n\geq 0}$ is irreducible if and only if

$$\mathbf{P} + \mathbf{P}^2 + \cdots + \mathbf{P}^N$$

is a strictly positive matrix.

Exercise 1.15. Consider a *k-color Pólya's urn* with distinct colors $\mathcal{C}_1, ..., \mathcal{C}_k$. The process starts with c_i balls of color $\mathcal{C}_i$, $i = 1, ..., k$, in the urn. At each step, a ball is drawn uniformly at random, its color noted and then, together with c additional balls of the same color, put back into the urn. Fix a color $\mathcal{C}_i$. Show that the probability of drawing $\mathcal{C}_i$ is constant in time. What is this probability?

Exercise 1.16. Prove that every finite-state Markov chain has at least one irreducible closed class.

Exercise 1.17. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$. Prove the following:

(a) Any irreducible, closed class E for the chain is *maximal* in the sense that if $E \subseteq F$ for an irreducible closed class F, then $E = F$.
(b) If E_1, E_2 are two irreducible closed classes for the chain and $E_1 \cap E_2 \neq \emptyset$, then $E_1 = E_2$.

Exercise 1.18. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and $\mathcal{A} = \{A_1, A_2, ...\}$ a partition of $\mathcal{S}$. Assume $(X_n)_{n\geq 0}$ is lumpable for $\mathcal{A}$.

(a) Show that if $(X_n)_{n\geq 0}$ is irreducible, then so is its lumped version.

(b) Is it true that if the lumped chain is irreducible, then the original chain $(X_n)_{n\geq 0}$ must also be irreducible? Prove your answer or give a counter example.

Chapter 2

Long-run Behavior of Markov Chains

2.1 Transience and Recurrence

Definition 2.1.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $x \in \mathcal{S}$. The* **first passage time** *T^x to state x is defined by*

$$T^x = \min\{n \geq 1 \,:\, X_n = x\}\,.$$

If the Markov chain starts in state x, i.e., if $X_0 = x$, we call T^x the **first return time**. *If $X_n \neq x$ for all $n \geq 1$, we say $T^x = \infty$.*

Note that the first passage time T^x is a stopping time for the Markov chain. As a consequence of the strong Markov property, the Markov chain "renews itself" after each visit to x, that is, the process $(Y_n)_{n\geq 0}$ defined by $Y_n = X_{T^x+n}$ is a probabilistic replica of the Markov chain $(X_n)_{n\geq 0}$ with $X_0 = x$.

Proposition 2.1.1. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with finite state space $\mathcal{S}$. Then for all $x, y \in \mathcal{S}$,*

$$\mathbb{E}_x(T^y) < \infty\,.$$

Proof. Since the Markov chain is irreducible, for any pair of states $x, y \in \mathcal{S}$ there exists an integer $n(x,y)$ and an $\epsilon(x,y) > 0$ such that $P_{xy}^{n(x,y)} > \epsilon(x,y)$. Since $\mathcal{S}$ is finite, there exist $n = \max\{n(x,y) : x, y \in \mathcal{S}\}$ and $\epsilon = \min\{\epsilon(x,y) : x, y \in \mathcal{S}\}$, and so

$$n(x,y) \leq n \;\text{ and }\; \epsilon(x,y) \geq \epsilon \quad \text{ for all } x, y \in \mathcal{S}\,.$$

Writing $\mathbb{P}_x(T^y > n)$ for $\mathbb{P}(T^y > n \,|\, X_0 = x)$, we have

$$\mathbb{P}_x(T^y > n) \leq (1 - \epsilon)\,,$$

and from this, by iteration and repeated application of the Markov property,

$$\mathbb{P}_x(T^y > kn) \leq (1 - \epsilon)^k \quad \text{for } k \geq 1\,.$$

The random variable T^y is positive and integer valued, and so

$$\mathbb{E}_x(T^y) = \sum_{m=0}^{\infty} \mathbb{P}_x(T^y > m)\,.$$

Note that the probability $\mathbb{P}_x(T^y > m)$ is a decreasing function of m. Thus we get the upper bound

$$\mathbb{E}_x(T^y) = \sum_{m=0}^{\infty} \mathbb{P}_x(T^y > m) \leq \sum_{k=0}^{\infty} n\mathbb{P}_x(T^y > kn) \leq n\sum_{k=0}^{\infty}(1-\epsilon)^k < \infty\,.$$

□

Definition 2.1.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $x \in \mathcal{S}$. We say*

- *state x is* **recurrent** *if* $\mathbb{P}(T^x < \infty \,|\, X_0 = x) = 1$,
- *state x is* **transient** *if* $\mathbb{P}(T^x < \infty \,|\, X_0 = x) < 1$.

We say a Markov chain is recurrent (resp. transient) if all of its states are recurrent (resp. transient).

Note that an absorbing state is a recurrent state.

Example 2.1.1. (a) Consider biased random walk on the integers $\mathcal{S} = \{0, 1, 2, 3, 4\}$ with *reflecting* boundary at 0 and at 4. See Figure 2.1 for its transition graph. This is an

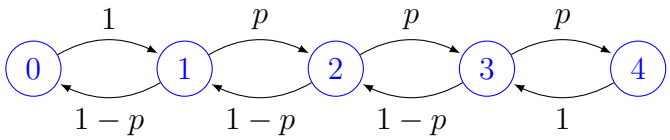


Figure 2.1

irreducible, finite state Markov chain. By Proposition 2.1.1, $\mathbb{E}_x(T^x) < \infty$ for all $x \in \mathcal{S}$. It

follows that $\mathbb{P}(T^x < \infty \,|\, X_0 = x) = 1$ for all $x \in \mathcal{S}$, and so all states are recurrent states. With probability 1, the Markov chain will return to its starting state in finite time. This is a recurrent Markov chain.

(b) Consider biased random walk on the integers $\mathcal{S} = \{0, 1, 2, 3, 4\}$ with *absorbing* boundary at 0 and at 4. See Figure 2.2. Both 0 and 4 are absorbing states and are therefore

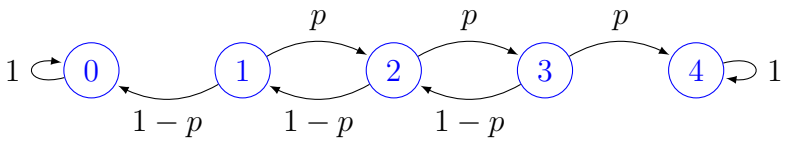


Figure 2.2

recurrent. States $1, 2$, and 3 lead into the absorbing boundary from which the Markov chain cannot return to its starting state, and therefore $1, 2$, and 3 are transient states. For example for state 2, we have $P^2_{2,4} = p^2 > 0$. Thus $\mathbb{P}(T^2 = \infty \,|\, X_0 = 2) \geq p^2$, and therefore

$$\mathbb{P}(T^2 < \infty \,|\, X_0 = 2) \leq 1 - p^2 < 1\,.$$

□

To reiterate, a state x is recurrent if the Markov chain, given that it starts in state x, will return to x in finite time T^x with probability 1. The return time T^x is a stopping time for the Markov chain. After the Markov chain has returned to its starting state x, by the strong Markov property, it will return a second time to x in finite time with probability 1. By successively invoking the strong Markov property, we prove that a Markov chain revisits a recurrent state x infinitely many times with probability 1. This last fact is often used as an equivalent characterization of recurrence of a state x. Theorem 2.1.3 below makes this precise.

Notation: (1) Let V^y denote the random variable "number of visits to state y (not including a possible initial visit at time 0, if the Markov chain starts in state y)" and $\mathbb{E}_x(V^y)$ the expected number of visits to state y, given that the Markov chain starts in state x.

(2) From now onwards, we will use the notation $f_{xy} = \mathbb{P}(T^y < \infty \,|\, X_0 = x)$.

The following lemma gives the distribution of the random variable V^y:

Lemma 2.1.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $x, y \in \mathcal{S}$. Then for $k \geq 1$,*

$$\mathbb{P}(V^y \geq k \,|\, X_0 = x) = f_{xy} f_{yy}^{k-1}\,. \tag{2.1}$$

Proof. The hitting (or return) time T^y is a stopping time for the Markov chain. Thus Formula (2.1) follows from the strong Markov property. The reader is asked to flesh out the proof in Exercise 2.2. □

Theorem 2.1.3. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$.*

(a) If $y \in \mathcal{S}$ is recurrent, then

$$\mathbb{P}(V^y = \infty \,|\, X_0 = y) = 1$$

and hence

$$\mathbb{E}_y(V^y) = \infty\,.$$

Furthermore, $\mathbb{P}(V^y = \infty \,|\, X_0 = x) = f_{xy}$ for all $x \in \mathcal{S}$.

(b) If y is transient, then

$$\mathbb{P}(V^y < \infty \,|\, X_0 = x) = 1$$

and

$$\mathbb{E}_x(V^y) = \frac{f_{xy}}{1 - f_{yy}} < \infty$$

for all $x \in \mathcal{S}$.

Proof. (a) By Lemma 2.1.2, $\mathbb{P}(V^y \geq k \,|\, X_0 = x) = f_{xy} f_{yy}^{k-1}$. If y is recurrent, then $f_{yy} = 1$. Consequently,

$$\mathbb{P}(V^y = \infty \,|\, X_0 = x) = \lim_{k\to\infty} \mathbb{P}(V^y \geq k \,|\, X_0 = x) = \lim_{k\to\infty} f_{xy} f_{yy}^{k-1} = f_{xy}\,.$$

Thus if $f_{xy} > 0$, then $\mathbb{E}_x(V^y) = \infty$.
(b) Assume y is transient, so $f_{yy} < 1$. Then

$$\mathbb{P}(V^y = \infty \,|\, X_0 = x) = \lim_{k\to\infty} \mathbb{P}(V^y \geq k \,|\, X_0 = x) = \lim_{k\to\infty} f_{xy} f_{yy}^{k-1} = 0\,.$$

So for a transient state y, the random variable V^y is finite with probability 1, no matter what state x the Markov chain starts in (if $f_{xy} = 0$, then $V^y \equiv 0$).
Recall that the expectation of a nonnegative, integer-valued random variable Y is

$$\mathbb{E}(Y) = \sum_{k=1}^{\infty} \mathbb{P}(Y \geq k)\,.$$

Thus we have for the expectation of V^y,

$$\mathbb{E}_x(V^y) = \sum_{k=1}^{\infty} \mathbb{P}(V^y \geq k \,|\, X_0 = x) = \sum_{k=1}^{\infty} f_{xy} f_{yy}^{k-1} = \frac{f_{xy}}{1 - f_{yy}} < \infty$$

for all $x \in \mathcal{S}$. This completes the proof of the theorem. □

Let $y \in \mathcal{S}$. Recall that the indicator function $\mathbb{1}_y$ on $\mathcal{S}$ is defined by

$$\mathbb{1}_y(z) = \begin{cases} 1 & \text{for } z = y \\ 0 & \text{for } z \neq y\,. \end{cases}$$

Then

$$V^y = \sum_{n=1}^{\infty} \mathbb{1}_y(X_n)\,,$$

and, since $\mathbb{E}_x(\mathbb{1}_y(X_n)) = \mathbb{P}(X_n = y \,|\, X_0 = x) = P^n_{xy}$, and by the Monotone Convergence Theorem (see Corollary C.3.2),

$$\mathbb{E}_x(V^y) = \sum_{n=1}^{\infty} P^n_{xy}\,. \tag{2.2}$$

Corollary 2.1.4. *If y is a transient state, then*

$$\lim_{n\to\infty} P^n_{xy} = 0$$

for all $x \in \mathcal{S}$.

The following proposition shows that recurrence is in some sense "contagious":

Proposition 2.1.5. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $x, y \in \mathcal{S}$. If x is recurrent and $x \longrightarrow y$, then*

(a) y is also recurrent, and

(b) $y \longrightarrow x$, and $f_{xy} = f_{yx} = 1$.

Proof. Assume $x \neq y$. Since $x \longrightarrow y$, there exists a $k \geq 1$ such that $P^k_{xy} > 0$. If we had $f_{yx} < 1$, then with probability $(1 - f_{yx}) > 0$, the Markov chain, once in state y, would never visit x at any future time. It follows that

$$\mathbb{P}(T^x = \infty \,|\, X_0 = x) = 1 - f_{xx} \geq P^k_{xy}(1 - f_{yx}) > 0\,.$$

However, since x is recurrent, $f_{xx} = 1$, and so it must be that $f_{yx} = 1$. In particular, $y \longrightarrow x$. Since $y \longrightarrow x$, there exists an $\ell \geq 1$ such that $P^\ell_{yx} > 0$. We have

$$\mathbb{E}_y(V^y) = \sum_{n=1}^{\infty} P^n_{yy} \geq \sum_{m=1}^{\infty} P^\ell_{yx} P^m_{xx} P^k_{xy} = P^\ell_{yx} P^k_{xy} \sum_{m=1}^{\infty} P^m_{xx} = \infty\,,$$

and thus

$$\mathbb{E}_y(V^y) = \infty\,,$$

which implies that y is recurrent by Proposition 2.1.3. Switching the roles of x and y in the argument, yields $f_{xy} = 1$. This completes the proof. □

Notes: (1) It follows from the previous proposition that a recurrent state can never lead into a transient state. However, a transient state can lead into transient and recurrent states.
(2) Furthermore, it follows that if two states communicate with each other, then either both states are recurrent or both states are transient. Hence **transience and recurrence are class properties.**

Proposition 2.1.6. *Let $(X_n)_{n\geq 0}$ be a Markov chain with* **finite** *state space $\mathcal{S}$. Then $\mathcal{S}$ contains at least one recurrent state.*

Proof. Assume all states in $\mathcal{S}$ are transient. Then by Theorem 2.1.3(b), for all $x, y \in \mathcal{S}$,

$$\mathbb{E}_x(V^y) = \sum_{n=1}^{\infty} P_{xy}^n < \infty\,.$$

Since $|\mathcal{S}| < \infty$, we have

$$\sum_{y\in\mathcal{S}} \sum_{n=1}^{\infty} P_{xy}^n < \infty\,. \tag{2.3}$$

Since the iterated double sum in (2.3) converges absolutely, any reordering of the terms in the summation yields the same answer. However, we have

$$\sum_{n=1}^{\infty} \sum_{y\in\mathcal{S}} P_{xy}^n = \sum_{n=1}^{\infty} 1 = \infty\,,$$

which is a contradiction. It follows that at least one element in $\mathcal{S}$ must be recurrent. □

Corollary 2.1.7. *An irreducible, finite state Markov chain is recurrent.*

As a consequence of Proposition 2.1.6, the state space $\mathcal{S}$ of a finite state Markov chain has at least one irreducible closed class consisting of recurrent states. Note however that a Markov chain with infinite state space $\mathcal{S}$ may be transient, that is, it may not have a single recurrent state.

Example 2.1.2 (Success runs)**.** Consider a Markov chain on state space $\mathcal{S} = \mathbb{N}_0$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} q_0 & p_0 & 0 & 0 & \cdots \\ q_1 & 0 & p_1 & 0 & \cdots \\ q_2 & 0 & 0 & p_2 & \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix} .$$

The transition graph is shown in Figure 2.3. We assume that $p_k \in (0,1)$ for al $k \in \mathbb{N}_0$.

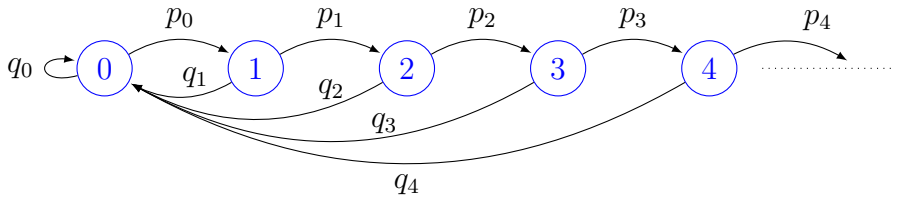


Figure 2.3

With this assumption, the chain is irreducible (and therefore either all states are recurrent or all states are transient). We will compute f_{00}. Due to the special structure of the transition matrix, for each $n \geq 1$, there is exactly one path that starts at 0 and returns back to 0 for the first time after n steps. That path is $0, 1, 2, ..., n-1, 0$, and so we have

$$\mathbb{P}_0(T^0 = n) = p_0\, p_1 \cdots p_{n-2}\, q_{n-1}$$

and

$$\mathbb{P}_0(T^0 > n) = p_0\, p_1 \cdots p_{n-1}$$

from which we get

$$\mathbb{P}_0(T^0 = \infty) = \lim_{n\to\infty} \mathbb{P}_0(T^0 > n) = \lim_{n\to\infty} \prod_{k=0}^{n-1} p_k\,.$$

By Lemma A.5.1,

$$\lim_{n\to\infty} \prod_{k=0}^{n-1} p_k = 0 \quad \Longleftrightarrow \quad \sum_{k=0}^{\infty} (1-p_k) = \sum_{k=0}^{\infty} q_k = \infty\,.$$

Since

$$f_{00} = 1 - \mathbb{P}_0(T^0 = \infty)\,,$$

it follows that the *success run chain is recurrent iff* $\sum_{k=0}^{\infty} q_k = \infty$ *and transient iff* $\sum_{k=0}^{\infty} q_k < \infty$. □

Example 2.1.3. Consider the Markov chain on state space $\mathbb{N}_0$ with transition probabilities $P_{00} = P_{01} = P_{10} = P_{12} = \frac{1}{2}$, and $P_{k,0} = \frac{1}{k^2}$, $P_{k,k+1} = \frac{k^2-1}{k^2}$ for $k \geq 2$, and zero otherwise. This is a success run chain for which

$$\sum_{k=0}^{\infty} q_k = 1 + \sum_{k=2}^{\infty} \frac{1}{k^2} < \infty \,.$$

All states of this chain are transient. □

If the Markov chain is reducible and the state space $\mathcal{S}$ contains at least one recurrent state, we consider the decomposition

$$\mathcal{S} = \mathcal{R} \cup \mathcal{T}$$

where $\mathcal{R}$ is the set of recurrent states and $\mathcal{T}$ is the set of transient states. The set of recurrent states $\mathcal{R}$ further partitions into the *disjoint* union of k irreducible closed classes $R_1, ..., R_k$ (since the relation $x \longleftrightarrow y$ is symmetric and transitive on $\mathcal{R}$). The decomposition

$$\mathcal{S} = (R_1 \cup \cdots \cup R_k) \cup \mathcal{T}$$

is called the *canonical decomposition* of the state space $\mathcal{S}$. Thus, under a reordering of the states in $\mathcal{S}$, the one-step transition matrix $\mathbf{P}$ of a finite state Markov chain can be written in *canonical form* as

$$\mathbf{P}_{can} = \begin{array}{c} \\ R_1 \\ \vdots \\ \vdots \\ R_k \\ \mathcal{T} \end{array} \begin{array}{c} \begin{array}{cccc} R_1 & \cdots & \cdots & R_k \end{array} \mathcal{T} \\ \left(\begin{array}{cccc|c} \mathbf{P}_1 & & & & \\ & \mathbf{P}_2 & & & \mathbf{0} \\ & & \ddots & & \\ & & & \mathbf{P}_k & \\ \hline \cdots & \mathbf{T} & \cdots & \cdots & \mathbf{Q} \end{array}\right) \end{array} \tag{2.4}$$

where the top left block is a block diagonal matrix consisting of k square-matrix blocks, each of which is made up of the one-step transition probabilities for one of the irreducible closed classes of recurrent states. The rest of the matrix entries (below the horizontal line) correspond to transition probabilities involving transient states. In this format, the n-step transition matrix $\mathbf{P}^n$ becomes

$$\mathbf{P}^n_{can} = \left(\begin{array}{cccc|c} \mathbf{P}_1^n & & & & \\ & \mathbf{P}_2^n & & & \mathbf{0} \\ & & \ddots & & \\ & & & \mathbf{P}_k^n & \\ \hline \cdots & \mathbf{T}_n & \cdots & \cdots & \mathbf{Q}^n \end{array}\right) . \tag{2.5}$$

Definition 2.1.3. *A square matrix* $\mathbf{P}$ *is called* **substochastic** *if*

- *all matrix entries are nonnegative, and*
- *each row sum is less than or equal to* 1.

Note that $\mathbf{Q}$ is a substochastic matrix. By Corollary 2.1.4 we have

$$\lim_{n\to\infty} \mathbf{Q}^n = \mathbf{0}\,.$$

Example 2.1.4. Recall Example 1.6.3. The Markov chain has state space $\mathcal{S} = \{1, 2, ..., 6\}$ and transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0.1 & 0 & 0.9 \\ 0.1 & 0.3 & 0.2 & 0 & 0.4 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.1 & 0 & 0.4 \\ 0 & 0.6 & 0.3 & 0.1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Its transition graph is shown in Figure 2.4.

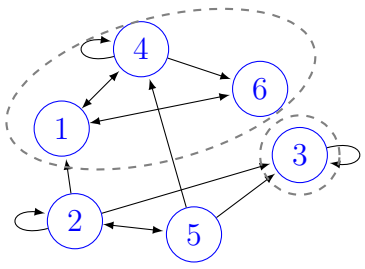


Figure 2.4

The irreducible closed classes (circled in gray) are $R_1 = \{3\}$ and $R_2 = \{1, 4, 6\}$. For this Markov chain the set of recurrent states is $\mathcal{R} = R_1 \cup R_2 = \{1, 3, 4, 6\}$, and the set of

transient states is $\mathcal{T} = \{2, 5\}$. The canonical form of the transition matrix is

$$\mathbf{P}_{can} = \begin{array}{c} \\ 3 \\ 1 \\ 4 \\ 6 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 3 & 1 & 4 & 6 & 2 & 5 \end{array} \\ \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0 & 0.5 & 0.1 & 0.4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0.2 & 0.1 & 0 & 0 & 0.3 & 0.4 \\ 0.3 & 0 & 0.1 & 0 & 0.6 & 0 \end{array}\right) \end{array}$$

from which we read off the sub-matrices

$$\mathbf{P_1} = (1) \qquad \mathbf{P_2} = \begin{pmatrix} 0 & 0.1 & 0.9 \\ 0.5 & 0.1 & 0.4 \\ 1 & 0 & 0 \end{pmatrix} \qquad \mathbf{Q} = \begin{pmatrix} 0.3 & 0.4 \\ 0.6 & 0 \end{pmatrix} .$$

$\mathbf{P_1}$ and $\mathbf{P_2}$ are stochastic matrices, and $\mathbf{Q}$ is a substochastic matrix. □

The canonical decomposition of the state space simplifies the study the dynamical behavior of the Markov chain in that it allows us to restrict our focus to smaller parts of the state space. If the Markov chain starts in a recurrent state x and $x \in R_k$, the chain will forever remain in R_k and will visit each state in R_k infinitely often with probability 1. If the Markov chain has only a finite number of transient states and it starts in one of these transient states, with probability 1 the Markov chain will enter one of the irreducible closed classes R_j of recurrent states in finite time. We call the time at which this happens the time of *absorption*. From the time of absorption onwards, the chain will remain within R_j and visit all states in R_j infinitely often with probability 1. We will discuss in detail questions surrounding absorption in Section 2.3.

2.2 Stationary distributions

In this section we focus on the question of existence and on basic properties of *stationary distributions* for a Markov chain. This special class of distributions plays an important role in the long-term behavior of the Markov chain. We have introduced the notion of a stationary distribution in Definition 1.3.2. Recall:

Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$. A probability distribution π on $\mathcal{S}$ is called a **stationary** or **invariant distribution** if

$$\pi(y) = \sum_{x\in\mathcal{S}} \pi(x) P_{xy} \quad \text{for all } y \in \mathcal{S}\,, \tag{2.6}$$

or equivalently (with π as a row vector), if

$$\pi \mathbf{P} = \pi \,. \tag{2.7}$$

Note that for a finite transition matrix $\mathbf{P}$, a stationary distribution π is a nonnegative left eigenvector corresponding to eigenvalue 1. Since $\mathbf{P}$ is stochastic, any constant column vector is a *right* eigenvector corresponding to eigenvalue 1, and so $\mathbf{P}$ is guaranteed to have a *left* eigenvector corresponding to eigenvalue 1. If $\mathbf{P}$ is a strictly positive matrix, then the Perron–Frobenius Theorem (Theorem A.6.1) guarantees that such a left eigenvector for $\mathbf{P}$ can be normalized to a (strictly positive) probability vector. However, for Markov chains with infinite state space $\mathcal{S}$, results from linear algebra such as the Perron–Frobenius Theorem are not available.

Since for any stationary distribution π we have $\pi \mathbf{P}^n = \pi$ for all $n \geq 1$, once the Markov chain is in stationary distribution π, it will remain *in stationarity* forever. We think of being in stationarity as a kind of *equilibrium* state for the chain. The notion of *probability flux* further explores this.

Definition 2.2.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and π a stationary distribution for the chain. Let $\mathcal{A}$ and $\mathcal{B}$ be two disjoint subsets of $\mathcal{S}$. The* **probability flux from $\mathcal{A}$ to $\mathcal{B}$** *is defined by*

$$\text{flux}(\mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} \pi(x) P_{xy} \tag{2.8}$$

Notice that by (2.7), we have

$$\pi(y) \sum_{x \in \mathcal{S}} P_{yx} = \sum_{x \in \mathcal{S}} \pi(x) P_{xy} \tag{2.9}$$

and consequently,

$$\sum_{x \in \mathcal{S} \setminus \{y\}} \pi(y) P_{yx} = \sum_{x \in \mathcal{S} \setminus \{y\}} \pi(x) P_{xy},$$

which says that

$$\text{flux}\left(\{y\}, \mathcal{S} \setminus \{y\}\right) = \text{flux}(\mathcal{S} \setminus \{y\}, \{y\})$$

for all $y \in \mathcal{S}$. In fact, more general *global balancing* equations hold for a Markov chain in stationarity:

Proposition 2.2.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and π a stationary distribution for the chain. Consider a proper subset $\mathcal{A} \subset \mathcal{S}$. Then*

$$\text{flux}(\mathcal{A}, \mathcal{A}^c) = \text{flux}(\mathcal{A}^c, \mathcal{A}). \tag{2.10}$$

Proof. Summing both sides of (2.9) over $y \in \mathcal{A}$ yields

$$\begin{aligned} \sum_{y\in\mathcal{A}}\sum_{x\in\mathcal{S}} \pi(y)P_{yx} &= \sum_{y\in\mathcal{A}}\sum_{x\in\mathcal{S}} \pi(x)P_{xy} \\ \sum_{y\in\mathcal{A}}\left(\sum_{x\in\mathcal{A}} \pi(y)P_{yx} + \sum_{x\in\mathcal{A}^c} \pi(y)P_{yx}\right) &= \sum_{y\in\mathcal{A}}\left(\sum_{x\in\mathcal{A}} \pi(x)P_{xy} + \sum_{x\in\mathcal{A}^c} \pi(x)P_{xy}\right). \end{aligned}$$

Subtracting $\sum_{y\in\mathcal{A}}\sum_{x\in\mathcal{A}} \pi(y)P_{yx}$ from both sides in the second line yields (2.10). □

Proposition 2.2.2. *Let π be a stationary distribution for the Markov chain $(X_n)_{n\geq 0}$. Then $\pi(y) = 0$ for any transient state $y \in \mathcal{S}$.*

Proof. Assume π is a stationary distribution and y is a transient state. We have

$$\sum_{x\in\mathcal{S}} \pi(x)P^n_{xy} = \pi(y) \quad \text{for all } n \geq 1\,.$$

By Corollary 2.1.4, we have $\lim_{n\to\infty} P^n_{xy} = 0$ for all $x \in \mathcal{S}$. By the Bounded Convergence theorem (see Appendix C), we can interchange limit and summation in the following sum and get

$$\lim_{n\to\infty}\sum_{x\in\mathcal{S}} \pi(x)P^n_{xy} = \sum_{x\in\mathcal{S}} \pi(x)\lim_{n\to\infty} P^n_{xy} = 0\,.$$

So $\pi(y) = 0$. □

Proposition 2.2.3. *Let π be a stationary distribution and $\pi(x) > 0$ for some (necessarily recurrent) state $x \in \mathcal{S}$. If $x \longrightarrow y$, then $\pi(y) > 0$. As a consequence, a stationary distribution π is either everywhere strictly positive or everywhere zero on an irreducible, closed class of recurrent states.*

Proof. Since $x \longrightarrow y$, there exists an n such that $P^n_{xy} > 0$. So

$$\pi(y) = \sum_{z\in\mathcal{S}} \pi(z)P^n_{zy} \geq \pi(x)P^n_{xy} > 0\,.$$

□

In praxis, how do we compute a stationary distribution? We can always take a brute force approach and attempt to directly solve the (finite or infinite) system of linear equations (1.11). Other, often computationally faster, approaches require more knowledge about the properties of stationary distributions which we will discuss in the following sections.

2.2.1 Existence and uniqueness of an invariant measure

In this section, we will consider more general measures on $\mathcal{S}$, not only probability measures. We will return to probability measures in the following section.

Definition 2.2.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain on a discrete state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. A* **nonnegative measure** *μ on $\mathcal{S}$ assigns a value $\mu(x) \in [0,\infty)$ to each $x \in \mathcal{S}$. We identify μ with a (finite or infinite) nonngegative row vector. A nonnegative measure μ is called an* **invariant measure** *if*

$$\mu(y) = \sum_{x\in\mathcal{S}} \mu(x) P_{xy} \quad \text{for all } y \in \mathcal{S}\,, \tag{2.11}$$

or equivalently,

$$\mu\mathbf{P} = \mu\,.$$

Note that a Markov chain can have a non-trivial invariant measure, but not a stationary probability distribution. An example for this is simple symmetric random walk on $\mathbb{Z}$. The constant measure $\mu = 1$ on $\mathbb{Z}$ is an invariant measure (in fact, it is the unique invariant measure, up to a multiplicative constant), but it cannot be normalized to a probability measure on $\mathbb{Z}$.

Let $x \in \mathcal{S}$. By the strong Markov property, the Markov chain "probabilistically renews" itself after each visit to x. Thus what happens in between two consecutive visits to x should, in some sense, be typical for the evolution of the Markov chain in the long run. The following results will make this precise for the average amount of time a Markov chain spends in each state. For the analysis to make sense, we will need to assume that the chain that starts in state x will return to state x in finite time with probability 1, in other words we need to assume recurrence.

Theorem 2.2.4 (Existence of an invariant measure)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, recurrent Markov chain with state space $\mathcal{S}$. Let $x \in \mathcal{S}$ be an arbitrary state, and assume $X_0 = x$. Consider the first return time T^x, and define $\mu(y)$ to be the expected number of visits to state y strictly before time T^x, that is,*

$$\mu(y) = \mathbb{E}_x\left(\sum_{n=0}^{T^x-1} \mathbb{1}_{\{X_n=y\}}\right) = \mathbb{E}_x\left(\sum_{n=0}^{\infty} \mathbb{1}_{\{X_n=y\}}\mathbb{1}_{\{T^x>n\}}\right) \quad \text{for } y \in \mathcal{S}\,. \tag{2.12}$$

Then $0 < \mu(y) < \infty$ for all $y \in \mathcal{S}$ and μ is an invariant measure for the Markov chain.

Proof. Note that the definition of μ implies that $\mu(x) = 1$ and

$$\sum_{y\in\mathcal{S}} \mu(y) = \mathbb{E}_x(T^x)\,. \tag{2.13}$$

Consider the following probabilities

$$\tilde{p}_y(n) = \mathbb{P}(\mathbb{1}_{\{X_n=y\}}\mathbb{1}_{\{T^x>n\}} = 1) = \mathbb{P}_x(X_1 \neq x, ..., X_{n-1} \neq x, X_n = y)$$

for $y \in \mathcal{S}$, $y \neq x$, and $n \geq 1$. Note that $\tilde{p}_y(1) = P_{xy}$, and that $\tilde{p}_y(n) = \mathbb{E}_x(\mathbb{1}_{\{X_n=y\}}\mathbb{1}_{\{T^x>n\}})$. It follows that

$$\mu(y) = \sum_{n=1}^{\infty} \tilde{p}_y(n) \qquad \text{for } y \neq x\,.$$

Conditioning on the state the chain visits at the $(n-1)$th step, we get for $n \geq 2$,

$$\tilde{p}_y(n) = \sum_{z\neq x} \tilde{p}_z(n-1)\, P_{zy}\,. \tag{2.14}$$

Summing both sides over $n \geq 2$ and adding $\tilde{p}_y(1)$ yields

$$\tilde{p}_y(1) + \sum_{n=2}^{\infty} \tilde{p}_y(n) = P_{xy} + \sum_{n=2}^{\infty}\sum_{z\neq x} \tilde{p}_z(n-1)\, P_{zy}\,,$$

and thus (after changing the order of summation on the right hand side), we get

$$\mu(y) = \sum_{z\in\mathcal{S}} \mu(z)\, P_{zy}\,.$$

This shows that μ is an invariant measure. Strict positivity of μ follows from essentially the same proof as for Proposition 2.2.3 (recall that $\mu(x) = 1 > 0$ and the chain is assumed to be irreducible). Lastly, we show that $\mu(y) < \infty$ for all $y \in \mathcal{S}$. Assume $\mu(y) = \infty$ for some $y \in \mathcal{S}$. But since

$$\mu(x) = 1 = \sum_{z\in\mathcal{S}} \mu(z) P^n_{zx} \quad \text{ for all } n \geq 1\,,$$

it follows that $P^n_{yx} = 0$ for all $n \geq 1$, and hence y does not lead to x. This contradicts the assumption that the Markov chain is irreducible. It follows that $0 < \mu(y) < \infty$ for all $y \in \mathcal{S}$. □

Remark 2.2.5. *Theorem 2.2.4 holds more generally (we omit the proof): If the Markov chain starts in initial distribution π_0 and if T is any a.s. finite stopping time, that is $\mathbb{P}(T < \infty) = 1$, such that $X_T \sim \pi_0$, then μ defined by*

$$\mu(y) = \mathbb{E}_{\pi_0}\left(\sum_{n=1}^{\infty} \mathbb{1}_{\{X_n=y\}}\mathbb{1}_{\{T>n\}}\right) \quad \text{for } y \in \mathcal{S} \tag{2.15}$$

defines a strictly positive, invariant measure for the Markov chain. Theorem 2.2.4 proves this statement for the special case $\pi_0 = \delta_x$ (unit mass at state x) and $T = T^x$.

Example 2.2.1. An example of a stopping time (other than T^x) to which Remark 2.2.5 applies is the **commute time** between distinct states x and y, denoted by $T^{x\leftrightarrow y}$. It is defined by

$$T^{x\leftrightarrow y} = \min\{n > T^y : X_n = x\}\,, \text{ given that } X_0 = x\,,$$

that is, $T^{x\leftrightarrow y}$ is the time of the first return to x after the first visit to y. Since the Markov chain does not revisit x after time T^y and before time $T^{x\leftrightarrow y}$, $\mu(x)$ in (2.15) becomes

$$\mu(x) = \mathbb{E}_x\left(\sum_{n=0}^{T^y-1} \mathbb{1}_{\{X_n=x\}}\right)\,,$$

and so $\mu(x)$ is the expected number of visits (including the visit at time $n = 0$) to x before time T^y. Let us use the notation $V^x_{T^y}$ for the number of visits to state x before time T^y. Note that $\mathbb{P}_x(V^x_{T^y} \geq 1) = 1$ since we are including the starting state x in $V^x_{T^y}$. Then

$$\mathbb{P}_x(V^x_{T^y} \geq 2) = \mathbb{P}_x(T^x < T^y)\,.$$

Invoking the Strong Markov property for the stopping time T^x, we get

$$\mathbb{P}_x(V^x_{T^y} \geq 3) = \left[\mathbb{P}_x(T^x < T^y)\right]^2\,,$$

and by induction,

$$\mathbb{P}_x(V^x_{T^y} \geq n) = \left[\mathbb{P}_x(T^x < T^y)\right]^{n-1}\,.$$

This shows that the random variable $V^x_{T^y}$ has a geometric distribution, and its expectation is

$$\mu(x) = \mathbb{E}_x(V^x_{T^y}) = \sum_{n\geq 1}[\mathbb{P}_x(T^x < T^y)]^{n-1} = \frac{1}{\mathbb{P}_x(T^y < T^x)}\,.$$

(See also Exercise 2.21.) We will return to a discussion of $T^{x\leftrightarrow y}$ in Chapter 8. □

Theorem 2.2.6 (Uniqueness of the invariant measure)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, recurrent Markov chain with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. The invariant measure μ constructed in Theorem 2.2.4 is the unique invariant measure up to a positive multiplicative factor.*

Proof. Our proof follows the proof given in [6]. Assume ν is an invariant measure for $(X_n)_{n\geq 0}$. Since the Markov chain is irreducible, we have $\nu(y) > 0$ for all $y \in \mathcal{S}$. Instead of working with the transition probabilities P_{xy}, we will now consider the following modified transition probabilities $\bar{P}_{xy}$ defined by

$$\bar{P}_{xy} = \frac{\nu(y)}{\nu(x)} P_{yx} \quad \text{for all } x, y, \in \mathcal{S}\,.$$

We recognize that $\bar{P}_{xy}$ are the transition probabilities of the time-reversed chain (see Section 7.1) whose transition matrix we denote by $\bar{\mathbf{P}}$. It is straightforward to verify that $\sum_{y\in\mathcal{S}} \bar{P}_{xy} = 1$ for all $x \in \mathcal{S}$ and that the corresponding n-step transition probabilities are

$$\bar{P}^n_{xy} = \frac{\nu(y)}{\nu(x)} P^n_{yx} \quad \text{for all } x, y, \in \mathcal{S} \text{ and } n \geq 1\,.$$

Furthermore, $\bar{\mathbf{P}}$ is irreducible and recurrent. Irreducibility follows from the fact that $\bar{P}^n_{xy} > 0 \iff P^n_{yx} > 0$. To show recurrence, note that $\bar{P}^n_{xx} = P^n_{xx}$ for all $n \geq 1$, and use (2.2) and Theorem 2.1.3. Fix state x. We introduce the following notation for the Markov chain with $\bar{\mathbf{P}}$ for the probability of visiting x for the first time at time n, given that the chain starts in state y:

$$\bar{f}^{(n)}_{yx} = \mathbb{P}_y(T^x = n)\,.$$

Conditioning on the state visited at time 1, we get the recurrence relation

$$\bar{f}^{(n+1)}_{yx} = \sum_{z\neq x} \bar{P}_{yz} \bar{f}^{(n)}_{zx}\,,$$

and from this, by using $\bar{P}_{yz} = \frac{\nu(z)}{\nu(y)} P_{zy}$,

$$\nu(y) \bar{f}^{(n+1)}_{yx} = \sum_{z\neq x} \nu(z) P_{zy} \bar{f}^{(n)}_{zx}\,. \tag{2.16}$$

Now recall equation (2.14). We can rewrite equation (2.14) as

$$\nu(x)\, \tilde{p}_y(n+1) = \sum_{z\neq x} (\nu(x) \tilde{p}_z(n))\, P_{zy}\,. \tag{2.17}$$

Take a closer look at (2.16) and (2.17): Both equations define a recurrence relation for sequences (indexed by elements of $\mathcal{S}$) dependent on n. These two (time n dependent) sequences are

$$(\nu(y)\,\bar{f}_{yx}^{(n)})_{y\in\mathcal{S}} \quad \text{and} \quad (\nu(x)\,\tilde{p}_y(n))_{y\in\mathcal{S}} \tag{2.18}$$

(remember that x is a fixed state). Equations (2.16) and (2.17) define the *same* recurrence relation for both sequences in (2.18). At time $n = 1$, (2.18) becomes

$$(\nu(y)\,\bar{f}_{yx}^{(1)})_{y\in\mathcal{S}} = (\nu(y)\bar{P}_{yx})_{y\in\mathcal{S}} \quad \text{and} \quad (\nu(x)\,\tilde{p}_y(1))_{y\in\mathcal{S}} = (\nu(x)\,P_{xy})_{y\in\mathcal{S}}\,.$$

But for all $y \in \mathcal{S}$, we have $\nu(y)\,\bar{P}_{yx} = \nu(x)\,P_{xy}$. So for the base case $n = 1$, the two sequences are equal. And since the two sequences are subject to the same recurrence relation, they are equal for all $n \geq 1$. Thus we get from (2.18),

$$\tilde{p}_y(n) = \frac{\nu(y)}{\nu(x)}\bar{f}_{yx}^{(n)} \quad \text{for all } n \geq 1\,. \tag{2.19}$$

As a last step, we sum both sides of (2.40) over n. We get

$$\sum_{n=1}^{\infty}\tilde{p}_y(n) = \mu(y) = \frac{\nu(y)}{\nu(x)}\sum_{n=1}^{\infty}\bar{f}_{yx}^{(n)} = \frac{\nu(y)}{\nu(x)}$$

where the last equation follows from the recurrence of the chain $\bar{\mathbf{P}}$ (which implies $\sum_{n=1}^{\infty}\bar{f}_{yx}^{(n)} = \mathbb{P}_y(T^x < \infty) = 1$). Altogether, this shows that $\nu = \nu(x)\,\mu$. Hence ν and μ differ only by the multiplicative factor $\nu(x)$. This completes the proof. □

Example 2.2.2 (Biased simple random walk on $\mathbb{Z}$)**.** Let $p \in (0,1)$ and $p \neq \frac{1}{2}$. Consider the Markov chain $(X_n)_{n\geq 0}$ on state space $\mathbb{Z}$ with transition probabilities

$$P_{x,x+1} = p\,, \quad P_{x,x-1} = 1-p \qquad \text{for all } x \in \mathbb{Z}\,.$$

It is called biased simple random walk on $\mathbb{Z}$. Clearly, $(X_n)_{n\geq 0}$ is irreducible. Set $1-p = q$. For a nonnegative measure μ on $\mathbb{Z}$ we will write μ_x for $\mu(x)$, $x \in \mathbb{Z}$. We compute an invariant measure $\mu = (..., \mu_{-1}, \mu_0, \mu_1, ...)$ for $(X_n)_{n\geq 0}$ by solving the system (2.11) which reads

$$\mu_x = \mu_{x-1}p + \mu_{x+1}q \qquad \text{for } x \in \mathbb{Z}\,. \tag{2.20}$$

The linear system (2.20) has at least two distinct solutions (each one up to a multiplicative constant): One solution is the constant measure $\mu = (..., 1, 1, 1, ...)$. A second solution is the measure $\tilde{\mu}$ defined by

$$\tilde{\mu}_x = \left(\frac{p}{q}\right)^x \qquad \text{for } x \in \mathbb{Z}\,.$$

Note that neither μ nor $\tilde{\mu}$ can be normalized to a probability measure on $\mathbb{Z}$. It follows from Theorem 2.2.6 (uniqueness of an invariant measure for an irreducible, recurrent chain) that **biased simple random walk on $\mathbb{Z}$ is transient**. □

2.2.2 Positive recurrence versus null recurrence

Let $(X_n)_{n\geq 0}$ be an irreducible, recurrent Markov chain. Let $x \in \mathcal{S}$, and consider the strictly positive measure μ on $\mathcal{S}$ as defined in (2.12). By Theorem 2.2.4 and Theorem 2.2.6, the measure μ is the unique invariant measure (up to a multiplicative constant) for the Markov chain. By (2.3.5), the measure μ can be normalized to a probability distribution (a stationary distribution) if and only if $\mathbb{E}_x(T^x) < \infty$. This motivates the following definition.

Definition 2.2.3 (Positive recurrence and null recurrence)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and assume $x \in \mathcal{S}$ is a recurrent state.*

(a) We say x is **positive recurrent** *if $\mathbb{E}_x(T^x) < \infty$.*
(b) We say x is **null recurrent** *if $\mathbb{E}_x(T^x) = \infty$.*

We will use the notation $m_x = \mathbb{E}_x(T^x)$ for the mean return time to state x. Recall that for the invariant measure defined in (2.12), we have $\mu(x) = 1$. Thus, if μ can be normalized to a stationary distribution π, that is, if $m_x < \infty$, we get

$$\pi(x) = \frac{1}{m_x}\,.$$

The choice of x for the construction of a strictly positive invariant measure μ in Theorem 2.2.4 was arbitrary. By Theorem 2.2.6 (uniqueness), if we had chosen $y \in \mathcal{S}$ (with $y \neq x$) instead of x, the resulting strictly positive invariant measure $\tilde{\mu}$ would differ from μ only by a positive multiplicative constant. It follows that either both μ and $\tilde{\mu}$ can be normalized to a probability measure or both μ and $\tilde{\mu}$ cannot be normalized. Hence either both x and y are positive recurrent or both x and y are null recurrent. We fomulate this result in the following proposition:

Proposition 2.2.7 (Positive recurrence and null recurrence are class properties)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and $R \subseteq \mathcal{S}$ an irreducible closed class of recurrent states. Then either all states in R are positive recurrent or all states in R are null recurrent.*

Recall that an irreducible Markov chain with *finite* state space is recurrent. By Proposition 2.1.1, we know it is *positive recurrent*. Hence an irreducible, finite state Markov chain has a unique stationary distribution.

The following theorem summarizes our findings:

Theorem 2.2.8. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on state space $\mathcal{S}$. The Markov chain has a stationary distribution π if and only if it is positive recurrent. In the positive recurrent case, the stationary distribution is unique, and it is given by*

$$\pi(x) = \frac{1}{m_x} \qquad \text{for } x \in \mathcal{S}\,.$$

2.2.3 Stationary distributions for reducible chains

We now consider reducible Markov chains $(X_n)_{n\geq 0}$. Recall the canonical decomposition $\mathcal{S} = (R_1 \cup R_2 \cup ...) \cup \mathcal{T}$ of the states space $\mathcal{S}$ into irreducible closed classes of recurrent states R_k and the set of transient states $\mathcal{T}$. For any stationary distribution π for $(X_n)_{n\geq 0}$, we have $\pi(x) = 0$ if $x \in \mathcal{T}$ (Proposition 2.2.2).
If the chain starts in a recurrent state $y \in R_k$, then the chain will remain in R_k forever. In this case it suffices to study the Markov chain restricted to R_k. Under this restriction, the Markov chain is irreducible. If R_k consists of positive recurrent states, then there exists a unique stationary distribution $\pi^{res}_{R_k}$ on R_k for the restricted chain. We can extend $\pi^{res}_{R_k}$ to a probability measure π_{R_k} on the entire state space $\mathcal{S}$ by defining

$$\pi_{R_k}(z) = \begin{cases} \pi^{res}_{R_k}(z) & \text{for } z \in R_k \\ 0 & \text{for } z \notin R_k\,. \end{cases} \tag{2.21}$$

Note that π_{R_k} is a stationary distribution for the unrestricted chain $(X_n)_{n\geq 0}$ on $\mathcal{S}$.
If the chain starts in a recurrent state $x \in R_j$ and the irreducible closed class R_j consist of null recurrent states, then the restricted chain on R_j does not have a stationary distribution.
In sum: Each irreducible closed class R_k of positive recurrent states contributes a unique stationary distribution π_{R_k} (concentrated on R_k) for the Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{S}$.
We will make use of the following fact about stationary distributions (the proof is straightforward and left as an exercise):

Lemma 2.2.9. *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and assume $\pi^1, \pi^2, ..., \pi^k$ are k stationary distributions for $(X_n)_{n\geq 0}$. Then any convex mixture of the $\pi^1, \pi^2, ..., \pi^k$ is also a stationary distribution for $(X_n)_{n\geq 0}$. That is, for any nonnegative constants $c_1, c_2, ..., c_k$ with $\sum_{i=1}^k c_i = 1$, the measure*

$$\pi = \sum_{i=1}^{k} c_i \pi^i$$

is a stationary distribution for $(X_n)_{n\geq 0}$.

The following proposition summarizes our discussion with a description of the stationary distributions for a reducible Markov chain.

Proposition 2.2.10. *Let $(X_n)_{n\geq 0}$ be a reducible Markov chain on state space $\mathcal{S}$. Assume the chain has at least one positive recurrent state and that $R_1, R_2, ..., R_k$ are the irreducible closed classes of positive recurrent states for the chain. Then the stationary distributions π are exactly the distributions of the form*

$$\pi = \sum_{i=1}^{k} c_i \pi_{R_i}$$

with π_{R_i} as defined in (2.21), and $c_i \geq 0$ for $1 \leq i \leq k$, and $\sum_{i=1}^k c_i = 1$.

It follows that a reducible (or irreducible) Markov chain with exactly one irreducible closed class of positive recurrent states has a unique stationary distribution. A reducible chain with two or more irreducible closed classes of positive recurrent states has infinitely many stationary distributions. All other Markov chains have no stationary distributions.

2.2.4 Steady state distributions

Definition 2.2.4. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$. Suppose there exists a probability distribution λ on $\mathcal{S}$ such that*

$$\lim_{n\to\infty} P_{xy}^n = \lambda(y) \quad \text{for all } x \in \mathcal{S}\,, \tag{2.22}$$

then λ is called the **limiting** *or* **steady state distribution** *for the Markov chain.*

Note that if a Markov chain $(X_n)_{n\geq 0}$ has a limiting distribution λ, then for *any* initial

distribution π_0 for the chain, we have

$$\lim_{n\to\infty} \pi_n(x) = \lambda(x) \quad \text{for all } x \in \mathcal{S}\,.$$

This follows from averaging (2.22) over all initial states x with respect to π_0. Not every Markov chain has a limiting distribution.

Proposition 2.2.11. *If a Markov chain has a limiting distribution λ, then λ is a stationary distribution, and it is the unique stationary distribution for the chain.*

Proof. Assume $\lim_{n\to\infty} P^n_{xy} = \lim_{n\to\infty} P^{n+1}_{xy} = \lambda(y)$ for all $x \in \mathcal{S}$. By the Bounded Convergence theorem (see Appendix C), we can interchange lim and summation in the following and get

$$\lim_{n\to\infty} P^{n+1}_{xy} = \lim_{n\to\infty} \sum_{z\in\mathcal{S}} P^n_{xz} P_{zy} = \sum_{z\in\mathcal{S}} \lim_{n\to\infty} P^n_{xz} P_{zy} = \sum_{z\in\mathcal{S}} \lambda(z) P_{zy} = \lambda(y)$$

which shows that λ is a stationary distribution. Assume there exists another stationary distribution π with $\pi \neq \lambda$. If the chain starts in distribution π, it will always stay in distribution π, so the limiting distribution λ must be equal to π. This shows the uniqueness of the stationary distribution λ. □

Example 2.2.3. Recall the 2-state chain from Example 1.3.3. The unique stationary distribution $\pi = (\frac{b}{a+b}, \frac{a}{a+b})$ is the limiting distribution for the general 2-state chain. □

Example 2.2.4. Consider the Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{0, 1, 2\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}.$$

We compute

$$\mathbf{P}^2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad \mathbf{P}^3 = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$$

from which we conclude that $\mathbf{P}^{2n+1} = \mathbf{P}$ for all $n \geq 0$, and $\mathbf{P}^{2n} = \mathbf{P}^2$ for all $n \geq 1$. Due to this *periodic* behavior of the higher order transition matrices, (2.22) cannot hold. This Markov chain does not have a limiting distribution. Note however that, since $(X_n)_{n\geq 0}$ is irreducible and has finite state space, it is positive recurrent and has a unique stationary distribution π. The stationary distribution is $\pi = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. □

2.3 Absorbing chains

Let $(X_n)_{n\geq 0}$ be a reducible Markov chain on state space $\mathcal{S}$. Recall the canonical decomposition $\mathcal{S} = (R_1 \cup R_2 \cup \cdots) \cup \mathcal{T}$.

Definition 2.3.1. *A Markov chain that has at least one recurrent state and at least one transient state is called an* **absorbing chain**.

This section concerns several aspects of the long-term behavior of absorbing chains. We will focus on the following questions:

1. Given that a Markov chain starts in one of its transient states $x \in \mathcal{T}$, what is the probability that it eventually gets absorbed into a specific irreducible closed class R_k?
2. What is the probability of eventual absorption into any one of the irreducible closed classes? If eventual absorption is certain, what is the expected time until absorption?

2.3.1 First step analysis

A *first step analysis* is a very useful approach for many computations of functionals for Markov chains, among them absorption probabilities and the expected time until absorption. It is based on the idea of conditioning on the first step the Markov chain takes. Let $(X_n)_{n\geq 0}$ be a Markov chain. Recall that, by the Markov property, the one step forward shifted chain

$$(Y_n)_{n\geq 0} = (X_{1+n})_{n\geq 0}$$

is a Markov chain that has the same transition probabilities as $(X_n)_{n\geq 0}$ and, conditional on $X_1 = x$, is started in x and independent of X_0. A first step analysis approach exploits this fact. It allows us to write quantities of interest for $(X_n)_{n\geq 0}$ in terms of quantities for $(Y_n)_{n\geq 0}$ and in the process establishes equations for these quantities.

We demonstrate the approach with the computation of absorption probabilities. Let $(X_n)_{n\geq 0}$ be a Markov chain. We assume that 0 is an absorbing state, that the Markov chain has other recurrent states, and that the Markov chain has a finite number of transient states. Let $\mathcal{R}$ be the set of recurrent states and $\mathcal{T}$ the set of transient states for the chain. We also consider the stopping time

$$T = \min\{n \geq 0 : X_n \in \mathcal{R}\}\,.$$

If the chain starts in a transient state (which we will assume), then T is called the time of absorption. We also consider the analogous random variable

$$\tilde{T}\,\min\{k \geq 0 : Y_k \in \mathcal{R}\}$$

for the shifted chain $(Y_n)_{n\geq 0}$. Note that $\tilde{T} = T - 1$.
We assume that the Markov chain starts in a transient state x, and we are interested in the probability that the chain will eventually get absorbed in state 0. That is, we would like to compute the probability

$$\mathbb{P}(X_T = 0 \,|\, X_0 = x)\,.$$

We will use the notation $a_y = \mathbb{P}(X_T = 0 \,|\, X_0 = y)$ for $y \in \mathcal{S}$. Although we may only be interested in the probability a_x, a first step analysis will establish equations involving *all* probabilities a_y, $y \in \mathcal{S}$. Note that $a_0 = 1$ and $a_z = 0$ for all $z \in \mathcal{R}$ and $z \neq 0$. We have

$$\begin{aligned} a_x &= \sum_{y\in\mathcal{S}} \mathbb{P}(X_T = 0,\, X_1 = y \,|\, X_0 = x) \\ &= \sum_{y\in\mathcal{S}} \mathbb{P}(X_1 = y \,|\, X_0 = x)\, \mathbb{P}(X_T = 0 \,|\, X_0 = x,\, X_1 = y)\,. \end{aligned} \tag{2.23}$$

By the Markov property,

$$\mathbb{P}(X_T = 0 \,|\, X_0 = x,\, X_1 = y) = \mathbb{P}(Y_{\tilde{T}} = 0 \,|\, Y_0 = y) = \mathbb{P}(X_T = 0 \,|\, X_0 = y) = a_y\,.$$

Thus the second sum in (2.23) becomes $\sum_{y\in\mathcal{S}} P_{xy} a_y$, and we get the system of equations

$$a_x = \sum_{y\in\mathcal{S}} P_{xy} a_y \qquad \text{for } x \in \mathcal{S}\,. \tag{2.24}$$

Solving this system of linear equations in the variables a_y will simultaneously compute all absorption probabilities $\mathbb{P}(X_T = 0 \,|\, X_0 = y)$ (for any starting state y).

Note: System (2.24) always has a unique solution if the number of transient states is finite. This will be evident from the discussion in Section 2.3.2. However, for an infinite number of transient states, (2.24) results in an infinite system of equations which may have multiple solutions. Section 2.3.3 below addresses this situation.

Example 2.3.1. Recall Example 2.1.1. Here we take $p = \frac{2}{3}$. Figure 2.5 shows the transition graph. States $1, 2$, and 3 lead into the absorbing boundary (consisting of the absorbing states 0 and 4) and are therefore transient states. As above, T denotes the time until absorption. We are interested in the probabilities

$$a_i = \mathbb{P}(X_T = 0 \,|\, X_0 = i) \qquad \text{for } i = 1, 2, 3\,.$$

Note that here $a_0 = 1$ and $a_4 = 0$. Equations (2.24) read

$$\begin{aligned} a_1 &= \tfrac{2}{3} a_2 + \tfrac{1}{3} \\ a_2 &= \tfrac{2}{3} a_3 + \tfrac{1}{3} a_1 \\ a_3 &= \tfrac{1}{3} a_2 \end{aligned}$$

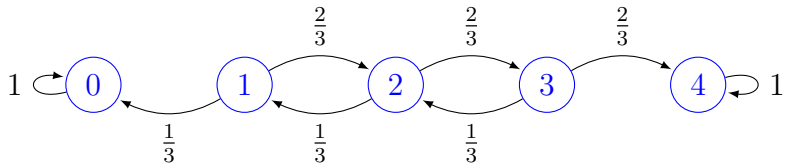


Figure 2.5

Solving this system yields the absorption probabilities (for absorption into state 0)

$$a_1 = \frac{7}{15} \qquad a_2 = \frac{1}{5} \qquad a_3 = \frac{1}{15}\,.$$

□

2.3.2 Finite number of transient states

In this section we will simplify the computations of absorption probabilities and expected hitting times with the use of linear algebra. Let $\mathbf{P}$ be the transition matrix of an absorbing Markov chain and assume that $|\mathcal{T}| < \infty$. Recall the canonical forms (2.4) and (2.5). By (2.2), given that the chain starts in state x, the expected number of visits to state $y \neq x$ is $\sum_{k=1}^{\infty} P_{xy}^k$. (If $x = y$, then the sum gives the expected number of returns to state x.) The matrix $\mathbf{G}$, defined by

$$\mathbf{G} = \sum_{k=0}^{\infty} \mathbf{P}^k\,,$$

is called the **potential matrix** for $\mathbf{P}$. Its (x, y)-entry is the expected number of visits to state y, given that the chain starts in state x. Note that if x or y is recurrent, then the (x, y)-entry in $\mathbf{G}$ can only be either 0 or ∞. If both x and y are transient, then the (x, y)-entry is finite and nonnegative. This is the most interesting case. The matrix $\mathbf{G}$ has the form

$$\mathbf{G} = \left(\begin{array}{cccc|c} \mathbf{G}_1 & & & & \\ & \mathbf{G}_2 & & & \mathbf{0} \\ & & \ddots & & \\ & & & \mathbf{G}_k & \\ \hline \cdots & \mathbf{T}^{(\infty)} & \cdots & \cdots & \mathbf{V} \end{array}\right) \tag{2.25}$$

where for all $1 \leq i \leq k$, the submatrix $\mathbf{G}_i = \sum_{k=0}^{\infty} \mathbf{P}_i^k$ is an $|R_i| \times |R_i|$-matrix all of whose entries are ∞, and the entries in the submatrix $\mathbf{T}^{(\infty)} = \sum_{k=1}^{\infty} \mathbf{T}_k$ are either 0 or ∞. Note that while throughout this section we assume $|\mathcal{T}| < \infty$, it is still possible that the Markov

chain has infinitely many irreducible closed classes R_i of recurrent states, or that at least one of the R_i is countably infinite. However, to simplify things, we will assume that this is not the case here. By Proposition 2.1.3(b), the entries in the finite square submatrix $\mathbf{V}$ in (2.25) are finite and nonnegative.

Definition 2.3.2. *The matrix* $\mathbf{V} = \sum_{k=0}^{\infty} \mathbf{Q}^k$ *in* (2.25) *is called the* **fundamental matrix** *of the Markov chain.*

Let $n = |\mathcal{T}|$ and let $\mathbf{I}$ denote the $n \times n$ identity matrix. Observe that for all $m \geq 0$, we have

$$(\mathbf{I} - \mathbf{Q}) \left(\sum_{k=0}^{m} \mathbf{Q}^k \right) = \left(\sum_{k=0}^{m} \mathbf{Q}^k \right) (\mathbf{I} - \mathbf{Q}) = \mathbf{I} - \mathbf{Q}^{m+1} . \tag{2.26}$$

Since (left or right) matrix multiplication by a constant matrix (here $(\mathbf{I} - \mathbf{Q})$) is a continuous operation, and since $\lim_{m \to \infty} \mathbf{Q}^m = \mathbf{0}$, taking the limit in (2.26) as $m \to \infty$ yields

$$(\mathbf{I} - \mathbf{Q})\mathbf{V} = \mathbf{V}(\mathbf{I} - \mathbf{Q}) = \mathbf{I} .$$

As a result, we have

$$\boxed{\mathbf{V} = (\mathbf{I} - \mathbf{Q})^{-1} .}$$

Note that for any $x, y \in \mathcal{T}$, the entry $v_{x,y}$ in the fundamental matrix $\mathbf{V}$ is the expected number of visits to state y *before absorption* into one of the irreducible closed classes of recurrent states, given that the Markov chain starts in state x. We summarize this result in the following proposition.

Proposition 2.3.1 (Expected time until absorption)**.** *Let* $(X_n)_{n \geq 0}$ *be a reducible Markov chain with transition matrix* $\mathbf{P}$ *and let* $\mathcal{T}$ *be the set of transient states. Assume* $|\mathcal{T}| < \infty$*. Consider the submatrix* $\mathbf{Q}$ *of* $\mathbf{P}$ *indexed by the elements in* $\mathcal{T}$*. Assume the Markov chain starts in a transient state* $x \in \mathcal{T}$*, and denote the time until absorption into one of the irreducible closed classes of recurrent states by* T^{abs}*. Then*

$$\mathbb{E}_x(T^{abs}) = \sum_{y \in \mathcal{T}} v_{x,y}$$

where the $v_{x,y}$ *are the matrix-entries in the fundamental matrix* $\mathbf{V} = (\mathbf{I} - \mathbf{Q})^{-1}$*.*

Example 2.3.2. We return to Example 2.1.4. The Markov chain has state space $\mathcal{S} = \{1, 2, ..., 6\}$ and its transition matrix (in canonical form) is

$$\mathbf{P}_{can} = \begin{array}{c} \\ 3 \\ 1 \\ 4 \\ 6 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 3 & 1 & 4 & 6 & 2 & 5 \end{array} \\ \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0 & 0.5 & 0.1 & 0.4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0.2 & 0.1 & 0 & 0 & 0.3 & 0.4 \\ 0.3 & 0 & 0.1 & 0 & 0.6 & 0 \end{array}\right) \end{array}. \tag{2.27}$$

Recall its transition graph from Figure 1.3.2. Here $\mathbf{Q} = \begin{pmatrix} 0.3 & 0.4 \\ 0.6 & 0 \end{pmatrix}$, from which we compute $\mathbf{V} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{pmatrix} 2.17 & 0.87 \\ 1.3 & 1.52 \end{pmatrix}$. Note that both transient states 2 and 5 lead into R_1 as well as into R_2. Therefore the potential matrix is

$$\mathbf{G} = \begin{array}{c} \\ 3 \\ 1 \\ 4 \\ 6 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 3 & 1 & 4 & 6 & 2 & 5 \end{array} \\ \left(\begin{array}{cccc|cc} \infty & 0 & 0 & 0 & 0 & 0 \\ 0 & \infty & \infty & \infty & 0 & 0 \\ 0 & \infty & \infty & \infty & 0 & 0 \\ 0 & \infty & \infty & \infty & 0 & 0 \\ \hline \infty & \infty & \infty & \infty & 2.17 & 0.87 \\ \infty & \infty & \infty & \infty & 1.3 & 1.52 \end{array}\right) \end{array}.$$

If the Markov chain starts in state 2, then the expected time until absorption (entry into either R_1 or R_2) is $2.17 + 0.87 = 3.04$. If the Markov chain starts in state 5, then the expected time until absorption is $1.3 + 1.52 = 2.82$. □

While Proposition 2.3.1 gives the expected time until absorption, we can say more about the distribution of T^{abs}. Again, consider the matrix $\mathbf{Q}$ which is indexed by the transient states $\mathcal{T}$. Its matrix entries are $Q_{x,y} = P_{xy}$ for $x, y \in \mathcal{T}$, the entries of $\mathbf{Q}^n$ are

$$(\mathbf{Q}^n)_{x,y} = \mathbb{P}_x(X_1 \in \mathcal{T}, X_2 \in \mathcal{T}, ..., X_{n-1} \in \mathcal{T}, X_n = y) \qquad \text{for } x, y \in \mathcal{T}\,.$$

This yields

$$\mathbb{P}_x(X_n \in \mathcal{T}) = \mathbb{P}_x(T^{abs} > n) = \sum_{y \in \mathcal{T}} (\mathbf{Q}^n)_{x,y} \qquad \text{for all } x, y \in \mathcal{T}\,. \tag{2.28}$$

Let us use the notation $a_x(n) = \mathbb{P}_x(T^{abs} > n)$. For the column vectors $\mathbf{a}(n) = (a_{x_1}(n), a_{x_2}(n), ...)^t$ and $\mathbf{1} = (1, 1, ...)^t$ (both indexed by the elements in $\mathcal{T}$), we can write (2.28) more compactly as

$$\mathbf{a}(n) = \mathbf{Q}^n \, \mathbf{1} \, . \tag{2.29}$$

Note that $\mathbb{P}_x(T^{abs} = n) = (\mathbf{a}(n-1) - \mathbf{a}(n))_x$ and that both (2.28) and (2.29) hold, regardless of $|\mathcal{T}|$ being *finite or infinite*. We can summarize this result in the following proposition.

Proposition 2.3.2 (Distribution of T^{abs})**.** *Let $(X_n)_{n\geq 0}$ be a reducible Markov chain with transition matrix $\mathbf{P}$. We assume that the chain has transient states as well as recurrent states. Let $\mathcal{T}$ be the (not necessarily finite) set of transient states and denote the restriction of $\mathbf{P}$ to $\mathcal{T}$ by $\mathbf{Q}$. Then for any $x \in \mathcal{T}$, the distribution of the time T^{abs} until absorption into one of the irreducible closed classes of recurrent states, given that the chain starts in state x, is given by*

$$\mathbb{P}_x(T^{abs} > n) = (\mathbf{Q}^n \, \mathbf{1})_x$$

and hence

$$\mathbb{P}_x(T^{abs} = n) = ((\mathbf{Q}^{n-1} - \mathbf{Q}^n) \, \mathbf{1})_x \, .$$

Next we address the question of how to compute the *absorption probabilities*

$$a_{x,R_i} = \mathbb{P}(T^{R_i} < \infty \,|\, X_0 = x) \qquad \text{for } x \in \mathcal{T}, \text{ and } 1 \leq i \leq k \, ,$$

that is, the probability that, if the Markov chain starts in transient state x, it will eventually end up in (be absorbed into) the irreducible closed class R_i. Notice that for the computation of the absorption probabilities a_{x,R_i} it suffices to only consider the case where all sets R_i are singletons: If R_i contains more than one state, we can group its states together by defining a new state $\hat{i}$ (replacing the elements of R_i) which is then an absorbing state for the modified Markov chain and for which the new transition probabilities are

$$\hat{P}_{x\hat{i}} = \mathbb{P}(X_1 \in R_i \,|\, X_0 = x) = \sum_{y \in R_i} P_{xy}$$

for all $x \in \mathcal{T}$. We then have

$$\hat{a}_{x,\hat{i}} = \mathbb{P}(T^{\hat{i}} < \infty \,|\, \hat{X}_0 = x) = a_{x,R_i} \, .$$

For this reason, we may assume that the Markov chain $(X_n)_{n\geq 0}$ has k absorbing states $\{1, 2, ..., k\}$ and no other irreducible closed classes of recurrent states. The canonical form of its transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ \vdots \\ k \\ \mathcal{T} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & \cdots & k & \mathcal{T} \end{array} \\ \left(\begin{array}{cccc|c} 1 & & & & \\ & \ddots & & & \mathbf{0} \\ & & 1 & & \\ \hline \cdots & \mathbf{T} & \cdots & & \mathbf{Q} \end{array}\right) \end{array} \tag{2.30}$$

for which we write, in short, $\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{T} & \mathbf{Q} \end{pmatrix}$. Since for $x \in \mathcal{T}$ and any absorbing state $j \in \{1, 2, ..., k\}$,

$$a_{x,j} = \lim_{n\to\infty} P^n_{xj}\,,$$

we need to understand $\lim_{n\to\infty} \mathbf{P}^n$. First, observe that

$$\mathbf{P}^2 = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{T}_2 & \mathbf{Q}^2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{T} + \mathbf{Q}\mathbf{T}) & \mathbf{Q}^2 \end{pmatrix}.$$

By induction on n, we get

$$\mathbf{P}^n = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{T}_n & \mathbf{Q}^n \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{I} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-1})\mathbf{T} & \mathbf{Q}^n \end{pmatrix}.$$

Hence

$$\lim_{n\to\infty} \mathbf{P}^n = \lim_{n\to\infty} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{I} + \mathbf{Q} + \cdots + \mathbf{Q}^{n-1})\mathbf{T} & \mathbf{Q}^n \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{V}\mathbf{T} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{T} & \mathbf{0} \end{pmatrix}.$$

We summarize this result in the following proposition.

Proposition 2.3.3 (Absorption probabilities)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain on finite state space $\mathcal{S}$. Assume that the chain has k absorbing states $\{1, 2, ..., k\}$, no other recurrent states, and a non-empty set of transient states $\mathcal{T}$. Then for every $x \in \mathcal{T}$, the probability $a_{x,j}$ that the chain starting in x will eventually be absorbed in state j for $j \in \{1, 2, ..., k\}$ is the (x, j)-entry of the matrix*

$$\mathbf{A} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{T}\,.$$

Example 2.3.3. We continue Example 2.3.2. Instead of working with the transition matrix (2.27), we will work with the modified transition matrix

$$\hat{\mathbf{P}} = \begin{array}{c} \\ 3 \\ r \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 3 & r & 2 & 5 \end{array} \\ \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0.2 & 0.1 & 0.3 & 0.4 \\ 0.3 & 0.1 & 0.6 & 0 \end{array}\right) \end{array}. \tag{2.31}$$

where states $1, 4, 6$ from the original state space $\mathcal{S}$ have been combined to a new state which we have called r. In (2.31) we have $\mathbf{T} = \begin{pmatrix} 0.2 & 0.1 \\ 0.3 & 0.1 \end{pmatrix}$ and $\mathbf{Q} = \begin{pmatrix} 0.3 & 0.4 \\ 0.6 & 0 \end{pmatrix}$, and so we compute

$$\mathbf{A} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{T} = \begin{array}{c} \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} 3 & r \end{array} \\ \begin{pmatrix} 0.7 & 0.3 \\ 0.72 & 0.28 \end{pmatrix} \end{array} \qquad \text{(the matrix entries have been rounded)}\,.$$

From the matrix $\mathbf{A}$ we read off that, for example, $a_{5,R_2} = \mathbb{P}_5(T^{R_2} < T^{R_1}) = 0.28$. This is the probability that the Markov chain, given that it starts in state 5, eventually enters the closed subset of states $R_2 = \{1, 4, 6\}$. □

2.3.3 Infinite number of transient states

Let $(X_n)_{n\geq 0}$ be a reducible Markov chain with at least one recurrent state and an *infinite* number of transient states $\mathcal{T}$. With $|\mathcal{T}| = \infty$, it is possible that, if the Markov chain starts in a transient state x, it will never leave the set $\mathcal{T}$, and so, possibly,

$$\mathbb{P}_x(T^{abs} = \infty) = \lim_{n\to\infty} \mathbb{P}_x(T^{abs} > n) > 0\,.$$

Recall (2.28) and (2.29). Using the notation $a_x(n) = \mathbb{P}_x(T^{abs} > n)$,

$$\mathbb{P}_x(T^{abs} = \infty) = \lim_{n\to\infty} a_x(n)\,.$$

For all $n \geq 0$, let $\mathbf{a}(n)$ be the row vector whose components are the $a_x(n)$, $x \in \mathcal{T}$, with respect to some fixed ordering of the elements in $\mathcal{T}$. Let $\mathbf{a} = \lim_{n\to\infty} \mathbf{a}(n)$ (component wise). From

$$\mathbf{a}(n)^t = \mathbf{Q}^n\,\mathbf{1}^t$$

(recall that the superscript t denotes the transpose) we get

$$\mathbf{a}(n+1)^t = \mathbf{Q}\,\mathbf{a}(n)^t\,, \tag{2.32}$$

and, by taking the limit as $n \to \infty$ in (2.32),

$$\mathbf{a}^t = \lim_{n\to\infty} [\mathbf{Q}\,\mathbf{a}(n)^t] = \mathbf{Q}\mathbf{a}^t \tag{2.33}$$

where the rightmost equation is justified by the Dominated Convergence theorem.

Proposition 2.3.4 (Escape probability / probability of no absorption)**.** *Let $(X_n)_{n\geq 0}$ be a reducible Markov chain with an infinite number of transient states $\mathcal{T}$ and at least one recurrent state. Let $\mathbf{Q}$ be the restriction of the transition matrix $\mathbf{P}$ to $\mathcal{T}$. For $x \in \mathcal{T}$, let $e_x = \mathbb{P}_x(T^{abs} = \infty)$ and let $\mathbf{e}^t$ be the column vector whose components are the e_x. Then $\mathbf{e}^t$ is the* **maximal** *solution to*

$$\mathbf{a}^t = \mathbf{Q}\mathbf{a}^t \qquad \text{with } \mathbf{0} \leq \mathbf{a} \leq \mathbf{1} \text{ (component wise)}\,.$$

Proof. The entries of $\mathbf{e}$ are probabilities, so it is clear that $\mathbf{0} \leq \mathbf{e} \leq \mathbf{1}$ must hold. The fact that $\mathbf{e}$ is a solution to $\mathbf{a}^t = \mathbf{Q}\mathbf{a}^t$ was shown in (2.33). Note that $\mathbf{a} = \mathbf{0}$ is always a solution (in fact the unique solution if $|\mathcal{T}| < \infty$), but for the case $|\mathcal{T}| = \infty$ there may be multiple solutions. We need to show the maximality property of $\mathbf{e}$. Let $\tilde{\mathbf{a}}$ be another solution, so $\tilde{\mathbf{a}}$ is a vector indexed by $\mathcal{T}$ and $\tilde{\mathbf{a}}^t = \mathbf{Q}\tilde{\mathbf{a}}^t$ and $\mathbf{0} \leq \tilde{\mathbf{a}} \leq \mathbf{1}$. By induction on n we have

$$\tilde{\mathbf{a}}^t = \mathbf{Q}^n\,\tilde{\mathbf{a}}^t \quad \text{for } n \geq 0\,.$$

But then

$$\tilde{\mathbf{a}}^t = \mathbf{Q}^n\,\tilde{\mathbf{a}}^t \leq \mathbf{Q}^n\,\mathbf{1}^t = \mathbf{e}(n)^t \quad \text{for } n \geq 0\,, \tag{2.34}$$

from which, after taking the limit as $n \to \infty$, we get

$$\tilde{\mathbf{a}} \leq \mathbf{e}\,.$$

□

Example 2.3.4 (Simple random walk on $\mathbb{N}_0$ with absorbing boundary at 0)**.** Consider the Markov chain on $\mathbb{N}_0$ whose transition graph is shown in Figure 2.6.
In order to avoid trivial cases, we assume $0 < p < 1$, and we write $q = 1 - p$. The transition matrix $\mathbf{P}$ (which is in canonical form) is

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & & & & \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & q & 0 & p & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} 0 & p & & & \\ q & 0 & p & & \\ & q & 0 & p & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

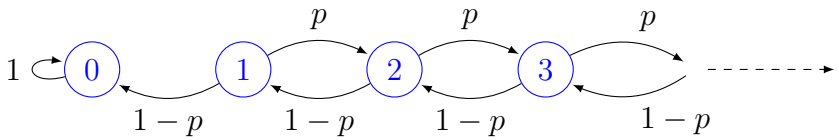


Figure 2.6

Note that 0 is the only recurrent state, and all other states $x \geq 1$ lead into 0. We need to solve $\mathbf{a} = \mathbf{Qa}$ which results in the system of equations

$$\begin{aligned} a_1 &= pa_2 \\ a_2 &= qa_1 + pa_3 \\ a_3 &= qa_2 + pa_4 \\ &\vdots \end{aligned} \tag{2.35}$$

This system is easiest solved by using the substitution $u_n = a_{n-1} - a_n$ for $n \geq 2$, and $u_1 = -a_1$. Thus the system (2.35) reads

$$pu_{n+1} = qu_{n-1} \qquad \text{for } n \geq 2\,.$$

Setting $u_1 = c$ (any constant), we get $u_2 = c\,\dfrac{q}{p}$ and, by induction on n,

$$u_n = c\,(\frac{q}{p})^{n-1} \qquad \text{for } n \geq 2\,.$$

Note that $a_n = -(u_1 + \cdots + u_n)$, so the general solution to (2.35) is

$$a_n = (-c)\sum_{k=0}^{n-1}(\frac{q}{p})^k \qquad \text{for } n \geq 1\,. \tag{2.36}$$

In order to find the *maximal* solution for which $0 \leq a_n \leq 1$ for all $n \geq 1$, we need to distinguish cases.

Case $p > q$ **:** Since for this case, the chain is biased towards stepping away from 0, we suspect that there is positive probability of remaining in the set $\mathbb{N}$ of transient states forever. Indeed, for this case the maximum solution $\mathbf{a}$ under the constraint $\mathbf{0} \leq \mathbf{a} \leq \mathbf{1}$ is achieved for the constant

$$-c = \left(\sum_{k=0}^{\infty}(\frac{q}{p})^k\right)^{-1} = 1 - \frac{q}{p}\,,$$

which results in the solution

$$a_n = \mathbb{P}_n(T^0 = \infty) = 1 - (\frac{q}{p})^n > 0 \qquad \text{for } n \geq 1\,. \tag{2.37}$$

Case $q \geq p$: The maximum solution $\mathbf{a}$ for (2.36) under the constraint $\mathbf{0} \leq \mathbf{a} \leq \mathbf{1}$ is achieved for the constant $c = 0$. Thus

$$a_n = 0 \qquad \text{for } n \geq 1\,, \tag{2.38}$$

or equivalently stated, *eventual absorption* of the Markov chain in 0 is *certain* for any starting state n. □

Example 2.3.5 (Simple random walk on $\mathbb{N}_0$ with reflecting boundary at 0)**.** We slightly modify the Markov chain from Example 2.3.4 by replacing $P_{01} = 0$ with $P_{01} = 1$ (but make no other changes). The resulting chain is irreducible. Its transition graph is shown in Figure 2.7.

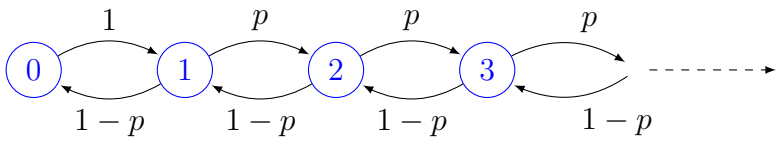


Figure 2.7

Let $q = 1 - p$.
Case $p > q$: For this case, (2.37) proves that the chain is *transient*. Indeed, we have $\mathbb{P}_0(T^0 < \infty) = \mathbb{P}_1(T^0 < \infty) = \frac{q}{p} < 1$.

Case $q \geq p$: For this case, (2.38) proves that the chain is *recurrent* since here we have $\mathbb{P}_0(T^0 < \infty) = \mathbb{P}_1(T^0 < \infty) = 1 - 0 = 1$. □

We now turn to the computation of the absorption probabilities

$$a_{x,R_i} = \mathbb{P}_x(T^{abs} < \infty\,, X_{T^{abs}} \in R_i)$$

where x is a transient state and R_i is an irreducible closed class of recurrent states. As discussed in the previous subsection, for the computation of a_{x,R_i} it suffices to assume that each irreducible closed class R_k is a singleton set (if not, we work with an appropriately modified chain). Thus, by some abuse of notation, we will identify R_i with the single state that represents the class R_i. Our goal is to compute the matrix $\mathbf{A}$ whose entries are the probabilities a_{x,R_i} for $x \in \mathcal{T}$ and $R_i \in \{R_1, R_2, ...\}$.

Lemma 2.3.5. *Let $(X_n)_{n\geq 0}$ be an absorbing Markov chain, let $\{R_1, R_2, ...\}$ be the set of absorbing states and $\mathcal{T}$ the set of transient states. Then the absorption probabilities $a_{x,R_i} = \mathbb{P}_x(T^{abs} < \infty\,, X_{T^{abs}} = R_i)$ for $x \in \mathcal{T}$ are the matrix entries of the matrix*

$$\mathbf{A} = \sum_{n=0}^{\infty} \mathbf{Q}^n\, \mathbf{T}\,.$$

Proof. Recall the format of the transition matrix $\mathbf{P}$ in canonical form (2.30). We have

$$\begin{aligned}
a_{x,R_i} &= \mathbb{P}_x(T^{abs} < \infty\,,\ X_{T^{abs}} = R_i) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x(T^{abs} = n\,, X_n = R_i) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x(X_1 \in \mathcal{T}, ..., X_{n-1} \in \mathcal{T},\ X_n = R_i) \\
&= \sum_{n=1}^{\infty} \sum_{y\in\mathcal{T}} \mathbb{P}_x(X_1 \in \mathcal{T}, ..., X_{n-1} = y,\ X_n = R_i) \\
&= \sum_{n=1}^{\infty} \sum_{y\in\mathcal{T}} Q_{xy}^{(n-1)}\, P_{yR_i} \\
&= \left(\sum_{n=0}^{\infty} \mathbf{Q}^n\, \mathbf{T}\right)_{x,R_i}.
\end{aligned}$$

□

Note that it may be difficult to compute $\mathbf{A} = \sum_{n=0}^{\infty} \mathbf{Q}^n\, \mathbf{T}$ directly. The following proposition gives an alternate way of computing $\mathbf{A}$ via solving an infinite system of linear equations.

Proposition 2.3.6 (Absorption probabilities)**.** *Let $(X_n)_{n\geq 0}$ be an absorbing Markov chain, let $\{R_1, R_2, ...\}$ be the set of absorbing states and $\mathcal{T}$ be the set of transient states. Then the absorbing probabilities $a_{x,R_i} = \mathbb{P}_x(T^{abs} < \infty\,, X_{T^{abs}} = R_i)$ for $x \in \mathcal{T}$ are the matrix entries of the matrix $\mathbf{A}$ which is the <u>minimal</u> solution to the matrix equation*

$$\mathbf{M} = \mathbf{Q}\mathbf{M} + \mathbf{T} \tag{2.39}$$

under the constraint $\mathbf{0} \leq \mathbf{A} \leq \mathbf{1}$ (entry wise).

Proof. Applying a first-step analysis to the computation of a_{x,R_i}, we get

$$\begin{aligned}
a_{x,R_i} &= \mathbb{P}_x(T^{abs} < \infty\,,\ X_{T^{abs}} = R_i) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x(T^{abs} = n\,, X_n = R_i) \\
&= \sum_{n=2}^{\infty} \sum_{y\in\mathcal{T}} \mathbb{P}_x(T^{abs} = n\,, X_1 = y, X_n = R_i) + P_{xR_i} \\
&= \sum_{n=2}^{\infty} \sum_{y\in\mathcal{T}} \mathbb{P}_x(X_2 \in \mathcal{T}, ..., X_{n-1} \in \mathcal{T}, X_1 = y, X_n = R_i) + P_{xR_i} \\
&= \sum_{n=2}^{\infty} \sum_{y\in\mathcal{T}} \mathbb{P}_x(X_2 \in \mathcal{T}, ..., X_{n-1} \in \mathcal{T}, X_n = R_i \,|\, X_1 = y)\, P_{xy} + P_{xR_i} \\
&= \sum_{n=2}^{\infty} \sum_{y\in\mathcal{T}} \mathbb{P}_y(X_1 \in \mathcal{T}, ..., X_{n-2} \in \mathcal{T}, X_{n-1} = R_i)\, P_{xy} + P_{xR_i} \\
&= \sum_{n=1}^{\infty} \sum_{y\in\mathcal{T}} \mathbb{P}_y(X_1 \in \mathcal{T}, ..., X_{n-1} \in \mathcal{T}, X_n = R_i)\, P_{xy} + P_{xR_i} \\
&= \sum_{n=1}^{\infty} \sum_{y\in\mathcal{T}} P_{xy}\, a_{y,R_i} + P_{xR_i} = \sum_{n=1}^{\infty} \sum_{y\in\mathcal{T}} Q_{xy}\, a_{y,R_i} + P_{xR_i} = (\mathbf{QA} + \mathbf{T})_{x,R_i}\ .
\end{aligned}$$

The above shows that the matrix of absorption probabilities $\mathbf{A}$ is a solution to the matrix equation $\mathbf{M} = \mathbf{QM} + \mathbf{T}$ (which may have multiple solutions). We need to show that the $\mathbf{A}$ is the *minimal* solution under the constraint $\mathbf{0} \le \mathbf{A} \le \mathbf{1}$. By Lemma 2.3.5, $\mathbf{A} = \sum_{n=0}^{\infty} \mathbf{Q}^n\, \mathbf{T}$. Assume the matrix $\tilde{\mathbf{A}}$ also solves $\mathbf{M} = \mathbf{QM} + \mathbf{T}$ and satisfies the inequalities $\mathbf{0} \le \tilde{\mathbf{A}} \le \mathbf{1}$. It follows that

$$\tilde{\mathbf{A}} = \mathbf{Q}\tilde{\mathbf{A}} + \mathbf{T} \ge \mathbf{T}$$

(all inequalities are to be understood component wise). So

$$\tilde{\mathbf{A}} \ge \mathbf{QT} + \mathbf{T}\,,$$

and thus

$$\tilde{\mathbf{A}} \ge \mathbf{Q}(\mathbf{QT} + \mathbf{T}) + \mathbf{T} = (\mathbf{Q}^2 + \mathbf{Q} + \mathbf{I})\mathbf{T}\,.$$

By induction, we have

$$\tilde{\mathbf{A}} \ge \sum_{k=0}^{n} \mathbf{Q}^k\mathbf{T} \quad \text{for all } n \ge 1\,.$$

Taking the limit as $n \to \infty$ yields

$$\tilde{\mathbf{A}} \ge \sum_{k=0}^{\infty} \mathbf{Q}^k\mathbf{T} = \mathbf{A}\,.$$

□

2.4 Periodicity

Some Markov chains exhibit a kind of cyclic behavior in terms of the set of states the chain can visit at certain times. Consider, for example, simple random walk on the integers $\mathbb{Z}$. If the random walk starts in state 0, say, then at even times, the state of the chain will always be an even integer. And at odd times, the state of the chain will always be an odd integer. This is typical behavior of a so-called *periodic* chain: Certain collections of states are periodically "forbidden". In this section we define the notion of periodicity of a Markov chain and discuss certain dynamical properties that arise if the chain is periodic.

Definition 2.4.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$. Let $x \in \mathcal{S}$ with the property that $P_{xx}^n > 0$ for some $n \geq 1$. Define*

$$c(x) := \gcd\{n \geq 1 \,:\, P_{xx}^n > 0\}$$

where gcd *stands for* greatest common divisor.

(a) If $c(x) \geq 2$, then state $x \in \mathcal{S}$ is called **periodic** *with* **period** *$c(x)$.*
(b) If $c(x) = 1$, then state x is called **aperiodic**.
(c) If all states in $\mathcal{S}$ have the same period c, we call the Markov chain periodic with period c.
(d) If all states in $\mathcal{S}$ are aperiodic, we call the Markov chain aperiodic.

Note that $1 \leq c(x) \leq \min\{n \,:\, P_{xx}^n > 0\}$. If $P_{xx} > 0$, then x is aperiodic.

Example 2.4.1. Consider simple random walk on $\mathbb{Z}$ with $P_{x,x+1} = p$ and $P_{x,x-1} = 1 - p$ for $0 < p < 1$ and $x \in \mathbb{Z}$. This Markov chain is periodic with period 2. If we modify the random walk by adding positive holding probability to each state, that is, if we use $P_{x,x+1} = p$, $P_{xx} = r$, and $P_{x,x-1} = q$ for positive p, r, q with $p + r + q = 1$ and for all $x \in \mathbb{Z}$, the Markov chain is aperiodic. □

Note: Often times, in order to avoid periodicity issues, we will work with a so-called **lazy version** of a given periodic Markov chain: If a periodic Markov chain has transition matrix $\mathbf{P}$, we will instead work with a lazy version $\tilde{\mathbf{P}} = p\mathbf{I} + (1-p)\mathbf{P}$ for some $0 < p < 1$. Adding positive holding probability p to each state guarantees that the modified chain with transition matrix $\tilde{\mathbf{P}}$ is aperiodic.

Example 2.4.2. Figure 2.8 shows the transition graph for an 8-state Markov chain. A directed edge indicates a positive one-step transition probability in the given direction.

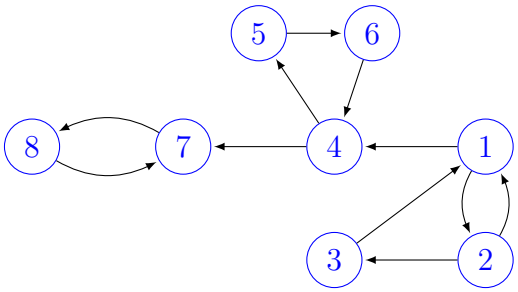


Figure 2.8

States 7, 8 have period 2, states 4, 5, 6 have period 3, and states 1, 2, 3 are aperiodic. Note that this Markov chain is not irreducible. □

Proposition 2.4.1 (Periodicity is a class property)**.** *Let $x, y \in \mathcal{S}$ and assume x communicates with y. Then $c(x) = c(y)$.*

Proof. Since $x \longrightarrow y$, there exists $n \geq 1$ such that $P^n_{xy} > 0$. Similarly, since $y \longrightarrow x$, there exists $m \geq 1$ such that $P^m_{yx} > 0$. Thus

$$P^{n+m}_{xx} \geq P^n_{xy} P^m_{yx} > 0\,,$$

so $c(x)$ divides $(n+m)$. Furthermore, for any $k \geq 1$ with $P^k_{yy} > 0$, we have

$$P^{n+k+m}_{xx} \geq P^n_{xy} P^k_{yy} P^m_{yx} > 0\,,$$

and so $c(x)$ divides $(n+k+m)$. It follows that $c(x)$ divides k, and so $c(x) \leq c(y)$. But the roles of x and y can be interchanged in this argument, so we also have $c(y) \leq c(x)$. Hence $c(x) = c(y)$. □

Corollary 2.4.2. *An irreducible Markov chain is either aperiodic or periodic with period $c > 1$.*

Example 2.4.3 (Simple random walk on a connected graph is either aperiodic or has period 2). Let $G(V, E)$ be a connected graph and consider simple random walk on G. We will show that simple random walk on a graph cannot be periodic with period $c > 2$. Recall Definition A.2.1 for the definition of a *bipartite* graphs. Figure 2.9 shows some examples. Proposition A.2.1 in Appendix A states that a connected graph $G(V, E)$

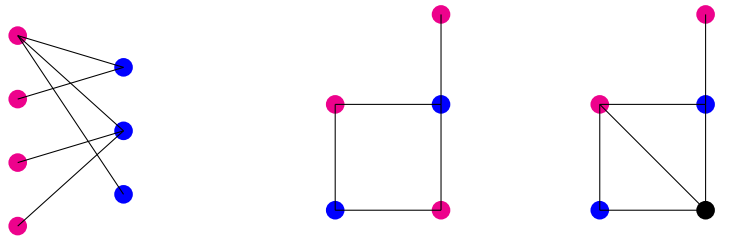


Figure 2.9

is bipartite if and only if it does not contain an odd-length cycle. First, assume the graph $G(V, E)$ is bipartite, and $V = V_1 \cup V_2$. Choose a vertex $v \in V$. Let us assume $v \in V_2$. Since G has no self-loops, $P_{vv} = 0$, and since G has no isolated points (by connectedness, at least one edge must emanate from each vertex v), $P^2_{vv} > 0$. Assume $P^n_{vv} > 0$ for some $n > 2$. Then there exists a sequence $v, v_1, v_2, ..., v_{n-1}, v$ of (not necessarily distinct) vertices such that $P_{vv_1} P_{v_1 v_2} \cdots P_{v_{n-1} v} > 0$. Since the graph is bipartite, we must have $v_1, v_3, ..., v_{n-1} \in V_1$ and $v_2, v_4, ..., v_{n-2} \in V_2$. Hence the number $(n-2)$ is even, and therefore n is also even. It follows that vertex v has period 2, and thus simple random walk on any bipartite graph has period 2.

We now assume $G(V, E)$ is not bipartite. By Proposition A.2.1, the graph contains an odd-length cycle, say a cycle of length m. Let v be a vertex in this cycle. Then clearly $P^m_{vv} > 0$. On the other hand, for any vertex $v \in V$ we also have $P^2_{vv} > 0$. Since $\gcd\{2, m\} = 1$, vertex v is aperiodic. Since the graph is connected, by Proposition 2.4.1, all vertices are aperiodic, and so simple random walk on a non-bipartite graph is aperiodic.

This proves that simple random walk on a connected graph is either aperiodic or has period 2 (in the latter case the graph is bipartite). □

Aside: Bipartite graphs are exactly the **2-colorable** graphs, that is, the graphs for which each vertex can be colored in one of two colors such that no adjacent vertices have the same color. (An n-coloring of a graph is defined analogously.) The **chromatic number** $\chi(G)$ of a graph G is the smallest number n for which there exists an n-coloring of the graph. Bipartite graphs are exactly the graphs G with $\chi(G) = 2$. □

Proposition 2.4.3. *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$.*

(a) A state $x \in \mathcal{S}$ is aperiodic if and only if there exists $M(x) \in \mathbb{N}$ such that for all $n \geq M(x)$, we have $P^n_{xx} > 0$.

(b) If $(X_n)_{n\geq 0}$ is irreducible and aperiodic, and if $\mathcal{S}$ is finite, then there exists $N \in \mathbb{N}$ such that for all $n \geq N$ and all $x, y \in \mathcal{S}$, $P^n_{xy} > 0$.

Proof. We need the following lemma which is an immediate corollary of Theorem A.3.1 (Schur's theorem). For a proof see Appendix A.

Lemma 2.4.4. *If J is a subset of the natural numbers $\mathbb{N}$ that is closed under addition, and if the greatest common divisor of the elements in J is c, then there exists $N \in \mathbb{N}$ such that $nc \in J$ for all $n \geq N$.*

(a) Note that for any state $x \in \mathcal{S}$, the set $J_x = \{n : P^n_{xx} > 0\}$ is closed under addition. In particular, since x is aperiodic, Lemma 2.4.4 states that there exists a number $M(x)$ such that $P^n_{xx} > 0$ for all $n \geq M(x)$. Conversely, if for state x there exists a natural number $M(x)$ such that $P^n_{xx} > 0$ for all $n \geq M(x)$, then $P^q_{xx} > 0$ and $P^r_{xx} > 0$ for two distinct prime numbers q and r. Hence $c(x) = 1$ and so x is aperiodic.

(b) Since the chain is irreducible, for each pair of states x, y there exists a number $m(x, y)$ such that $P^{m(x,y)}_{xy} > 0$ Therefore, for all $n \geq M(x) + m(x, y)$, we have $P^n_{xy} > 0$. Since the state space $\mathcal{S}$ is finite, we can take the maximum

$$N := \max\{M(x) + m(x, y) \,|\, x, y \in \mathcal{S}\}\,.$$

It follows that $P^N_{xy} > 0$ for all $x, y \in \mathcal{S}$, and therefore also $P^n_{xy} > 0$ for all $x, y \in \mathcal{S}$ and $n \geq N$. □

Definition 2.4.2. *(a) A matrix $\mathbf{P}$ is called positive, if all of its entries are strictly positive. (b) A square matrix $\mathbf{P}$ is called* **regular***, if there exists an $N \in \mathbb{N}$ such that $\mathbf{P}^N$ is a positive matrix.*

It follows that a stochastic matrix $\mathbf{P}$ is regular if and only if there exists an $N \in \mathbb{N}$ such that $\mathbf{P}^n$ is positive for all $n \geq N$.

Corollary 2.4.5. *A finite transition matrix $\mathbf{P}$ is regular if and only if the corresponding Markov chain is irreducible and aperiodic.*

We now take a closer look at the cyclic structure of an irreducible, **periodic** Markov chain.

Theorem 2.4.6 (Cyclic classes)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, periodic Markov chain with state space $\mathcal{S}$, transition matrix $\mathbf{P}$, and period $c \geq 2$. Then*

(a) for all $x, y \in \mathcal{S}$ there exists an integer r with $0 \leq r \leq c-1$ such that

$$P_{xy}^n > 0 \;\Rightarrow\; n \equiv r \mod c\,,$$

and furthermore, there exists an integer N such that $P_{xy}^{kc+r} > 0$ for all $k \geq N$.

(b) There exists a partition

$$\mathcal{S} = S_0 \cup S_1 \cup \ldots \cup S_{c-1}$$

into c so-called **cyclic classes** *S_r such that for all $0 \leq r \leq c-1$,*

$$x \in S_r \quad \Rightarrow \quad \sum_{y \in S_{r+1}} P_{xy} = 1 \qquad (\text{with } S_c = S_0)\,.$$

Proof. (a) Let $x, y \in \mathcal{S}$ and consider $n_0 = \min\{n \,|\, P_{xy}^n > 0\}$. Then $n_0 = k_0 c + r$ for some $0 \leq r \leq c-1$ and some $k_0 \geq 0$.
For all m with $P_{yx}^m > 0$, we have $P_{xx}^{n_0+m} \geq P_{xy}^{n_0} P_{yx}^m > 0$, and so $(n_0 + m)$ is a multiple of c. Furthermore, for any n with $P_{xy}^n > 0$, the sum $(n+m)$ is also multiple of c. Hence

$$(n+m) - (n_0+m) = n - n_0 = kc \qquad \text{for some } k \geq 0\,,$$

and

$$n \equiv n_0 \equiv r \mod c \qquad \text{for all } n \text{ with } P_{xy}^n > 0\,.$$

By Lemma 2.4.4, there exists $N' \in \mathbb{N}$ such that $P_{yy}^{cn} > 0$ for all $n \geq N'$. Thus

$$P_{xy}^{n_0+cn} \geq P^{n_0} P_{yy}^{nc} > 0 \qquad \text{for } n \geq N'\,.$$

Setting $k = k_0 + n$ and $N = k_0 + N'$, we get $P_{xy}^{kc+r} > 0$ for all $k \geq N$.
(b) Let $x \in \mathcal{S}$. For $0 \leq r \leq c-1$, define the sets S_r, called the *cyclic classes*, by

$$S_r = \{z \in \mathcal{S} \,:\, \exists n \geq 0 \; s.t. \; P_{xz}^n > 0 \text{ and } n \equiv r \mod c\}\,. \tag{2.40}$$

The sets S_r are non-empty by construction. By part (a), the sets S_r are disjoint, and by irreducibility, their union is the entire state space $\mathcal{S}$. Let $y \in S_r$ for some $0 \leq r \leq c-1$.

Then, by (2.40),

$$\begin{aligned} S'_s &= \{z \in \mathcal{S} : \exists n \text{ s.t. } P^n_{yz} > 0 \text{ and } n \equiv s \mod c\} \\ &= \{z \in \mathcal{S} : \exists m \text{ s.t. } P^m_{xz} > 0 \text{ and } m \equiv (r+s) \mod c\} = S_{r+s}\,. \end{aligned}$$

(If $(r+s) \geq c$ and $(r+s) \equiv p \mod c$, then we set $S_{r+s} = S_p$.) Thus the cyclic classes S_r do not depend on the state x that was used for their construction in (2.40). The above also shows that all $0 \leq r \leq c-1$,

$$x \in S_r \quad \Rightarrow \quad \sum_{y \in S_{r+1}} P_{xy} = 1\,.$$

□

Example 2.4.4. Figure 2.10 shows the transition graph of an irreducible Markov chain on eight states. The chain is periodic with period 3. The cyclic classes are $S_0 = \{4, 7\}$, $S_1 = \{1, 3, 8\}$, and $S_2 = \{2, 5, 6\}$. □

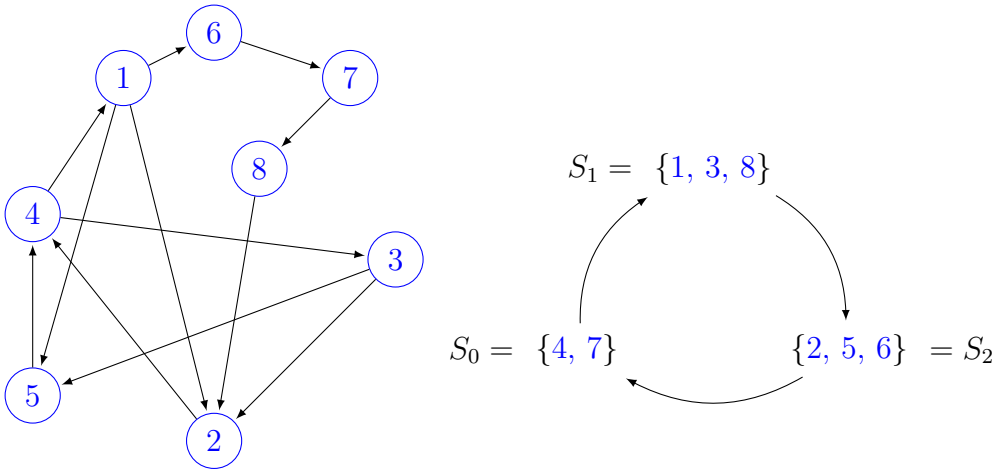


Figure 2.10

Proposition 2.4.7. *Let $(X_n)_{n\geq 0}$ be an irreducible, periodic Markov chain with state space $\mathcal{S}$ and period $c \geq 2$. Then the Markov chain $(Y_n)_{n\geq 0}$ defined by $Y_n = X_{cn}$ for $n \geq 0$ is aperiodic and reducible. Its irreducible closed classes are $S_0, S_1, ..., S_{c-1}$ as introduced in Theorem 2.4.6.*

Proof. The process $(Y_n)_{n\geq 0}$ is a Markov chain and has transition matrix $\mathbf{M} = \mathbf{P}^c$ (see Exercise 1.3). Let $x \in \mathcal{S}$. Then

$$d = \gcd\{k \,|\, M^k_{xx} > 0\} = \gcd\{k \,|\, P^{kc}_{xx} > 0\}$$

which must be equal to 1, since $d > 1$ leads to the contradiction

$$c = \gcd\{n \,|\, P^n_{xx} > 0\} = d\, c > c\,.$$

Hence $(Y_n)_{n\geq 0}$ is aperiodic. By construction of S_r, for all $0 \leq r \leq c-1$ and for all $x \in S_r$, we have $M_{xy} = P^c_{xy} = 0$ if $y \notin S_r$. Hence the cyclic classes S_r are closed. Let $y, z \in S_r$. By irreducibility of $\mathbf{P}$, there exists an $n \geq 0$ such that $P^n_{yz} > 0$. Since y, z belong to the same cyclic class, n must be a multiple of c. Hence there exists $k \geq 0$ such that $P^{kc}_{yz} = M^k_{yz} > 0$ which shows that S_r is irreducible for $\mathbf{M}$. □

Corollary 2.4.8. *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent, and periodic Markov chain with state space $\mathcal{S}$ and period $c \geq 2$. Let π be the unique stationary distribution and consider the cyclic classes $S_0, S_1, ..., S_{c-1}$. Then*

$$\frac{1}{c} = \pi(S_0) = \pi(S_1) = \cdots = \pi(S_{c-1})\,.$$

Furthermore, the unique stationary distribution for the Markov chain $(Y_n)_{n\geq 0}$ (defined, as in the previous proposition, by $Y_n = X_{cn}$ for $n \geq 0$) when restricted to the cyclic class S_r is

$$c\,\pi|_{S_r} \qquad \text{for } 0 \leq r \leq c-1$$

where $c\,\pi|_{S_r}(y) = c\,\pi(y)$ if $y \in S_r$, and zero otherwise.

Proof. (a) Consider the indicator function $f = \mathbb{1}_{S_r}$. We have $f(X_k) = \mathbb{1}_{\{X_k \in S_r\}}$ and $\mathbb{E}_\pi(f) = \pi(S_k)$. By Theorem 3.1.1,

$$\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} \mathbb{1}_{\{X_k \in S_r\}} = \pi(S_k) \qquad \text{with probability } 1\,. \tag{2.41}$$

Because of periodicity, for any sample path ω, we have $\mathbb{1}_{\{X_k \in S_r\}}(\omega) = 1$ exactly once every c steps and zero otherwise. This makes the limit on the left-hand side in (2.41) equal to $\frac{1}{c}$ for each $0 \leq r \leq c-1$.

(b) By Proposition 2.4.7, the Markov chain $(Y_n)_{n\geq 0}$, when restricted to S_r, is irreducible and aperiodic. Let $y \in S_r$ and denote by $T^y_{(c)}$ the first return time to state y for the chain

$(Y_n)_{n\geq 0}$, and by T^y the first return time to state y for the chain $(X_n)_{n\geq 0}$. Clearly, because of periodicity, we have the equality of the events $\{T^y_{(c)} = k\} = \{T^y = c\,k\}$ for $k \geq 0$. It follows that the mean return time for state y for the chain $(Y_n)_{n\geq 0}$ is $\dfrac{1}{c\,\pi(y)}$, and therefore the unique stationary distribution π^{S_r} of $(Y_n)_{n\geq 0}$ restricted to S_r is given by

$$\pi^{S_r}(y) = c\,\pi(y) \qquad \text{for } y \in S_r\,.$$

□

Remark 2.4.9. *Periodicity of a finite-state Markov chain is closely related to the number of eigenvalues of modulus* 1 *of its transition matrix* $\mathbf{P}$*. If the Markov chain has period* $c \geq 2$*, then the* c*th roots of unity* $1, e^{2\pi i/c}, ..., e^{2\pi i(c-1)/c}$ *are simple eigenvalues of* $\mathbf{P}$*, an* $\mathbf{P}$ *has no other eigenvalues of modulus* 1*. See Theorem A.6.2(e).*

Exercises

Exercise 2.1. Consider a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S} = \mathbb{N}_0$ and transition probabilities

$$P_{xy} = \begin{cases} 1/2 & \text{if } x = y \\ 1/2 & \text{if } x > 0 \text{ and } y = x - 1 \\ (1/2)^{(y+1)} & \text{if } x = 0 \text{ and } y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Is this a recurrent Markov chain or a transient Markov chain?

Exercise 2.2. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$. Let $x, y \in \mathcal{S}$, and recall the notation $f_{xy} = \mathbb{P}(T^y < \infty \,|\, X_0 = x)$. Consider the random variable V^y which gives the long-run number of visits (not including the initial visit, in case $X_0 = y$) to state y. Prove that for $k \geq 1$,

$$\mathbb{P}(V^y \geq k \,|\, X_0 = x) = f_{xy} f_{yy}^{k-1}.$$

Exercise 2.3. Consider a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S} = \mathbb{N}_0$ and transition probabilities

$$P_{xy} = \begin{cases} 3/5 & \text{if } y = 0 \\ 1/5 & \text{if } y \in \{x+3, x+6\} \end{cases}$$

for $x \in \mathbb{N}_0$. Classify each state of this Markov chain as positive recurrent, null recurrent, or transient, and find all irreducible closed classes of $\mathcal{S}$ (if any).

Exercise 2.4. Consider the Markov chain with state space $\mathcal{S} = \mathbb{N}_0$ and transition probabilities

$$P_{x0} = \frac{3}{x+3} \qquad P_{x,x+1} = \frac{x}{x+3} \qquad \text{for } x \geq 0\,.$$

Show that the chain is irreducible and determine whether it is positive recurrent, null recurrent, or transient. Does it have a stationary distribution?

Exercise 2.5. Let $Y_1, Y_2, ...$ be an infinite sequence of i.i.d. Bernoulli random variables with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = 0) = 1 - p$. Consider the Markov chain $(X_n)_{n\geq 0}$ that tracks the number of consecutive 1s in the last run. More precisely, let $X_0 = 0$, and for $n > 0$ and $1 \leq k \leq n$, let $X_n = k$ if $Y_{n-k} = 0$ and $Y_i = 1$ for $(n-k) < i \leq n$. Show that $(X_n)_{n\geq 0}$ is irreducible and positive recurrent and compute its stationary distribution π.

Exercise 2.6. Consider the Markov chain $(X_n)_{n\geq 0}$ from Exercise 2.5. It tracks the number of consecutive 1s in the last run of a sequence of i.i.d. Bernoulli random variables for which 1 occurs with probability p. Let $k \geq 1$. Compute a formula for $\mathbb{E}_0(T^k)$, the expected time until we see a sequence of k 1s in a row for the first time.

Exercise 2.7. Consider an infinite sequence $(Y_n)_{n\geq 1}$ of i.i.d. coin flips of a fair coin, so we have $\mathbb{P}(Y_i = T) = \mathbb{P}(Y_i = H) = 1/2$. Let τ be the time until we see the pattern TT appear for the first time.

(a) Compute $\mathbb{E}(\tau)$.

(b) Recall the *Fibonacci sequence* $1, 1, 2, 3, 5, 8, 13, ...$ It is described by the recurrence

$$f_n = f_{n-1} + f_{n-2}$$

for $n \geq 3$, with $f_1 = f_2 = 1$. Prove that for $n \geq 2$,

$$\mathbb{P}(\tau = n) = \frac{f_{n-1}}{2^n}.$$

(*Hint:* For $n \geq 4$, any sequence of coin flips that is an element of the event $\{\tau = n\}$ must start with either H or TH.)

(c) Use your result from part (b) to prove the following identity for Fibonacci numbers:

$$\sum_{n=1}^{\infty} \frac{f_n}{2^n} = 2.$$

Exercise 2.8. Prove that every finite-state Markov chain has at least one recurrent state.

Exercise 2.9. Consider a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ with transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 3/4 & 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/3 & 0 & 0 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \end{pmatrix} \end{array}.$$

Describe *all* stationary distributions for this Markov chain.

Exercise 2.10. Consider the Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{0, 1, ..., 6\}$ with transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1/4 & 3/4 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2/7 & 0 & 3/7 & 0 & 2/7 \\ 0 & 1/5 & 0 & 1/5 & 2/5 & 0 & 1/5 \\ 0 & 0 & 5/7 & 0 & 2/7 & 0 & 0 \\ 0 & 0 & 0 & 3/8 & 1/4 & 1/8 & 1/4 \\ 0 & 0 & 0 & 0 & 6/7 & 0 & 1/7 \end{pmatrix} \end{array}.$$

(a) Classify each state as transient or recurrent and find all irreducible closed classes.

(b) Compute $\lim_{n\to\infty} P^n_{3,0}$.

(c) Assume $X_0 = 3$. Compute the expected number of visits to state 3.

Exercise 2.11. Consider simple random walk on the vertices of the infinite tree shown in Figure 2.11. The walk starts at the *root* of the tree labeled 0. At each step, the walk chooses its next location uniformly at random from its neighboring vertices.

(a) Is this Markov chain recurrent, or transient, or neither?

(b) Now assume that the walk starts at vertex x (marked in the picture). What is the probability that the walk will never visit vertex 0?

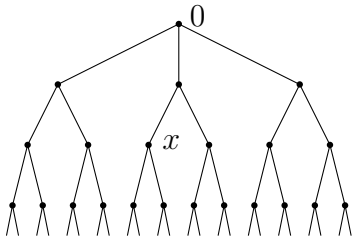


Figure 2.11: The graph continues to infinity in the same manner.

Exercise 2.12. Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be independent Markov chains with state space $\mathcal{S} = \{1, 2\}$, each with transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix} \end{array}.$$

Assume $X_0 = 1$ and $Y_0 = 2$. Consider $T = \min\{n : X_n = Y_n\}$, i.e., the first time both chains are in the same state.

(a) Compute $\mathbb{E}(T)$. (b) Compute $\mathbb{P}(X_T = 2)$.

Exercise 2.13. Let $(X_n)_{n\geq 0}$ be a finite-state Markov chain that has at least two transient states. Consider its fundamental matrix $\mathbf{V}$. Recall that the rows and columns of $\mathbf{V}$ are labeled by the transient states $\mathcal{T}$ of $(X_n)_{n\geq 0}$. We denote the matrix entries of $\mathbf{V}$ by v_{xy} for $x, y \in \mathcal{T}$. Also recall the notation $f_{xy} = \mathbb{P}_x(T^y < \infty)$.

(a) Show that for any transient state $x \in \mathcal{T}$,

$$f_{xx} = \frac{v_{xx} - 1}{v_{xx}}\,.$$

(b) Let x and y be distinct transient states. Derive an analogous formula for f_{xy} in terms of the matrix entries of $\mathbf{V}$.

Exercise 2.14. Consider the following Markov chains. For each chain, determine the periodicity of its states.

(a) $\mathcal{S} = \{0, 1, 2\}$, and the transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix} \end{array}.$$

(b) $\mathcal{S} = \{0, 1, 2, 3, 4\}$, and the transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 0 & 3/4 & 1/4 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Exercise 2.15. Consider simple random walk on a finite graph $G(V, E)$ (see page 27).

(a) Verify that π on V defined by

$$\pi(v) = \frac{\deg(v)}{2|E|} \quad \text{for } v \in V$$

is a stationary distribution.

(b) Consider simple random walk on a k-dimensional hypercube $\mathbb{Z}_2^k$ (see Example 1.5.6). Choose any vertex $\mathbf{x} \in \mathbb{Z}_2^k$ and assume the random walk starts in $\mathbf{x}$. What is the expected number of steps until the walk returns to $\mathbf{x}$ for the first time?

Exercise 2.16. Consider a standard 8×8 chessboard as shown in Figure 2.12 below.

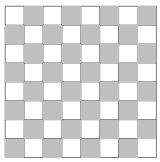


Figure 2.12: Standard chessboard

A king can move one square at a time in any direction (horizontally, vertically, or diagonally) that is available from his current location. Assume the king chooses each move uniformly at random from all permissible moves and that he starts out in the bottom left corner of the board.

(a) How many times, on average, will he spend on a border or corner square (i.e. a square located along the perimeter of the board) before he returns to the bottom left corner for the first time?

(b) Suppose the king, before each attempt, flips a coin and moves only if the coin shows heads, otherwise waits for that time period in his current spot. The probability that

the coin lands heads is $\mathbb{P}(H) = p$ for some fixed p with $0 < p < 1$. All coin flips are independent. Find the expected duration for the same journey (bottom left corner to bottom left corner).

Exercise 2.17. Recall that a birth/death chain is a Markov chain whose state space $\mathcal{S}$ is either $\{0, 1, ..., N\}$ or $\mathbb{N}_0$ and whose transition probabilities are

$$P_{xy} = \begin{cases} q_x & \text{if } y = x - 1 \\ p_x & \text{if } y = x + 1 \\ r_x & \text{if } y = x \end{cases}$$

with $p_x + q_x + r_x = 1$ for all $x \in \mathcal{S}$. Consider an irreducible birth/death chain $(X_n)_{n\geq 0}$ and a measure ν on $\mathcal{S}$ defined by $\nu_0 = 1$ and

$$\nu_k = \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k}$$

for $k \in \mathcal{S}$ with $k \geq 1$. Show that ν defines an invariant measure for $(X_n)_{n\geq 0}$ and that $(X_n)_{n\geq 0}$ has a stationary distribution if and only if $\sum_{k\in\mathcal{S}} \nu_k < \infty$.

Exercise 2.18. For each of the following irreducible birth/death chains, determine whether or not it has a stationary distribution. Compute the stationary distribution if it exists.

(a) Simple biased random walk on $\mathbb{N}_0$ with reflecting boundary at 0. Assume $p = 1/3$ and $q = 2/3$.

(b) A birth/death chain on $\mathbb{N}_0$ with $r_0 = \frac{2}{3}$, $p_0 = \frac{1}{3}$ and

$$q_k = \frac{k+1}{k+2}, \quad p_k = \frac{1}{k+3}, \quad \text{and } r_k = (1 - p_k - q_k) \text{ for } k \geq 1\,.$$

Exercise 2.19. Consider an irreducible, positive recurrent Markov chain $(X_n)_{n\geq 0}$ with stationary distribution π. Let x and y be two distinct states. Find a formula for the expected number of visits to y that occur between two consecutive visits to x.

Exercise 2.20. Consider simple random walk on the graph in Figure 2.13. Assume the walk starts in vertex a. Use a first-step analysis to find $\mathbb{E}_a(T^b)$. (*Hint*: Make use of symmetries in the graph.)

Exercise 2.21. Consider an irreducible, positive recurrent Markov chain $(X_n)_{n\geq 0}$ with stationary distribution π. Let x and y be two distinct states.

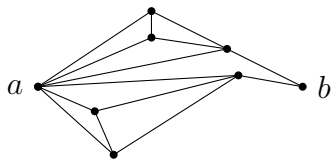


Figure 2.13

(a) Show that

$$\mathbb{E}_x(V_{T^y}^x) = \pi(x)(\mathbb{E}_x(T^y) + \mathbb{E}_y(T^x))$$

where $V_{T^y}^x$ is the number of visits to x between (including) times 0 and $T^y - 1$.

(b) Show that

$$\mathbb{P}_x(T^x < T^y) = \frac{1}{\pi(x)(\mathbb{E}_x(T^y) + \mathbb{E}_y(T^x))}.$$

(*Hint:* Recall Example 2.2.1.)

Exercise 2.22. Consider a connected graph $G(V, E)$ and assume there exists an edge $e = \{x_0, y_0\}$ with $e \in E$ such that the removal of e results in two disjoint components G_{x_0} and G_{y_0} for the remaining graph where $x_0 \in G_{x_0}$ and $y_0 \in G_{y_0}$. Denote the edge set of G_{x_0} by E_{x_0}. Prove that for simple random walk on $G(V, E)$ we have

$$\mathbb{E}_{x_0}(T^{y_0}) = 2|E_{x_0}| + 1\,.$$

(*Hint*: Consider simple random walk on the subgraph $\tilde{G}$ that results from adding vertex y_0 and edge e to the subgraph G_{x_0}.)

Exercise 2.23. Consider a finite, connected graph $G(V, E)$ and simple random walk on the graph.

(a) Show that for all $x \in V$,

$$\mathbb{E}_x(T^x) = 1 + \frac{1}{\deg(x)} \sum_{y:\{x,y\}\in E} \mathbb{E}_y(T^x).$$

(b) Use the result from part (a) to prove that for all vertices $x, y \in V$, we have

$$\mathbb{E}_x(T^y) \leq 2|E| - 1\,.$$

Exercise 2.24. Consider a finite, connected graph $G(V, E)$ and simple random walk on the graph. Assume there exists a vertex $y \in V$ for which $deg(y) = 1$. Use the result from Exercise 2.22 to show that there exists a vertex $x \in V$ for which the upper bound in Exercise 2.23(b) is sharp. That is, show that $\mathbb{E}_x(T^y) = 2|E| - 1$.

Exercise 2.25. Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain on $\mathcal{S}$ with stationary distribution π. Given a proper subset $\mathcal{B} \subset \mathcal{S}$, consider the process $(B_m)_{m\geq 0}$ with state space $\mathcal{B}$ obtained by observing $(X_n)_{n\geq 0}$ when in $\mathcal{B}$. That is, with

$$T_0 = \min\{n \geq 0 : X_n \in \mathcal{B}\}\,,$$

and

$$T_m = \min\{n > T_{m-1} : X_n \in \mathcal{B}\}$$

for $m \geq 1$, set

$$B_m = X_{T_m}\,.$$

Show that $(B_m)_{m\geq 0}$ is positive recurrent and find its stationary distribution.

Chapter 3

Limit Theorems for Markov Chains

3.1 The Ergodic Theorem

Let $\mathcal{S}$ be a discrete state space, μ a probability distribution on $\mathcal{S}$, and let f be a function $f : \mathcal{S} \to \mathbb{R}$. We will use the notation

$$\mathbb{E}_\mu(f) = \sum_{y \in \mathcal{S}} f(y)\mu(y)\,.$$

Theorem 3.1.1 (Ergodic theorem for Markov chains)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain with state space $\mathcal{S}$ and stationary distribution π. Assume the chain has initial distribution π_0. Let $f : \mathcal{S} \to \mathbb{R}$ be a bounded function. Then*

$$\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} f(X_k) = \mathbb{E}_\pi(f) \qquad \text{with probability 1}\,.$$

Proof. We first assume that $f \geq 0$ and that the chain starts in state x. (We will cover the general case later.) For $k \geq 0$, we consider the kth-return times $T^{x,k}$ to state x defined by $T^{x,0} = 0$ and

$$T^{x,k} = \min\{n :> T^{x,k-1} \text{ and } X_n = x\} \quad \text{ for } k \geq 1\,.$$

Note that $T^{x,1} = T^x$, the first return time to state x as previously defined. Because of the strong Markov property, the waiting times $(T^{x,k} - T^{x,k-1})$, for $k \geq 1$, between two consecutive visits to state x are i.i.d. random variables with the same distribution as T^x. Furthermore, the random variables $X_{T^{x,k-1}}$, $X_{T^{x,k-1}+1}, ..., X_{T^{x,k}-1}$ for the times in between

two consecutive visits to x form mutually independent "clusters" of random variables for distinct $k \geq 1$. As a consequence, the random variables Z_k, $k \geq 1$, defined by

$$Z_k = \sum_{m=T^{x,k-1}}^{T^{x,k}-1} f(X_m)$$

form an i.i.d. sequence. Since f is a bounded function, there exists $M < \infty$ such that $|f(y)| \leq M$ for all $y \in \mathcal{S}$. Since x is positive recurrent, $\mathbb{E}_x(T^x) < \infty$. Thus

$$\mathbb{E}_x(|Z_1|) \leq M\, \mathbb{E}_x(T^x) < \infty\,,$$

and we can apply the Strong Law of Large numbers to the sequence $Z_1, Z_2, ...$ which yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Z_k = \mathbb{E}_x(Z_1) \qquad \text{with probability } 1\,.$$

Setting $S_{T^{x,n}} = \displaystyle\sum_{m=0}^{T^{x,n}-1} f(X_m) = \sum_{k=1}^{n} Z_k$, we thus have

$$\lim_{n\to\infty} \frac{S_{T^{x,n}}}{n} = \mathbb{E}_x(Z_1) \qquad \text{with probability } 1\,. \tag{3.1}$$

We now compute $\mathbb{E}_x(Z_1)$.

$$\begin{aligned} \mathbb{E}_x(Z_1) = \mathbb{E}_x\left(\sum_{m=0}^{T^x-1} f(X_m)\right) &= \mathbb{E}_x\left(\sum_{y\in\mathcal{S}} f(y) \sum_{m=0}^{T^x-1} \mathbb{1}_{\{y\}}(X_m)\right) \\ &= \sum_{y\in\mathcal{S}} f(y)\, \mathbb{E}_x\left(\sum_{m=0}^{T^x-1} \mathbb{1}_{\{y\}}(X_m)\right)\,. \end{aligned}$$

Recall from Section 2.2.1 that $\mathbb{E}_x\left(\sum_{m=0}^{T^x-1} \mathbb{1}_{\{y\}}(X_m)\right) = \pi(y)\, \mathbb{E}_x(T^x)$. So altogether we get

$$\mathbb{E}_x(Z_1) = \sum_{y\in\mathcal{S}} f(y)\pi(y)\, \mathbb{E}_x(T^x) = \mathbb{E}_\pi(f)\, \mathbb{E}_x(T^x)\,, \tag{3.2}$$

and together with (3.1),

$$\lim_{n\to\infty} \frac{S_{T^{x,n}}}{n} = \mathbb{E}_\pi(f)\, \mathbb{E}_x(T^x) \qquad \text{with probability } 1\,. \tag{3.3}$$

Applying the Strong Law of Large Numbers to the i.i.d. random variables $(T^{x,k} - T^{x,k-1})$ and since $T^{x,n} = \sum_{k=1}^{n}(T^{x,k} - T^{x,k-1})$, we get

$$\lim_{n\to\infty} \frac{T^{x,n}}{n} = \mathbb{E}_x(T^x) \qquad \text{with probability } 1\,. \tag{3.4}$$

Combining the results from (3.3) and (3.4) yields

$$\lim_{n\to\infty} \frac{S_{T^{x,n}}}{T^{x,n}} = \mathbb{E}_\pi(f) \qquad \text{with probability 1}\,. \tag{3.5}$$

The limit in (3.5) is close to what we need to prove, but not exactly. Using the notation $S_m = \sum_{k=0}^{m-1} f(X_k)$, we really need

$$\lim_{m\to\infty} \frac{S_m}{m} = \mathbb{E}_\pi(f) \qquad \text{with probability 1}\,.$$

We now introduce the new random variables V_m^x defined by

$$V_m^x = \sum_{s=1}^{m} \mathbb{1}_{\{x\}}(X_s) \quad \text{for } m \geq 1\,. \tag{3.6}$$

For any $m \geq 1$, V_m^x is the number of returns to state x by time m. The most recent (up to time m) return to x happens at time $T^{x,V_m^x} \leq m$. After that time, the Markov chain will visit x again at time $T^{x,V_m^x+1} > m$. See the illustration in Figure 3.1.

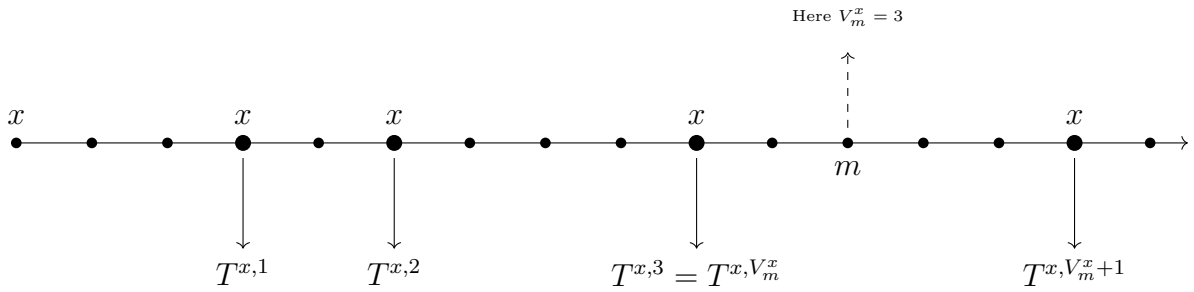


Figure 3.1

So we have

$$T^{x,V_m^x} \leq m \leq T^{x,V_m^x+1}\,,$$

and since we are (for now) assuming $f \geq 0$,

$$S_{T^{x,V_m^x}} \leq S_m \leq S_{T^{x,V_m^x+1}}\,.$$

This yields

$$\frac{S_m}{m} \leq \frac{S_m}{T^{x,V_m^x}} \leq \frac{S_{T^{x,V_m^x+1}}}{T^{x,V_m^x}} = \frac{S_{T^{x,V_m^x+1}}}{T^{x,V_m^x+1}} \frac{T^{x,V_m^x+1}}{T^{x,V_m^x}}\,. \tag{3.7}$$

Since x is recurrent, $\lim_{m\to\infty} T^{x,V_m^x} = \infty$ with probability 1, and

$$\mathbb{P}(T^{x,V_m^x+1} - T^{x,V_m^x} < \infty) = 1\,.$$

Thus

$$\lim_{m\to\infty} \frac{T^{x,V_m^x+1}}{T^{x,V_m^x}} = 1 \qquad \text{with probability } 1\,,$$

and (3.7) in combination with (3.5) yields

$$\lim_{m\to\infty} \frac{S_m}{m} \leq \mathbb{E}_\pi(f) \qquad \text{with probability } 1\,.$$

By a similar argument, using

$$\frac{S_{T^{x,V_m^x}}}{T^{x,V_m^x}} \frac{T^{x,V_m^x}}{T^{x,V_m^x+1}} \leq \frac{S_m}{T^{x,V_m^x+1}} \leq \frac{S_m}{m}\,,$$

we arrive at

$$\mathbb{E}_\pi(f) \leq \lim_{m\to\infty} \frac{S_m}{m}\,.$$

Altogether, we have

$$\lim_{m\to\infty} \frac{S_m}{m} = \mathbb{E}_\pi(f)\,. \tag{3.8}$$

Recall that we have assumed that $f \geq 0$ and that the Markov chain starts in state x. As the last step, we will now show that (3.8) holds without these restrictions. Assume f is a bounded real-valued function. We can write

$$f = \max(f, 0) - \max(-f, 0)\,.$$

Set $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. Then $f^+ \geq 0$ and $f^- \geq 0$, and we can apply (3.8) to f^+ and f^- separately, and then take the difference of the two results (which is allowed since $\infty - \infty$ is not involved), which yields (3.8) for any general, bounded f. Lastly, (3.8) holds for any starting state x. Thus if the Markov chain starts with initial distribution π_0, then we can average both sides of (3.8) with respect to the distribution π_0, and so

$$\lim_{m\to\infty} \frac{S_m}{m} = \mathbb{E}_\pi(f)$$

holds for any initial distribution π_0 and any bounded, real-valued function f on the state space $\mathcal{S}$. □

Remark 3.1.2. *What are the essential components that make the proof of Theorem 3.1.1 work?* **Recurrence** *allows us to split the trajectories into "disjoint blocks" which are of finite length with probability* 1. *The* **Strong Markov property** *guarantees that certain sequences of random variables $Z_1, Z_2, \ldots$ where Z_k is defined on block k for $k \geq 1$ are i.i.d. Probabilistically speaking, what happens on one of the blocks is typical for the behavior of the process overall, and it repeats itself on each block. The* **Strong Law of Large Numbers** *is the third main ingredient in the proof.*

Theorem 3.1.3 (Long-run fraction of time spent in a state)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with state space $\mathcal{S}$ and initial distribution π_0. For $x \in \mathcal{S}$, consider $m_x = \mathbb{E}_x(T^x)$, the mean return time to state x. Then the long-run fraction of time the Markov chain spends in state x is*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x\}}(X_k) = \frac{1}{m_x} \qquad \text{with probability 1}\,. \tag{3.9}$$

Proof. We will use the notation $\tilde{V}_n^x = \sum_{k=0}^{n-1} \mathbb{1}_{\{x\}}(X_k)$ for $n \geq 0$ (note that this slightly differs from V_n^x in (3.6)). The random variable $\tilde{V}_n^x$ is the number of visits to state x by time $(n-1)$ (taking the initial state into account), and

$$\lim_{n\to\infty} \tilde{V}_n^x = \tilde{V}^x$$

is the long-run total number of visits to state x.

Case 1: Assume the Markov chain is transient. For all $x \in \mathcal{S}$, we have $\mathbb{P}_x(T^x = \infty) > 0$, and so $m_x = \mathbb{E}_x(T^x) = \infty$. By Theorem 2.1.3,

$$\mathbb{P}(\tilde{V}^x < \infty) = 1$$

for any transient state x and any initial distribution of the Markov chain. Thus, with probability 1,

$$\lim_{n\to\infty} \frac{\tilde{V}_n^x}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x\}}(X_k) = 0 = \frac{1}{\infty} = \frac{1}{m_x}\,.$$

Case 2: We assume the Markov chain is recurrent. Let $x \in \mathcal{S}$ and consider T^x. Since the Markov chain is also irreducible, for any initial distribution, we habe $\mathbb{P}(T^x < \infty) = 1$. Hence by the strong Markov property, the process $(Y_n)_{n\geq 0} = (X_{T^x+n})_{n\geq 0}$ is a Markov chain that starts in state x and has the same transition probabilities as the original Markov chain $(X_n)_{n\geq 0}$. The long-run number of visits to state x for trajectories for $(X_n)_{n\geq 0}$ and corresponding trajectories for $(X_{T^x+n})_{n\geq 0}$ can differ at most by one (accounting for the initial state X_0). So the long-run fraction of time spent in state x is the same for both Markov chains. As a consequence we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x\}}(X_k) = \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n} \mathbb{1}_{\{x\}}(X_s)\,. \tag{3.10}$$

For the rest of the proof we will assume that the Markov chain starts in state x. What follows is similar to the proof of Theorem 3.1.1. We consider the kth-return times $T^{x,k}$ to state x defined by $T^{x,0} = 0$ and

$$T^{x,k} = \min\{n : n > T^{x,k-1} \text{ and } X_n = x\} \quad \text{for } k \geq 1\,.$$

Applying the Strong Law of Large Numbers to the i.i.d. random variables $(T^{x,k} - T^{x,k-1})$, we get

$$\lim_{n\to\infty} \frac{T^{x,n}}{n} = \mathbb{E}_x(T^x) \qquad \text{with probability } 1\,. \tag{3.11}$$

Here also we consider the random variables V_m^x defined by

$$V_m^x = \sum_{s=1}^{m} \mathbb{1}_{\{x\}}(X_s) \quad \text{for } m \geq 1$$

which give the number of returns to state x by time m. The most recent (up to time m) return to x happens at time $T^{x,V_m^x} \leq m$. After that time, the Markov chain will next visit x at time $T^{x,V_m^x+1} > m$. So we have

$$T^{x,V_m^x} \leq m \leq T^{x,V_m^x+1}\,,$$

and hence, assuming m is large enough so $V_m^x \geq 1$,

$$\frac{T^{x,V_m^x}}{V_m^x} \leq \frac{m}{V_m^x} \leq \frac{T^{x,V_m^x+1}}{V_m^x}\,. \tag{3.12}$$

Taking the limit as $m \to \infty$ on all three sides of (3.12) and applying (3.11) and (3.10), we get

$$\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} \mathbb{1}_{\{x\}}(X_k) = \frac{1}{m_x} \qquad \text{with probability } 1\,,$$

which completes the proof. □

Remark 3.1.4. *For the positive recurrent case, we can also derive (3.9) as a corollary to Theorem 3.1.1, applied to the indicator function $f = \mathbb{1}_{\{x\}}$ on $\mathcal{S}$. The result follows from $\mathbb{E}_\pi(\mathbb{1}_{\{x\}}) = \pi(x)$ and $\pi(x) = \frac{1}{m_x}$ (recall Theorem 2.2.8).*

Corollary 3.1.5 (Expected long-run fraction of time spent in a state)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with state space $\mathcal{S}$ and initial distribution π_0. For $x \in \mathcal{S}$, consider $m_x = \mathbb{E}_x(T^x)$, the mean return time to state x. Then the expected long-run fraction of time the Markov chain spends in state x is $\dfrac{1}{m_x}$.*
In particular, if $X_0 = z$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P_{zx}^k = \frac{1}{m_x}\,. \tag{3.13}$$

Proof. Note that for all $n \geq 1$,

$$0 \leq \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x\}}(X_k) \leq 1\,. \tag{3.14}$$

Hence we can apply the Dominated Convergence Theorem (Theorem C.3.3) to the sequence of random variables

$$Y_n = \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x\}}(X_k) \quad \text{for } n \geq 1\,,$$

which yields the result.

□

Remark 3.1.6. *(a) (3.13) tells us that for any states $x, z \in \mathcal{S}$, the sequence $\{P_{zx}^k\}_{k\geq 1}$ of k-step transition probabilities converges in* **Cesàro means**. *Convergence in Cesàro means is a weaker form of convergence than (regular) convergence of a sequence (see Exercise 3.1). We will address results surrounding the stronger form of convergence of the k-step transition probabilities in the next section.*
(b) If we do not assume irreducibility of the Markov chain, then (3.13) becomes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P_{zx}^k = \frac{f_{zx}}{m_x} \tag{3.15}$$

where $f_{zx} = \mathbb{P}_z(T^x < \infty)$.

Remark 3.1.7. *For $f_{zx} > 0$ and x a positive recurrent state, the limit of the Cesàro means in (3.15) is positive. Hence the expected long-run number of visits to state x is $O(n)$. On the other hand, if x is either transient or null recurrent (in both cases the Cesàro limit is 0), the expected long-run number of visits to x is of strictly smaller order than n. For the transient case, this is not a new result, since we already know that the expected number of visits to x is finite. However for the null recurrent case this does give us a new insight. A null recurrent state, being recurrent, is revisited infinitely many times with probability 1. But by (3.15), this number of revisits is of strictly smaller order than n, i.e., it is of order $o(n)$ (which is due to the "relatively long return time"; although the return time is finite with probability 1, it has infinite expectation).*

3.2 Convergence

Theorem 3.2.1 (Convergence theorem)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent, aperiodic Markov chain on discrete state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. Let π be the unique stationary distribution for the chain. Then π is also the limiting distribution for the chain, that is,*

$$\lim_{n\to\infty} P^n_{xy} = \pi(y) \qquad \text{for all } x, y \in \mathcal{S}\,.$$

As a consequence, for any initial distribution π_0 for the chain, we have

$$\lim_{n\to\infty} \pi_n(y) = \pi(y) \qquad \text{for all } y \in \mathcal{S}$$

where π_n denotes the distribution of X_n for $n \geq 0$.

Proof. The proof we present is due to Doeblin[1] and uses a *coupling argument*. Briefly put, coupling is a method by which one constructs two or more Markov chains on the same probability space. One can then use properties of the joint distribution of the coupled Markov chains to prove results about the distributions of the individual Markov chains. We will discuss the method of coupling in more detail in Section 11.3.

We consider two Markov chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ that have the same (finite or infinite) state space $\mathcal{S}$ and the same transition matrix $\mathbf{P}$. We will specify their initial distributions later in the proof. We then construct the special *independent coupling* of the two

[1] Wolfgang Doeblin, German-Jewish mathematician (1915–1940)

chains that results in the Markov chain $(X_n, Y_n)_{n\geq 0}$ with state space $\mathcal{S}^2$ and transition probabilities

$$\tilde{P}_{(x,r),(y,s)} := P_{xy}P_{rs}\,. \tag{3.16}$$

The underlying common probability space for the chain $(X_n, Y_n)_{n\geq 0}$ is the direct product of the probability space that $(X_n)_{n\geq 0}$ is defined on and the probability space that $(Y_n)_{n\geq 0}$ is defined on. We now show that $(X_n, Y_n)_{n\geq 0}$ is an irreducible, positive recurrent Markov chain.

Let $(x, r), (y, s) \in \mathcal{S}^2$. Since the transition matrix $\mathbf{P}$ is irreducible and aperiodic, there exists an $N > 0$ such that

$$P^n_{xy} > 0 \quad \text{and} \quad P^n_{rs} > 0 \quad \text{for } n \geq N\,.$$

Therefore

$$\tilde{P}^n_{(x,r),(y,s)} = P^n_{xy}P^n_{rs} > 0 \quad \text{for } n \geq N\,.$$

It follows that the coupling $(X_n, Y_n)_{n\geq 0}$ is irreducible and aperiodic.

Next we show positive recurrence of $(X_n, Y_n)_{n\geq 0}$. Consider the distribution $\tilde{\pi}$ on $\mathcal{S}^2$ defined by $\tilde{\pi}(x, r) := \pi(x)\pi(r)$. We have

$$\sum_{(x,r)\in\mathcal{S}^2} \tilde{\pi}(x,r)\tilde{P}_{(x,r),(y,s)} = \sum_{x\in\mathcal{S}}\sum_{r\in\mathcal{S}} \pi(x)\pi(r)P_{xy}P_{rs}$$

$$= \left(\sum_{x\in\mathcal{S}} \pi(x)P_{xy}\right)\left(\sum_{r\in\mathcal{S}} \pi(r)P_{rs}\right) = \pi(y)\pi(s) = \tilde{\pi}(y,s)\,.$$

This shows that $\tilde{\pi}$ is a stationary distribution for $(X_n, Y_n)_{n\geq 0}$, and so the Markov chain $(X_n, Y_n)_{n\geq 0}$ is positive recurrent.

Let $\Delta = \{(a, a) : a \in \mathcal{S}\}$ be the *diagonal* in $\mathcal{S}^2$. Consider the random variable

$$T^\Delta = \min\{n \geq 1 : (X_n, Y_n) \in \Delta\}\,,$$

that is, the first time the chain visits (any point on) the diagonal. For simplicity, we will denote $T = T^\Delta$. Since $(X_n, Y_n)_{n\geq 0}$ is irreducible and recurrent, we have $\mathbb{P}(T < \infty) = 1$. This implies

$$\lim_{n\to\infty} \mathbb{P}(T > n) = 0\,.$$

It follows from the Markov property that

$$\mathbb{P}(X_n = y \,|\, T \leq n) = \mathbb{P}(Y_n = y \,|\, T \leq n) \quad \text{for } n \geq 1\,, \tag{3.17}$$

since from time T onwards, after $(X_n, Y_n)_{n\geq 0}$ has hit Δ, the laws of the random variables X_n and Y_n will be the same. Multiplying both sides of (3.17) by $\mathbb{P}(T \leq n)$ gives

$$\mathbb{P}(X_n = y, T \le n) = \mathbb{P}(Y_n = y, T \le n) \quad \text{for } n \ge 1\,.$$

By the law of total probability,

$$\mathbb{P}(X_n = y) = \mathbb{P}(X_n = y, T \le n) + \mathbb{P}(X_n = y, T > n)\,,$$

$$\mathbb{P}(Y_n = y) = \mathbb{P}(Y_n = y, T \le n) + \mathbb{P}(Y_n = y, T > n)\,.$$

It follows that

$$\mathbb{P}(X_n = y) - \mathbb{P}(X_n = y, T > n) = \mathbb{P}(Y_n = y) - \mathbb{P}(Y_n = y, T > n)\,.$$

Note that

$$\lim_{n\to\infty} \mathbb{P}(X_n = y, T > n) \le \lim_{n\to\infty} \mathbb{P}(T > n) = 0\,,$$
$$\lim_{n\to\infty} \mathbb{P}(Y_n = y, T > n) \le \lim_{n\to\infty} \mathbb{P}(T > n) = 0\,.$$

It follows that

$$\lim_{n\to\infty} [\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)] = 0\,. \tag{3.18}$$

Finally, let us specify the initial distributions for the chains $(X_n)_{n\ge0}$ and $(Y_n)_{n\ge0}$. We let chain $(X_n)_{n\ge0}$ start in state x and chain $(Y_n)_{n\ge0}$ start in the stationary distribution π. Then $\mathbb{P}(Y_n = y) = \pi(y)$ for all $n \ge 0$, and thus (3.18) yields

$$\lim_{n\to\infty} \mathbb{P}(X_n = y) = \lim_{n\to\infty} P^n_{xy} = \pi(y) \qquad \text{for all } x, y \in \mathcal{S}\,.$$

This completes the proof of the convergence theorem. □

As a consequence of Theorem 3.2.1, for an irreducible, aperiodic Markov chain with *finite* state space $\mathcal{S}$ and $|\mathcal{S}| = k$, transition matrix $\mathbf{P}$, and stationary distribution π, we have

$$\mathbf{P}^n \xrightarrow{n\to\infty} \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(k) \\ \pi(1) & \pi(2) & \cdots & \pi(k) \\ \vdots & \vdots & & \vdots \\ \pi(1) & \pi(2) & \cdots & \pi(k) \end{pmatrix}.$$

Example 3.2.1. Let $0 < p < 1$. Consider the Markov chain $(X_n)_{n\ge0}$ with $\mathcal{S} = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 1-p & 0 & p \\ 0 & 1 & 0 \end{pmatrix}$$

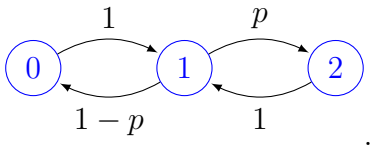


Figure 3.2

and whose transition graph is shown in Figure 3.2. The Markov chain is irreducible and **periodic** with period 2. The Markov chain $(X_n, Y_n)_{n\geq 0}$ as constructed in the proof of Theorem 3.2.1 has state space $\mathcal{S} \times \mathcal{S}$ (with 9 states). Figure 3.3 shows the transition graph for $(X_n, Y_n)_{n\geq 0}$ (the arrows represent positive one-step transition probabilities).

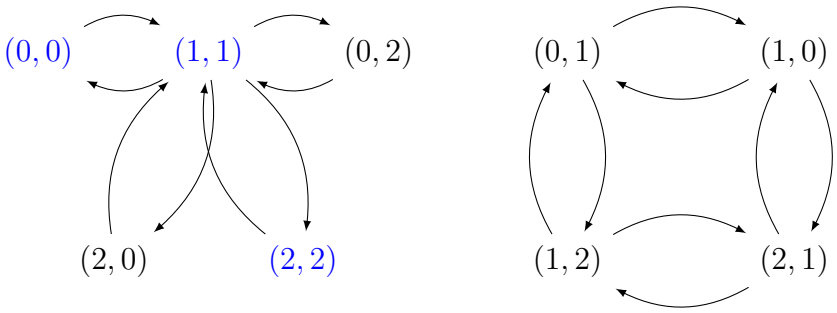


Figure 3.3

The set of blue states in the above diagram is the diagonal $\Delta = \{(0,0), (1,1), (2,2)\}$ of $\mathcal{S} \times \mathcal{S}$. The Markov chain $(X_n, Y_n)_{n\geq 0}$ has two irreducible closed classes and, as shown in the proof of Theorem 3.2.1, cannot have transient states. Since $(X_n, Y_n)_{n\geq 0}$ is *reducible*, the coupling argument in the proof of Theorem 3.2.1 breaks down. Indeed, the Markov chain $(X_n)_{n\geq 0}$ does not converge due to its periodicity. □

Theorem 3.2.2 (Periodic case)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain with transition probabilities P_{xy}, $x, y \in \mathcal{S}$ and unique stationary distribution π. Assume the Markov chain is periodic with period $c \geq 2$. Then for any $x, y \in \mathcal{S}$ there exists a unique integer r with $0 \leq r \leq c-1$ such that*

$$\lim_{n\to\infty} P_{xy}^{r+nc} = c\,\pi(y)\,.$$

Proof. Let $x, y \in \mathcal{S}$. Consider the partition of $\mathcal{S}$ into its cyclic classes $S_0, S_1, ..., S_{c-1}$. Without loss of generality, let us assume that $x \in S_0$ and $y \in S_r$. For δ_x (i.e., unit mass

at state x), let $\mu = \delta_x \mathbf{P}^r$. Note that μ is a probability distribution concentrated on S_r. We now consider the Markov chain $(Y_n)_{n\geq 0}$ with $Y_n = X_{cn},\ n \geq 0$, with initial distribution μ. It can be viewed as a Markov chain on reduced state space S_r. By Proposition 2.4.7, this Markov chain is irreducible and aperiodic. By Corollary 2.4.8, its unique stationary distribution is $\pi^{S_r}(y) = c\,\pi(y)$ for $y \in S_r$. Hence, by Theorem 3.2.1, we have

$$\lim_{n\to\infty} P_{xy}^{r+nc} = \lim_{n\to\infty} \mathbb{P}_x(X_{r+nc} = y) = \lim_{n\to\infty} \mathbb{P}_\mu(Y_n = y) = c\,\pi(y)\,.$$

□

Example 3.2.2. We return to Example 3.2.1. This is an irreducible Markov chain on $\mathcal{S} = \{0, 1, 2\}$, hence it is positive recurrent. Its stationary distribution is $\pi = (\frac{1-p}{2}, \frac{1}{2}, \frac{p}{2})$. The Markov chain is periodic with period 2. Its cyclic classes are $S_0 = \{0, 2\}$ and $S_1 = \{1\}$. A direct computation yields

$$\mathbf{P}^2 = \begin{pmatrix} 1-p & 0 & p \\ 0 & 1 & 0 \\ 1-p & 0 & p \end{pmatrix} \quad \text{and} \quad \mathbf{P}^3 = \begin{pmatrix} 0 & 1 & 0 \\ 1-p & 0 & p \\ 0 & 1 & 0 \end{pmatrix} = \mathbf{P} \tag{3.19}$$

from which we conclude that $\mathbf{P}^{1+2n} = \mathbf{P}$ for all $n \geq 0$, and $\mathbf{P}^{2n} = \mathbf{P}^2$ for all $n \geq 1$.

Case 1: Let x and y be in different cyclic classes. Then by Theorem 3.2.2,

$$\lim_{n\to\infty} P_{xy}^{1+2n} = 2\pi(y)\,.$$

We verify that this is indeed the case by reading off from (3.19) that

$$\lim_{n\to\infty} P_{0,1}^{1+2n} = 1 = 2\pi(1), \quad \lim_{n\to\infty} P_{2,1}^{1+2n} = 1 = 2\pi(1),$$

$$\lim_{n\to\infty} P_{1,0}^{1+2n} = 1-p = 2\pi(0), \quad \lim_{n\to\infty} P_{1,2}^{1+2n} = p = 2\pi(2)\,.$$

Case 2: Let x and y be in the same cyclic class. Then by Theorem 3.2.2,

$$\lim_{n\to\infty} P_{xy}^{2n} = 2\pi(y)\,.$$

We verify that this is indeed the case by reading off from (3.19) that

$$\lim_{n\to\infty} P_{0,0}^{2n} = 1-p = 2\pi(0), \quad \lim_{n\to\infty} P_{0,2}^{2n} = p = 2\pi(2), \quad \lim_{n\to\infty} P_{1,1}^{2n} = 1 = 2\pi(1)\,,$$

and similarly for the rest. This illustrates Theorem 3.2.2. □

We quote the following result for *null recurrent* states. For a reference see [29].

Theorem 3.2.3 (Orey's theorem)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and assume $y \in \mathcal{S}$ is a null recurrent state. Then*

$$\lim_{n\to\infty} P_{xy}^n = 0 \qquad \text{for all } x \in \mathcal{S}\,.$$

As a consequence, for any initial distribution π_0 for the chain, we have

$$\lim_{n\to\infty} \pi_n(y) = 0$$

where π_n denotes the distribution of X_n for $n \geq 0$.

3.3 Long-run behavior of reducible chains

Here we combine the results from the previous sections to understand the asymptotic behavior of reducible Markov chains. Let $(X_n)_{n\geq 0}$ be a reducible Markov chain on (finite or infinite) state space $\mathcal{S}$. In general, for given $y \in \mathcal{S}$, whether or not $\lim_{n\to\infty} \pi_n(y)$ exists, and if so, its value, will depend on the initial distribution π_0 of the chain.

Case 1: $y \in \mathcal{S}$ is either null recurrent or transient. Then by Theorem 3.2.3 and by Corollary 2.1.4 we have

$$\lim_{n\to\infty} P_{xy}^n = 0 \qquad \text{for all } x \in \mathcal{S}\,,$$

and hence

$$\lim_{n\to\infty} \pi_n(y) = 0 \qquad \text{for any initial distribution } \pi_0 \text{ of the chain}\,.$$

Case 2: $y \in \mathcal{S}$ is positive recurrent. Then $y \in R_k$ for a unique irreducible closed class R_k of (positive) recurrent states for the chain. There exists a unique stationary distribution π_{R_k} that is concentrated on R_k.

(a) If $x \in R_j$ where R_j is an irreducible closed class that is distinct from R_k, then x does not lead to y and therefore

$$\lim_{n\to\infty} P_{xy}^n = 0\,.$$

(b) If $x \in R_k$ and the states in R_k are aperiodic, then

$$\lim_{n\to\infty} P_{xy}^n = \pi_{R_k}(y)\,.$$

If the states in R_k are periodic with period $c \geq 2$, then the above limit does not exist.

(c) If x is a transient state and the states in R_k are aperiodic, then

$$\lim_{n\to\infty} P_{xy}^n = (a_{x,R_k})\pi_{R_k}(y)$$

where a_{x,R_k} is the (absorption) probability that, given that the chain starts in state x, it will eventually enter R_k. If the states in R_k are periodic with period $c \geq 2$, then the above limit does not exist.

Example 3.3.1. Recall Exemples 2.3.2 and 2.3.3. The Markov chain has state space $\mathcal{S} = \{1, 2, ..., 6\}$ and transition matrix (in canonical form)

$$\mathbf{P} = \begin{array}{c} \\ 3 \\ 1 \\ 4 \\ 6 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 3 & 1 & 4 & 6 & 2 & 5 \end{array} \\ \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0 & 0.5 & 0.1 & 0.4 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0.2 & 0.1 & 0 & 0 & 0.3 & 0.4 \\ 0.3 & 0 & 0.1 & 0 & 0.6 & 0 \end{array}\right) \end{array}.$$

The irreducible closed classes are $R_1 = \{3\}$ and $R_2 = \{1, 4, 6\}$ and are aperiodic. States $3, 1, 4, 6$ are positive recurrent states (finite-state Markov chain do not have null recurrent states). We compute

$$\pi_{R_1} = (1, 0, ..., 0)$$

and

$$\pi_{R_2} = (0, \frac{18}{37}, \frac{2}{37}, \frac{17}{37}, 0, 0)\,.$$

In Example 2.3.3 we have computed for the absorption probabilities (written as a matrix):

$$\mathbf{A} = \begin{array}{c} \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cc} R_1 & R_2 \end{array} \\ \begin{pmatrix} 0.7 & 0.3 \\ 0.72 & 0.28 \end{pmatrix} \end{array}.$$

Here the n-step transition matrices do approach a limit $\mathbf{L}$ as $n \to \infty$. We have

$$\mathbf{P}^n \xrightarrow{n\to\infty} \begin{array}{c} \\ 3 \\ 1 \\ 4 \\ 6 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 3 & 1 & 4 & 6 & 2 & 5 \end{array} \\ \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{18}{37} & \frac{2}{37} & \frac{17}{37} & 0 & 0 \\ 0 & \frac{18}{37} & \frac{2}{37} & \frac{17}{37} & 0 & 0 \\ 0 & \frac{18}{37} & \frac{2}{37} & \frac{17}{37} & 0 & 0 \\ \hline 0.7 & (0.3)\frac{18}{37} & (0.3)\frac{2}{37} & (0.3)\frac{17}{37} & 0 & 0 \\ 0.72 & (0.28)\frac{18}{37} & (0.28)\frac{2}{37} & (0.28)\frac{17}{37} & 0 & 0 \end{array}\right) \end{array} = \mathbf{L}\,.$$

Assuming the Markov chain starts with initial distribution π_0, we have

$$\lim_{n\to\infty} \pi_n = \lim_{n\to\infty} \pi_0 \mathbf{P}^n = \pi_0 \mathbf{L}\,.$$

□

Exercises

Exercise 3.1 (Cesàro means)**.** (a) Consider a sequence $\{a_n\}_{n\geq 1}$ of real numbers with $\lim_{n\to\infty} a_n = a$ and consider its Cesàro means

$$c_n = \frac{1}{n}\sum_{k=1}^{n} a_k$$

for $n \geq 1$. Prove that $\{c_n\}_{n\geq 1}$ is also convergent and $\lim_{n\to\infty} c_n = a$.

(b) Show that the converse of the statement of part (a) is not true. That is, show that if a sequence $\{a_n\}_{n\geq 1}$ converges in Cesàro means, it does not imply that $\{a_n\}_{n\geq 1}$ converges.

Exercise 3.2. Consider a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S} = \{1, 2, ..., 7\}$ and transition matrix

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.1 & 0 & 0.2 & 0.1 & 0.2 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0 & 0.1 & 0.2 & 0.4 \end{pmatrix} \end{array}.$$

Recall the notation $f_{x,y} = \mathbb{P}(T^y < \infty \,|\, X_0 = x)$. Compute the following probabilities: $f_{4,5}$, $f_{6,2}$, $f_{6,5}$, $f_{6,6}$, and $f_{6,7}$.

Exercise 3.3. Consider a so-called 1-*server queue*, where at time n, a number X_n of customers are in queue. The first in line of these X_n customers is being served, while the rest of the customers are waiting. During each time interval, the probability that a new customer joins the queue is $p = 1/5$, and the probability that the customer currently being served finishes their service and leaves the queue is $q = 1/3$. This system defines a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathbb{N}_0$. What is the long-run expected fraction of time the queue is empty?

Exercise 3.4. Consider a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S} = \{1, 2, ..., 7\}$ and the transition matrix $\mathbf{P}$ from Exercise 3.2. Compute the following limits (if they exist). If a limit does not exist, explain why.

(a) $\lim_{n\to\infty} P^n_{1,2}$, $\lim_{n\to\infty} P^n_{6,2}$, $\lim_{n\to\infty} P^n_{6,7}$, $\lim_{n\to\infty} P^n_{6,5}$

(b) $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^k_{6,5}$ (c) $\lim_{n\to\infty} \mathbf{P}^n$

Exercise 3.5. Consider a Markov chain on state space $\mathcal{S} = \{1, 2, ...\}$ whose transition probabilities are

$$P_{xy} = \frac{1}{x+1} \text{ for } y = 1, 2, ..., x+1.$$

Compute the mean return time to state 1.

Exercise 3.6. Consider the Markov chain on state space $\mathcal{S} = \mathbb{N}_0$ with transition probabilities

$$P_{x,x+1} = \frac{1}{x+1} \quad \text{and} \quad P_{x,0} = \frac{x}{x+1} \quad \text{for all } x \geq 0\,.$$

(a) Will this Markov chain converge to a unique limiting distribution? If so, compute this distribution.

(b) Let $x > 0$. What is the expected number of visits to state 0 between two consecutive visits to state x?

Exercise 3.7. Each morning a student takes one of the three books he owns from his shelf. The probability that he chooses Book i is α_i with $\alpha_1 = 1/3$, $\alpha_2 = 1/2$, $\alpha_3 = 1/6$. We assume that his choices on successive days are independent. In the evening, he replaces the book at the left-hand end of the shelf. Let p_n denote the probability that on day n the student finds the books in order $1, 2, 3$ (from left to right). Show that, irrespective of the initial arrangement of the books, $\lim_{n\to\infty} p_n$ exists, and determine the limit. (This type of process is a sorting scheme known as a **Tsetlin library**.)

Exercise 3.8. A fair 6-sided die is rolled repeatedly. Consider the Markov chain $(X_n)_{n\geq 1}$ where X_n denotes the sum of the first n rolls. Does

$$\lim_{n\to\infty} \mathbb{P}(X_n \text{ is a multiple of } 7)$$

exist? Why or why not? If the limit exists, compute it.

Exercise 3.9. A Math graduate student owns 2 bikes which she keeps either at home or on a bike rack near her campus office. Every day she makes two journeys: In the morning she travels to her campus office, and in the evening she returns home. When it rains, she always walks. When the weather is clear, she bikes, provided that at least one of her bikes is parked at her location of departure (her home or her campus office). Suppose that it rains on each journey with probability p independently of past or future weather. What is the long-run proportion of journeys on which she has to walk although the weather is clear?

Chapter 4

Random Walks on $\mathbb{Z}$

4.1 Basics

We have introduced general random walk on $\mathbb{Z}$ in Example 1.5.3. In this chapter, we will mainly (but not exclusively) focus on *simple* random walk on $\mathbb{Z}$. We begin with the basic definition.

Definition 4.1.1 (Simple random walk on $\mathbb{Z}$)**.** *Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables taking values in $\{-1, 1\}$.*

- *The stochastic process $(S_n)_{n\geq 0}$ defined by $S_0 = 0$ (unless otherwise noted) and*

$$S_n = \sum_{k=1}^{n} X_k$$

  for $n \geq 1$ is called **simple random walk on $\mathbb{Z}$.**

- *If $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = \frac{1}{2}$, we call the process* **simple symmetric random walk** *on $\mathbb{Z}$. Otherwise, we call the process* **simple biased random walk** *on $\mathbb{Z}$.*

Simple random walk on $\mathbb{Z}$ is a Markov chain, more specifically it is a birth/death chain with transition probabilities $P_{x,x+1} = p$ and $P_{x,x-1} = 1 - p$ for some fixed $p \in (0, 1)$ and for all $x \in \mathbb{Z}$. The underlying probability space for $(S_n)_{n\geq 0}$ can be identified with $\Omega = \{-1, 1\}^{\mathbb{N}}$, the space of all infinite binary sequences. Elements $\omega \in \Omega$ are in one-to-one correspondence with trajectories, also called sample paths, for $(S_n)_{n\geq 0}$ where S_n is the location of the random walk at time n. The easiest way to visualize a trajectory is via a graph that plots location against time (and has line segments connecting neighboring points (n, S_n) and $(n+1, S_{n+1})$). Figure 4.1 shows a sample path for simple random walk

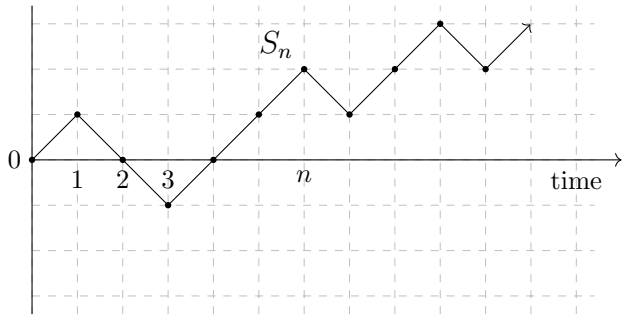


Figure 4.1: A sample path for simple random walk on $\mathbb{Z}$

on $\mathbb{Z}$. For simple symmetric random walk, for all $n \geq 1$, we have

$$\mathbb{E}(S_n) = 0 \quad \text{and} \quad \text{Var}(S_n) = n\,.$$

By the Central Limit Theorem,

$$\frac{S_n}{\sqrt{n}} \quad \xrightarrow{n\to\infty} \quad N(0,1) \qquad \text{in Distribution}$$

where $N(0,1)$ is a standard normal random variable. Recall that for a normal random variable centered at 0, the probability that its absolute value exceeds three standard deviations is very small, only about 0.003. Therefore for large n, we have

$$\mathbb{P}(-3\sqrt{n} \leq S_n \leq 3\sqrt{n}) \approx 0.997\,. \tag{4.1}$$

The green shaded region in Figure 4.2 marks the region bounded by $-3\sqrt{n}$ and $3\sqrt{n}$ in the vertical direction. The two red lines mark the boundaries for any trajectory. For large times n, the point (n, S_n) on a trajectory will fall inside the green region for an overwhelming majority of trajectories.

Simple random walk on $\mathbb{Z}$ is periodic with period 2. Given that the walk starts at 0, at even times it can only be at an even integer, and at odd times at an odd integer. Hence if x and y have unequal parity, then $P^{2n}_{xy} = 0$ for $n \geq 1$.

Notice that for random walk on $\mathbb{Z}$, due to the translation invariance of the one-step transition probabilities, we have

$$P^n_{xy} = P^n_{0,(y-x)} \qquad \text{for all } x, y \in \mathbb{Z} \text{ and } n \geq 1.$$

It therefore suffices to study the transition probabilities P^n_{0m} for $m \in \mathbb{Z}$.

Let $m \geq 0$ and $n \geq 1$ with $m \leq n$. We assume that either both m and n are even or both m and n are odd. For the walk to end up at m in n steps, it must have taken x steps to the left and $x + m$ steps to the right (in any order), with $2x + m = n$. There are

$$\binom{n}{\frac{n-m}{2}} = \binom{n}{\frac{n+m}{2}} \tag{4.2}$$

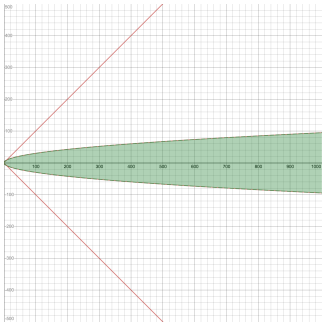


Figure 4.2

such binary sequences (sequences of left/right steps) of length n. We denote the number of paths of length n that end up at m by $N_n(0,m)$. The one-to-one correspondence between binary sequences and paths yields

$$N_n(0,m) = \binom{n}{\frac{n+m}{2}} \quad \text{for } m > 0\,.$$

Now let $m < 0$. For the walk to end up at m in n steps, it must have taken x steps to the right and $x + |m|$ steps to the left (in any order), with $2x + |m| = n$. Because of symmetry, the number of such sequences is also (4.2). For simple symmetric random walk, each binary sequence of length n occurs with the same probability $\frac{1}{2^n}$. This yields the following transition probabilities:

Proposition 4.1.1. *For simple symmetric random walk on $\mathbb{Z}$,*

$$\mathbb{P}(S_n = m) = P^n_{0m} = \binom{n}{\frac{n+m}{2}} \frac{1}{2^n} \quad \text{for } m \in \{-n, -n+2, ..., n-2, n\}\,. \tag{4.3}$$

Notice for simple symmetric random walk, the distribution of S_n is symmetric, i.e., $\mathbb{P}(S_n = m) = \mathbb{P}(S_n = -m)$. If n is even, the mass function in (4.3) takes its maximum at $m = 0$, and if n is odd, it takes its maximum at -1 and 1.

Now assume that we have biased random walk with $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_k = -1) = 1 - p$. Set $q = 1 - p$. For $m \geq 0$, any path of length n that ends up at m, occurs with probability

$$p^{\frac{n+m}{2}} q^{\frac{n-m}{2}}. \tag{4.4}$$

And for $m < 0$, any path of length n that ends up at m, occurs with probability $p^{\frac{n-|m|}{2}} q^{\frac{n+|m|}{2}}$, which can be rewritten as (4.4). Hence we arrive at the following result:

Proposition 4.1.2. *For simple biased random walk on $\mathbb{Z}$ with $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_k = -1) = q$ and $p + q = 1$,*

$$\mathbb{P}(S_n = m) = P^n_{0m} = \binom{n}{\frac{n+m}{2}} p^{\frac{n+m}{2}} q^{\frac{n-m}{2}} \qquad \text{for } m \in \{-n, -n+2, ..., n-2, n\}\,.$$

4.2 Pólya's Random Walk Theorem

Here we present Pólya's famous theorem about recurrence / transience of simple symmetric random walk on the integer lattice $\mathbb{Z}^d$. A number of alternate proofs are available. We will return to Pólya's theorem in Section 8.6, where we will reprove the result by taking the electric network approach to the study of random walks on graphs.

Theorem 4.2.1 (Pólya, 1921)**.** *Simple symmetric random walk on $\mathbb{Z}^d$ is recurrent for $d = 1$ and $d = 2$ and transient for $d > 2$. Simple biased random walk on $\mathbb{Z}^d$ is transient for all $d \geq 1$.*

Proof. We separate the proof by the dimension d.

Case $d = 1$.

Clearly, simple random walk on $\mathbb{Z}$ is irreducible, so it suffices to prove that the expected number of returns to a given state is infinite. Let this state be 0. We will compute $\sum_{n=1}^{\infty} P^n_{00}$. Note that $P^n_{00} = 0$ if n is odd. So we only need to consider $n = 2m$. Returning to the starting point in $2m$ steps means that the walk must have taken m steps to the right and m steps to the left (in any order). Thus we have

$$P^{2m}_{00} = \binom{2m}{m} p^m (1-p)^m\,.$$

Convergence or divergence of the series $\sum_{m=1}^{\infty} \binom{2m}{m} p^m (1-p)^m$ will of course be determined by the asymptotics of $\binom{2m}{m} p^m (1-p)^m$. Using Stirling's approximation (A.1) we get, for large m,

$$\binom{2m}{m} p^m (1-p)^m \approx \frac{(2m)^{2m} e^{-2m} \sqrt{2\pi 2m}}{\left(m^m e^{-m} \sqrt{2\pi m}\right)^2} p^m (1-p)^m = \frac{(4p(1-p))^m}{\sqrt{\pi m}}\,.$$

Note that $4p(1-p) < 1$ for $p \neq \frac{1}{2}$. Thus

$$\sum_{m=1}^{\infty} P_{00}^{2m} < \infty \quad \text{for } p \neq \frac{1}{2}\,,$$

and so biased simple random walk on $\mathbb{Z}$ is transient. For $p = \frac{1}{2}$, we have $P_{00}^{2m} \approx \frac{1}{\sqrt{\pi m}}$, and so

$$\sum_{m=1}^{\infty} P_{00}^{2m} = \infty \quad \text{for } p = \frac{1}{2}\,,$$

which implies that simple symmetric random walk on $\mathbb{Z}$ is recurrent.

Case $d = 2$.

First simple symmetric random walk on $\mathbb{Z}^2$. The walk takes steps north, south, east, or west on the 2-dimensional integer lattice, each with probability $\frac{1}{4}$. Equivalently, we can consider simple symmetric random walk on the diagonal lattice which is the green lattice in Figure 4.3. Simple symmetric random walk on the green lattice arises from two *independent* steps of simple symmetric random walk on $\mathbb{Z}$. Hence for large m, we can use the approximation

$$P_{00}^{2m} \approx \left(\frac{1}{\sqrt{\pi m}}\right)^2 = \frac{1}{\pi m}\,,$$

and thus

$$\sum_{m=1}^{\infty} P_{00}^{2m} = \infty\,.$$

This shows that simple symmetric random walk on $\mathbb{Z}^2$ is recurrent. For simple biased

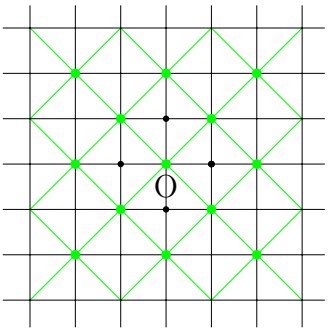


Figure 4.3

random walk on $\mathbb{Z}^2$ consider its projection onto the east-west or the north-south axis (choose a direction in which the walk is biased). under this projection, we have biased random walk on $\mathbb{Z}$ with positive holding probability. Adding fixed holding probability to each state and (and appropriately rescaling the transition probabilities) does not change

transience or recurrence of a Markov chain. Thus the projected one-dimensional (assumed biased) random walk is transient. It must follow that the original biased random walk on $\mathbb{Z}^2$ is also transient, since returning to 0 infinitely often requires returning to 0 infinitely often (and simultaneously) in both directions.

Case $d = 3$.

Simple symmetric random walk on $\mathbb{Z}^3$ takes steps east, west, north, south, up, or down with equal probability $\frac{1}{6}$. Note that we cannot extend the argument from the 2 dimensional case to the 3-dimensional case: Combining three independent one-dimensional random walks would create 8 possible directions for each step for the alternative random walk, however simple random walk on $\mathbb{Z}^3$ progresses in one of 6 possible directions in each step. Instead, we explicitly calculate the return probabilities. If the walk returns to 0 at time $2m$, then it must have taken i steps east (and hence also i steps west), j steps north (and hence also j steps south), and $m - i - j$ steps up (and hence also $m - i - j$ steps down). Thus

$$\begin{aligned} P_{00}^{2m} &= \sum_{\substack{i,j\geq 0 \\ i+j\leq m}} \frac{(2m)!}{(i!\,j!\,(m-i-j)!)^2} \left(\frac{1}{6}\right)^{2m} \\ &= \left(\frac{1}{2}\right)^{2m} \binom{2m}{m} \sum_{\substack{i,j\geq 0 \\ i+j\leq m}} \left(\frac{m!}{i!\,j!\,(m-i-j)!} \left(\frac{1}{3}\right)^m\right)^2 . \end{aligned}$$

Since

$$\sum_{\substack{i,j\geq 0 \\ i+j\leq m}} \frac{m!}{i!\,j!\,(m-i-j)!} \left(\frac{1}{3}\right)^m = 1\,,$$

we get

$$P_{00}^{2m} \leq \left(\frac{1}{2}\right)^{2m} \binom{2m}{m} \max_{\substack{i,j\geq 0 \\ i+j\leq m}} \left(\frac{m!}{i!\,j!\,(m-i-j)!} \left(\frac{1}{3}\right)^m\right) .$$

The multinomial coefficient $\frac{m!}{i!\,j!\,(m-i-j)!}$ is a maximum when i and j are $m/3$ (or as close as possible, since they are integers), so

$$P_{00}^{2m} \leq \left(\frac{1}{2}\right)^{2m} \binom{2m}{m} \frac{m!}{((m/3)!)^3} \left(\frac{1}{3}\right)^m .$$

Applying Stirling's approximation to the right hand side yields

$$\left(\frac{1}{2}\right)^{2m} \binom{2m}{m} \frac{m!}{((m/3)!)^3} \left(\frac{1}{3}\right)^m \approx C\frac{1}{m^{3/2}}\,,$$

for some constant C that does not depend on m. And so for large m we have

$$P_{00}^{2m} \leq C\frac{1}{m^{3/2}}$$

which implies

$$\sum_{m=1}^{\infty} P_{00}^{2m} < \infty \,.$$

This shows that simple symmetric random walk on $\mathbb{Z}^3$ is transient. By a similar projection argument that we have used in lower dimensions, we also know that simple biased walk on $\mathbb{Z}^3$ is transient.

Case $d \geq 4$.
Simple random walk in dimension 4 or higher is transient: Consider the projected random walk onto the first 3 dimensions, which is transient. □

Remark 4.2.2. *Consider simple symmetric random walk on $\mathbb{Z}^d$. The asymptotics for the return probabilities P_{00}^{2m} for any $d \geq 1$ are well known:*

$$P_{00}^{2m} \sim 2\left(\frac{d}{4m\pi}\right)^{d/2} \qquad \text{as } m \to \infty \,. \tag{4.5}$$

Note that this formula matches with our computations for $d = 1$ and $d = 2$. For a proof of (4.5), see [34].

Pólya's stroll in the park:

Pólya was motivated to study random walks, more precisely collisions of two independent walks, when he noticed that he frequently ran into the same couple during his daily walks in the woods near his hotel where he was staying during a conference. How often will two random walkers meet? Assume both parties start at the origin **O** at time 0. See figure **??** for an illustration. We keep track of the two walkers' *location relative to each other* and interpret these changing locations as a single simple symmetric random walk of only one of the walkers and for which we record only every other step. With this viewpoint, Walker 1, say, is the "moving origin", and Walker 2 performs a simple symmetric random walk on a (moving) $\mathbb{Z}^2$ lattice for which only every other step (and hence the location at even times only) is being recorded. Since simple symmetric random walk on $\mathbb{Z}^2$ is recurrent, and returns to the starting point can only occur at even times, it follows that Walker 2 will "return" to the moving origin (= the location of Walker 1) infinitely many times with probability 1. Walker 1 and Walker 2 will meet on their random walk in the park **infinitely many times with probability one** (unless they get tired around dinner time and return to their hotels).

Note that this is not to say that the two walkers will meet at the (fixed) origin **O** infinitely many times. In fact, the expected number of times at which Walker 1 and Walker 2 meet at **O** is finite. This follows since their individual walks are independent, and so the probability that both are at **O** at a specific time n is the product of their individual probabilities of being at **O** at that time. Based on this, one shows that the expected number meetings at **O** is finite.

Also note that a parity issue can arise: If the two walkers start at different locations, for example one step apart from each other, then they can never meet. If they start at an even number of steps apart from each other, then they *will* meet infinitely many times with probability one.

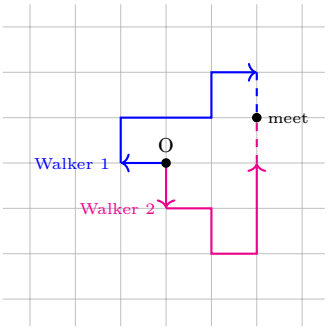


Figure 4.4: Two independent random walkers meet

4.3 Wald's Equations

For a random walk $(S_n)_{n\geq 0}$, the expectation of the location of the walk at time n is of obvious interest. For a *fixed* time n, the expectation $\mathbb{E}(S_n) = n\,\mathbb{E}(X_i)$ is immediate. But what if time itself is a random variable T, is it then true that $\mathbb{E}(S_T) = \mathbb{E}(T)\,\mathbb{E}(X_i)$, as one might guess? The answer is *no* in general, as the following example illustrates.

Example 4.3.1. Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0. Consider the random time T defined by

$$T = \min\{n : X_{n+1} = -1\}$$

which is the number of steps the walk takes to the right before its first step to the left. For an illustration, see Figure 4.5. Then $T + 1 \sim \text{Geom}(\frac{1}{2})$, and we compute $\mathbb{E}(T) = 1$. Clearly, $S_T = T \cdot 1 = T$, and so $\mathbb{E}(S_T) = 1$ as well. But

$$\mathbb{E}(T)\,\mathbb{E}(X_i) = 1 \cdot 0 \neq 1 = \mathbb{E}(S_T)\,.$$

□

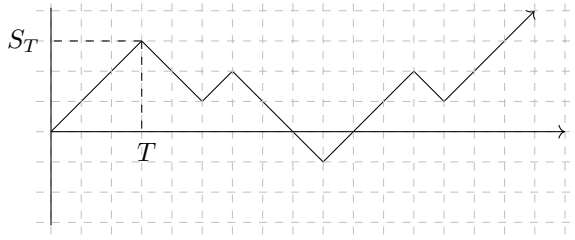


Figure 4.5

Theorem 4.3.1 gives conditions on the random time T and on the type of dependence of T and the random variables X_i, $i \geq 1$, that guarantee that $\mathbb{E}(S_T) = \mathbb{E}(X_i)\,\mathbb{E}(T)$ holds.

Theorem 4.3.1 (Wald's equations)**.** *Let $X_1, X_2, \ldots.$ be i.i.d. random variables with $\mathbb{E}(|X|_i) < \infty$. Consider $(S_n)_{n\geq 1}$ with $S_n = \sum_{i=1}^n X_i$, and let T be a stopping time for $(S_n)_{n\geq 1}$.*

(a) Wald's first equation: If $\mathbb{E}(T) < \infty$, then

$$\mathbb{E}(S_T) = \mathbb{E}(X_i)\,\mathbb{E}(T)\,. \tag{4.6}$$

(b) Wald's second equation: If $\mathbb{E}(T) < \infty$ and, in addition, $\mathbb{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then

$$\mathrm{Var}(S_T) = \sigma^2\,\mathbb{E}(T)\,. \tag{4.7}$$

Proof. (a) First, note that we can write

$$S_T = \sum_{i=1}^{T} X_i = \sum_{i=1}^{\infty} X_i \mathbb{1}_{\{T\geq i\}}\,.$$

Taking expectations, we get

$$\mathbb{E}(S_T) = \mathbb{E}\left(\sum_{i=1}^{\infty} X_i \mathbb{1}_{\{T\geq i\}}\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i \mathbb{1}_{\{T\geq i\}}) \tag{4.8}$$

where the interchange of $\mathbb{E}$ and $\sum$ is justified as we will now show: For all $n \geq 1$,

$$\sum_{i=1}^{n} X_i \mathbb{1}_{\{T\geq i\}} \leq \sum_{i=1}^{n} |X_i| \mathbb{1}_{\{T\geq i\}} \leq \sum_{i=1}^{\infty} |X_i| \mathbb{1}_{\{T\geq i\}}\,.$$

By the Monotone Convergence theorem, we have

$$\mathbb{E}(\sum_{i=1}^{\infty} |X_i| \mathbb{1}_{\{T\geq i\}}) = \mathbb{E}(\lim_{n\to\infty} \sum_{i=1}^{n} |X_i| \mathbb{1}_{\{T\geq i\}}) = \lim_{n\to\infty} \mathbb{E}(\sum_{i=1}^{n} |X_i| \mathbb{1}_{\{T\geq i\}})\,.$$

Since the event $\{T \geq i\} = \{T \leq i-1\}^c$ is defined by the random variables $X_1, X_2, ..., X_{i-1}$, the random variables $\mathbb{1}_{\{T \geq i\}}$ and X_i are independent, and we have $\mathbb{E}(|X_i|\mathbb{1}_{\{T \geq i\}}) = \mathbb{E}(|X_i|)\mathbb{P}(T \geq i)$. Thus

$$\mathbb{E}(\sum_{i=1}^{\infty} |X_i|\mathbb{1}_{\{T \geq i\}}) = \mathbb{E}(|X_i|) \sum_{i=1}^{\infty} \mathbb{P}(T \geq i) = \mathbb{E}(|X_i|)\mathbb{E}(T) < \infty \,. \tag{4.9}$$

It follows that the random variable $\sum_{i=1}^{\infty} |X_i|\mathbb{1}_{\{T \geq i\}}$ is integrable, and it dominates the random variables $\sum_{i=1}^{n} X_i \mathbb{1}_{\{T \geq i\}}$ for all $n \geq 1$. By the Dominated Convergence theorem, the interchange of lim and $\sum$ in (4.8) is justified.

From (4.8) we then get

$$\mathbb{E}(S_T) = \sum_{i=1}^{\infty} \mathbb{E}(X_i)\mathbb{P}(T \geq i) = \mathbb{E}(X_i) \sum_{i=1}^{\infty} \mathbb{P}(T \geq i) = \mathbb{E}(X_i)\mathbb{E}(T) \,,$$

which proves (a). (For an alternate proof of Wald's first equation using martingale techniques, see Corollary 6.2.3.)

(b) Our proof follows the outline of the proof presented in [12]. By (4.6), $\mathbb{E}(S_T) = 0$, and so $\mathrm{Var}(S_T) = \mathbb{E}(S_T^2)$. For each $i \geq 1$, define the new (thus bounded) stopping time

$$T \wedge i := \min\{T, i\}.$$

Since $\mathbb{P}(T < \infty) = 1$ by assumption, we have

$$\lim_{i \to \infty} S_{T \wedge i} = S_T \quad a.s. \tag{4.10}$$

In the following we will compute $\mathbb{E}(S_{T \wedge i}^2)$, and then argue that $\lim_{i \to \infty} \mathbb{E}(S_{T \wedge i}^2)$ yields (4.7). We can rewrite $S_{T \wedge i}^2$ as

$$S_{T \wedge i}^2 = S_{T \wedge (i-1)}^2 + (2X_i S_{i-1} + X_i^2)\mathbb{1}_{\{T \geq i\}} \,. \tag{4.11}$$

As explained above, the random variables $\mathbb{1}_{\{T \geq i\}}$ and X_i are independent. Hence the random variables $S_{i-1}\mathbb{1}_{\{T \geq i\}}$ and X_i are also independent. Taking expectations of both sides of (4.11) (recall that we assume $\mathbb{E}(X_i) = 0$, and hence $\mathbb{E}(X_i^2) = \sigma^2$) yields the recursion

$$\mathbb{E}(S_{T \wedge i}^2) = \mathbb{E}(S_{T \wedge (i-1)}^2) + \sigma^2 \mathbb{P}(T \geq i)$$

from which, by induction, we get

$$\mathbb{E}(S_{T \wedge i}^2) = \sigma^2 \sum_{k=1}^{i} \mathbb{P}(T \geq k) \,. \tag{4.12}$$

Note that for the sum on the right-hand side of (4.12), we have

$$\sum_{k=1}^{i} \mathbb{P}(T \geq k) = \mathbb{E}(T \wedge i)\,,$$

which establishes (4.7) for the finite random variables $T \wedge i$. We then compute the limit as $i \to \infty$ on both sides of (4.12). For the right-hand side of (4.12), clearly

$$\lim_{i\to\infty} \sigma^2 \sum_{k=1}^{i} \mathbb{P}(T \geq k) = \sigma^2 \sum_{k=1}^{\infty} \mathbb{P}(T \geq k) = \sigma^2 \mathbb{E}(T)\,. \tag{4.13}$$

As a last step, we need to prove that $\lim_{i\to\infty} \mathbb{E}(S_{T\wedge i}^2) = \mathbb{E}(S_T^2)$. Let $j < i$ and consider the random variable

$$S_{T\wedge i} - S_{T\wedge j}\,.$$

Similarly to the expression in (4.11), we can write

$$(S_{T\wedge i} - S_{T\wedge j})^2 = (S_{T\wedge(i-1)} - S_{T\wedge j})^2 + (2X_i(S_{i-1} - S_j) + X_i^2)\mathbb{1}_{T\geq i}$$

from which, by taking expectations on both sides, we get the recursion

$$\mathbb{E}[(S_{T\wedge i} - S_{T\wedge j})^2] = \mathbb{E}[(S_{T\wedge(i-1)} - S_{T\wedge j})^2] + \sigma^2 \mathbb{P}(T \geq i)\,,$$

and by induction,

$$\mathbb{E}[(S_{T\wedge i} - S_{T\wedge j})^2] = \sigma^2 \sum_{k=j+1}^{i} \mathbb{P}(T \geq i)\,. \tag{4.14}$$

Since $\mathbb{E}(T) < \infty$, the sequence of probabilities $\{\mathbb{P}(T \geq i)\}_{i\geq 0}$ is a Cauchy sequence. Therefore, by (4.14), the sequence of random variables $\{S_{T\wedge i}\}_{i\geq 0}$ is a Cauchy sequence with respect to the L^2-norm and (by completeness of the L^2-space) converges to a limit random variable Y in L^2 (recall Definition B.4.4). By (4.10), we also have a.s. convergence of $\{S_{T\wedge i}\}_{i\geq 0}$ to S_T, and so $Y = S_T$ with probability 1. Since

$$\lim_{i\to\infty} \sqrt{\mathbb{E}[(S_{T\wedge i} - S_T)^2]} = 0\,,$$

it follows from the reverse triangle inequality that

$$\lim_{i\to\infty} \sqrt{\mathbb{E}(S_{T\wedge i}^2)} = \sqrt{\mathbb{E}(S_T^2)}\,,$$

and, consequently, we get

$$\lim_{i\to\infty} \mathbb{E}(S_{T\wedge i}^2) = \mathbb{E}(S_T^2)\,. \tag{4.15}$$

Applying (4.13) and (4.15) to (4.12), we get (4.7). This completes the proof of Wald's second equation. □

We finish this section by quoting a theorem for (general) random walk $(S_n)_{n\geq 0}$ whose step distribution has finite mean. The result of the theorem may be useful for checking that in a given situation the conditions for Wald's equations are satisfied. We omit the proof (for a reference see [34]).

Theorem 4.3.2. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(|X_i|) < \infty$ and $(S_n)_{n\geq 0}$ random walk with $S_0 = 0$ and $S_n = \sum_{i=1}^{n} X_i$ for $n \geq 1$. Consider the first hitting time T of the half-line $[1, \infty)$, that is, $T = \min\{n : S_n \in [1, \infty)\}$. If $\mathbb{E}(X_i) \geq 0$, then*

$$\mathbb{P}(T < \infty) = 1,$$

and

(a) if $\mathbb{E}(X_i) = 0$, then $\mathbb{E}(T) = \infty$;

(b) if $\mathbb{E}(X_i) > 0$, then $\mathbb{E}(T) < \infty$.

4.4 Gambler's Ruin

In the classical gambler's ruin problem, a gambler starts with a fortune of x dollars and makes successive 1 dollar bets against the house. The game ends when either the gambler is ruined (his fortune is 0 dollars) or the gambler's fortune has reached N dollars. The probability of wining 1 dollar is p and the probability of losing 1 dollar is $1 - p$. If $p = \frac{1}{2}$, we call it a fair game, otherwise a biased game. The process that models the evolution of the gambler's fortune over time (until the time the game ends) is simple random walk on $\mathbb{Z}$. Figure 4.6 shows the transition graph for this process.

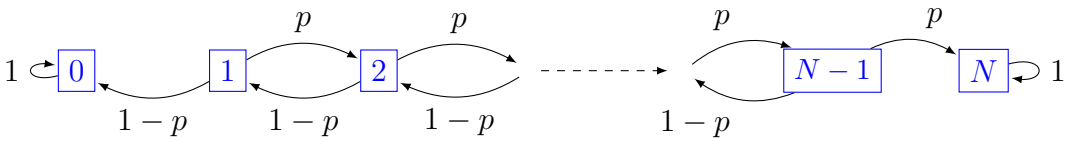


Figure 4.6: Transition graph for the gambler's ruin process

Questions of interest related to the gambler's ruin chain, and which we will address in this section, include:

- Will the game eventually end?
- What is the probability that the gambler will end up ruined?

- If the game does eventually end, what is the expected duration of the game?

Note that the gambler's ruin chain is a birth/death chain. Since it is also a random walk, Wald's equations will be useful for certain computations. Because of the obvious translation invariance of the problem, we will study questions about the gambler's ruin problem for starting state 0 and $a < 0 < b$, rather than for starting state x and $0 < x < N$.

FAIR GAME:

Let $(S_n)_{n\geq 0}$ be simple random walk on $\mathbb{Z}$ with $X_0 = 0$ and $S_n = \sum_{k=1}^n X_k$ where the random variables X_k are i.i.d with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}$. Let $a \in \mathbb{Z}^-$ and $b \in \mathbb{Z}^+$. States a and b are absorbing states and consitute the *boundary*, and all other states lead into $\{a, b\}$. By the general theory of finite-state absorbing chains (Section 2.3), the process will be absorbed in $\{a, b\}$ in finite time with probability 1, that is, $\mathbb{P}(T^{\{a,b\}} < \infty) = 1$, and furthermore $\mathbb{E}(T^{\{a,b\}}) < \infty$. We can also see this more directly: Without loss of generality assume $b \geq |a|$. Then from any starting point $x \in (a, b)$,

$$\mathbb{P}_x(T^{\{a,b\}} \leq (b-a)) \geq (\frac{1}{2})^{b-a}$$

since taking a direct path from x to either point of the boundary $\{a, b\}$ takes at most $b - a$ steps. So

$$\mathbb{P}_x(T^{\{a,b\}} > (b-a)) \leq 1 - (\frac{1}{2})^{b-a}$$

and

$$\mathbb{P}(T^{\{a,b\}} > n(b-a)) \leq (1 - (\frac{1}{2})^{b-a})^n \quad \text{for all } n \geq 1\,.$$

It follows that the series $\sum_{i=1}^\infty \mathbb{P}(T^{\{a,b\}} \geq i) = \mathbb{E}(T^{\{a,b\}})$ converges.

We apply Wald's first equation for the stopping time $T = T^{\{a,b\}}$ to compute the *ruin probability* $r = \mathbb{P}(T^a < T^b)$. This yields

$$\mathbb{E}(S_T) = 0 = b\,(1-r) + a\,r$$

from which we compute

$$r = \frac{b}{|a| + b}\,. \tag{4.16}$$

Next we apply Wald's second equation to compute $\mathbb{E}(T)$. Since $\mathbb{E}(X_k) = 0$ and $\mathrm{Var}(X_k) = 1$, Wald's second equation and (4.16) yield

$$\mathbb{E}(S_T^2) = a^2 \frac{b}{|a| + b} + b^2 \frac{|a|}{|a| + b} = \mathbb{E}(T)$$

from which we compute

$$\mathbb{E}(T) = |a|\, b\,. \tag{4.17}$$

We have proved the following result:

Proposition 4.4.1. *Consider simple symmetric random walk on $\mathbb{Z}$, and let $a < x < b$. Then*

$$\mathbb{P}_x(T^a < T^b) = \frac{b-x}{b-a} \quad \text{and} \quad \mathbb{P}_x(T^b < T^a) = \frac{x-a}{b-a}\,, \tag{4.18}$$

and

$$\mathbb{E}_x(T^{\{a,b\}}) = (x-a)(b-x). \tag{4.19}$$

Let $a < x$. We can use Proposition 4.4.1 to compute $\mathbb{P}_x(T^a < \infty)$ and the expected hitting time $\mathbb{E}_x(T^a)$ for simple symmetric random walk by passing to the limit $b \to \infty$ in (4.18) and (4.19). Indeed, since

$$\{T^a < T^b \,|\, S_0 = x\} \subseteq \{T^a < T^{b+1} \,|\, S_0 = x\} \quad \text{for all } b > x,$$

and

$$\{T^a < \infty \,|\, S_0 = x\} = \bigcup_{b=x+1}^{\infty} \{T^a < T^b \,|\, S_0 = x\},$$

by the continuity property of probability (recall Lemma B.1.1(a)), we have

$$\mathbb{P}_x(T^a < \infty) = \lim_{b\to\infty} \mathbb{P}_x(T^a < T^b)\,. \tag{4.20}$$

Also, under the same assumption $a < x < b$, the random variables $T^{\{a,b\}}$ are nondecreasing with respect to b, and

$$\lim_{b\to\infty} T^{\{a,b\}} = T^a \quad a.s.,$$

hence by the Monotone Convergence theorem (recall Theorem C.3.1)), we have

$$\mathbb{E}_x(T^a) = \lim_{b\to\infty} \mathbb{E}_x(T^{\{a,b\}})\,. \tag{4.21}$$

Example 4.4.1. Consider a compulsive gambler who plays against an infinitely rich adversary and who will quit the game only at his ruin. Assume the gambler starts with a fortune of x. From (4.18), we get

$$\begin{aligned} \mathbb{P}_x(T^0 < \infty) &= \lim_{b\to\infty} \mathbb{P}_x(T^0 < T^b) \\ &= \lim_{b\to\infty} \frac{b-x}{b} = 1\,. \end{aligned}$$

And from (4.19), we compute

$$\mathbb{E}_x(T^0) = \lim_{b\to\infty} x\,(b-x) = \infty\,.$$

So with probability 1, the game will end with bancruptcy for the gambler in finite time. But the expected time until the gambler's bancruptcy is infinite.
The distribution of T^0 can be computed using the results from Proposition 4.5.2 or from Corollary 4.5.9. □

Remark 4.4.2. *Simple random walk with zero holding probability is periodic with period 2. To avoid periodicity, one often adds positive holding probability h to each move. That is, for simple symmetric random walk, the distribution of the i.i.d. random variables X_k will be given by $\mathbb{P}(X_k = 0) = h$ and $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = p$ with $h, p > 0$ and $2p+h = 1$. Notice that this modification does not change the ruin probabilities r and $1-r$. However, since in this case $\mathrm{Var}(X_k) = 2p < 1$, the expected time $\mathbb{E}(T)$ until absorption becomes $\mathbb{E}(T) = \frac{1}{2p}|a|b$. It should be intuitive that adding positive holding probability to the walk will "slow things down" and, on average, the game will last longer (here by a factor $\frac{1}{2p}$). For example, if we take holding probability $h = \frac{1}{2}$, then on average the walk moves either left or right only half the time, and the expected time until absorption doubles.*

We now use the gambler's ruin probabilities (4.18) to answer a few additional questions about simple symmetric random walk. Let $a < x < b$. We assume that the walk starts in x. What is the probability that the walk will return to x *before absorption* in $\{a, b\}$? We denote this probability by $P_{xx}^{\{a,b\}}$. Conditioning on the first step of the random walk, we get

$$P_{xx}^{\{a,b\}} = \frac{1}{2}\frac{x-1-a}{x-a} + \frac{1}{2}\frac{b-x-1}{b-x}$$

which simplifies to

$$P_{xx}^{\{a,b\}} = 1 - \frac{1}{2}\frac{b-a}{(x-a)(b-x)}\,.$$

Let $V_{xx}^{\{a,b\}}$ be the random variable "number of returns to x before absorption in $\{a, b\}$". By the strong Markov property,

$$\mathbb{P}(V_{xx}^{\{a,b\}} = n) = \left(P_{xx}^{\{a,b\}}\right)^n (1 - P_{xx}^{\{a,b\}}) \quad \text{for } n \geq 0\,,$$

which shows that the random variable $V_{xx}^{\{a,b\}}$ has a geometric distribution with parameter $(1 - P_{xx}^{\{a,b\}})$. The expected number of returns to x before absorption in the boundary $\{a, b\}$ is therefore

$$\mathbb{E}(V_{xx}^{\{a,b\}}) = \frac{P_{xx}^{\{a,b\}}}{1 - P_{xx}^{\{a,b\}}}\,.$$

We can ask similar questions for visits to other states $y \neq a, b$ before absorption. Again, assume the walk starts in state x. Assume $a < y < b$ and, without loss of generality, assume $x < y$. Let $P_{xy}^{\{a,b\}}$ denote the probability that the walk visits state y before absorption. Then, from (4.18), we get

$$P_{xy}^{\{a,b\}} = \frac{x-a}{y-a}\,.$$

Let $V_{xy}^{\{a,b\}}$ be the random variable "number of visits to y before absorption in $\{a,b\}$". By the strong Markov property,

$$\mathbb{P}(V_{xy}^{\{a,b\}} = n) = \begin{cases} 1 - P_{xy}^{\{a,b\}} & \text{for } n = 0 \\ P_{xy}^{\{a,b\}} \left(P_{yy}^{\{a,b\}}\right)^{n-1} (1 - P_{yy}^{\{a,b\}}) & \text{for } n \geq 1\,. \end{cases}$$

From this we compute the expectation of the random variable $V_{xy}^{\{a,b\}}$ as

$$\begin{aligned} \mathbb{E}(V_{xy}^{\{a,b\}}) &= P_{xy}^{\{a,b\}} \sum_{n=1}^{\infty} n \left(P_{yy}^{\{a,b\}}\right)^{n-1} (1 - P_{yy}^{\{a,b\}}) \\ &= \frac{P_{xy}^{\{a,b\}}}{1 - P_{yy}^{\{a,b\}}}\,. \end{aligned}$$

□

BIASED GAME:

For a biased game we have $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_i = -1) = 1 - p$ with $p \neq \frac{1}{2}$. We assume $0 < p < 1$ and set $q = 1 - p$. As before, let $a < 0 < b$. Since this is a finite-state absorbing chain, here also we have $\mathbb{P}(T^{\{a,b\}} < \infty) = 1$ and $\mathbb{E}(T^{\{a,b\}}) < \infty$. Wald's identities aren't useful in this case towards computing the ruin probability $\mathbb{P}_x(T^a < T^b)$, but we will take a different approach which will involve solving a system of equations. We use the following notation: For $a < x < b$, let $r_x = \mathbb{P}_x(T^a < T^b)$ be the probability that the random walk gets absorbed in boundary state a, given that the walk starts in state x. The idea is to set up a recurrence relation for the r_x by conditioning on the outcome of the first step of the random walk: From its starting state x, the walk either moves to $x+1$ and eventually gets absorbed in state a with probability r_{x+1}, or it moves to $x-1$ and eventually gets absorbed in state a with probability r_{x-1}. This results in the system of equations

$$r_x = p\, r_{x+1} + q\, r_{x-1} \quad \text{for } a < x < b \tag{4.22}$$

with boundary conditions $r_a = 1$ and $r_b = 0$. We then need to solve this system. Note that we can rewrite (4.22) as

$$(p+q)r_x = p\, r_{x+1} + q\, r_{x-1}$$

and so get

$$q(r_{x-1} - r_x) = p(r_x - r_{x+1})\,,$$
$$r_x - r_{x+1} = \frac{q}{p}(r_{x-1} - r_x)\,,$$

and by iteration,

$$r_x - r_{x+1} = (\frac{q}{p})^{x-a}(r_a - r_{a+1})\,.$$

Setting $c = r_a - r_{a+1} = 1 - r_{a+1}$, we have

$$r_x - r_{x+1} = c\,(\frac{q}{p})^{x-a} \qquad \text{for } a \le x < b\,.$$

So

$$r_x = (r_x - r_{x+1}) + (r_{x+1} - r_{x+2}) + \cdots + (r_{b-1} - r_b) = c\sum_{i=x}^{b-1}(\frac{q}{p})^{i-a}\,,$$

from which we get for $a \le x < b$,

$$\begin{aligned} r_x = c\sum_{i=x}^{b-1}(\frac{q}{p})^{i-a} &= c\,(\frac{q}{p})^{x-a}\sum_{j=0}^{b-x-1}(\frac{q}{p})^{j} \\ &= c\,(\frac{q}{p})^{x-a}\frac{1-(\frac{q}{p})^{b-x}}{1-\frac{q}{p}}\,. \end{aligned}$$

Setting $r_a = 1$, we compute the constant c as

$$c = \frac{1-\frac{q}{p}}{1-(\frac{q}{p})^{b-a}}\,.$$

Altogether, we have proved the following result:

Proposition 4.4.3. *Let $0 < p < 1$, $p \neq \frac{1}{2}$, and $q = 1-p$. For simple biased random walk with $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_k = -1) = q$, we have*

$$\mathbb{P}_x(T^a < T^b) = r_x = \frac{(\frac{q}{p})^{x-a} - (\frac{q}{p})^{b-a}}{1-(\frac{q}{p})^{b-a}} \qquad \text{for } a \le x < b\,, \tag{4.23}$$

and

$$\mathbb{P}_x(T^b < T^a) = 1 - r_x = \frac{1-(\frac{q}{p})^{x-a}}{1-(\frac{q}{p})^{b-a}} \qquad \text{for } a \le x < b\,. \tag{4.24}$$

Notes: (1) In Section 6.6.2, we will compute the absorption probabilities (4.23) and (4.24) in an alternate way using martingale techniques.

(2) Formulas (4.23) and (4.24) for the absorption probabilities remain the same, if the biased random walk has positive holding probability h. That is, if $\mathbb{P}(X_k = 0) = h$, $\mathbb{P}(X_k = 1) = p$, and $\mathbb{P}(X_k = -1) = q$ with $h, p, q > 0$, $p \neq q$, and $h + p + q = 1$. This can be seen by modifying the system of equations (4.22) accordingly and by solving the system. Intuitively, absorption probabilities only depend on the evolution of the trajectories as far as their changes in locations over time are concerned. However, a positive holding probability will have an effect on the expected duration until absorption (see Proposition 4.4.4 below). The same was true for symmetric random walk (see Remark 4.4.2).

Now that we have computed the absorption probabilities r_x and $1 - r_x$, we can use Wald's first equation to compute the expected time until absorption $\mathbb{E}_x(T)$. We have $\mathbb{E}(X_k) = p - q$. And we compute

$$\begin{aligned} \mathbb{E}_x(S_T) &= a(1 - \frac{1 - (\frac{q}{p})^{x-a}}{1 - (\frac{q}{p})^{b-a}}) + b\frac{1 - (\frac{q}{p})^{x-a}}{1 - (\frac{q}{p})^{b-a}} \\ &= a + (b-a)\frac{1 - (\frac{q}{p})^{x-a}}{1 - (\frac{q}{p})^{b-a}} \,. \end{aligned}$$

Thus, by Theorem 4.3.1(a), we get the following result for simple biased random walk (with or without positive holding probability):

Proposition 4.4.4. *Consider simple biased random walk with $\mathbb{P}(X_k = 0) = h$, $\mathbb{P}(X_k = 1) = p$, and $\mathbb{P}(X_k = -1) = q$. Assuming $h \geq 0$, $p, q > 0$, $p \neq q$, and $h + p + q = 1$, we have*

$$\mathbb{E}_x(T) = \frac{a - x}{p - q} + \left(\frac{b - a}{p - q}\right) \frac{1 - (\frac{q}{p})^{x-a}}{1 - (\frac{q}{p})^{b-a}} \,. \tag{4.25}$$

Example 4.4.2. Again, we consider the case of the compulsive gambler who plays against an infinitely rich adversary. Here the individual bets are assumed to be not fair. The limits stated in (4.20) and (4.21) are also valid for biased random walk, and so (4.23) yields

$$\mathbb{P}_x(T^0 < \infty) = \lim_{b \to \infty} \frac{(\frac{q}{p})^x - (\frac{q}{p})^b}{1 - (\frac{q}{p})^b} = \begin{cases} 1 & \text{for } q > p \\ (\frac{q}{p})^x < 1 & \text{for } q < p \,. \end{cases}$$

For the case $q > p$, for which bancruptcy of the gambler will occur with probability 1, we

get from (4.25) for the expected time until bancruptcy,

$$\mathbb{E}_x(T^0) = \lim_{b\to\infty}\left[\frac{-x}{p-q} + \left(\frac{b}{p-q}\right)\frac{1-(\frac{q}{p})^x}{1-(\frac{q}{p})^b}\right] = \frac{x}{q-p}\,.$$

The distribution of T^0 can be computed using the result from Corollary 4.5.9. □

4.5 Reflection Principle and Duality

In order to compute probabilities for simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$, we often need to count paths of a fixed length and with specific properties. Towards this end, it is often helpful to apply to paths simple geometric operations, such as reflection, rotation, or cutting and pasting, which can establish a one-to-one correspondence between the paths in a set of interest and the paths in a set that is easier to count or understand. In this spirit, two approaches have proven to be particularly useful. They are **reflection and time-reversal of paths**. First, we will first discuss the *Reflection Principle*. It is based on a one-to-one correspondence of paths via the reflection of certain portions of a path across a horizontal line.

4.5.1 The Reflection Principle

Let $a \in \mathbb{Z}$. Recall that T^a denotes the first hitting time of the random walk, that is, $T^a = \min\{n \geq 1 : S_n = a\}$.

Lemma 4.5.1 (Reflection Principle)**.** *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk on $\mathbb{Z}$.*
(a) **Reflecting the portion of the trajectory between times T^a and n:** *Assume $S_0 = 0$ and $a, k \geq 1$. Then*

$$\mathbb{P}(S_n = a + k) = \mathbb{P}(S_n = a - k,\ T^a \leq n)\,.$$

(b) **Reflecting the portion of the trajectory between times 0 and T^0:** *Assume $c, b > 0$. Then*

$$\mathbb{P}(S_n = c \,|\, S_0 = -b) = \mathbb{P}(S_n = c,\ T^0 \leq n \,|\, S_0 = b)\,.$$

Proof. (a) Fix $n \geq 1$. For simple symmetric random walk, any n-length path ω_n occurs with the same probability $\frac{1}{2^n}$. Any $\omega_n \in \{S_n = a + k\}$ must have visited a at a time prior to n. Consider $T^a = T^a(\omega_n)$ and reflect the portion of ω_n between times T^a and n

about the horizontal line $y = a$ (see Figure 4.7). In this way, we establish a one-to-one correspondence between n-length paths ending in $a + k$ and paths that first visit a at a time prior to n and end up in $a - k$.

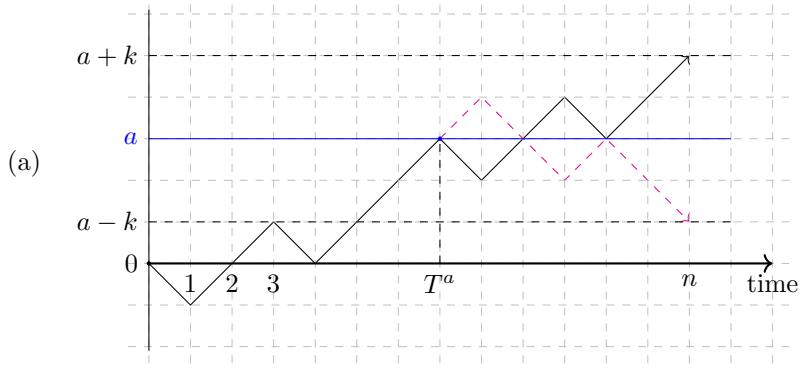


Figure 4.7

(b) An analogous one-to-one correspondence of n-length paths, using reflection of the portion of ω_n between times 0 and T^0 about the horizontal line $y = 0$, establishes the second part of the lemma (see Figure 4.8). □

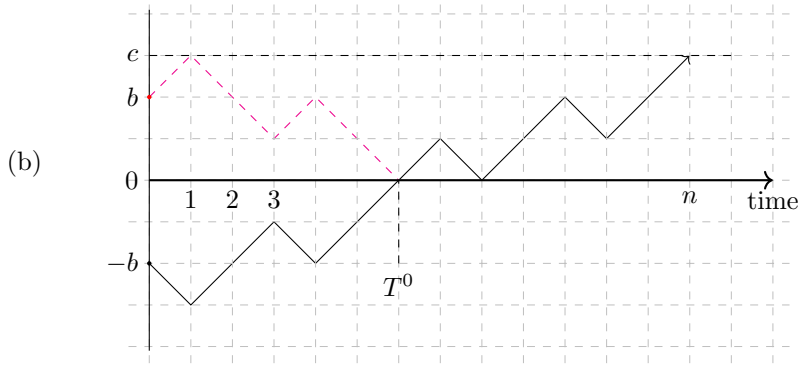


Figure 4.8

As a first application of the Reflection Principle, we compute the distribution of the first hitting time T^a.

Proposition 4.5.2. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0 and $a \in \mathbb{Z}^+$. Then*

$$\begin{aligned}\mathbb{P}(T^a \leq n) &= 2\mathbb{P}(S_n > a) + \mathbb{P}(S_n = a) \\ &= \mathbb{P}(S_n \notin [-a,\, a-1])\,.\end{aligned}$$

By symmetry, we have $\mathbb{P}(T^a \leq n) = \mathbb{P}(T^{-a} \leq n)$.

Proof. We have

$$\begin{aligned}\mathbb{P}(T^a \leq n) &= \sum_{b\in\mathbb{Z}} \mathbb{P}(S_n = b,\, T^a \leq n) \\ &= \sum_{b\geq a} \mathbb{P}(S_n = b) + \sum_{b<a} \mathbb{P}(S_n = b,\, T^a \leq n) \\ &= \mathbb{P}(S_n \geq a) + \mathbb{P}(S_n > a) \\ &= \mathbb{P}(S_n \notin [-a, a-1])\end{aligned}$$

where the second term in the third equality is a consequence of the Reflection Principle. □

Corollary 4.5.3. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0. For $n \geq 1$, we have*

(a) $\mathbb{P}(T^0 > 2n) = P_{00}^{2n}$

(b) $\mathbb{P}(S_1 \neq 0, S_2 \neq 0, ..., S_{2n} \neq 0) = P_{00}^{2n}$

(c) $\mathbb{P}(S_1 > 0, S_2 > 0, ..., S_{2n} > 0) = \frac{1}{2} P_{00}^{2n}$

(d) $\mathbb{P}(S_1 \geq 0, S_2 \geq 0, ..., S_{2n} \geq 0) = P_{00}^{2n}\,.$

By symmetry, equalities (c) and (d) also hold if we reverse all inequality signs in the probabilities on the left-hand side.

Proof. (a)-(c): We have

$$\begin{aligned}
\mathbb{P}(T^0 > 2n) &= \mathbb{P}(S_1 \neq 0, S_2 \neq 0, ..., S_{2n} \neq 0) \\
&= 2\mathbb{P}(S_1 > 0, S_2 > 0, ..., S_{2n} > 0) \\
&= 2\mathbb{P}(S_2 > 0, ..., S_{2n} > 0 \,|\, S_1 = 1)\tfrac{1}{2} \\
&= \mathbb{P}(S_1 > -1, ..., S_{2n-1} > -1 \,|\, S_0 = 0) \\
&= \mathbb{P}(S_1 < 1, ..., S_{2n-1} < 1 \,|\, S_0 = 0) \\
&= \mathbb{P}(T^1 > 2n-1) = \mathbb{P}(S_{2n-1} \in \{-1, 0\}) = \mathbb{P}(S_{2n-1} = -1)\,.
\end{aligned}$$

Note that $\mathbb{P}(S_{2n-1} = -1) = \binom{2n-1}{n}\frac{1}{2^{2n-1}} = \binom{2n}{n}\frac{1}{2^n} = \mathbb{P}(S_{2n} = 0)$, which is due to the identity $2\binom{2n-1}{n} = \binom{2n}{n}$. Thus

$$\mathbb{P}(T^0 > 2n) = \mathbb{P}(S_1 \neq 0, S_2 \neq 0, ..., S_{2n} \neq 0) = 2\mathbb{P}(S_1 > 0, S_2 > 0, ..., S_{2n} > 0) = P_{00}^{2n}\,.$$

To prove (d), note that

$$\begin{aligned}
\mathbb{P}(S_1 > 0, S_2 > 0, ..., S_{2n} > 0) &= \mathbb{P}(S_1 = 1)\,\mathbb{P}(S_2 > 0, ..., S_{2n} > 0 \,|\, S_1 = 1) \\
&= \tfrac{1}{2}\mathbb{P}(S_1 \geq 0, ..., S_{2n-1} \geq 0)\,.
\end{aligned}$$

Since $2n-1$ is odd, we have $\mathbb{P}(S_{2n} \geq 0 \,|\, S_{2n-1} \geq 0) = 1$, and thus

$$\mathbb{P}(S_1 \geq 0, ..., S_{2n-1} \geq 0) = \mathbb{P}(S_1 \geq 0, ..., S_{2n} \geq 0)\,.$$

It follows that

$$\mathbb{P}(S_1 \geq 0, ..., S_{2n} \geq 0) = 2\mathbb{P}(S_1 > 0, S_2 > 0, ..., S_{2n} > 0) = P_{00}^{2n}\,,$$

which completes the proof. □

Corollary 4.5.4. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0. Then for $n \geq 1$,*

$$\begin{aligned}
\mathbb{P}(T^0 = 2n) &= P_{00}^{2n-2} - P_{00}^{2n} \\
&= \frac{1}{2n-1}P_{00}^{2n} \\
&= \frac{1}{2n}P_{00}^{2n-2}\,.
\end{aligned}$$

Proof. Note that the event of first return to 0 at time $2n$ is

$$\{S_2 \neq 0, ..., S_{2n-2} \neq 0, S_{2n} = 0\} = \{S_2 \neq 0, ..., S_{2n-2} \neq 0\} \backslash \{S_2 \neq 0, ..., S_{2n-2} \neq 0, S_{2n} \neq 0\}\,.$$

Since $\{S_2 \neq 0, ..., S_{2n-2} \neq 0, S_{2n} \neq 0\} \subset \{S_2 \neq 0, ..., S_{2n-2} \neq 0\}$, by Corollary 4.5.3,

$$\mathbb{P}(T^0 = 2n) = P_{00}^{2n-2} - P_{00}^{2n}\,.$$

A straightforward calculation yields

$$P_{00}^{2n-2} = \binom{2n-2}{n-1}\frac{1}{2^{2n-2}} = \frac{4n^2}{(2n-1)2n}\binom{2n}{n}\frac{1}{2^{2n}} = \frac{2n}{2n-1}P_{00}^{2n}\,,$$

and so

$$\mathbb{P}(T^0 = 2n) = P_{00}^{2n-2} - P_{00}^{2n} = (\frac{2n}{2n-1} - 1)P_{00}^{2n} = \frac{1}{2n-1}P_{00}^{2n}\,.$$

Equivalently, since

$$P_{00}^{2n} = \frac{2n-1}{2n}P_{00}^{2n-2}\,,$$

we get

$$\mathbb{P}(T^0 = 2n) = \frac{1}{2n}P_{00}^{2n-2}\,.$$

□

Corollary 4.5.5. *Simple symmetric random walk on $\mathbb{Z}$ is null recurrent.*

Proof. The probability that the random walk returns to 0 in finite time is

$$\mathbb{P}(T^0 < \infty) = \sum_{n=1}^{\infty}\mathbb{P}(T^0 = 2n) = \sum_{n=1}^{\infty}(P_{00}^{2n-2} - P_{00}^{2n}) = \lim_{n\to\infty}(1 - P_{00}^{2n}) = 1\,,$$

and so simple symmetric random walk on $\mathbb{Z}$ is recurrent. For the expectation $\mathbb{E}(T^0)$, we have

$$\mathbb{E}(T^0) = \sum_{m\geq 1}^{\infty}\mathbb{P}(T^0 \geq m) = 2 + 2\sum_{n=1}^{\infty}P_{00}^{2n} = \infty\,.$$

The last equality follows since $\sum_{n=1}^{\infty}P_{00}^{2n}$ is the expected number of returns to 0, which is infinite by the recurrence of the random walk (see Proposition 2.1.3). It follows that simple symmetric random walk on $\mathbb{Z}$ is *null* recurrent. □

4.5.2 The ballot problem

Consider the following question which is known as the *ballot problem*: In an election between two candidates, Candidate 1 receives c votes and Candidate 2 receives d votes, with $c > d$. What is the probability that throughout the election, Candidate 1 was always ahead of Candidate 2? The answer to this question is stated in Corollary 4.5.7 below.

Consider simple (symmetric or biased) random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$. We will use the following notation:
$N_n(a,b)$ = "number of paths of length n that start at a and end at b"
$\widetilde{N}_n(a,b)$ = "number of paths of length n that start at a and end at b for which $S_k \neq 0$ for $0 < k < n$"

Proposition 4.5.6. *Consider simple random walk on $\mathbb{Z}$. Let $b \in \mathbb{Z}^+$. We have*

$$\widetilde{N}_n(0,b) = \frac{b}{n} N_n(0,b)\,.$$

Proof. Any such path must take its first step to 1. So

$$\widetilde{N}_n(0,b) = \widetilde{N}_{n-1}(1,b)\,.$$

By the Reflection Principle Part (b), the number of paths of length $n-1$ that start at 1 and end at b and for which $S_k = 0$ for some $0 < k < n$ is equal to $N_{n-1}(-1,b)$. Thus

$$\widetilde{N}_{n-1}(1,b) = N_{n-1}(1,b) - N_{n-1}(-1,b).$$

We have

$$\begin{aligned} N_{n-1}(1,b) - N_{n-1}(-1,b) &= \binom{n-1}{\frac{n+b-2}{2}} - \binom{n-1}{\frac{n+b}{2}} \\ &= \binom{n-1}{k-1} - \binom{n-1}{k} \end{aligned}$$

where we have set $k = \frac{n+b}{2}$. We compute

$$\begin{aligned} \binom{n-1}{k-1} - \binom{n-1}{k} &= \frac{[k(n-1)\cdots(n-k+1)] - [(n-1)\cdots(n-k)]}{k!} \\ &= \frac{(2k-n)(n-1)\cdots(n-k+1)}{k!} = \frac{c}{n}\frac{n(n-1)\cdots(n-k+1)}{k!} \\ &= \frac{b}{n}\binom{n}{k} = \frac{b}{n}\binom{n}{\frac{n+b}{2}} = \frac{b}{n} N_n(0,b)\,. \end{aligned}$$

□

Corollary 4.5.7 (Ballot problem)**.** *Let $c > d > 0$. Consider an election between two candidates, Candidate 1 and Candidate 2, in which Candidate 1 receives c votes and Candidate 2 receives d votes. The probability that, throughout the election, Candidate 1 was always ahead of Candidate 2 is*

$$\frac{c-d}{c+d}\,. \tag{4.26}$$

Proof. A number of $c+d$ votes were cast. The election process can be modeled as a simple random walk of length $c+d$ for which we set $X_k = 1$ if Candidate 1 receives the kth vote that was cast, and $X_k = -1$ if Candidate 2 receives the kth vote that was cast. We have

$$S_{c+d} = c - d > 0\,.$$

The event "Candidate 1 is ahead of Candidate 2 throughout the election" is the same as the event $\{S_k > 0 \,:\, 0 < k < c+d\}$. It follows that the probability that Candidate 1 was always ahead of Candidate 2 is

$$\frac{\widetilde{N}_{c+d}(0, c-d)}{N_{c+d}(0, c-d)} \;=\; \frac{c-d}{c+d}\,.$$

□

For an alternate proof of (4.26) that uses martingale theory, see Exercise 6.11.
We have a second immediate corollary to Proposition 4.5.6:

Corollary 4.5.8. *Let $(S_n)_{n\geq 0}$ be simple (symmetric or biased) random walk starting at 0. Let $b \in \mathbb{Z}, b \neq 0$. Then*

$$\mathbb{P}(S_1 \neq 0, ..., S_{n-1} \neq 0, S_n = b) = \frac{|b|}{n}\,\mathbb{P}(S_n = b)\,.$$

4.5.3 Dual walks

Another useful concept for computations is the concept of *duality* which results from *time reversal* of paths. Let $(S_n)_{n\geq 0}$ be simple (symmetric or biased) random walk starting at 0, and recall that $S_n = \sum_{k=1}^{n} X_k$ for $n \geq 1$. Since the random variables $X_1, X_2, ...$ are i.i.d., the distributions of the random vectors $(X_1, ..., X_n)$ and $(X_n, ..., X_1)$ are the same. We will use this fact to construct a new walk. We define

$$X_1^* = X_n,\; X_2^* = X_{n-1},\; ...,\; X_n^* = X_1\,.$$

From this we define the **dual walk** S_n^* of length n by

$$S_k^* = \sum_{i=1}^{k} X_i^*$$

for $k = 1, ..., n$ and $S_0^* = 0$. Figure 4.9 shows an n-length sample path ω_n for $(S_k)_{k\geq 0}$ and its corresponding dual path ω_n^* for $(S_k^*)_{0\leq k\leq n}$.

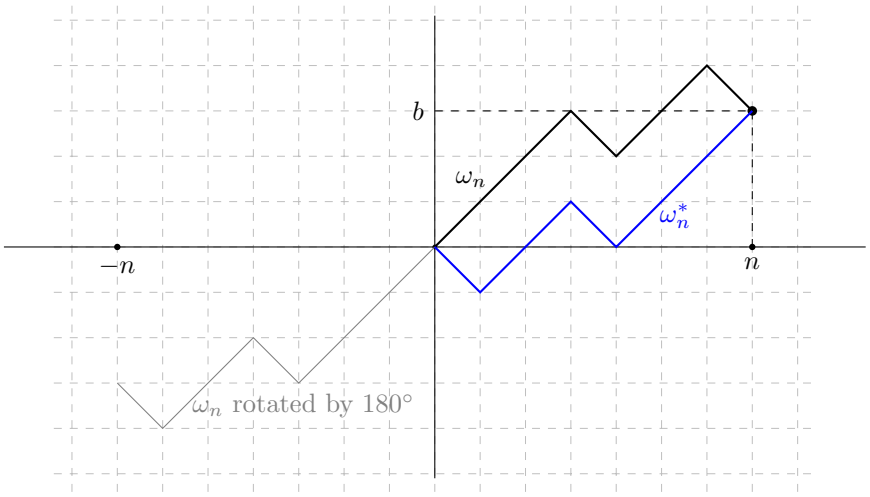


Figure 4.9

Figure 4.10 (with the same sample path ω_n as in Figure 4.9) is an illustration of the fact

$$(\omega_n^*)^* = \omega_n.$$

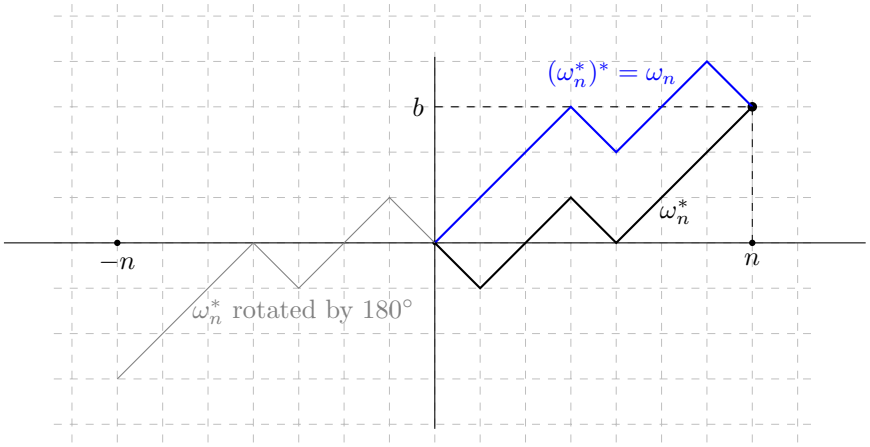


Figure 4.10

The paths ω_n and ω_n^* start and end at the same points. We get ω_n^* by letting time "run backwards". Geometrically, this is accomplished by rotating ω_n by 180° about the origin (which accomplishes the time reversal) and then translating the resulting path so that its starting point coincides with the origin.
Since $(\omega_n^*)^* = \omega_n$, there is a one-to-one correspondence between n-length paths ω_n for $(S_k)_{0\leq k\leq n}$ and n-length paths ω_n^* for $(S_k^*)_{0\leq k\leq n}$. We also have $\mathbb{P}(\omega_n) = \mathbb{P}(\omega_n^*)$, no matter whether the random walk is symmetric or biased. Hence any event E for $(S_k)_{0\leq k\leq n}$ has a corresponding dual event E^* for $(S_k^*)_{0\leq k\leq n}$ and $\mathbb{P}(E) = \mathbb{P}(E^*)$. Applying duality, we get the following corollary to Corollary 4.5.8.

Corollary 4.5.9. *Let $(S_n)_{n\geq 0}$ be simple (symmetric or biased) random walk starting at 0. Let $b \in \mathbb{Z}, b \neq 0$. Then*

$$\mathbb{P}(T^b = n) = \frac{|b|}{n}\,\mathbb{P}(S_n = b)\,. \tag{4.27}$$

Proof. Let ω_n be an n-length path of the event $\{S_1 \neq 0, ..., S_{n-1} \neq 0, S_n = b\}$. Its dual path ω_n^* is an element of the event $\{S_1^* \neq b, ..., S_{n-1}^* \neq b, S_n^* = b\}$. There is a one-to-one correspondence between such paths ω_n and ω_n^*. Since $\mathbb{P}(\omega_n) = \mathbb{P}(\omega_n^*)$, we have

$$\mathbb{P}(S_1 \neq 0, ..., S_{n-1} \neq 0, S_n = b) = \mathbb{P}(S_1^* \neq b, ..., S_{n-1}^* \neq b, S_n^* = b)\,.$$

But $\mathbb{P}(S_1^* \neq b, ..., S_{n-1}^* \neq b, S_n^* = b) = \mathbb{P}(T^b = n)$, and so by Corollary 4.5.8, we get

$$\mathbb{P}(T^b = n) = \frac{|b|}{n}\,\mathbb{P}(S_n = b)\,.$$

□

As an interesting consequence for the case of simple symmetric random walk, we get the following result for the expected number of visits to a state $b \neq 0$, before the walk returns to its starting point 0 for the first time.

Proposition 4.5.10. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0. Let $b \in \mathbb{Z}, b \neq 0$. Consider the random variable $V_{T^0}^b$ = "number of visits to state b before the first return to 0". Then*

$$\mathbb{E}(V_{T^0}^b) = 1\,.$$

Proof. Because of symmetry, it suffices to consider the case $b > 0$. Consider the events $A_n(b) = \{S_1 > 0, ..., S_{n-1} > 0, S_n = b\}$. The random variable $V_{T^0}^b$ can be written as

$$V_{T^0}^b = \sum_{n=1}^{\infty} \mathbb{1}_{A_n(b)} \,.$$

By the Monotone Convergence theorem (Theorem C.3.1), we have

$$\mathbb{E}(V_{T^0}^b) = \sum_{n=1}^{\infty} \mathbb{E}(\mathbb{1}_{A_n(b)}) \,.$$

Since $\mathbb{E}(\mathbb{1}_{A_n(b)}) = \mathbb{P}(A_n(b)) = \mathbb{P}(T^b = n)$, we get

$$\begin{aligned} \mathbb{E}(V_{T^0}^b) &= \sum_{n=1}^{\infty} \mathbb{P}(T^b = n) \\ &= \mathbb{P}(T^b < \infty) = 1 \end{aligned}$$

since simple symmetric random walk on $\mathbb{Z}$ is recurrent. □

Remark 4.5.11. *By Corollary 4.5.5 (and from earlier results), we know that simple symmetric random walk on $\mathbb{Z}$ is recurrent. By Theorems 2.2.4 and 2.2.6, we know that there exists, up to a multiplicative constant, a unique invariant measure μ for simple symmetric random walk on $\mathbb{Z}$. This invariant measure μ is given by $\mu(b) = \mathbb{E}(V_{T^0}^b)$ for $b \in \mathbb{Z}$. Clearly, the constant measure $\lambda = c$ (for any positive constant c) is an invariant measure for simple symmetric random walk, hence the only invariant measure, up to a multiplicative constant. It follows that there exists a constant $c_0 > 0$ such that*

$$\mathbb{E}(V_{T^0}^b) = c_0 \qquad \text{for all } b \in \mathbb{Z} \,.$$

Proposition 4.5.10 tells us that $c_0 = 1$.

4.5.4 Maximum and minimum

Often times, one is interested in the extreme values a random walk attains, either by a fixed time n or in the long run.

Maximum (or minimum) attained by time n:

Here we compute the distribution of the maximum value that simple symmetric random walk attains by time n. An analogous result for the minimum can be deduced by symmetry. We define the random variable

$$M_n = \max\{S_k \,:\, 1 \le k \le n\}\,.$$

Clearly, $M_n \ge S_n$. Assuming the random walk starts at 0, we have $M_n \ge 0$.

Proposition 4.5.12. *Let $(S_n)_{n\ge 0}$ be simple symmetric random walk starting at 0 and M_n the maximum the walk attains by time n. Then*

$$\mathbb{P}(M_n \ge a) = \mathbb{P}(S_n \notin [-a,\, a-1])\,,$$

and

$$\mathbb{P}(M_n = a) = \max(\mathbb{P}(S_n = a),\, \mathbb{P}(S_n = a+1))\,.$$

Proof. Because of the equality of the events

$$\{M_n \ge a\} = \{T^a \le n\}\,,$$

we have by Proposition 4.5.2,

$$\mathbb{P}(M_n \ge a) = \mathbb{P}(S_n \notin [-a,\, a-1])\,.$$

Note that $\mathbb{P}(M_n = a) = \mathbb{P}(M_n \ge a) - \mathbb{P}(M_n \ge a+1)$. Thus we get

$$\mathbb{P}(M_n = a) = \mathbb{P}(S_n = -a-1) + \mathbb{P}(S_n = a) = \max(\mathbb{P}(S_n = a+1),\, \mathbb{P}(S_n = a))\,.$$

The conclusion follows from symmetry and from the fact that either $\mathbb{P}(S_n = a+1) = 0$ or $\mathbb{P}(S_n = a) = 0$. □

The distribution of M_n for simple biased random walk involves more complicated expressions. For a reference see [17].

Maximum (or minimum) attained in the long run:

Recall that simple symmetric random walk on $\mathbb{Z}$ is recurrent, and therefore the random walk visits every state in $\mathbb{Z}$ infinitely often with probability 1. There is no maximum or minimum in the long run. For this reason we only consider biased simple random walk. Assume $q > p$. The random walk is transient and has negative drift. By the Strong Law of Large Numbers, we have

$$\lim_{n\to\infty} S_n = -\infty \qquad \text{with probability } 1\,,$$

and so there is no global minimum. Let

$$M = \max\{S_n \,:\, n \geq 0\}$$

be the *maximum random variable.*

Proposition 4.5.13. *Let $(S_n)_{n\geq 0}$ be simple, biased random walk starting at 0 with $q > p$ and let M be the* **maximum random variable**. *Then M has a geometric distribution with parameter $(1 - \frac{p}{q})$, that is,*

$$\mathbb{P}(M = k) = (1 - \frac{p}{q})(\frac{p}{q})^k \quad \text{for } k \geq 0\,, \tag{4.28}$$

and the expected long-run maximum is

$$\mathbb{E}(M) = \frac{p}{q-p}\,.$$

Proof. We have $\mathbb{P}(M \geq 0) = 1$. For $k \geq 1$,

$$\mathbb{P}(M \geq k) = \mathbb{P}(T^k < \infty)\,.$$

Recall the gambler's ruin formula (4.5.9) for $p \neq q$. For $k \geq 1$, we use the formula to compute

$$\mathbb{P}(T^k < \infty) = \lim_{a\to -\infty} \mathbb{P}(T^k < T^a) = \lim_{a\to -\infty} \frac{1 - (\frac{q}{p})^{-a}}{1 - (\frac{q}{p})^{k-a}} = (\frac{p}{q})^k\,,$$

which yields

$$\mathbb{P}(M = k) = (\frac{p}{q})^k - (\frac{p}{q})^{k+1} = (1 - \frac{p}{q})(\frac{p}{q})^k\,.$$

For $k = 0$, we have

$$\mathbb{P}(M = 0) = \mathbb{P}(S_n \leq 0 \text{ for all } n \geq 0) = 1 - \mathbb{P}(M \geq 1) = 1 - \frac{p}{q} > 0$$

This establishes (4.28) and consequently,

$$\mathbb{E}(M) = \frac{p}{q-p}\,.$$

□

4.6 Arcsine Law

4.6.1 Last returns to Zero

Consider simple symmetric random walk $(S_n)_{n\geq 0}$ starting at 0. While in the previous section we were concerned with first hitting times of states, we now ask the question of what is the *last return time* to state 0 up to a fixed time $2n$. In a fair game of coin tossing of fixed length $2n$, this will be the last time the two players have equal fortunes before one of the two players takes the lead and remains in the lead until the end of the game. We introduce the random variable

$$Y_{2n} = \max\{2k \leq 2n \,:\, S_{2k} = 0\},$$

that is, the last hitting time of state 0 up to (including) time $2n$. See Figure 4.11 for an illustration.

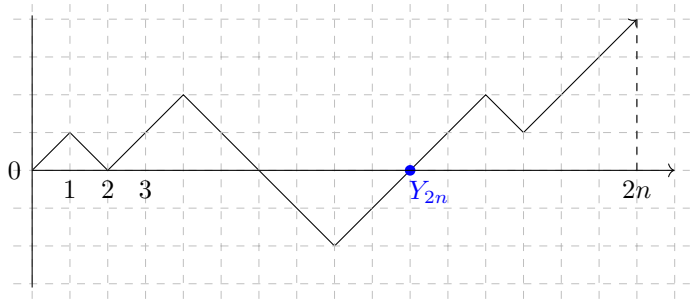


Figure 4.11

Figure 4.12 shows a histogram of $40,000$ repetitions of a simulation of the time of last visit to 0 for simple symmetric random walk of length $2n = 1000$. (The simulation was produced with the statistical software package R.) It is noteworthy that the distribution of the random variable Y_{2n} (in the above simulation, Y_{1000}) seems to be symmetric about the midpoint time n and has rather large spikes near the endpoints (near time 0 and time $2n$) of the time interval. The following proposition gives a precise formula for the distribution of Y_{2n}.

Proposition 4.6.1. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0. Then for all $0 \leq k \leq n$, we have*

$$\mathbb{P}(Y_{2n} = 2k) = P_{00}^{2k}\, P_{00}^{2n-2k}\,.$$

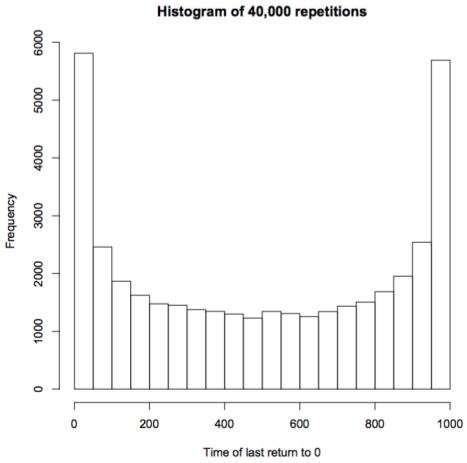


Figure 4.12

Proof. We have

$$\begin{aligned}
\mathbb{P}(Y_{2n} = 2k) &= \mathbb{P}(S_{2k} = 0, S_{2k+1} \neq 0, ..., S_{2n} \neq 0) \\
&= \mathbb{P}(S_{2k} = 0)\, \mathbb{P}(S_{2k+1} \neq 0, ..., S_{2n} \neq 0 \,|\, S_{2k} = 0) \\
&= \mathbb{P}(S_{2k} = 0)\, \mathbb{P}(S_1 \neq 0, ..., S_{2n-2k} \neq 0) \\
&= \mathbb{P}(S_{2k} = 0)\, \mathbb{P}(S_{2n-2k} = 0) \\
&= P_{00}^{2k}\, P_{00}^{2n-2k}
\end{aligned}$$

where the next to last equality follows from Corollary 4.5.3. □

For fixed n, the distribution of Y_{2n} is called the **discrete arcsine distribution**. It is defined by

$$\mathbb{P}(Y_{2n} = 2k) = \binom{2k}{k}\binom{2n-2k}{n-k}\frac{1}{2^{2n}} \quad \text{for } k = 0, 1, ..., n\,,$$

and zero otherwise.

Here is the reason for the name *discrete arcsine* distribution. Recall from Section 4.2 that for large m,

$$P_{00}^{2m} \approx \frac{1}{\sqrt{\pi m}}\,.$$

Thus for large n and values k that are neither close to 0 nor close to n, we have

$$\mathbb{P}(Y_{2n} = 2k) \approx \frac{1}{\pi\sqrt{k(n-k)}} = \frac{1}{n\pi} \cdot \frac{1}{\sqrt{\frac{k}{n}(1-\frac{k}{n})}}\,.$$

Set $x = \frac{k}{n}$ and consider the function $f(s) = \dfrac{1}{\pi\sqrt{s(1-s)}}$ on the interval $(0,1)$. Then

$$\mathbb{P}(Y_{2n} \le 2k) = \sum_{i=0}^{nx} \mathbb{P}(Y_{2n} = 2i) \approx \int_0^x f(s)\, ds = \frac{2}{\pi} \arcsin\sqrt{x}\,.$$

For an illustration, see Figure 4.13. It shows the appropriately rescaled histogram from Figure 4.12 and overlayed arcsine density. The result may appear counterintuitive. One

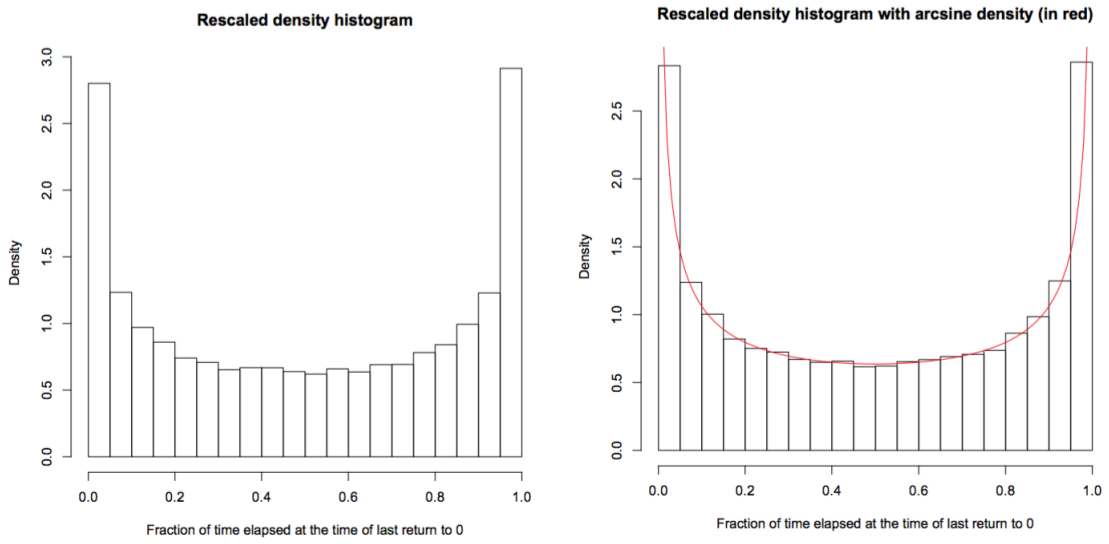


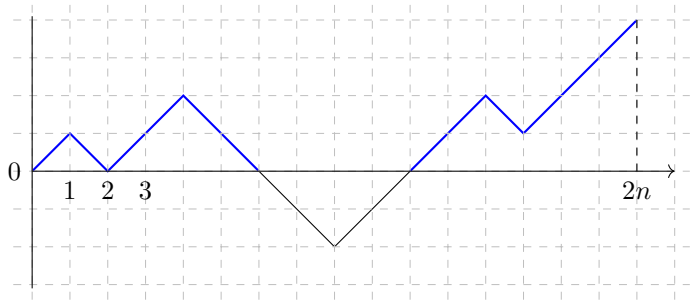
Figure 4.13

would perhaps expect that for a fair game, equalizations occur fairly evenly distributed in the course of the game. However, the arcsine law for the time of last equalization tells us that with high probability, a gambler either takes the lead early on in the game (and remains in lead until the end) or takes the lead close towards the end of the game. With probability $\frac{1}{2}$ the winner of the game is determined during the first half of the game.

4.6.2 How often in the lead?

Next, we look into the distribution of a special **occupancy time** for simple symmetric random walk. For fixed n, we are interested in the distribution of the number of time intervals between time 0 and time $2n$ during which the line segment of the trajectory of the random walk lies above the x-axis. Applied to a fair game of coin tossing of length $2n$ for two players, the question is, what is the distribution of the number of time intervals during which Player 1 (say) is in the lead?

We say the time interval $(m, m+1)$ is a *positive time interval* for the random walk if $S_m > 0$ or $S_{m+1} > 0$. Consider the random variable L_{2n}= "number of positive time intervals between 0 and $2n$". For an illustration see Figure 4.14. As Proposition 4.6.2 will show, the random variable L_{2n} has a discrete ascsine distribution. In fact, we will show that $L_{2n} \sim Y_{2n}$.



Here $L_{2n} = 12$

Figure 4.14

Proposition 4.6.2. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk starting at 0. Then for all $0 \leq k \leq n$, we have*

$$\mathbb{P}(L_{2n} = 2k) = P_{00}^{2k}\, P_{00}^{2n-2k}\,.$$

Proof. To simplify our writing, we will use the notation $\ell_{2n}(2k) = \mathbb{P}(L_{2n} = 2k)$. By Corollary 4.5.3,

$$\ell_{2n}(2n) = P_{00}^{2n} = P_{00}^{2n}\, P_{00}^{0}\,,$$

and by symmetry,

$$\ell_{2n}(0) = \mathbb{P}(S_1 \leq 0, ..., S_{2n} \leq 0) = P_{00}^{2n}\, P_{00}^{0}\,.$$

This shows that the statement holds for $k = 0$ and $k = n$.
Now assume $1 \leq k \leq n-1$. With this assumption on k, there must exist a time t with $1 \leq t \leq n-1$ such that $S_{2t} = 0$. Consider the first time T^0 the random walk returns to 0. Note that there is equal probability that all time intervals between 0 and T^0 are positive and that all time intervals between 0 and T^0 are negative. We can condition on T^0 and, with the use of the strong Markov property, get

$$\begin{aligned}
\ell_{2n}(2k) &= \sum_{t=1}^{n-1} \mathbb{P}(T^0 = 2t)\mathbb{P}(L_{2n} = 2k \,|\, T^0 = 2t) \\
&= \sum_{t=1}^{n-1} \mathbb{P}(T^0 = 2t)\frac{1}{2}\ell_{2n-2t}(2k) + \sum_{t=1}^{n-1} \mathbb{P}(T^0 = 2t)\frac{1}{2}\ell_{2n-2t}(2k-2t)\,.
\end{aligned}$$

Observe that in the above expression we need to set $\ell_m(s) = 0$ if $s > m$ or if $s < 0$. This reduces the two summations in the last expression, and we get

$$\ell_{2n}(2k) = \frac{1}{2}\sum_{t=1}^{n-k} \mathbb{P}(T^0 = 2t)\ell_{2n-2t}(2k) + \frac{1}{2}\sum_{t=1}^{k} \mathbb{P}(T^0 = 2t)\ell_{2n-2t}(2k-2t)\,.$$

We will now prove the statement $\ell_{2n}(2k) = P_{00}^{2k}\, P_{00}^{2n-2k}$ by induction on n. Clearly, the statement holds for $n = 1$. Assume the statement holds for $m < n$. Then

$$\begin{aligned}
\ell_{2n}(2k) &= \frac{1}{2}\sum_{t=1}^{n-k} \mathbb{P}(T^0 = 2t)\ell_{2n-2t}(2k) + \frac{1}{2}\sum_{t=1}^{k} \mathbb{P}(T^0 = 2t)\ell_{2n-2t}(2k-2t) \\
&= \frac{1}{2}\sum_{t=1}^{n-k} \mathbb{P}(T^0 = 2t)P_{00}^{2k}\, P_{00}^{2n-2t-2k} + \frac{1}{2}\sum_{t=1}^{k} \mathbb{P}(T^0 = 2t)P_{00}^{2k-2t}\, P_{00}^{2n-2k} \\
&= \frac{1}{2}P_{00}^{2k}\sum_{t=1}^{n-k} \mathbb{P}(T^0 = 2t)\, P_{00}^{2n-2t-2k} + \frac{1}{2}P_{00}^{2n-2k}\sum_{t=1}^{k} \mathbb{P}(T^0 = 2t)P_{00}^{2k-2t} \\
&= \frac{1}{2}P_{00}^{2k}P_{00}^{2n-2k} + \frac{1}{2}P_{00}^{2n-2k}P_{00}^{2k} \\
&= P_{00}^{2k}P_{00}^{2n-2k}\,.
\end{aligned}$$

This completes the proof. □

Example 4.6.1. Compute the probability that in an infinite sequence of independent fair coin tosses, heads is in the lead at least 80% of the time.
Answer: For large n,

$$\mathbb{P}(L_{2n} \geq (0.8)2n) = (1 - \mathbb{P}(L_{2n} < (0.8)2n)) \approx 1 - \frac{2}{\pi}\arcsin\sqrt{0.8} = 0.295\,.$$

Hence

$$\lim_{n\to\infty} \mathbb{P}(L_{2n} \geq (0.8)2n) = 0.295\,.$$

□

4.7 The Range of a Random Walk

Definition 4.7.1. *Let $(S_n)_{n\geq 0}$ be random walk on $\mathbb{Z}$. The* **range** *R_n at time n is the random variable*

$$R_n = \text{card}\{S_0, S_1, ..., S_n\}\,.$$

The range R_n is the number of *distinct* points the random walk has visited by time n. Clearly, $1 \leq R_n \leq (n+1)$. For simple random walk on $\mathbb{Z}$ starting at 0,

$$\{S_0, S_1, ..., S_n\} = \{-a, -a+1, ..., 0, ..., b-1, b\}$$

for some integers a and b with $0 \leq a, b \leq n$.

The time until a random walk visits the nth new state:

We define the following sequence $T^{(n)}$ of random times. $T^{(n)}$ is the time at which the random walk visits its nth new (i.e., up to time $T^{n)}-1$ unvisited) state. So $R_{T^{(n)}-1} = n-1$ and $R_{T^{(n)}} = n$. Note that the times $T^{(n)}$ are stopping times for the random walk. Simple random walk reaches a new *extreme* value at time $T^{(n)}$. Also note that $T^{(1)} \equiv 0$ and $T^{(2)} \equiv 1$.

Example 4.7.1. Consider simple symmetric random walk on $\mathbb{Z}$ and the following finite-length sample path ω_{10}:

$$\omega_{10} = (0, 1, 0, -1, -2, -1, 0, 1, 2, 1, 2)\,.$$

Here $R_{10}(\omega) = 5$ and $T^{(1)}(\omega) = 0,\ T^{(2)}(\omega) = 1,\ T^{(3)}(\omega) = 3,\ T^{(4)}(\omega) = 4,\ T^{(5)}(\omega) = 8$ for any sample path ω that matches the given ω_{10} up to time 10. □

Proposition 4.7.1. *Let $(S_n)_{n\geq 0}$ be simple symmetric random walk (with any starting point). Then*

$$\mathbb{E}(T^{(n)}) = \frac{1}{2}n(n-1)\,.$$

Proof. Because of translation invariance of the transition probabilities, we can assume the walk starts at 0. Let a and b be two positive integers. We will make repeated use of formula (4.19) for the expected duration of the game in the gambler's ruin problem. If the gambler starts with 0 dollars, the expected time until the gambler either reaches a fortune of b or has incurred a loss of a is $\mathbb{E}(T^{\{-a,b\}}) = ab$. By time $T^{(i)}$, the random walk has visited exactly i distinct, consecutive integers, and $S_{T^{(i)}}$ is either the smallest integer or the largest integer of this set. Assume the walk is at the largest integer at time $T^{(i)}$. Then the expected time until the walk visits the next *new* state, that is, $\mathbb{E}(T^{(i+1)} - T^{(i)})$ is the same as $\mathbb{E}(T^{\{-i,1\}})$ in the gambler's ruin problem. Similarly, if at time $T^{(i)}$ the walk is at the smallest integer, then then $\mathbb{E}(T^{(i+1)} - T^{(i)}) = \mathbb{E}(T^{\{-1,i\}})$. Either way we get

$$\mathbb{E}(T^{(i+1)} - T^{(i)}) = i\,,$$

and from this,

$$\begin{aligned}\mathbb{E}(T^{(n)}) &= \mathbb{E}(T^{(1)}) + \sum_{i=1}^{n-1} \mathbb{E}(T^{(i+1)} - T^{(i)}) \\ &= 0 + 1 + 2 + \cdots + (n-1) = \frac{1}{2} n(n-1)\,.\end{aligned}$$

□

$\mathbb{E}(R_n)$ and asymptotic results for R_n:

The following results more broadly apply to random walks on $\mathbb{Z}^d$ for $d \geq 1$, not only to simple random walk.

Theorem 4.7.2. *Consider random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}^d$ with $S_0 = \mathbf{0}$ and $S_n = \sum_{k=1}^n X_k$, $n \geq 1$, for i.i.d. random variables X_k taking values in $\mathbb{Z}_d$. Let $T^{\mathbf{0}}$ be the first return time to $\mathbf{0}$. Then*

$$\mathbb{E}(R_n) = \sum_{k=0}^{n} \mathbb{P}(T^{\mathbf{0}} > k) \tag{4.29}$$

and

$$\lim_{n\to\infty} \frac{\mathbb{E}(R_n)}{n} = \mathbb{P}(\text{no return to } \mathbf{0})\,. \tag{4.30}$$

Proof. Consider the events $E_k = \{S_k \neq S_i \,:\, 0 \leq i \leq k-1\}$ for $k \geq 1$ and their corresponding indicator random variables

$$\mathbb{1}_{E_k} = \begin{cases} 1 & \text{if } S_k \neq S_i \text{ for all } i = 0, 1, ..., k-1 \\ 0 & \text{otherwise}\,. \end{cases}$$

Then

$$R_n = \sum_{k=0}^{n} \mathbb{1}_{E_k}$$

and

$$\mathbb{E}(R_n) = \sum_{k=0}^{n} \mathbb{P}(E_k)\,.$$

We compute

$$\begin{aligned}\mathbb{P}(E_k) &= \mathbb{P}(S_k - S_{k-1} \neq \mathbf{0},\ S_k - S_{k-2} \neq \mathbf{0}, ...,\ S_k \neq \mathbf{0}) \\ &= \mathbb{P}(X_k \neq \mathbf{0},\ X_k + X_{k-1} \neq \mathbf{0}, ...,\ X_k + X_{k-1} + \cdots + X_1 \neq \mathbf{0}) \\ &= \mathbb{P}(X_1 \neq \mathbf{0},\ X_1 + X_2 \neq \mathbf{0}, ...,\ X_1 + X_2 + \cdots + X_k \neq \mathbf{0}) \\ &= \mathbb{P}(T^{\mathbf{0}} > k)\,.\end{aligned}$$

It follows that

$$\mathbb{E}(R_n) = \sum_{k=0}^{n} \mathbb{P}(T^{\mathbf{0}} > k)\,.$$

We have

$$\lim_{k\to\infty} \mathbb{P}(T^{\mathbf{0}} > k) = \mathbb{P}(T^{\mathbf{0}} = \infty) = \mathbb{P}(\text{no return to } \mathbf{0})\,.$$

Hence

$$\lim_{k\to\infty} \mathbb{P}(E_k) = \mathbb{P}(\text{no return to } \mathbf{0})\,,$$

and as a consequence, the sequence of Césaro averages $\dfrac{1}{n}\displaystyle\sum_{k=0}^{n} \mathbb{P}(E_k) = \dfrac{\mathbb{E}(R_n)}{n}$ converges to the same limit. It follows that

$$\lim_{n\to\infty} \frac{\mathbb{E}(R_n)}{n} = \mathbb{P}(\text{no return to } \mathbf{0})$$

which completes the proof. □

In 1964, Kesten, Spitzer, and Whitman proved an analogous result to (4.30) for *almost sure convergence*:

Theorem 4.7.3 (Kesten–Spitzer–Whitman)**.** *Consider a random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}^d$ with $S_0 = \mathbf{0}$ and $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$, with i.i.d. random variables X_k. Then*

$$\lim_{n\to\infty} \frac{R_n}{n} = \mathbb{P}(\text{no return to } \mathbf{0}) \quad \text{with probability } 1\,.$$

The proof of Theorem 4.7.3 involves material beyond the scope of this text and is omitted. For a reference see [34]. Note that it follows from Theorem 4.7.3 that the **range of a random walk on $\mathbb{Z}^d$ grows sub-linearly with probability 1 if and only if the random walk is recurrent**.

The following corollary to Theorem 4.7.3 applies to random walk on $\mathbb{Z}$ in *one dimension*:

Corollary 4.7.4. *Let $(S_n)_{n\geq 0}$ be an irreducible random walk on $\mathbb{Z}$ with $S_0 = 0$ and $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$, with i.i.d. random variables X_k. If $\mathbb{E}(X_k) = 0$, then the random walk is recurrent.*

Proof. By the Strong Law of Large Numbers,

$$\lim_{n\to\infty} \frac{S_n}{n} = 0 \quad \text{with probability } 1\,.$$

Hence for a fixed $\epsilon > 0$, there is a *random variable* N such that

$$\frac{|S_n|}{n} < \epsilon \qquad \text{for } n > N \text{ with probability } 1\,.$$

In other words, for almost all sample paths ω, all but finitely many of the values $\frac{S_n(\omega)}{n}$, $n \geq 1$, lie in the interval $(-\epsilon, \epsilon)$. For such a sample path ω we have

$$|S_n(\omega)| < n\epsilon \quad \text{for all } n > N(\omega) \tag{4.31}$$

and possibly

$$|S_n(\omega)| \geq n\epsilon \quad \text{for some } n \text{ with } 1 \leq n \leq N(\omega)\,.$$

As a consequence,

$$R_n(\omega) \leq 2n\epsilon + N(\omega) \qquad \text{for all } n > N(\omega)\,.$$

and

$$\lim_{n\to\infty} \frac{R_n(\omega)}{n} \leq 2\epsilon\,.$$

Since ϵ is arbitrarily small, we conclude

$$\lim_{n\to\infty} \frac{R_n(\omega)}{n} = 0\,. \tag{4.32}$$

Since (4.31) and the subsequent discussion throughout (4.32) hold for all sample paths ω in a set of probability 1, we have

$$\lim_{n\to\infty} \frac{R_n}{n} = 0 \qquad \text{with probability } 1\,.$$

By Theorem 4.7.3, the return probability $\mathbb{P}(T^0 < \infty)$ is 1, and so the irreducible random walk is recurrent. □

4.8 Law of the Iterated Logarithm

Here we are interested in the **asymptotic growth rate** of the location of simple symmetric random walk $(S_n)_{n\geq 0}$ as as $n \to \infty$. What can we say about the size of the excursions the walk takes (away from its mean 0) in the long run? So far we know that by the Strong Law of Large Numbers (SLLN), with probability 1,

$$\frac{S_n}{n} \xrightarrow{n\to\infty} 0\,. \tag{4.33}$$

However, the denominator n in (4.33) is too large to give us precise information about the size of fluctuations of S_n, it "overpowers" the numerator S_n.

The Central Limit Theorem improves the information we get from the SLLN. It states

$$\frac{S_n}{\sqrt{n}} \xrightarrow{n\to\infty} N(0,1) \tag{4.34}$$

in distribution. A normal random variable, with high probability, takes values within three standard deviations from its mean. So for large n, with high probability, the location of the random walk S_n will lie in the interval $[-3\sqrt{n}, 3\sqrt{n}]$ (recall (4.1) and the surrounding discussion). But there will also be large but rare fluctuations outside this interval. One shows (we omit the proof) that (4.34) implies

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} = \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{S_n}{\sqrt{n}} = -\infty\,. \tag{4.35}$$

We see from (4.35) that the denominator $\sqrt{n}$ is too small to give us any details about the size of such fluctuations.

Theorem 4.8.1 below, which is known as the **Law of the Iterated Logarithm**, settles the question. Its statement involves a denominator that lies between $\sqrt{n}$ and n and is "exactly right" for giving information about the long-term fluctuations of the random walk. The theorem is due to Khinchine[1]. Its proof is beyond the scope of this text (for a reference see [21]). Note that the Law of the Iterated Logarithm applies quite generally. It applies to *any* random walk whose step distribution has mean 0 and variance 1.

Theorem 4.8.1 (Law of the Iterated Logarithm)**.** *Let $(S_n)_{n\geq 0}$ be a random walk where $S_n = \sum_{k=1}^n X_k$ and $X_1, X_2, \ldots$ are i.i.d. random variables with mean $\mu = 0$ and variance $\sigma^2 = 1$. Then*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1\right) = 1,$$

and furthermore (by applying the statement to $-X_1, -X_2, \ldots$),

$$\mathbb{P}\left(\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1\right) = 1.$$

Note that Theorem 4.8.1 says that with probability 1, for any $\epsilon > 0$, there exist *infinitely many* n such that

$$S_n > (1-\epsilon)\sqrt{2n\log\log n}$$

and *at most finitely many* n such that

$$S_n > (1+\epsilon)\sqrt{2n\log\log n}\,.$$

[1] Aleksandr Khinchine (1894-1959), Russian mathematician.

Also, correspondingly for negative fluctuations, there exist infinitely many n such that

$$S_n < (-1+\epsilon)\sqrt{2n \log\log n}$$

and at most finitely many n such that

$$S_n < (-1-\epsilon)\sqrt{2n \log\log n}\,.$$

Exercises

Exercise 4.1. Consider simple symmetric random walk $(S_n)_{\geq 0}$ on $\mathbb{Z}$ that starts at 0. Show that for all $n \geq 1$,

(a) $\mathbb{P}(S_{2n} = 0) = \mathbb{P}(S_{2n-1} = 1)$.

(b) $\mathbb{P}(S_{2n} = 0) = \max\{\mathbb{P}(S_{2n} = a) : a \in \mathbb{Z}\}$ and
$\mathbb{P}(S_{2n-1} = 1) = \max\{\mathbb{P}(S_{2n-1} = a) : a \in \mathbb{Z}\}$.

Exercise 4.2. Consider simple symmetric random walk $(S_n)_{\geq 0}$ on $\mathbb{Z}$ that starts at 0 and let $b < c$. Show that for all $n \geq 1$,

$$\mathbb{P}(S_n \in [b, c]) \leq (c - b + 1)\mathbb{P}(S_n \in \{0, 1\})$$

and conclude that for all finite intervals $[b, c]$,

$$\lim_{n\to\infty} \mathbb{P}(S_n \in [b, c]) = 0\,.$$

Exercise 4.3. Consider simple (symmetric or biased) random walk on $\mathbb{Z}$ that starts at 0. Compute $\mathbb{P}(S_n \leq 0 \text{ for all } n \geq 0)$, that is, the probability that the random walk never visits a positive integer.

Exercise 4.4. A standard American roulette wheel has 38 slots numbered $1 - 36$ and 0 and 00. The two slots labeled 0 and 00 are green, half of the numbers between 1 and 36 are black, and the other half of the numbers are red. The wheel is spun, and any of the 38 numbers is equally likely to come up. You start with \$50 and make a sequence of bets on red. For each bet, if you win, you gain \$1, and if you lose, you have to pay \$1 to the house. Your plan is to quit the game as soon as you either have reached a fortune of \$100 or have lost your entire initial fortune and are down to \$0, whichever happens first. What is the probability that you will quit the game with a fortune of \$100?

Exercise 4.5. Consider simple symmetric random walk $(X_n)_{n\geq 0}$ on the integers $\{0, 1, ..., 5\}$ with *partially reflecting boundary* at the two endpoints 0 and 5. More precisely, we have $P_{0,1} = P_{0,0} = \frac{1}{2}$ and $P_{5,4} = P_{5,5} = \frac{1}{2}$ and $P_{x,x+1} = P_{x,x-1} = \frac{1}{2}$ for $x = 2, 3, 4$. Assume the process starts in state 0. Compute $\mathbb{E}_0(T^5)$. (*Hint*: View $(X_n)_{n\geq 0}$ as a lumped version of another process. Figure 1.14 may be helpful.)

Exercise 4.6. Consider simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ that starts at 0.

(a) Show that for $n \geq 1$, we have $\mathbb{P}(S_1 \neq 0, ..., S_n \neq 0) = \frac{1}{n}\mathbb{E}(|S_n|)$.

(b) Assume $(S_n)_{n\geq 0}$ is simple *symmetric* random walk and $k \geq 2$ is even. Use part (a) to show that the expect displacement $\mathbb{E}(|S_k|)$ of the walk from its starting point 0 is

$$\mathbb{E}(|S_k|) = kP_{00}^k \, .$$

Exercise 4.7. Consider simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ with $P_{x,x+1} = p$ and $P_{x,x-1} = 1 - p = q$. Assume the walk starts at 0. Consider T^1, the first hitting time of State 1. Show that for all $n \geq 1$,

$$\mathbb{P}(T^1 = 2n - 1) = \frac{(2n-2)!}{n!(n-1)!} p^n q^{n-1} \, .$$

Exercise 4.8. Consider simple biased random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ with $P_{x,x+1} = p$, $P_{x,x-1} = 1 - p = q$, and $p < q$. Assume the walk starts at 0. Let $c \in \mathbb{Z}^+$ and consider the first hitting time T^c. Notice that here, since $p < q$, the event $\{T^c = \infty\}$ has positive probability, and therefore $\mathbb{E}(T^c) = \infty$. Prove that

$$\mathbb{E}(T^c \,|\, T^c < \infty) = \frac{c}{|p - q|} \, .$$

(*Hint:* The result from Exercise 4.15 may be useful.)

Exercise 4.9. Consider simple (symmetric or biased) random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ that starts at 0. Prove that

$$\mathbb{E}(S_n \,|\, S_{n+1}) = \frac{n}{n+1} S_{n+1} \, .$$

Exercise 4.10. Consider simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ that starts at 0. Fix a time $n > 0$. The random variable $M_n = \max\{S_k : 0 \leq k \leq n\}$ gives the maximum level the walk attains by time n. Consider two levels a and b with $0 < b < a$ and $n \geq 2a - b$. Use the Reflection Principle to show that

$$\mathbb{P}(M_n \geq a \text{ and } S_n = b) = \mathbb{P}(S_n = 2a - b).$$

Exercise 4.11. Consider simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ that starts at 0, and for it the first hitting times T^3 and T^5. Let

$$p_n = \mathbb{P}(T^3 \leq n \text{ and } T^5 > n).$$

Compute p_n for $n \geq 1$.

Exercise 4.12. Consider simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ that starts at 0, and let $M_n = \max\{S_k : 0 \leq k \leq n\}$. Compute

$$\mathbb{P}(T^4 \leq 6 \text{ and } M_6 < 6 \,|\, S_2 = 2).$$

Exercise 4.13. Consider biased random walk on $\mathbb{Z}$. At each time, the walk takes a step according to a random variable X_k with $\mathbb{P}(X_k = 1) = \frac{1}{2}$, $\mathbb{P}(X_k = -1) = \frac{1}{4}$, and $\mathbb{P}(X_k = 0) = \frac{1}{4}$. Assume that the walk starts in state 2. Compute the probability $P_{2,2}^{\{0,5\}}$, that is, the probability that the process returns to State 2 before it reaches either State 0 or State 5 for the first time.

Exercise 4.14. Consider simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ starting at 0. Fix $n > 0$ and define the random variable $Z_{2n} = \min\{2k \leq 2n : S_{2k} = S_{2n}\}$. Find the distribution of Z_{2n}.

Exercise 4.15. Consider simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ with $P_{x,x+1} = p$, $P_{x,x-1} = q$, and $P_{xx} = h$ with $p + q + h = 1$. The random walk starts at 0. The probability of *no return* is $\mathbb{P}_0(T^0 = \infty)$. Prove that

$$\mathbb{P}_0(T^0 = \infty) = |p - q|\,.$$

Exercise 4.16. Consider simple biased random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ with $P_{x,x+1} = p$, $P_{x,x-1} = q$, and $P_{xx} = h$ with $p + q + h = 1$. Assume $p > q$. The random walk starts in state 0. Let

$$\widetilde{M} = \min\{S_n : n \geq 0\}\,.$$

Compute $\mathbb{E}_0(\widetilde{M})$.

Exercise 4.17. Two players, Player A and Player B, play against each other in a series of fair \$1 bets. Let $C \in \mathbb{N}$. Each player starts with a fortune of \$$C$, and the game ends if either of the players has lost their \$$C$.

(a) What is the expected number of returns to the initial state of the game (where each player has a fortune of \$$C$), before the game ends?

(b) What is the expected number of times Player A is ahead of Player B, before the game ends?

Exercise 4.18. Let $(S_n)_{n\geq 0}$ be simple biased random walk on $\mathbb{Z}$ with $S_0 = 5$, $p = 1/4$, and $q = 3/4$. Let V^y be the random variable "number of visits to state y".

(a) Compute the probability that this random walk never visits 7.

(b) Compute $\mathbb{P}_5(V^7 = 3)$.

(c) Compute $\mathbb{P}_5(V^2 = 3)$.

Exercise 4.19. Consider simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ with an *absorbing wall* at 0. That is, $P_{x,x+1} = p$ and $P_{x,x-1} = 1 - p = q$ for all integers $x \neq 0$, and $P_{00} = 1$. Assume the walk starts at State c for some $c > 0$. We define the maximum random variable M by

$$M = \max\{S_n : n \geq 0\}.$$

For which values of c, p, q is $\mathbb{E}(M)$ finite, and for which values of c, p, q is $\mathbb{E}(M)$ infinite?

Chapter 5

Branching Processes

In this chapter we focus on the *Galton-Watson branching process* as a model for population growth. The process is named after Frances Galton[1] and Henry W. Watson[2] who introduced the model around 1875 to study the survival of family names in England. The Galton-Watson branching process is a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathbb{N}_0$. The random variables X_n, $n \geq 0$, give the size of the population at time n (i.e., the size of Generation n). At each time interval, each particle, *independently* of all other particles and of any past, present, or future state of the system, splits into k particles or dies, according to a fixed **offspring distribution** μ. See Figure 5.1 for an illustration. The size of Generation 0 and the offspring distribution μ fully determine the evolution of the process.

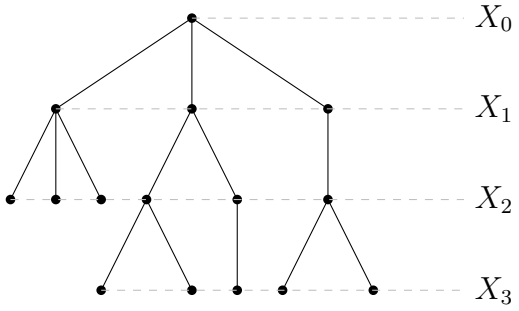


Figure 5.1: Branching process

Let Z be a random variable with distribution μ and taking values in $\mathbb{N}_0$. We assume Z is a.s. finite, so $\mathbb{P}(Z < \infty) = 1$. For $x \geq 1$, the one-step transition probabilities for the

[1]Sir Frances Galton (1822-1911), English anthropologist and statistician, cousin of Charles Darwin.
[2]Henry William Watson (1827-1903), English mathematician.

Galton-Watson branching process with offspring distribution μ are

$$P_{xy} = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) = \mathbb{P}(Z_1 + \cdots + Z_x = y)$$

where $Z_1, ..., Z_x$ are i.i.d. random variables with $Z_1 \sim Z$. Equivalently,

$$P_{xy} = \mu^{*x}(y)$$

where μ^{*x} is the xth **convolution power** of μ. Note that 0 is an absorbing state for this process.
The next section introduces *generating functions* which are a useful tool in the study of branching processes.

5.1 Generating functions

Definition 5.1.1. *Let Z be a random variable taking values in $\mathbb{N}_0$ with $\mathbb{P}(Z = k) = \mu(k)$ for $k \geq 0$. The* **probability generating function** *f_Z of Z is defined by*

$$f_Z(t) = \mathbb{E}(t^Z) = \sum_{k=0}^{\infty} \mu(k) t^k \qquad \text{for } -1 \leq t \leq 1\,. \tag{5.1}$$

Note that the power series in (5.1) has radius of convergence $r \geq 1$.
Definition 5.1.1 also applies if Z is not finite with probability 1. In this case, we compute

$$\mathbb{P}(Z < \infty) = f_Z(1).$$

Uniqueness of f_Z: The probability generating function f_Z encodes information about the distribution of Z. It fully determines the distribution. The coefficients in the power series representation of f_Z are the probabilities of the probability mass function μ of Z. Since two functions that are represented by a power series are equal if and only if the coefficients in their power series are equal, we have $f_Z = f_Y$ if and only if $Z \sim Y$. Note that we can recover the probability mass function μ for a random variable Z from its probability generating function f_Z by taking derivatives:

$$\mu(0) = f_Z(0), \quad \mu(1) = f_Z'(0), \quad \mu(2) = \frac{1}{2} f_Z''(0), \quad \ldots\ldots \,, \mu(k) = \frac{1}{k!} f_Z^{(k)}(0)\,.$$

Example 5.1.1. Let Z be a Bernoulli random variable with $\mathbb{P}(Z = 1) = p$ and $\mathbb{P}(Z = 0) = 1 - p$. Then

$$f_Z(t) = 1 - p + pt\,.$$

□

Example 5.1.2. Let Z have the geometric distribution with parameter $p < 1$. That is, $\mathbb{P}(Z = k) = (1-p)p^k$ for $k \geq 0$. Then

$$f_Z(t) = \sum_{k=0}^{\infty} (1-p)p^k t^k = \frac{1-p}{1-pt}\,.$$

□

Example 5.1.3. Let Z have a Poisson distribution with parameter λ. So $\mathbb{P}(Z = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, and

$$f_Z(t) = e^{-\lambda} \sum_{k=0}^{\infty} (\lambda t)^k / k! = e^{-\lambda} e^{\lambda t} = e^{\lambda(t-1)}\,.$$

□

Proposition 5.1.1. *Let Y and Z be independent random variables taking values in $\mathbb{N}_0$. Then*

$$f_{Y+Z}(t) = f_Y(t) f_Z(t)\,.$$

Proof. Since Y and Z are independent, $f_{Y+Z}(t) = \mathbb{E}(t^{Y+Z}) = \mathbb{E}(t^Y)\mathbb{E}(t^Z) = f_Y(t) f_Z(t)$.

□

By induction, if $Z_1, ..., Z_n$ are n independent random variables taking values in $\mathbb{N}_0$ and $Z = Z_1 + \cdots + Z_n$, then

$$f_Z(t) = \prod_{k=1}^{n} f_{Z_k}(t)\,.$$

As a consequence, if $(X_n^{(k)})_{n \geq 0}$ denotes a branching process starting with k particles, we have

$$f_{X_n^{(k)}}(t) = (f_{X_n}(t))^k\,.$$

Example 5.1.4. Let Z have a binomial distribution with parameters p and n. Then $Z \sim X_1 + \cdots + X_n$ where the X_k are i.i.d Bernoulli random variables with parameter p. Thus

$$f_Z(t) = (1 - p + pt)^n\,.$$

Alternatively, in computing $f_Z(t)$ directly, we have

$$f_Z(t) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{1-p} t^k = (1 - p + pt)^n\,.$$

□

Example 5.1.5. By Example 5.1.3, the probability generating function for a Poisson random variable $Y \sim \text{Poisson}(\lambda)$ is $f_Y(t) = e^{\lambda(t-1)}$. Let $X \sim \text{Poisson}(\nu)$ and assume X and Y are independent. Then

$$f_{X+Y}(t) = e^{\lambda(t-1)}e^{\nu(t-1)} = e^{(\lambda+\nu)(t-1)}$$

which we identify as the probability generating function of a Poisson random variable with parameter $\lambda + \nu$. Since $f_{X+Y}(t)$ uniquely determines the distribution of $X + Y$, this proves that the sum of two independent Poisson random variables is also a Poisson random variable (whose parameter is the sum of the individual parameters). □

Theorem 5.1.2 (Continuity theorem)**.** *Consider a sequence of random variables $X_1, X_2, ...$ that take values in $\mathbb{N}_0$. For each $n \geq 1$, we denote its probability mass function by $\mu^{(n)}$, that is, $\mathbb{P}(X_n = k) = \mu^{(n)}(k)$ for $k \geq 0$. We assume $\sum_{k=0}^{\infty} \mu^{(n)}(k) = 1$ for all $n \geq 1$, and denote the probability generating function of $\mu^{(n)}$ by f_n. Then the following holds.*

$$\lim_{n\to\infty} \mu^{(n)}(k) = \mu(k) \quad \text{for all } k \geq 0 \tag{5.2}$$

for a sequence of nonnegative numbers $\mu(k)$, $k \geq 0$, if and only if

$$\lim_{n\to\infty} f_n(t) = f(t) \quad \text{for all } 0 < t < 1$$

for a function $f : (0, 1) \to [0, \infty)$. In this case, $\sum_{k=0}^{\infty} \mu(k) \leq 1$, and f is the probability generating function of μ.

Note that (5.2) describes **convergence in distribution** of the random variables $X_1, X_2, ...$ to a (possibly not a.s. finite) random variable $X \sim \mu$. We have $\mathbb{P}(X < \infty) = \sum_{k=0}^{\infty} \mu(k)$. For a proof of Theorem 5.1.2, see [30].

Example 5.1.6. Fix a number $\lambda > 0$. Consider a sequence $X_1, X_2, ...$ of binomial random variables with $X_n \sim \text{Bin}(n, p_n)$ and $p_n = \lambda/n$ for $n \geq 1$. Then

$$f_{X_n}(t) = \left(1 - \frac{\lambda}{n} + t\frac{\lambda}{n}\right)^n = \left(1 + \frac{\lambda}{n}(t-1)\right)^n$$

and for all $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \left(1 + \frac{\lambda}{n}(t-1)\right)^n = e^{\lambda(t-1)}$$

which we recognize as the probability generating function of a Poisson random variable with parameter λ. Hence $X_1, X_2, ...$ converge in distribution to a random variable $X \sim \text{Poisson}(\lambda)$. □

Proposition 5.1.3. *Let $Z_1, Z_2, \ldots$ be i.i.d. random variables with $Z_i \sim Z$ and N a random variable independent of the $Z_1, Z_2, \ldots$. The Z_i's as well as N are a.s. finite, nonnegative, and integer-valued random variables. Consider the random sum*

$$S_N = Z_1 + \cdots + Z_N \,.$$

Then the probability generating function of S_N is

$$f_{S_N}(t) = f_N(f_Z(t)) \,.$$

Proof. We have

$$\mathbb{E}(t^{S_N}) = \mathbb{E}(\mathbb{E}(t^{S_N} \mid N)) = \sum_{k=0}^{\infty} \mathbb{P}(N = k)\, \mathbb{E}(t^{S_N} \mid N = k) \,.$$

Recall that $\mathbb{E}(t^{S_N} \mid N = k) = (f_Z(t))^k$, and so

$$\begin{aligned} f_{S_N}(t) &= \mathbb{E}(t^{S_N}) = \textstyle\sum_{k=0}^{\infty} \mathbb{P}(N = k)(f_Z(t))^k \\ &= f_N(f_Z(t)) \,. \end{aligned}$$

□

We can compute moments of Z via differentiation of f_Z:

Proposition 5.1.4. *Let Z be a random variable taking values in $\mathbb{N}_0$ and $f_Z(t) = \sum_{k=0}^{\infty} \mu(k) t^k$ its probability generating function. Then*

$$\mathbb{E}(Z) = \lim_{t \to 1^-} f_Z'(t)$$

(the limit may be infinite), and if $\mathbb{E}(Z) < \infty$,

$$\mathrm{Var}(Z) = \lim_{t \to 1^-} f_Z''(t) + \mathbb{E}(Z) - (\mathbb{E}(Z))^2 \,.$$

Proof. First, assume the radius of convergence r of the power series is strictly greater than 1. Then all derivatives of $f_Z(t)$ exist and are finite at $t = 1$ and can be computed by term-by-term differentiation of $\sum_{k=0}^{\infty} \mu(k) t^k$. We have

$$f_Z'(t) = \sum_{k=1}^{\infty} \mu(k) k t^{k-1}$$

and thus $f'_Z(1) = \mathbb{E}(Z)$. Furthermore,

$$f''_Z(t) = \sum_{k=2}^{\infty} \mu(k)k(k-1)t^{k-2}$$

and thus $f''_Z(1) = \mathbb{E}(Z^2 - Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)$ from which the given expression for $\mathrm{Var}(Z)$ follows.

Now assume $r = 1$. Since $\mu(k)k \geq 0$ for all $k \geq 1$, the function $f'_Z(t)$ on $[0, 1)$ is nondecreasing, and we have

$$f'_Z(t) \leq \sum_{k=1}^{\infty} \mu(k)k = \mathbb{E}(Z)\,.$$

Because of the monotonicity of the function $f'_Z(t)$ on $[0, 1)$, either $\lim\limits_{t\to 1^-} f'_Z(t) = \infty$ or $\lim\limits_{t\to 1^-} f'_Z(t) = L < \infty$. In the former case, if $\lim\limits_{t\to 1^-} f'_Z(t) = \infty$, it follows that

$$\mathbb{E}(Z) = \infty\,.$$

If $\lim\limits_{t\to 1^-} f'_Z(t) = L < \infty$, then for any nondecreasing sequence $(t_n)_{n\geq 1}$ of numbers in $[0, 1)$ with $\lim\limits_{n\to\infty} t_n = 1$ we must have

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} \mu(k)kt_n^{k-1} = L\,. \tag{5.3}$$

Applying the Monotone Convergence theorem (Theorem C.3.1) to (5.3), we get

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} \mu(k)kt_n^{k-1} = \sum_{k=1}^{\infty} \mu(k)k \left[\lim_{n\to\infty} t_n^{k-1}\right] = L$$

which yields

$$\lim_{t\to 1^-} f'_Z(t) = L = \mathbb{E}(Z)\,.$$

If $\mathbb{E}(Z) < \infty$, we can apply a similar reasoning as above to the function $f''_Z(t)$ on $[0, 1)$. This yields

$$\lim_{t\to 1^-} f''_Z(t) = \mathbb{E}(Z^2) - \mathbb{E}(Z),$$

from which we get

$$\mathrm{Var}(Z) = \lim_{t\to 1^-} f''_Z(t) + \mathbb{E}(Z) - (\mathbb{E}(Z))^2\,.$$

□

Note: We can compute higher order moments of Z in a similar fashion.

We return to branching processes. In the following, assume that the branching process starts with one particle. We will now compute the probability generating function

$$f_{X_n}(t) = \mathbb{E}(t^{X_n}) = \sum_{k=0}^{\infty} \mathbb{P}(X_n = k) t^k$$

of the size X_n of the n-th generation. Let Z be the offspring random variable for the process. For ease of notation, we will write $f(t)$ instead of $f_Z(t)$ for its generating function. Thus $f_{X_1}(t) = f(t)$. We will also use the following notation for the n-fold composition of the function f with itself:

$$f_2(t) = f(f(t))$$

and

$$f_n(t) = f(f_{n-1}(t)) = f(f(...f(t)...) = f_{n-1}(f(t)) \quad \text{for } n \geq 2\,.$$

Proposition 5.1.5. *Let $(X_n)_{n\geq 0}$ be a branching process starting with one particle, and let $f(t)$ be the probability generating function of its offspring random variable Z. Then for any $k, l > 0$ with $k + l = n$, the probability generating function for X_n, i.e., for the size of the n-th generation n, is*

$$f_{X_n}(t) = f_{X_k}(f_{X_l}(t))\,,$$

and in particular,

$$f_{X_n}(t) = f_n(t) = f(f(...f(t)...))\,,$$

the n-fold composition of f with itself.

Proof.

$$\begin{aligned} f_{X_n}(t) &= \sum_{m\geq 0} P^n_{1m} t^m \\ &= \sum_{m\geq 0} \left(\sum_{s\geq 0} P^k_{1s} P^l_{sm} \right) t^m = \sum_{s\geq 0} P^k_{1s} \left(\sum_{m\geq 0} P^l_{sm} t^m \right) \\ &= \sum_{s\geq 0} P^k_{1s} f_{X_l^{(s)}}(t) = \sum_{s\geq 0} P^k_{1s} (f_{X_l}(t))^s \\ &= f_{X_k}(f_{X_l}(t))\,. \end{aligned}$$

Since $f_{X_1}(t) = f(t)$, we have $f_{X_2}(t) = f_2(t)$. Since $f_{X_n}(t) = f(f_{X_{n-1}}(t))$, the formula $f_{X_n}(t) = f_n(t)$ follows by induction on n. □

Corollary 5.1.6. *Let $(X_n)_{n\geq 0}$ be a branching process starting with one particle, and let $f(t)$ be the probability generating function for its offspring random variable Z. Set $m = \mathbb{E}(Z)$ and $\sigma^2 = \mathrm{Var}(Z)$ and assume both $m, \sigma^2 < \infty$. Then*

(a) $\mathbb{E}(X_n) = m^n$, and

(b)

$$\mathrm{Var}(X_n) = \begin{cases} n\sigma^2 & \text{if } m = 1\,, \\ \dfrac{\sigma^2(m^n-1)m^{n-1}}{m-1} & \text{if } m \neq 1\,. \end{cases}$$

Proof. (a) Since $m, \sigma^2 < \infty$, the generating function $f(t)$ of Z is twice (left) differentiable at 1, and therefore the same is true for any n-fold composition of f. We have $\mathbb{E}(Z) = m = f'(1)$ and $\mathbb{E}(X_n) = f_n'(1)$. From this we get

$$\mathbb{E}(X_2) = f_2'(1) = f'(f(1))f'(1) = m^2\,,$$

and, by induction,

$$\mathbb{E}(X_n) = f'(f_{n-1}(1))f_{n-1}'(1) = m \cdot m^{n-1} = m^n\,.$$

(b) By Proposition 5.1.4, we have

$$\mathrm{Var}(X_n) = f_n''(1) + f_n'(1) - (f_n'(1))^2 = f_n''(1) + m^n - m^{2n}\,.$$

From $f_n'(t) = f'(f_{n-1}(t))f_{n-1}'(t)$ we get

$$f_n''(t) = f_{n-1}''(f(t))[f'(t)]^2 + f_{n-1}'(f(t))f''(t)$$

and so

$$f_n''(1) = f_{n-1}''(1)m^2 + m^{n-1}f''(1) = f_{n-1}''(1)m^2 + m^{n-1}(\sigma^2 - m + m^2)\,.$$

We denote $\mathrm{Var}(X_n) = \sigma_n^2$ and get the recursive formula

$$\begin{aligned} \sigma_n^2 &= f_n''(1) + m^n - m^{2n} \\ &= f_{n-1}''(1)m^2 + m^{n-1}(\sigma^2 - m + m^2) + m^n - m^{2n} \\ &= \sigma^2 m^{n-1} + m^2\sigma_{n-1}^2\,. \end{aligned}$$

Thus $\sigma_1^2 = \sigma^2$, $\sigma_2^2 = \sigma^2(m + m^2)$, $\sigma_3^2 = \sigma^2(m^2 + m^3 + m^4)$, and, by induction,

$$\sigma_n^2 = \sigma^2(m^{n-1} + \cdots + m^{2n-2}) = \sigma^2 m^{n-1}(1 + m + \cdots + m^{n-1})$$

from which we conclude

$$\mathrm{Var}(X_n) = \sigma_n^2 = \begin{cases} n\sigma^2 & \text{if } m = 1\,, \\ \dfrac{\sigma^2(m^n - 1)m^{n-1}}{m-1} & \text{if } m \neq 1\,. \end{cases}$$

□

Note that if the branching process starts with k particles, we have

$$\mathbb{E}(X_n^{(k)}) = k\, m^n$$

and, by independence,

$$\mathrm{Var}(X_n^{(k)}) = \begin{cases} k\, n\sigma^2 & \text{if } m = 1\,, \\ \dfrac{k\sigma^2(m^n - 1)m^{n-1}}{m-1} & \text{if } m \neq 1\,. \end{cases}$$

Definition 5.1.2. *Let $(X_n)_{n\geq 0}$ be a branching process and Z its offspring random variable. Set $m = \mathbb{E}(Z)$.*

- *If $m > 1$, we call the process* **supercritical**.
- *If $m < 1$, we call the process* **subcritical**.
- *If $m = 1$, we call the process* **critical**.

The above three cases behave very differently in terms of their long-term growth behavior of their population. In the supercritical case, we expect the population size to "explode", whereas in the subcritical case, the population will a.s. eventually die out. The following section, as well as Section 6.6.3, provide more details about the growth behavior of branching processes.

5.2 Extinction

Will the population whose size is modeled by a given branching process eventually die out? Clearly, $\mu(0) > 0$ is a necessary condition for this event to have positive probability. Throughout this section we will assume $\mu(0) > 0$. Assume the branching process starts with one particle, and let $T = \min\{n : X_n = 0\}$ be the *time until extinction*. We define the *extinction probability* e_0 by

$$e_0 = \mathbb{P}(T < \infty)\,.$$

Note that state 0 is absorbing. All other states $x \in \mathbb{N}$ are transient, since they lead into 0 due to $P_{x0} = (\mu(0))^x > 0$. By independence, the extinction probability $e_0^{(k)}$ for a process

$(X_n^{(k)})_{n\geq 0}$ that starts with k particles is

$$e_0^{(k)} = (e_0)^k \,.$$

Example 5.2.1. Assume the process starts with one particle. Consider an offspring distribution μ with $\mu(0) + \mu(1) = 1$ and $\mu(0), \mu(1) \neq 1$. Then the random variable T has a geometric distribution. Indeed, since the offspring distribution allows for only one or zero particles, we have $\mathbb{P}(T > n) = (\mu(1))^n$. Thus

$$e_0 = 1 - \lim_{n\to\infty} \mathbb{P}(T > n) = 1 \,.$$

We compute

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(T > 0) = \frac{1}{1-\mu(1)} \,.$$

□

In the following we will assume $\mu(0) > 0$ and $\mu(0) + \mu(1) < 1$.

Consider the sequence of events $\{T \leq n\}$ = "population goes extinct *by* Generation n", $n \geq 0$. Note that $\{T \leq n\}$ is the event that the populations dies out in Generation n or in an earlier generation. Clearly, $\{T \leq n-1\} \subseteq \{T \leq n\}$. So

$$\{T < \infty\} = \bigcup_{n\geq 0} \{T \leq n\} \,.$$

Using the notation $u_n = \mathbb{P}(T \leq n)$, we have

$$\lim_{n\to\infty} \uparrow u_n = e_0 \,.$$

Note that

$$u_n = \mathbb{P}(X_n = 0) = f_n(0) \,.$$

Also note that $\mathbb{P}(X_n = 0, X_{n-1} \neq 0) = u_n - u_{n-1}$, which is the probability that the population becomes extinct *in* Generation n.
We have $f_n(0) = f(f_{n-1}(0)) = f(\mathbb{P}(T \leq n-1))$. It follows that e_0 is the solution to the recursion $u_n = f(u_{n-1})$ with $u_0 = 0$. Since f is continuous, we can interchange lim and f and get

$$\lim_{n\to\infty} \uparrow u_n = f(\lim_{n\to\infty} \uparrow u_{n-1}) \,,$$

and thus

$$e_0 = f(e_0) \,.$$

This proves the following proposition:

Proposition 5.2.1. *Let $(X_n)_{n\geq 0}$ be a branching process that starts with one particle and let f be the probability generating function of its offspring distribution μ. We assume $0 < \mu(0) < 1$. Let $T = \min\{n : X_n = 0\}$. Then the extinction probability $e_0 = \mathbb{P}(T < \infty)$ is the* **smallest fixed point of the function f on the interval** $[0, 1]$.

Note that 1 is always a fixed point of f. But depending on the offspring distribution μ, there *may* be an additional smaller fixed point $e_0 \in [0, 1)$. [Of course f may have fixed points that lie outside the interval $[0, 1]$ as well, but they are not of interest here.]

We will now closer investigate the graph of f on the interval $[0, 1]$. Throughout we assume $\mu(0) = f(0) > 0$, and $\mu(0) + \mu(1) < 1$. These conditions imply $f'(t) > 0$ and $f''(t) > 0$ on $[0, 1)$, and hence the **continuous** function f is **strictly increasing and strictly convex** on $[0, 1)$. As a consequence of these properties of f, we have the following possible scenarios for the graph of f:

For the **supercritical** case $f'(1) = m > 1$, the graph of f crosses the diagonal $y = t$ exactly once on the interval $[0, 1)$. See Figure 5.2. This single fixed point e_0 of f on the interval $[0, 1)$ is the extinction probability.

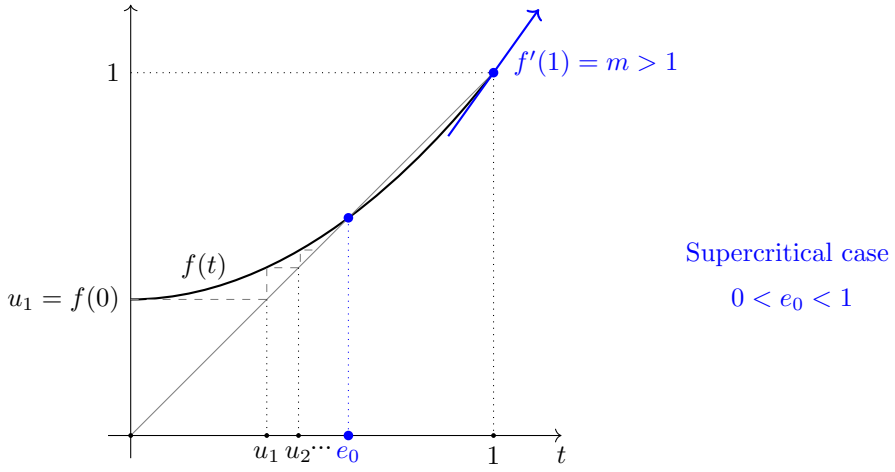


Figure 5.2: Graph of the generating function f for the supercritical case

For the **subcritical** case $f'(1) = m < 1$ or the **critical** case $f'(1) = m = 1$, the graph of f does not cross the diagonal $y = t$ on the interval $[0, 1)$. See Figure 5.3. Hence the extinction probability e_0 is 1.

Example 5.2.2. Consider a branching process $(X_n)_{n\geq 0}$ with offspring random variable $Z \sim \text{Bin}(\frac{1}{2}, 3)$. Assume the process starts with one particle.

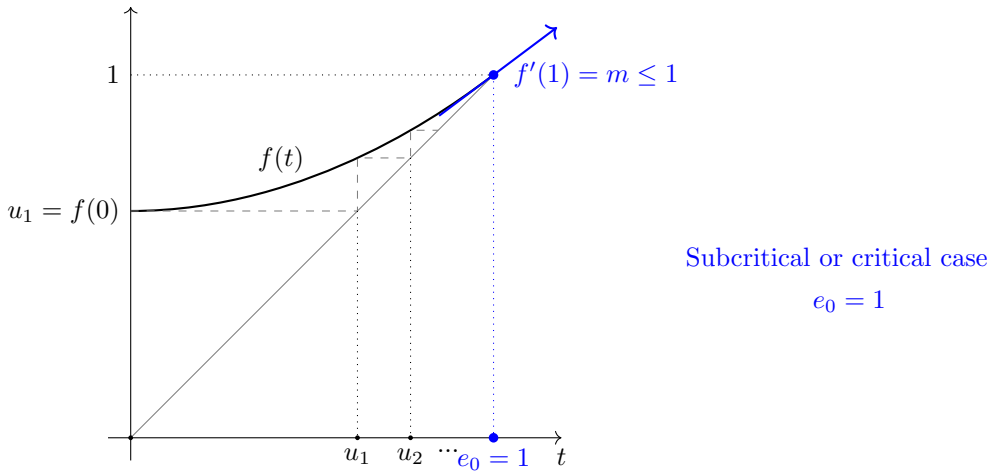


Figure 5.3: Graph of f for the subcritical or critical case

(a) Find the extinction probability e_0.
(b) Find the probability that the population goes extinct in the second generation.

Solution: Note that $\mathbb{E}(Z) = \frac{3}{2} > 1$. The generating function f_Z of Z is

$$f_Z(t) = (\frac{1}{2} + \frac{1}{2}\,t)^3 = \frac{1}{8}\,t^3 + \frac{3}{8}\,t^2 + \frac{3}{8}\,t + \frac{1}{8}\,.$$

(a) We need to find the smallest positive solution to $f_Z(t) = t$. This yields the equation

$$t^3 + 3t^2 - 5t + 1 = 0\,.$$

Factoring out $(t-1)$ on the left-hand side results in $(t-1)(t^2+4t-1) = 0$. We then set $t^2 + 4t - 1 = 0$ and get the solutions $t = -2 \pm \sqrt{5}$. Hence the smallest positive fixed point of f_Z is $-2 + \sqrt{5}$, and so we have

$$e_0 = -2 + \sqrt{5}\,.$$

(b) We need to find $\mathbb{P}(T = 2) = f(f(0)) - f(0)$ (we have written f for f_Z). A straightforward computation yields $f(0) = \frac{1}{8}$ and

$$f(f(0)) = f(\frac{1}{8}) = \frac{729}{8^4} \approx 0.18\,.$$

Thus we get

$$\mathbb{P}(T = 2) \approx 0.18 - 0.125 = 0.055\,.$$

□

Example 5.2.3. Consider a branching process $(X_n)_{n\geq 0}$ with offspring random variable $Z \sim \text{Geom}(p)$. Assume $0 < p < 1$, and set $q = 1 - p$. Recall that $\mathbb{E}(Z) = q/p$. Also, recall from Example 5.1.2 that

$$f_Z(t) = \sum_{k=0}^{\infty} q\, p^k t^k = \frac{q}{1-pt}\,.$$

We compute the extinction probability by solving

$$\frac{q}{1-pt} = t\,,$$

or equivalently,

$$p\,t^2 - t + q = p(t-1)\left(t - \frac{q}{p}\right) = 0\,.$$

It follows that the extinction probability is

$$e_0 = \min(1, q/p) = \begin{cases} q/p & \text{for } p > \frac{1}{2} \\ 1 & \text{for } p \leq \frac{1}{2}\,. \end{cases}$$

□

In the following we take a closer look at the **distribution of the time T until extinction**. Recall that $u_n = \mathbb{P}(T \leq n)$ and $\lim\limits_{n\to\infty} \uparrow u_n = e_0$. Because of the strict convexity of the function f, we have

$$f'(u_{n-1}) < \frac{f(e_0) - f(u_{n-1})}{e_0 - u_{n-1}} < f'(e_0)\,,$$

and so

$$f'(u_{n-1})(e_0 - u_{n-1}) < e_0 - u_n < f'(e_0)(e_0 - u_{n-1})\,. \tag{5.4}$$

Supercritical case

For the supercritical case $m > 1$, we have $e_0 < 1$. Because of the strict convexity of f, we have $f'(e_0) < 1$. The second inequality in (5.4) reads

$$e_0 - \mathbb{P}(T \leq n) < f'(e_0)(e_0 - \mathbb{P}(T \leq n-1) \quad \text{for all } n \geq 1\,. \tag{5.5}$$

From (5.5) we derive by induction,

$$e_0 - \mathbb{P}(T \leq n) < (f'(e_0))^n e_0 \quad \text{for all } n \geq 1$$

or equivalently

$$\mathbb{P}(n < T < \infty) < (f'(e_0))^n e_0 \quad \text{for all } n \geq 1\,.$$

Since $\mathbb{P}(T = \infty) = 1 - e_0 > 0$, we have $\mathbb{E}(T) = \infty$.

Subcritical case

For the subcritical case $m < 1$, we have $e_0 = 1$. By (5.4),

$$1 - \mathbb{P}(T \leq n) < m\,(1 - \mathbb{P}(T \leq n-1)\,.$$

hence

$$\mathbb{P}(T > n) < m\mathbb{P}(T > n-1)\,,$$

which, by induction, yields

$$\mathbb{P}(T > n) < m^n \quad \text{for all } n \geq 0\,.$$

Thus we get the following upper bound for the expected time until extinction

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(T > n) < \sum_{n=0}^{\infty} m^n = \frac{1}{1-m}\,.$$

Critical case

For the critical case $m = 1$, we have $e_0 = 1$, i.e., with probability 1, a sample path of the process will eventually reach the absorbing state 0 (extinction). To determine the expected time $\mathbb{E}(T)$ until extinction, a more subtle analysis of the convergence rate of $f_n(0) \uparrow 1$ is needed. Towards this end, we quote the following result whose proof can be found in [4].

Theorem 5.2.2. *Let $(X_n)_{n\geq 0}$ be a branching process with $X_0 = 1$, Z its offspring random variable, and f_n the probability generating function of X_n. If $\mathbb{E}(Z) = 1$ and $\mathrm{Var}(Z) = \sigma^2 < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n}\left(\frac{1}{1-f_n(t)} - \frac{1}{1-t}\right) = \frac{\sigma^2}{2}$$

uniformly on the interval $[0, 1)$.

We get the following corollary for an asymptotic estimate of the tail probabilities of T:

Corollary 5.2.3. *Let $(X_n)_{n\geq 0}$ be a branching process with $X_0 = 1$ and $\mathbb{E}(Z) = 1$ and $\sigma^2 < \infty$. Then, as $n \to \infty$,*

$$\mathbb{P}(T > n) \sim \frac{2}{n\sigma^2}\,. \tag{5.6}$$

Proof. Recall that $f_n(0) = \mathbb{P}(X_n = 0) = \mathbb{P}(T \leq n)$. Hence

$$1 - f_n(0) = \mathbb{P}(T > n)\,.$$

By Theorem 5.2.2,

$$\lim_{n\to\infty} \frac{2}{n\sigma^2}\left(\frac{1}{1-f_n(0)} - 1\right) = \lim_{n\to\infty} \frac{2/n\sigma^2}{1-f_n(0)} = 1$$

which proves the corollary. □

As a consequence of (5.6), for the critical case $\mathbb{E}(Z) = 1$, we have

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(T > n) = \infty\,.$$

Although extinction happens with probability 1, the expected time until extinction is infinite.

Exercises

Exercise 5.1. Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. Bernoulli random variables with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = 0) = 1 - p$. Furthermore, let N be an a.s. finite, nonnegative, integer-valued random variable independent of the $Y_1, Y_2, \ldots$. Consider the random sum $S_N = Y_1 + \cdots + Y_N$. For each of the following distributions of N, find the probability generating function $f_{S_N}(t)$ and from it, determine the distribution of S_N.

(a) $N \sim \text{Bin}(n, \tilde{p})$ (b) $N \sim \text{Poisson}(\lambda)$

Exercise 5.2. Consider a branching process $(X_n)_{n\geq 0}$ with offspring distribution μ for which $\mu(0) \neq 0$. Assume $X_0 = 1$.

(a) Show that every non-zero state is transient.

(b) Let f_n denote the probability generating function of X_n for $n \geq 1$. Show that there exists a function $f : (0,1) \to [0,\infty)$ such that

$$\lim_{n\to\infty} f_n(t) = f(t) \quad \text{for all } t \in (0,1).$$

Find f.

Exercise 5.3. Let $(X_n)_{n\geq 0}$ be a branching process with geometric offspring distribution μ given by $\mu(n) = \frac{1}{3}(\frac{2}{3})^n$ for $n \geq 0$. Assume $X_0 = 1$.

(a) Compute $\mathbb{P}(X_2 = 1)$.

(b) Compute $\mathbb{P}(X_1 = k \,|\, X_2 = 1)$.

Exercise 5.4. In a branching process $(X_n)_{n\geq 0}$ *with immigration*, a random number of immigrants I_n is independently added to the population at the n^{th} generation for $n \geq 1$. At time 0, the process starts with one individual. The offspring distribution for each individual is given by the offspring random variable Z with probability generating function f_Z. Denote the probability generating function of I_n by f_{I_n}. If f_{X_n} denotes the probability generating function of the size of the n^{th} generation, show that

$$f_{X_n}(t) = f_{X_{n-1}}(f_Z(t)) f_{I_n}(t)\,.$$

Exercise 5.5. Consider a branching process for which the offspring distribution μ is given by $\mu(0) = 1/4$, $\mu(1) = 2/5$, $\mu(2) = 7/20$, and $\mu(n) = 0$ otherwise.

(a) Assume that the process starts in Generation 0 with one individual. What is the probability that the population ultimately dies out?

(b) Assume the process starts with 3 individuals. What is the probability that the population survives forever?

(c) Assume that the process starts with 5 individuals. Compute the probability that the population dies out in the 3^{rd} generation.

(d) Assume that the process starts with one individual. Compute the probability that the population dies out in the 3^{rd} generation, given that it is not already extinct in the 2^{nd} generation.

Exercise 5.6. Consider a branching process $(X_n)_{n\geq 0}$ with offspring random variable Z whose probability generating function is f_Z. Let

$$Y_n = X_0 + X_1 + \cdots + X_n$$

be the total number of individuals up through Generation n.

(a) Prove that the probability generating functions f_{Y_n} satisfy the recurrence relation

$$f_{Y_n}(t) = t f_Z(f_{Y_{n-1}}(t)) \quad \text{for} \quad n \geq 1\,.$$

(b) Assume the branching process is subcritical. Consider $Y = \lim_{n\to\infty} Y_n$, the *total progeny* of the branching process. Show that

$$f_Y(t) = t f_Z(f_Y(t))$$

and compute $\mathbb{E}(Y)$. From the value of $\mathbb{E}(Y)$, conclude (once more) that for a subcritical branching process, extinction happens in finite time with probability 1.

Exercise 5.7. Let $(X_n)_{n\geq 0}$ be a supercritical branching process with geometric offspring distribution μ given by $\mu(n) = (1-p)p^n$ for $n \geq 0$. Assume $X_0 = 1$. Consider the following variation of this process: Each individual, independently of all other individuals, is given a survival probability of $\tilde{p}$. That is, each individual will die with probability $(1-\tilde{p})$ before it has a chance to produce offspring for the next generation. This defines a new branching process $(Y_n)_{n\geq 0}$. Find a condition on $\tilde{p}$ (in terms of p) under which the population of $(Y_n)_{n\geq 0}$ will ultimately die out with probability 1.

Exercise 5.8. Consider a branching process $(X_n)_{n\geq 0}$ with offspring distribution μ defined by $\mu(k) = (1/2)^{k+1}$ for $k \geq 0$. The process starts with one individual.

(a) Show that the probability generating function of the size of the n^{th} generation is

$$f_{X_n}(t) = \frac{n-(n-1)t}{n+1-nt}\,.$$

(b) Let T be the time of extinction. Is T an a.s. finite random variable? Compute the distribution of T.

Exercise 5.9. Consider the same branching process as in Exercise 5.8. Let N_k denote the total number of generations for which the population size is exactly k individuals. Compute $\mathbb{E}(N_1)$. (*Hint:* $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$)

Exercise 5.10. Again, consider a branching process $(X_n)_{n\geq 0}$ with offspring distribution μ defined by $\mu(k) = (1/2)^{k+1}$ for $k \geq 0$, and assume the process starts with one individual. In this case however, if the population ever dies out, then a single new individual will be added from outside. Note that this change makes the branching process an irreducible Markov chain. Is this Markov chain transient, positive recurrent, or null recurrent?

Exercise 5.11. Consider the (explosive) branching process $(X_n)_{n\geq 0}$ with offspring distribution ν defined by $\nu(k) = (1/2)^k$ for $k \geq 1$.

(a) For $n \geq 1$, compute a formula for the probability generating function $f_{X_n}(t)$.

(b) Does $\lim_{n\to\infty} f_{X_n}(t)$ exist for all $t \in (0,1)$? If so, what is the limit function?

Exercise 5.12. Fix $0 < p < 1$. Consider a branching process $(X_n)_{n\geq 0}$ with offspring distribution μ defined by $\mu(0) = 1 - p$ and $\mu(1) = p$. Assume the process starts with $k_0 > 0$ particles. Let T be the time of extinction of the process. Compute the distribution of T.

Exercise 5.13. Consider a branching process $(X_n)_{n\geq 0}$ whose offspring distribution has probability mass function $\mu(0) = (1 - q)$, and $\mu(1) = \mu(2) = \frac{q}{2}$ for some $q \in (0, 1)$.

(a) For which values of q will the population eventually die out with probability one?

(b) Assume $q = \frac{4}{5}$ and, furthermore, the process does not start with a fixed number of individuals, but the distribution of X_0 follows a Poisson distribution with parameter $\lambda = 2$. Find $\mathbb{P}(X_1 = 1)$.

Exercise 5.14. Let $(X_n)_{n\geq 0}$ be a supercritical branching process with offspring distribution μ for which $\mu(0) \neq 0$. Denote the probability of ultimate extinction of the process by e_0. Now consider the sequence

$$\tilde{\mu}(k) = e_0^{k-1}\mu(k) \quad \text{for } k \geq 0.$$

(a) Show that $\tilde{\mu}(k)$, $k \geq 0$, defines a probability distribution on $\mathbb{N}_0$.

(b) Show that the expectation of of the distribution $\tilde{\mu}$ on $\mathbb{N}_0$ is strictly less than 1.

Exercise 5.15. Consider a supercritical branching process $(X_n)_{n\geq 0}$ whose offspring distribution is Poisson(λ) with $\lambda > 1$. Let e_0 be the probability of ultimate extinction for the process.

(a) Let $\tilde{\mu}(k) = e_0^{k-1}\mu(k)$, $k \geq 0$, as in Exercise 5.14. What kind of distribution is $\tilde{\mu}$?

(b) Prove that

$$e_0 < \frac{1}{\lambda}.$$

(*Hint*: Use the result from Exercise 5.14 part (b).)

Exercise 5.16. Let $(X_n)_{n\geq 0}$ be a supercritical branching process with offspring distribution μ for which $\mu(0) \neq 0$ and denote its probability of ultimate extinction by e_0. Consider a second branching process $(\tilde{X}_n)_{n\geq 0}$ with offspring distribution $\tilde{\mu}$ defined by $\tilde{\mu}(k) = e_0^{k-1}\mu(k)$, $k \geq 0$. By the result from Exercise 5.14 part (b), the branching process $(\tilde{X}_n)_{n\geq 0}$ is subcritical.

(a) Let f_n denote the probability generating function of X_n, and let $\tilde{f}_n$ denote the probability generating function of $\tilde{X}_n$ for $n \geq 1$. Prove by induction that for all $n \geq 1$,

$$\tilde{f}_n(t) = \frac{1}{e_0} f_n(e_0 t).$$

(b) Let A be the event "the population of process $(X_n)_{n\geq 0}$ ultimately becomes extinct". So $\mathbb{P}(A) = e_0$. Use your result from part (a) to show that for all $n \geq 1$ and $k \geq 0$,

$$\mathbb{P}(\tilde{X}_n = k) = \mathbb{P}(X_n = k \mid A).$$

In other words, the distribution of the branching process $(\tilde{X}_n)_{n\geq 0}$ is the same as the distribution of the branching process $(X_n)_{n\geq 0}$, conditioned on the event that its population ultimately becomes extinct.

Chapter 6

Martingales

6.1 Definition of a Martingale

Martingales constitute an important class of stochastic processes. They are defined in terms of certain conditional expectations for their variables. We give a precise definition below. Although the dependence structure for the variables of a martingale is very different from that of a Markov chain, there is also some overlap between these two types of processes, with discrete harmonic functions playing a role in connecting the two structures. Under certain conditions, martingales can also be viewed as a generalization of random walks on $\mathbb{R}$ with mean-zero step distribution (see Proposition 6.1.2 below). Often times, as we will illustrate in examples, a question of interest for a given Markov chain can be rephrased as a question for a suitably chosen, related martingale and thus solved using the rich mathematical theory that is available for martingales.

Definition 6.1.1. *Let $(M_n)_{n\geq 0}$ and $(X_n)_{n\geq 0}$ be two real-valued stochastic processes on the same probability space. We say that $(M_n)_{n\geq 0}$ is a* **martingale with respect to** *$(X_n)_{n\geq 0}$ if for all $n \geq 0$,*

(a) M_n is a function of $X_0, ..., X_n$, and

(b) $\mathbb{E}(|M_n|) < \infty$, and

(c) $\mathbb{E}(M_{n+1} \,|\, X_0, ..., X_n) = M_n \quad a.s.$

We say that $(M_n)_{n\geq 0}$ is a **supermartingale** *if (c) is replaced by*

$$\mathbb{E}(M_{n+1} \,|\, X_0, ..., X_n) \leq M_n \qquad a.s.,$$

and $(M_n)_{n\geq 0}$ is a **submartingale** *if (c) is replaced by*

$$\mathbb{E}(M_{n+1} \,|\, X_0, ..., X_n) \geq M_n \qquad a.s.$$

If a process $(M_n)_{n\geq 0}$ satisfies condition (a) in Definition 6.1.1, we say $(M_n)_{n\geq 0}$ is **adapted to the process** $(X_n)_{n\geq 0}$. The following is a special case of Definition 6.1.1.

Definition 6.1.2. *Let $(M_n)_{n\geq 0}$ be a real-valued stochastic process. We say that $(M_n)_{n\geq 0}$ is a* **martingale** *if for all $n \geq 0$,*

(a) $\mathbb{E}(|M_n|) < \infty$, and

(b) $\mathbb{E}(M_{n+1} \,|\, M_0, ..., M_n) = M_n \quad a.s.$

Example 6.1.1. Consider a sequence $X_1, X_2, ...$ of *independent* random variables with $\mathbb{E}(|X_k|) < \infty$ and $\mathbb{E}(X_k) = 0$ for all $k \geq 1$. Then the process $(M_n)_{n\geq 1}$ defined by the partial sums $M_n = \sum_{k=1}^{n} X_k$ for $n \geq 1$ is a martingale with respect to $(X_n)_{n\geq 1}$: Clearly, for all $n \geq 1$, M_n is a function of $X_1, ..., X_n$. By the triangle inequality and the integrability assumption on the X_k,

$$\mathbb{E}(|M_n|) \leq \sum_{k=1}^{n} \mathbb{E}(|X_k|) < \infty\,,$$

and by independence of the X_k,

$$\begin{aligned} \mathbb{E}(M_{n+1} \,|\, X_1, ..., X_n) &= \mathbb{E}(M_n + X_{n+1} \,|\, X_1, ..., X_n) \\ &= M_n + \mathbb{E}(X_{n+1}) \\ &= M_n \end{aligned}$$

for all $n \geq 1$.

Example 6.1.2 (Random walk)**.** Fix $x \in \mathbb{R}$. As a special case of Example 6.1.1, consider the constant random variable $X_0 \equiv x$ and an *i.i.d.* sequence of random variables $X_1, X_2, ...$ with $\mathbb{E}(X_k) = m < \infty$. Set $\tilde{X}_k = X_k - m$. The process $(S_n)_{n\geq 0}$ with $S_n = X_0 + \tilde{X}_1 + \cdots + \tilde{X}_n$ is random walk on $\mathbb{R}$ with mean-zero step size. The process $(M_n)_{n\geq 0}$ defined by

$$M_n = S_n = \sum_{k=0}^{n} X_k - n\,m$$

is a martingale with respect to $(X_n)_{n\geq 0}$. □

Example 6.1.3. We consider random walk on $\mathbb{Z}$. Assume $X_0 = x$, $\mathbb{E}(X_k) = 0$, and $\mathrm{Var}(X_k) = \sigma^2 < \infty$. Then the process $(M_n)_{n\geq 0}$ defined by

$$M_n = (S_n)^2 - \sigma^2\, n$$

is a martingale with respect to $(X_n)_{n\geq 0}$. Clearly, for all $n \geq 0$, M_n is a function of $X_0, X_1, ..., X_n$. By the triangle inequality, and since $\mathbb{E}(X_k) = 0$,

$$\mathbb{E}(|M_n|) \leq \mathbb{E}(S_n^2) + \sigma^2\, n = x^2 + n\sigma^2 + \sigma^2\, n = x^2 + 2n\sigma^2 < \infty\,.$$

We have

$$\begin{aligned}
\mathbb{E}(S_{n+1}^2 \,|\, X_0, X_1, ..., X_n) &= \mathbb{E}\left((\sum_{i=0}^{n+1} X_i)^2 \,|\, X_0, X_1, ..., X_n \right) \\
&= \mathbb{E}\left(\sum_{i=0}^{n+1} X_i^2 + \sum_{0\leq i<j\leq n+1} 2X_iX_j \,|\, X_0, X_1, ..., X_n \right) \\
&= \sigma^2 + \sum_{i=0}^{n} X_i^2 + \sum_{0\leq i<j\leq n} 2X_iX_j = \sigma^2 + S_n^2\,.
\end{aligned}$$

It follows that

$$\mathbb{E}(M_{n+1} \,|\, X_0, X_1, ..., X_{n+1}) = \sigma^2 + S_n^2 - (n+1)\sigma^2 = S_n^2 - n\,\sigma^2 = M_n\,,$$

and so $(M_n)_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$. □

Example 6.1.4. Consider simple random walk on $\mathbb{Z}$ with step random variable X_i with $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = -1) = q$, and $\mathbb{P}(X_i = 0) = r$. Assume $0 < p < 1$, and $p+q+r = 1$. The process $(M_n)_{n\geq 0}$ defined by

$$M_n = (\frac{q}{p})^{S_n} \quad \text{for } n \geq 0$$

is a martingale with respect to $(X_n)_{n\geq 1}$. When $p = q$, this is the constant process $M_n \equiv 1$. Assume $p \neq q$. It is clear that parts (a) and (b) from Definition 6.1.1 hold. To check (c), consider

$$\mathbb{E}(M_{n+1} \,|\, X_1, ..., X_n) = \mathbb{E}\left((\frac{q}{p})^{S_n} (\frac{q}{p})^{X_{n+1}} \,|\, X_1, ..., X_n \right) .$$

Invoking the independence of the X_i, we get

$$\begin{aligned} \mathbb{E}(M_{n+1} \,|\, X_1, ..., X_n) &= (\tfrac{q}{p})^{S_n} \mathbb{E}\left((\tfrac{q}{p})^{X_{n+1}} \right) \\ &= (\tfrac{q}{p})^{S_n} \left[p\tfrac{q}{p} + q(\tfrac{q}{p})^{-1} + r \cdot 1 \right] = (\tfrac{q}{p})^{S_n} = M_n \,. \end{aligned}$$

□

Note that a martingale is both a submartingale and a supermartingale.

Proposition 6.1.1. *If $(M_n)_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$, then for all $n, k \geq 0$,*

$$\mathbb{E}(M_{n+k} \,|\, X_0, ..., X_n) = M_n$$

and, consequently,

$$\mathbb{E}(M_n) = \mathbb{E}(M_0)$$

for all $n \geq 0$.

Proof. Let $(M_n)_{n\geq 0}$ be a martingale with respect to $(X_n)_{n\geq 0}$. For any $k \geq 1$, by the tower property of conditional expectation,

$$\mathbb{E}(M_{n+k} \,|\, X_0, ..., X_n) = \mathbb{E}[\mathbb{E}(M_{n+k} \,|\, X_0, ..., X_{n+k-1}) \,|\, X_0, ..., X_n] \,. \tag{6.1}$$

By Definition 6.1.1(c), the right-hand side of (6.1) is $\mathbb{E}(M_{n+k-1} \,|\, X_0, ..., X_n)$. By iteration, we get

$$\mathbb{E}(M_{n+k-1} \,|\, X_0, ..., X_n) = ... = \mathbb{E}(M_{n+1} \,|\, X_0, ..., X_n) = M_n \,.$$

Thus $\mathbb{E}(M_{n+k} \,|\, X_0, ..., X_n) = M_n$ for all $n, k \geq 0$. Taking expectations yields

$$\mathbb{E}(M_n) = \mathbb{E}(M_0) \quad \text{for all } n \geq 0 \,.$$

□

Example 6.1.5. Let $(X_n)_{n\geq 0}$ be *finite* state Markov chain with state space $\mathcal{S} \subset \mathbb{R}$ and assume $(X_n)_{n\geq 0}$ is also a martingale. Thus we have

$$\sum_{k\in\mathcal{S}} P_{ik} k = i \quad \text{for all } i \in \mathcal{S} \,. \tag{6.2}$$

Since $\mathcal{S}$ is finite, there exists a smallest state a and a largest state b in $\mathcal{S}$. It follows from (6.2) that $P_{aa} = 1$ and $P_{bb} = 1$, so a and b are absorbing states. Assume that all other states in $\mathcal{S}$ are transient. With probability 1, the process will get absorbed in either state a or state b. By Proposition 6.1.1, we have

$$\sum_{k\in\mathcal{S}} P_{ik}^n k = i \quad \text{for all } i \in \mathcal{S} \text{ and for all } n \geq 0\,. \tag{6.3}$$

Consider the absorption probability $a_{i,b} = \mathbb{P}(T^b < \infty \,|\, X_0 = i)$, that is, the probability that the process ultimately gets absorbed in state b, given that the Markov chain started in state i. Recall that for any transient state j, we have $\lim_{n\to\infty} P_{ij}^n = 0$. Taking the limit of the sum in (6.3) yields

$$\lim_{n\to\infty} \sum_{k\in\mathcal{S}} P_{ik}^n k = a(1 - a_{i,b}) + b\, a_{i,b} = i\,,$$

and from this we get

$$a_{i,b} = \frac{i-a}{b-a}\,. \tag{6.4}$$

Notice that formula (6.4) matches formula (4.18) for the gambler's ruin probabilities. □

Example 6.1.6 (Fixation probabilities for the Wright-Fisher model)**.** Recall the Wright-Fisher model for genetic drift that was introduced in Section 1.5. It is a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{0, 1, ..., N\}$ where N is the population size. X_n is the count of allele A in Generation n. The transition probabilities P_{kl} follow a binomial distribution:

$$P_{kl} = \binom{N}{l} (\frac{k}{N})^l (1 - \frac{k}{N})^{N-l} \quad \text{for } k, l = 0, 1, ..., N\,.$$

States 0 and N are absorbing, and all other states lead into 0 and N. The Wright-Fisher model is also a martingale:

$$\mathbb{E}(|X_n|) \leq N\,,$$

and

$$\mathbb{E}(X_{n+1} \,|\, X_0, X_1, ..., X_n) = \mathbb{E}(X_{n+1} \,|\, X_n) = N\frac{X_n}{N} = X_n\,.$$

Thus formula (6.4) applies. Given that the chain starts in state i, the **fixation probability** $a_{iN} = \mathbb{P}(T^N < \infty \,|\, X_0 = i)$ is

$$a_{iN} = \frac{i}{N}\,.$$

It is the probability that eventually the gene pool gets fixated in A, that is, all N individuals in the population have genotype A. The complementary probability $a_{i0} = 1 - a_{iN}$ is

$$a_{i0} = \frac{N-i}{N}\,.$$

It is the probability that eventually the gene pool gets fixated in a, that is, all N individuals in the population have genotype a.
If a new mutation arises in a population, its original count is 1. And so the probability that this newly arisen genotype eventually fixates is $1/N$. □

Example 6.1.7 (Fixation probabilities for the Moran model)**.** Recall the Moran model for genetic drift that was introduced in Section 1.5. It is a Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{0, 1, ..., N\}$. X_n gives the allele A count in the population at time n. The transition probabilities are

$$P_{x,x+1} = P_{x,x-1} = \frac{(N-x)x}{N^2} \qquad P_{xx} = 1 - 2\frac{(N-x)x}{N^2}\,,$$

and 0 and N are absorbing states. The Moran model $(X_n)_{n\geq 0}$ is a martingale. Indeed,

$$\mathbb{E}(|X_n|) \leq N\,,$$

and

$$\begin{aligned}\mathbb{E}(X_{n+1}\,|\,X_0, X_1, ..., X_n) &= \mathbb{E}(X_{n+1}\,|\,X_n)\\ &= (X_n+1)p + (X_n-1)p + X_n(1-2p) = X_n\end{aligned}$$

where we have used $p = \frac{(N-X_n)X_n}{N^2}$.
Thus we get the same **fixation probabilities** $a_{iN} = \mathbb{P}(T^N < \infty\,|\,X_0 = i)$ as for the Wright-Fisher model:

$$a_{iN} = \frac{i}{N} \quad \text{and} \quad a_{i0} = \frac{N-i}{N}\,.$$

□

The simplest example of a martingale is random walk $(S_n)_{n\geq 0}$ where $S_0 = x$ and $S_n = \sum_{k=0}^{n} X_k$ for i.i.d. random variables $X_1, X_2, ...$ with $\mathbb{E}(X_1) = 0$. See Example 6.1.2. The following shows that under the condition of *square-integrability*, a martingale $(M_n)_{n\geq 0}$ retains some essential features of random walk with mean-zero steps.
Let $(M_n)_{n\geq 0}$ be a martingale. Using the notation $D_n = M_n - M_{n-1}$, we have

$$M_n = M_0 + \sum_{k=1}^{n} D_k \qquad \text{for } n \geq 1\,.$$

By Proposition 6.1.1, $\mathbb{E}(D_k) = 0$ for $k \geq 1$.

Definition 6.1.3. *We say the martingale $(M_n)_{n\geq 0}$ is square-integrable if*

$$\mathbb{E}(M_n^2) < \infty$$

for all $n \geq 0$.

Proposition 6.1.2. *Let $(M_n)_{n\geq 0}$ be a square-integrable martingale. Then its increments D_k are uncorrelated, that is,*

$$\mathrm{Cov}(D_k D_l) = 0 \qquad \text{for } k \neq l$$

and therefore

$$\mathrm{Var}(M_n) = \mathrm{Var}(M_0) + \sum_{k=1}^{n} \mathrm{Var}(D_k)\,.$$

Proof. Since $\mathbb{E}(D_k) = 0$ for $k \geq 1$, we only need to show that $\mathbb{E}(D_k D_l) = 0$ for $k < l$. Write

$$\mathbb{E}(D_k D_l) = \mathbb{E}\left[\mathbb{E}(D_k D_l \,|\, X_0, ..., X_{l-1})\right]\,.$$

But

$$\mathbb{E}(D_k D_l \,|\, X_0, ..., X_{l-1}) = D_k \mathbb{E}(D_l \,|\, X_0, ..., X_{l-1}) = 0$$

since, by the martingale property, $\mathbb{E}(D_l \,|\, X_0, ..., X_{l-1}) = M_{l-1} - M_{l-1} = 0$. It follows that $\mathbb{E}(D_k D_l) = 0$ for $k < l$. □

6.2 Optional Stopping Theorem

For convenience, we recall the definition of a stopping time T for a stochastic process $(X_n)_{n\geq 0}$.

Definition 6.2.1. *Let $(X_n)_{n\geq 0}$ be a stochastic process. A random variable T defined on the same probability space as $(X_n)_{n\geq 0}$ and taking values in $\mathbb{N}_0 \cup \{\infty\}$ is called a* **stopping time** *T if for all $m \in \mathbb{N}_0$, the event $\{T = m\}$ can be determined by the values of $X_0, X_1, ..., X_m$.*

Whether or not a stopping time T has occurred at time m only depends on the history of the process $(X_n)_{n\geq 0}$ up to (including) time m. The event $\{T = m\}$ is independent of

any event determined by the future of the process, that is, by $X_{m+1}, X_{m+2}, ...$ Note that the definition of T implies that the events $\{T \leq m\}$ and $\{T > m\}$ are also determined by the values of $X_0, X_1, ..., X_m$. In the following we will use the notation $a \wedge b = \min(a, b)$.

Theorem 6.2.1. *Let $(M_n)_{n\geq 0}$ be a martingale with respect to $(X_n)_{n\geq 0}$ and T a stopping time. Then the* **stopped process** *$(M_{n\wedge T})_{n\geq 0}$ defined by*

$$M_{n\wedge T} = \begin{cases} M_n & \text{if } n < T\,, \\ M_T & \text{if } n \geq T \end{cases}$$

is also a martingale with respect to $(X_n)_{n\geq 0}$.

Proof. Fix $n > 0$. We have

$$|M_{n\wedge T}| \leq \max_{0\leq k\leq n} |M_k| \leq |M_0| + \cdots + |M_n|\,,$$

so $\mathbb{E}(|M_{n\wedge T}|) < \infty$.
Note that for $\omega \in \{n < T\}$, we have $M_n(\omega) = M_{n\wedge T}(\omega)$ as well as $M_{n+1}(\omega) = M_{(n+1)\wedge T}(\omega)$. And for $\omega \in \{n \geq T\}$, we have $M_{n\wedge T}(\omega) = M_{(n+1)\wedge T}(\omega)$. So we can write

$$M_{(n+1)\wedge T} = M_{n\wedge T} + (M_{n+1} - M_n)\mathbb{1}_{\{n<T\}}\,. \tag{6.5}$$

Taking the conditional expectation with respect to $X_0, ..., X_n$ on both sides in (6.5) yields

$$\mathbb{E}(M_{(n+1)\wedge T} \,|\, X_0, ..., X_n) = \mathbb{E}(M_{n\wedge T} \,|\, X_0, ..., X_n) + \mathbb{E}((M_{n+1} - M_n)\mathbb{1}_{\{n<T\}} \,|\, X_0, ..., X_n)\,.$$

Since T is a stopping time, the event $\{n < T\}$ is determined by the values of $X_0, ..., X_n$. Using the martingale property for $(M_n)_{n\geq 0}$, we get

$$\begin{aligned} \mathbb{E}(M_{(n+1)\wedge T} \,|\, X_0, ..., X_n) &= M_{n\wedge T} + \mathbb{1}_{\{n<T\}}\mathbb{E}((M_{n+1} - M_n) \,|\, X_0, ..., X_n) \\ &= M_{n\wedge T} + \mathbb{1}_{\{n<T\}}0 = M_{n\wedge T}\,. \end{aligned}$$

□

It follows from Theorem 6.2.1 that $\mathbb{E}(M_{n\wedge T}) = \mathbb{E}(M_0)$ for all $n \geq 0$. However, in general, it is *not* true that $\mathbb{E}(M_T) = \mathbb{E}(M_0)$. This is illustrated by the following example.

Example 6.2.1. A gambler starts with a fortune of $\$N$ and makes a sequence of fair \$1 bets against an infinitely rich adversary. The gambler's fortune can be described by simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{N}_0$ with absorbing boundary at 0. It is a

martingale. Let T be the first hitting time of state 0 (bancruptcy) which is a stopping time for the random walk. Since $(S_n)_{n\geq 0}$ is recurrent, $\mathbb{P}(T<\infty)=1$. We have

$$\mathbb{E}(S_0) = N \neq 0 = \mathbb{E}(S_T)\,.$$

□

Given a martingale $(M_n)_{n\geq 0}$ and a stopping time T, we would like to know conditions that guarantee $\mathbb{E}(M_T)=\mathbb{E}(M_0)$. This is the content of the Optional Stopping Theorem (OST). The following version of the OST is not its most general version, but it is sufficient for our purposes for applications to Markov chains. It gives a variety of (often fairly easy to check) sufficient conditions under which $\mathbb{E}(M_T)=\mathbb{E}(M_0)$ holds.

Theorem 6.2.2 (Optional Stopping Theorem)**.** *Let $(M_n)_{n\geq 0}$ be a martingale with respect to $(X_n)_{n\geq 0}$ and T a stopping time for $(X_n)_{n\geq 0}$. Assume at least one of the conditions hold:*

(a) There exists an N_0 such that $\mathbb{P}(T\leq N_0)=1$;

(b) $\mathbb{P}(T<\infty)=1$ and there exists an $K_0<\infty$ such that $\mathbb{P}(|M_n|\leq K_0)=1$ if $n\leq T$;

(c) $\mathbb{E}(T)<\infty$ and there exists $K_0<\infty$ such that

$$\mathbb{E}(|M_n - M_{n-1}|\,|X_0, X_1, ..., X_{n-1}) \leq K_0$$

for $n\leq T$.

Then

$$\mathbb{E}(M_T)=\mathbb{E}(M_0)\,.$$

Proof. (a) Since $\mathbb{P}(T\leq N_0)=1$, we can write

$$M_T = M_0 + \sum_{k=0}^{N_0-1}(M_{k+1}-M_k)\mathbb{1}_{\{k<T\}}\,. \tag{6.6}$$

Since T is a stopping time, the event $\{k<T\}$ is determined by $X_0, ..., X_k$. So

$$\begin{aligned}\mathbb{E}[(M_{k+1}-M_k)\mathbb{1}_{\{k<T\}}] &= \mathbb{E}[\mathbb{E}[(M_{k+1}-M_k)\mathbb{1}_{\{k<T\}}\,|\,X_0,...,X_k]]\\ &= \mathbb{E}[\mathbb{1}_{\{k<T\}}\mathbb{E}[(M_{k+1}-M_k)\,|\,X_0,...,X_k]]\,.\end{aligned}$$

By the martingal property, $\mathbb{E}[(M_{k+1}-M_k)\mid X_0,...,X_k]]=0$. So (6.6) yields

$$\mathbb{E}(M_T)=\mathbb{E}(M_0)\,.$$

(b) For $n \in \mathbb{N}$ consider the stopping time $T \wedge n = \min(n, T)$. Since $T \wedge n$ is a bounded stopping time, the conditions from part (a) are fulfilled, and we have

$$\mathbb{E}(M_{T\wedge n}) = \mathbb{E}(M_0) \quad \text{for all } n \in \mathbb{N}\,.$$

Since $\mathbb{P}(|M_n| \leq K_0) = 1$, if $n \leq T$, we have

$$|\mathbb{E}(M_T) - \mathbb{E}(M_0)| = |\mathbb{E}(M_T) - \mathbb{E}(M_{T\wedge n})| \leq 2K_0\mathbb{P}(T > n)\,.$$

Since $\mathbb{P}(T < \infty) = 1$, we have $\lim_{n\to\infty} \mathbb{P}(T > n) = 0$. This proves

$$\mathbb{E}(M_T) = \mathbb{E}(M_0)\,.$$

(c) As in part (b), consider the bounded stopping time $T \wedge n = \min(n, T)$. By part (a) we have

$$\mathbb{E}(M_{T\wedge n}) = \mathbb{E}(M_0) \quad \text{for all } n \in \mathbb{N}\,.$$

Recall that we can write the random variable M_T as

$$M_T = M_0 + \sum_{k=0}^{T-1}(M_{k+1} - M_k) = M_0 + \sum_{k=0}^{\infty}(M_{k+1} - M_k)\mathbb{1}_{\{k<T\}}\,.$$

So

$$\lim_{n\to\infty} M_{T\wedge n} = \lim_{n\to\infty}\left[M_0 + \sum_{k=0}^{n-1}(M_{k+1} - M_k)\mathbb{1}_{\{k<T\}}\right] = M_T \qquad a.s.$$

We now introduce the random variable

$$\tilde{M}_T := |M_0| + \sum_{k=0}^{\infty}|M_{k+1} - M_k|\,\mathbb{1}_{\{k<T\}}\,.$$

Note that

$$|M_{T\wedge n}| \leq \tilde{M}_T \quad a.s.$$

By the Monotone Convergence Theorem (see Appendix C),

$$\mathbb{E}(\tilde{M}_T) = \mathbb{E}(|M_0|) + \sum_{k=0}^{\infty}\mathbb{E}\left(|M_{k+1} - M_k|\,\mathbb{1}_{\{k<T\}}\right)\,.$$

For each individual term in the previous sum we have

$$\begin{aligned}\mathbb{E}(|M_{k+1} - M_k|\mathbb{1}_{\{k<T\}}) &= \mathbb{E}[\mathbb{E}(|M_{k+1} - M_k|\mathbb{1}_{\{k<T\}}\,|\,X_0, ..., X_k)] \\ &= \mathbb{E}(\mathbb{1}_{\{k<T\}})\mathbb{E}[|M_{k+1} - M_k|\,|\,X_0, ..., X_k)] \\ &\leq \mathbb{P}(k < T)\,K_0\,.\end{aligned}$$

Thus, since $\mathbb{E}(T) = \sum_{k=0}^{\infty} \mathbb{P}(k < T) < \infty$, we get

$$\mathbb{E}(\tilde{M}_T) \leq \mathbb{E}(M_0) + K_0\, \mathbb{E}(T) < \infty\,.$$

Using the Dominated Convergence Theorem (using $\tilde{M}_T$ as the dominating random variable), we conclude

$$\lim_{n\to\infty} \mathbb{E}(M_{T\wedge n}) = \mathbb{E}(M_T)\,.$$

Since $\mathbb{E}(M_{T\wedge n}) = \mathbb{E}(M_0)$ for all n, we arrive at

$$\mathbb{E}(M_T) = \mathbb{E}(M_0)$$

which concludes the proof of part (c). □

As an immediate corollary to the OST, we will reprove Wald's first equation (see Theorem 4.3.1(a)).

Corollary 6.2.3 (Wald's first equation, revisited)**.** *Let $X_1, X_2, ...$ be i.i.d. random variables with $\mathbb{E}(|X_i|) < \infty$. Consider $(S_n)_{n\geq 1}$ with $S_n = \sum_{i=1}^n X_i$, and let T be a stopping time for $(S_n)_{n\geq 1}$. Set $m = \mathbb{E}(X_i)$. If $\mathbb{E}(T) < \infty$, then*

$$\mathbb{E}(S_T) = m\,\mathbb{E}(T)\,.$$

Proof. We have shown in Example 6.1.2 that the process

$$M_n = S_n - n\,m \quad \text{for } n \geq 1$$

is a martingale. We will show that conditions (c) of the OST hold. Indeed, since

$$|M_n - M_{n-1}| = |S_n - S_{n-1} - m| = |X_n - m|\,,$$

we have

$$\mathbb{E}(|M_n - M_{n-1}|\,|\,X_1, ..., X_{n-1}) = \mathbb{E}(|X_n - m|) \leq \mathbb{E}(|X_1|) + m < \infty$$

for all $n \geq 1$. Applying the OST, we get

$$\mathbb{E}(M_T) = \mathbb{E}(S_T - Tm) = \mathbb{E}(M_1) = \mathbb{E}(X_1) - m = 0\,,$$

and so

$$\mathbb{E}(S_T) = m\,\mathbb{E}(T)\,.$$

□

6.3 Martingale transforms

An alternate title for this section could be *There is no system for beating a fair game* (if funds are limited). Let $(M_n)_{n\geq 0}$ be a martingale. We can think of M_n as the capital at time n of a player in a fair game. As above, we use the notation $D_k = M_k - M_{k-1}$, and can write

$$M_n = \sum_{k=1}^{n} D_k \, .$$

Definition 6.3.1. *Let $(M_n)_{n\geq 0}$ be a martingale with respect to a process $(X_n)_{n\geq 0}$. A sequence of random variables $(H_n)_{n\geq 1}$ is called a* **predictable process** *if, for each $n \geq 1$, H_n is a function of $X_0, ..., X_{n-1}$.*

For $n \geq 1$, we can think of the random variable H_n as representing the stake the gambler puts on Game n, knowing the outcomes of Games $1, ..., n-1$ and the original capital M_0. So $H_n(M_n - M_{n-1}) = H_n \, D_n$ is the profit (or loss) of the gambler for the nth game, and

$$C_n = M_0 + \sum_{k=1}^{n} H_n \, D_n$$

is the total capital of the gambler immediately after the nth game.

Notation: We will write

$$(M \cdot H)_n = M_0 + \sum_{k=1}^{n} H_k \, D_k \, ,$$

and

$$(M \cdot H) = M_0 + \sum_{k=1}^{\infty} H_k \, D_k \, .$$

Definition 6.3.2. *The process $(M \cdot H)_n, n \geq 0$, is called the* **martingale transform** *or* **discrete-time stochastic integral** *of $(M_n)_{n\geq 0}$ by $(H_n)_{n\geq 1}$.*

Proposition 6.3.1. *Let $(M_n)_{n\geq 0}$ be a martingale and $(H_n)_{n\geq 1}$ a predictable process. If there exists $K_0 > 0$ such that $|H_n| \leq K_0$ for all $n \geq 1$, then*

$$(M \cdot H)_n, \ n \geq 0$$

is a martingale.

Proof. Clearly, $(M \cdot H)_n = M_0 + \sum_{k=1}^n H_k\, D_k$ is measurable with respect to $\sigma(X_0, ..., X_n)$ for all $n \geq 0$. We also have

$$\begin{aligned} \mathbb{E}(|(M \cdot H)_n|) &\leq \mathbb{E}\left[|M_0| + \sum_{k=1}^n |H_k|\,|D_k|\right] \\ &\leq \mathbb{E}(|M_0|) + \textstyle\sum_{k=1}^n K_0\ (|M_n| + |M_{n-1}|) < \infty\,. \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathbb{E}((M \cdot H)_n \,|\, X_0, ..., X_{n-1}) &= \mathbb{E}\left(M_0 + \sum_{k=1}^n H_k\,(M_k - M_{k-1}) \,|\, X_0, ..., X_{n-1}\right) \\ &= M_0 + \sum_{k=1}^{n-1} H_k\,(M_k - M_{k-1}) + H_n\,\mathbb{E}(M_n - M_{n-1} \,|\, X_0, ..., X_{n-1}) \\ &= M_0 + \sum_{k=1}^{n-1} H_k\,(M_k - M_{k-1}) + H_n\,0 = (M \cdot H)_{n-1} \end{aligned}$$

which proves that $(M \cdot H)_n,\ n \geq 0$, is a martingale. □

As a corollary, we can recover part (a) from the Optional Stopping Theorem.

Corollary 6.3.2. *Let $(M_n)_{n\geq 0}$ be a martingale and T a stopping time. Then for all $n_0 \in \mathbb{N}$,*

$$\mathbb{E}(M_{T\wedge n_0}) = \mathbb{E}(M_0)\,.$$

In particular, if T is bounded, we have $\mathbb{E}(M_T) = \mathbb{E}(M_0)$.

Proof. Let $(H_n)_{n\geq 1}$ be the process defined by

$$H_n = \mathbb{1}_{\{T\geq n\}} = \begin{cases} 1 & \text{if } T \geq n \\ 0 & \text{if } T < n\,. \end{cases}$$

Since the event $\{T = n\} \in \sigma(X_0, ..., X_n)$, the event

$$\{T \geq n\} = \bigcap_{k=0}^{n-1} \{T = k\}^c$$

and therefore

$$\{T \geq n\} \in \sigma(X_0, ..., X_{n-1})\,.$$

Thus $(H_n)_{n\geq 1}$ is a predictable process. Note that

$$\begin{aligned} (H \cdot M)_{n_0} &= M_0 + \sum_{k=1}^{n_0} \mathbb{1}_{\{T\geq n_0\}}(M_k - M_{k-1}) \\ &= M_0 + \sum_{k=1}^{T\wedge n_0} (M_k - M_{k-1}) \\ &= M_{T\wedge n_0}\,. \end{aligned}$$

Since $(H \cdot M)_n$ is a martingale, we have $\mathbb{E}((H \cdot M)_{n_0}) = \mathbb{E}((H \cdot M)_0) = \mathbb{E}(M_0)$, and so we arrive at

$$\mathbb{E}(M_{T \wedge n_0}) = \mathbb{E}(M_0)\,.$$

□

6.4 Martingale Convergence Theorem

The Martingale Convergence Theorem is one of the main results in martingale theory. It can be understood as the probabilistic counterpart to the convergence of a bounded, non-decreasing sequence of real numbers. The theorem is very useful in applications. We state the theorem but do not include its proof (which requires tools outside the scope of this book). For a reference see [36].

Theorem 6.4.1 (Martingale Convergence Theorem)**.** *Let $(M_n)_{n\geq 0}$ be a submartingale with respect to $(X_n)_{n\geq 0}$. If $\sup_n \mathbb{E}(|M_n|) < \infty$, then with probability 1,*

$$\lim_{n\to\infty} M_n = M_\infty \tag{6.7}$$

exists. Furthermore, the limit M_∞ is finite with probability 1, and $\mathbb{E}(|M_\infty|) < \infty$.

Notice that from (6.7), the rest of the statement of Theorem 6.4.1 follows from Fatou's Lemma (see Appendix C): We have

$$\begin{aligned} \mathbb{E}(|M_\infty|) &= \mathbb{E}(\liminf_{n\to\infty} |M_n|) \leq \liminf_{n\to\infty} \mathbb{E}(|M_n|) \\ &\leq \sup_n \mathbb{E}(|M_n|) < \infty\,, \end{aligned}$$

and so $\mathbb{P}(-\infty < M_\infty < \infty) = 1$.

Corollary 6.4.2. *Let $(M_n)_{n\geq 0}$ be a non-negative supermartingale with respect to $(X_n)_{n\geq 0}$. Then with probability 1,*

$$\lim_{n\to\infty} M_n = M_\infty$$

exists and is finite, and $\mathbb{E}(M_\infty) \leq \mathbb{E}(M_0) < \infty$.

Proof. Since $(M_n)_{n\geq 0}$ is a supermartingale, the process $(-M_n)_{n\geq 0}$ is a submartingale. Furthermore,

$$\mathbb{E}(M_n) = \mathbb{E}(|-M_n|) \leq \mathbb{E}(M_0) < \infty \quad \text{for all } n \geq 0\,,$$

and so the Martingale Convergence Theorem applies to $(-M_n)_{n\geq 0}$. □

6.5 Transience/Recurrence of MCs via martingales

Definition 6.5.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with discrete state space S and one-step transition probabilities P_{xy}, $x, y \in \mathcal{S}$. A function f on S is called* **harmonic** *at x if*

$$f(x) = \sum_{y\in\mathcal{S}} P_{xy} f(y)\,. \tag{6.8}$$

We say f is harmonic on S if f is harmonic at x for all $x \in \mathcal{S}$.
If "$=$" in (6.8) is replaced by "$\leq$" (resp. by "$\geq$"), we say f is **subharmonic** *(resp.* **superharmonic***) at x.*

Recall that not every martingale is a Markov chain, and not every Markov chain is a martingale. The following proposition gives a class of martingales derived from Markov chains via harmonic functions.

Proposition 6.5.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain with discrete state space $\mathcal{S}$.*

(a) Let f be a bounded harmonic function on $\mathcal{S}$. Then $(f(X_n))_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$.
(b) Conversely, if $(X_n)_{n\geq 0}$ is irreducible and f a function on $\mathcal{S}$ such that $(f(X_n))_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$, then f is harmonic on $\mathcal{S}$.

Proof. (a) Clearly, $f(X_n)$ satisfies part (a) of Definition 6.1.1. Since f is bounded, $\mathbb{E}(|f(X_n)|) < \infty$. By the Markov property,

$$\mathbb{E}(f(X_{n+1}) \,|\, X_0 = x_0, X_1 = x_1, ..., X_n = x) = \mathbb{E}(f(X_{n+1}) \,|\, X_n = x)$$

and, since f is harmonic,

$$\mathbb{E}(f(X_{n+1}) \,|\, X_n = x) = \sum_{y\in\mathcal{S}} P_{xy} f(y) = f(x)\,.$$

So $\mathbb{E}(f(X_{n+1}) \,|\, X_0, X_1, ..., X_n) = f(X_n)$.
(b) Conversely, let f be a function on $\mathcal{S}$ such that $(f(X_n))_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$. For any $x \in \mathcal{S}$, there exists an $n \geq 0$ such that $\mathbb{P}(X_n = x) > 0$. We then have

$$f(x) = \mathbb{E}(f(X_{n+1}) \,|\, X_n = x) = \sum_{y\in S} P_{xy} f(y)\,.$$

This shows that f is harmonic at x. □

More generally, we can show that if the state space $\mathcal{S}$ is finite and f is a right eigenvector of the transition matrix $\mathbf{P}$ corresponding to eigenvalue $\lambda \neq 0$, then $(f(X_n)/\lambda^n)_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$. See Exercise 6.5.
Similarly to the above, we can show that a bounded subharmonic (resp. superharmonic) function defines a submartingale (resp. supermartingale) $(f(X_n))_{n\geq 0}$. And vice versa, if $(f(X_n))_{n\geq 0}$ is a submartingale (resp. supermartingale), then f must be subharmonic (resp. superharmonic).

Theorem 6.5.2. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with discrete state space $\mathcal{S}$. If the Markov chain is recurrent, then $(X_n)_{n\geq 0}$ has no nonnegative superharmonic or bounded subharmonic functions except for the constant functions.*

Proof. Let f be a nonnegative superharmonic (resp. bounded subharmonic) function for the Markov chain. It follows that $(f(X_n))_{n\geq 0}$ is a nonnegative supermartingale (resp. bounded submartingale). By the Martingale Convergence Theorem, with probability 1, the process $(f(X_n))_{n\geq 0}$ converges to a finite random variable M_∞. But since $(X_n)_{n\geq 0}$ is recurrent, with probability 1, the Markov chain visits every state $x \in \mathcal{S}$ infinitely often. Thus we must have $M_\infty = f(x)$ for all $x \in \mathcal{S}$. It follows that f must be constant on $\mathcal{S}$.
□

Theorem 6.5.2 gives a criterion for transience for a Markov chain:

Proposition 6.5.3. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with discrete state space $\mathcal{S}$. Choose a state $x_0 \in \mathcal{S}$. Then $(X_n)_{n\geq 0}$ is transient if and only if there exists a bounded non-constant function f on $\mathcal{S} \setminus \{x_0\}$ with the property*

$$f(x) = \sum_{y\neq x_0} P_{xy} f(y) \quad \text{for } x \neq x_0\,. \tag{6.9}$$

Proof. Assume $(X_n)_{n\geq 0}$ is transient and consider the function f on $\mathcal{S}$ defined by

$$f(x) = \begin{cases} \mathbb{P}(T^{x_0} = \infty \,|\, X_0 = x) & \text{for } x \neq x_0 \\ 0 & \text{for } x = x_0\,. \end{cases}$$

Because of transience, f is not equal to zero on $\mathcal{S} \setminus \{x_0\}$. Using a first-step analysis, we have

$$f(x) = \sum_{y\neq x_0} P_{xy} f(y) \quad \text{for } x \neq x_0\,.$$

and so (6.9) holds. Conversely, assume (6.9) holds. Let $\tilde{a} = \sum_{y \in \mathcal{S}} P_{x_0 y} f(y)$. If $\tilde{a} \geq 0$, then we define $f(x_0) = 0$ which makes f subharmonic. (If $\tilde{a} < 0$, then we work with the function $-f$ instead.) Clearly f is bounded on $\mathcal{S}$. Assume $(X_n)_{n \geq 0}$ is recurrent. Then, by Theorem 6.5.2, f is constant and therefore, here, equal to zero. But this contradicts our assumption that f is non-constant on $\mathcal{S} \setminus \{x_0\}$. Hence $(X_n)_{n \geq 0}$ must be transient. □

Example 6.5.1 (Simple random walk on $\mathbb{N}_0$ with reflecting boundary at 0)**.** The state space is $\mathcal{S} = \mathbb{N}_0$. Let $0 < p < 1$, and set $q = 1 - p$. The transition probabilities are

$$\begin{aligned} P_{x,x+1} &= p \qquad \text{for } x \geq 0 \\ P_{x,x-1} &= q \qquad \text{for } x > 0 \\ P_{00} &= q\,. \end{aligned}$$

Figure 6.1 shows the transition graph. Clearly, this is an irreducible Markov chain.

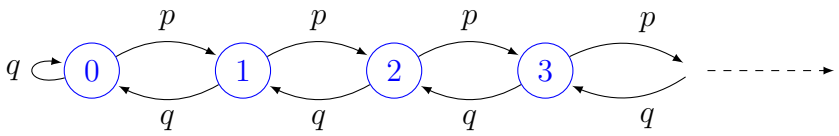


Figure 6.1

With $x_0 = 0$, (6.9) yields the system of equations

$$\begin{aligned} f(1) &= pf(2) \\ f(2) &= qf(1) + pf(3) \\ f(3) &= qf(2) + pf(4) \\ &\vdots \\ f(n) &= qf(n-1) + pf(n+1) \\ &\vdots \end{aligned}$$

We can set $f(1) = 1$. This yields $f(2) = 1/p$. We rewrite the second equation in the above system as

$$q[f(2) - f(1)] = p[f(3) - f(2)]\,,$$

which yields

$$\left(\frac{q}{p}\right)^2 = f(3) - f(2)\,.$$

By induction,

$$\left(\frac{q}{p}\right)^n = f(n+1) - f(n) \qquad \text{for } n \geq 1\,.$$

We set $f(0) = 0$ and compute

$$\begin{aligned} f(1) &= 1 \\ f(2) &= 1/p \\ f(3) &= 1/p + (\frac{q}{p})^2 \\ &\vdots \\ f(n) &= 1/p + \sum_{k=2}^{n-1} (\frac{q}{p})^k \\ &\vdots \end{aligned}$$

From the above we see that the function f is bounded and non-constant if and only if $p > q$. Thus simple random walk on the nonnegative integers with reflecting boundary at 0 is transient if and only if $p > q$. □

Example 6.5.2 (Simple random walk on $\mathbb{Z}$)**.** We now use the result from Example 6.5.1 to prove that simple (unrestricted) random walk on $\mathbb{Z}$ is transient if and only if $p \neq q$. If $p > q$, then use the above computed function f from Example 6.5.1 and extend it to a function on all of $\mathbb{Z}$ by setting $f(x) = 0$ for $x \leq 0$. If $p < q$, then we show by a very similar argument that simple random walk on the negative integers with reflecting boundary at 0 is transient. In this case, extent the analogous function f on the negative integers to a function on all of $\mathbb{Z}$ by setting $f(x) = 0$ for $x \geq 0$. If $p = q = \frac{1}{2}$, then the system (6.9) does not yield a bounded non-constant function f on the positive integers (or on the negative integers). Hence, by Proposition 6.5.3, simple symmetric random walk on $\mathbb{Z}$ is recurrent. □

Example 6.5.3 (Criterion for transience for general birth/death chains)**.** We assume $\mathcal{S} = \mathbb{N}_0$ and the process is irreducible. The transition probabilities are

$$\begin{aligned} P_{x,x+1} &= p_x \quad \text{for } x \geq 0 \\ P_{x,x-1} &= q_x \quad \text{for } x > 0 \\ P_{x,x} &= r_x \quad \text{for } x \geq 0 \end{aligned}$$

with $p_x + q_x + r_x = 1$ and $p_x, q_x > 0$ for all x. Figure 6.2 shows the transition graph. Equations (6.9) read

$$\begin{aligned} f(1) &= p_1 f(2) + r_1 f(1) \\ f(2) &= q_2 f(1) + p_2 f(3) + r_2 f(2) \\ f(3) &= q_3 f(2) + p_3 f(4) + r_3 f(3) \\ &\vdots \\ f(n) &= (q_n f(n-1) + p_n f(n+1) + r_n f(n) \\ &\vdots \end{aligned}$$

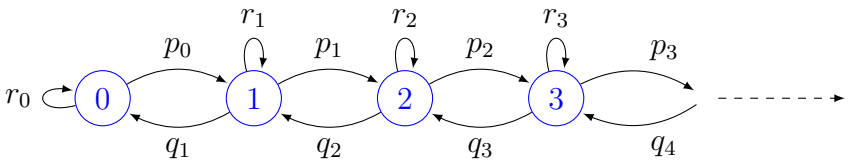


Figure 6.2

Setting $f(1) = 1$, we get from the first equation $f(2) = \frac{1-r_1}{p_1}$. The rest of the equations can be written as

$$q_n\,[f(n) - f(n-1)] = p_n\,[f(n+1) - f(n)] \qquad \text{for } n \geq 2\,,$$

from which we compute by induction,

$$f(n+1) - f(n) = \prod_{j=1}^{n} \frac{q_j}{p_j} \qquad \text{for } n \geq 1\,.$$

This yields

$$f(n) = \frac{1-r_1}{p_1} + \sum_{k=2}^{n-1} \left(\prod_{j=1}^{k} \frac{q_j}{p_j} \right) \qquad \text{for } n \geq 3\,.$$

We conclude that the birth/death chain on $\mathbb{N}_0$ is transient if and only if

$$\boxed{\sum_{k=2}^{\infty} \left(\prod_{j=1}^{k} \frac{q_j}{p_j} \right) < \infty\,.}$$

□

6.6 Applications

6.6.1 Waiting times for sequence patterns

Consider a sequence $(X_n)_{n\geq 1}$ of i.i.d. discrete random variables taking values in a finite set V. Suppose we are given a *pattern of length* k, i.e., a fixed sequence $(x_1, x_2, ..., x_k)$ of elements in V. If we observe the outcomes of the random sequence $(X_n)_{n\geq 1}$, one at a time, how long on average do we have to wait until we see the pattern $(x_1, x_2, ..., x_k)$ appear for the first time? As an example, consider $V = \{0, 1, 2, 3\}$ and let T be the first time at which we see the pattern $(0, 2, 3, 3)$. If we observe the outcomes

$$2, 3, 2, 0, 0, 1, 3, 1, 0, 0, 0, 3, \overbrace{0, 2, 3, 3}^{\text{pattern}}, ...$$

then $T = 16$. We could model this type of problem as an absorbing Markov chain $(Y_n)_{n\geq 0}$ whose state space $\mathcal{S}$ consists of all k-length patterns from V. The unique absorbing state is $(x_1, x_2, ..., x_k)$. We start by generating the initial $X_1, ..., X_k$. The resulting pattern is Y_0, and thus the initial distribution is uniform distribution on $\mathcal{S}$. At time $n > 0$, Y_n is the most recent k-length pattern $X_{n+1}, X_{n+2}, ..., X_{k+n}$. We can set up the transition matrix for $(Y_n)_{n\geq 0}$ and from it compute the expected time $\mathbb{E}(T)$ until absorption. Since $(Y_n)_{n\geq 0}$ has finite state space and is irreducible, $\mathbb{E}(T) < \infty$.

Here we present an alternative approach by introducing a martingale and making use of the Optional Stopping Theorem. Imagine a sequence of gamblers, each starting with an initial capital of \$1 and playing a sequence of fair games until their first loss, upon which the gambler leaves the casino. Let $(x_1, ..., x_k)$ be the desired pattern in the sequence of i.i.d. random variables $(X_n)_{n\geq 1}$. The game proceeds as follows. Gambler j enters the game at time j and bets his \$1 on x_1 in a fair game. If he loses the bet, he quits (and has lost \$1). If he wins, he continues with game $j + 1$. He bets his total capital (that is, his initial \$1 plus his winnings from game j) on x_2. If he loses game $j + 1$, he quits the game with a net loss of \$1. If he wins, he moves on to game $j + 2$ betting on x_3 with his total capital (\$1 plus his winnings from games j and $j + 1$), and so on. In the meantime, in Game $j + 1$, Gambler $j + 1$ has entered and bets her initial \$1 on x_1 in game $j + 1$. If she loses this bet, she quits. If she wins this bet, she continues on to game $j + 2$ at which she bets her total capital on x_2, and so on. The game ends at time T.

Here is a concrete example. Consider a sequence of i.i.d. fair coin tosses $X_1, X_2, ...$ Let D (D for duration of the game) be the first time at which we see the pattern THT for the first time. If, for example, the tosses result in $\omega = TTHHTHT$ we have $D(\omega) = 7$. What is $\mathbb{E}(D)$? Since $p = q = \frac{1}{2}$, each game has payoff equal to the gamblers stake. Gambler 1 starts with Game 1 and bets \$1 on T. If he wins, his fortune is now \$2. He moves on to Game 2 and bets \$2 on H. If he wins, he has now \$4 and moves on to Game 3, at which he bets \$4 on T. If he wins Game 3, we have $T = 3$, and the casino starts a new game. Gambler 1 has made a net profit of \$7. However, if Gambler 1 loses anytime at or before time 3, he quits the game with a loss of \$1.

Let W_n be the total winnings of the casino by time n. Since all bets are fair, $(W_n)_{n\geq 1}$ is a martingale. The random time D is a stopping time for $(W_n)_{n\geq 1}$. As mentioned above, $\mathbb{E}(D) < \infty$. Furthermore it is clear that $|W_n - W_{n-1}|$ is bounded for all $n \geq 2$, since at any given time n at most k (where k is the length of the pattern sequence) players are in the game, and the size of their bets is uniformly bounded. It follows that part (c) of the Optional Stopping Theorem applies. We have

$$\mathbb{E}(W_D) = \mathbb{E}(W_1) = 0\,.$$

Here is a summary for the game that resulted in $D = 7$ with $\omega = TTHHTHT$:

Flip	T	T	H	H	T	H	T
Player # entering the game	1	2	3	4	5	6	7
Total player payoff	$-\$1$	$-\$1$	$-\$1$	$-\$1$	\$7	$-\$1$	\$1

The desired pattern HTH has length $k = 3$. So only the last 3 gamblers in the game can possibly win (and some of them will lose \$1). All gamblers who entered before Game $D - k + 1$ have lost \$1.

Example 6.6.1. A sequence of i.i.d. Bernoulli flips with a fair coin. The game stops as soon as the pattern THT has appeared for the first time. Denote this random time by D.

Game #	1 to $(D-3)$	$D-2$	$D-1$	D
Flip	$\cdots$	T	H	T
Player # entering the game	1 to $(D-3)$	$D-2$	$D-1$	D
Total player payoff	$-\$1$ each	\$7	$-\$1$	\$1

Here $W_D = D - 3 - 7 + 1 - 1$. Since $\mathbb{E}(W_D) = 0 = \mathbb{E}(D) - 10$, we get

$$\mathbb{E}(D) = 10\,.$$

□

We now turn to more general case of i.i.d. (possibly) biased coin flips. All games are still fair games. Note that this means that the gambler either loses \$1 (at which point he quits the game), or his fortune grows by a factor of p^{-1} (in which case he continues the game and bets his total fortune on the next game). Indeed, say the gambler's fortune is currently $\$x$. He bets $\$x$ on the next game in which he wins with probability p. Since the game is fair, if he wins, the casino has to pay him $\$y$ which is computed from

$$-x(1-p) + yp = 0\,.$$

If the gambler wins, his new fortune is therefore $\$(x+y) = \$xp^{-1}$. As before, we assume that the gambler starts with an initial fortune of $x = \$1$.

Example 6.6.2. Consider a sequence of i.i.d. biased coin flips with $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = 1 - p$. Assume $0 < p < 1$, and set $q = 1 - p$. The game stops as soon as the pattern THTHT has appeared for the first time. Denote this random time by D.

Game #	1 to $(D-5)$	$D-4$	$D-3$	$D-2$	$D-1$	D
Flip	$\cdots$	T	H	T	H	T
Player #	1 to $(D-5)$	$D-4$	$D-3$	$D-2$	$D-1$	D
Player payoff	$-\$1$ each	$\$(p^{-2}q^{-3}-1)$	$-\$1$	$\$(p^{-1}q^{-2}-1)$	$-\$1$	$\$(q^{-1}-1)$

Here

$$W_D = (D-5) - (p^{-2}q^{-3} - 1) + 1 - (p^{-1}q^{-2} - 1) + 1 - (q^{-1} - 1)\,,$$

and so

$$\mathbb{E}(D) = p^{-2}q^{-3} + p^{-1}q^{-2} + q^{-1}\,.$$

□

Example 6.6.3. As in Example 6.6.2, consider a sequence of i.i.d. coin flips with $\mathbb{P}(H) = p$ with $0 < p < 1$. The expected time $\mathbb{E}(D)$ until we see n heads in a row is

$$\mathbb{E}(D) = (\frac{1}{p})^n + (\frac{1}{p})^{n-1} + \cdots + \frac{1}{p} = \frac{1}{p}\frac{(\frac{1}{p})^n - 1}{\frac{1}{p} - 1} = \frac{1-p^n}{(1-p)p^n}\,.$$

As a specific numerical example, for a fair coin with $p = \frac{1}{2}$, the expected time $\mathbb{E}(D)$ until we see $n = 5$ heads in a row is $\mathbb{E}(D) = 62$. □

6.6.2 Gambler's ruin, revisited

We can solve the gambler's ruin problem from Section 4.4 with the use of the OST.

Fair game.
Consider simple symmetric random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ with $S_n = \sum_{k=0}^n X_k$ with $X_0 = x$ and i.i.d. $X_k \sim \text{Unif}(\{-1, 1\})$ for $k \geq 1$. Let $a < x < b$. From Example 6.1.2, we know that $(S_n)_{n\geq 0}$ is a martingale. The first time $T = T^{\{a,b\}}$ the random walk hits one of the boundary points a or b is a stopping time. We have $\mathbb{P}(T < \infty) = 1$. Clearly, $P(|S_n| \leq K_0) = 1$ for $n \leq T$ and $K_0 = \max\{(|a|, |b|\}$. Therefore conditions (b) of the Optional Stopping Theorem are satisfied, and we have

$$\mathbb{E}(S_T) = \mathbb{E}(S_0) = x\,.$$

Using the notation $r_x = \mathbb{P}_x(T^a < T^b)$, we get

$$r_x a + (1 - r_x)b = x\,,$$

from which we compute

$$r_x = \frac{b-x}{b-a} \tag{6.10}$$

and

$$\mathbb{P}_x(T^b < T^a) = 1 - r_x = \frac{x-a}{b-a}\,.$$

Note that since $(S_n)_{n\geq 0}$ is a martingale *and* a Markov chain on finite state space $\mathcal{S} = \{a, a+1, ..., b\}$, the gambler's ruin probabilities r_x follow from (6.4) in Example 6.1.5 as well. We will now compute $\mathbb{E}(T)$. From Example 6.1.3, we know that the process $M_n = (S_n)^2 - n\,\sigma^2$ for $n \geq 0$ defines a martingale with respect to $(X_n)_{n\geq 0}$. Here $\sigma^2 = 1$. We will show that conditions (c) in the OST apply to $(M_n)_{n\geq 0}$, and then use the OST to compute $\mathbb{E}(T)$. Condition $\mathbb{E}(T) < \infty$ holds since for any irreducible, finite state Markov chain, the expected hitting time of a subset of states, here $\{a, b\}$, is finite (recall Proposition 2.1.1). The expression $|M_n - M_{n-1}| = |S_n^2 - n - (S_{n-1}^2 - (n-1))|$ is equal to

$$|2X_n(X_0 + \cdots X_{n-1}) + X_n^2 - 1|\,.$$

Thus

$$\begin{aligned}
\mathbb{E}(|M_n - M_{n-1}| \,|\, X_0, ..., X_{n-1}) &= \mathbb{E}(|2X_n(X_0 + \cdots X_{n-1}) + X_n^2 - 1| \,|\, X_0, ..., X_{n-1}) \\
&\leq 2\mathbb{E}(|X_n||X_0 + \cdots X_{n-1}| \,|\, X_0, ..., X_{n-1}) + \mathbb{E}(|X_n^2 - 1| \,|\, X_0, ..., X_{n-1}) \\
&= 2|X_0 + \cdots X_{n-1}|\,\mathbb{E}(|X_n|) = 2|X_0 + \cdots X_{n-1}|\,.
\end{aligned}$$

For $n \leq T$, the last expression is bounded above by $K_0 = 2\max\{|a|, |b|\}$, and so conditions (c) of the OST are satisfied for $(M_n)_{n\geq 0}$, and we have

$$\mathbb{E}(M_T) = \mathbb{E}(S_T^2) - \mathbb{E}(T) = \mathbb{E}(M_0) = x^2\,. \tag{6.11}$$

From (6.11) and (6.10), we have

$$x^2 = a^2\,\frac{b-x}{b-a} + b^2\,\frac{x-a}{b-a} - \mathbb{E}(T)\,,$$

from which we compute

$$\begin{aligned}
\mathbb{E}(T) &= a^2\,\frac{b-x}{b-a} + b^2\,\frac{x-a}{b-a} - x^2 \\
&= \frac{(b-x)(x-a)(b-a)}{(b-a)} = (b-x)(x-a)\,,
\end{aligned}$$

which confirms our result from Section 4.4.

Biased game.

We now consider biased simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ starting at x. Recall Example 6.1.4. The process $(M_n)_{n\geq 0}$ defined by $M_n = (\frac{q}{p})^{S_n}$ is a martingale. Let $a < x < b$. As for the symmetric case, the time $T = T^{\{a,b\}}$ until the walk hits one of the boundary points a or b is a stopping time, and $\mathbb{P}(T < \infty) = 1$. We also have $\mathbb{P}(|M_n| \leq K_0) = 1$ for $n \leq T$

and $K_0 = \max\{(\frac{q}{p})^a, (\frac{q}{p})^b\}$. Thus conditions (b) of the Optional Stopping Theorem are satisfied, and

$$\mathbb{E}(M_T) = \mathbb{E}(M_0) = (\frac{q}{p})^x \,.$$

Setting $r_x = \mathbb{P}_x(T^a < T^b)$, we get

$$\mathbb{E}(M_T) = r_x(\frac{q}{p})^a + (1 - r_x)(\frac{q}{p})^b = (\frac{q}{p})^x \,,$$

from which we compute

$$r_x \left((\frac{q}{p})^a - (\frac{q}{p})^b \right) = (\frac{q}{p})^x - (\frac{q}{p})^b$$

and

$$r_x = \frac{(\frac{q}{p})^x - (\frac{q}{p})^b}{(\frac{q}{p})^a - (\frac{q}{p})^b} = \frac{(\frac{q}{p})^{x-a} - (\frac{q}{p})^{b-a}}{1 - (\frac{q}{p})^{b-a}} \,. \tag{6.12}$$

For the computation of $\mathbb{E}(T)$, we apply the OST to the martingale $(Y_n)_{n\geq 0}$ defined by $Y_n = S_n - n(p-q)$ (and for which we verify that conditions (c) of the OST hold). Thus we have

$$\mathbb{E}(Y_T) = \mathbb{E}(S_T) - \mathbb{E}(T)(p-q) = \mathbb{E}(Y_0) = x \,.$$

This yields

$$a\, r_x + b\,(1 - r_x) - x = (p-q)\mathbb{E}(T)$$

from which, using (6.12), we compute

$$\mathbb{E}(T) = \frac{a-x}{p-q} + \left(\frac{b-a}{p-q}\right) \frac{1 - (\frac{q}{p})^{x-a}}{1 - (\frac{q}{p})^{b-a}}$$

which confirms our results from Section 4.4.

6.6.3 Branching process, revisited

We return to the Galton-Watson branching process $(X_n)_{n\geq 0}$ from Chapter 5. It is determined by an $\mathbb{N}_0$-valued *offspring random variable* Z. At each time interval, each individual in the current generation gives rise to a number of offspring according to the distribution of Z. It is assumed that all individuals reproduce independently. For simplicity, we assume that the branching process $(X_n)_{n\geq 0}$ starts with one individual. Then the size of Generation 1 is $X_1 \sim Z$, and the size of Generation n is

$$X_n = Z_1^{(n)} + \cdots + Z_{X_{n-1}}^{(n)}$$

where $Z_1^{(n)}, Z_2^{(n)}, ...$, $n \geq 1$, are i.i.d. random variables with $Z_k^{(n)} \sim Z$ for all $n, k \geq 1$.

To avoid trivial cases for the branching process, we assume $\mathbb{P}(Z \geq 2) > 0$. Furthermore, since we will be interested in the probability of ultimate extinction of such a process, we assume $\mathbb{P}(Z = 0) > 0$. Let $m = \mathbb{E}(Z)$. We assume $0 < m < \infty$.
The process $(M_n)_{n\geq 0}$ defined by

$$M_n = X_n/m^n$$

is a non-negative martingale with respect to $(X_n)_{n\geq 0}$. Indeed,

$$\mathbb{E}(|M_n|) = \frac{1}{m^n}\mathbb{E}(X_n) = 1 < \infty$$

and

$$\mathbb{E}(M_{n+1} \,|\, X_0, ..., X_n) = \frac{1}{m^{n+1}}\mathbb{E}(X_{n+1} \,|\, X_n) = \frac{1}{m^{n+1}} X_n m = M_n \,.$$

By the Martingale Convergence Theorem, with probability 1, the limit

$$M_\infty = \lim_{n\to\infty} M_n$$

exists and is finite. We can now ask about the distribution of M_∞. Since all states for the branching process $(X_n)_{n\geq 0}$ lead into state 0, all states other than state 0 are transient. It follows that, with probability 1,

$$\lim_{n\to\infty} X_n \in \{0, \infty\} \,.$$

For the subcritical and critical cases (i.e., for the case $m \leq 1$), we have

$$\lim_{n\to\infty} X_n = 0$$

with probability 1. Since the state space is discrete, this means that with probability 1, for any trajectory of the branching process, there exists an n_0 such that $X_n = 0$ for all $n \geq n_0$. It follows that for if $m \leq 1$, with probability 1,

$$\lim_{n\to\infty} M_n = M_\infty = 0$$

Note that in this case we have

$$\mathbb{E}(M_\infty) = 0 \neq \lim_{n\to\infty} \mathbb{E}(M_n) = 1 \,.$$

The **supercritical case** $m > 1$ presents interesting questions about the **growth rate of the population size for the branching process**. In this case the event $\{M_\infty > 0\}$ may have positive probability, depending on properties of the offspring distribution. Note that since $M_n = X_n/m^n$, on the event $\{M_\infty > 0\}$, the branching process $(X_n)_{n\geq 0}$ will have *exponential growth rate* due to the a.s. finiteness of M_∞. Let e_0 denote the probability of

ultimate extinction of the population, and recall that in the supercritical case, $0 < e_0 < 1$. Thus we have

$$\mathbb{P}(M_\infty = 0) \geq e_0 \,.$$

The following theorem explains the growth rate of the population size for a supercritical branching process. The proof of the theorem lies outside the scope of this text. For a reference see [4].

Theorem 6.6.1 (Kesten-Stigum)**.** *Let $(X_n)_{n\geq 0}$ be a supercritical branching process with $X_0 = 1$. Let Z be its offspring random variable and assume $\mathbb{P}(Z = k) \neq 1$ for all $k \in \mathbb{N}_0$. Let e_0 be its extinction probability. The following are equivalent:*

(a) $\mathbb{E}(Z \ln Z) < \infty$,

(b) $\mathbb{P}(M_\infty = 0) = e_0$,

(c) $\mathbb{E}(M_\infty) = 1$.

If $\mathbb{E}(Z \ln Z) = \infty$, *then* $\mathbb{P}(M_\infty = 0) = 1$, *and thus* $\mathbb{E}(M_\infty) = 0$.

Theorem 6.6.1 settles the question about the conditions under which a supercritical branching process $(X_n)_{n\geq 0}$ has *exponential growth rate*: Exponential growth rate of the population happens precisely when $\mathbb{E}(Z \ln Z) < \infty$, and in this case it happens with probability $1 - e_0$.

When $\mathbb{E}(Z \ln Z) = \infty$, then, here also, the population will not die out and instead grow to infinity with probability $1 - e_0$. But the growth rate of the population size will be *sub-exponential*.

6.6.4 Pólya's Urn, revisited

Recall Pólya's urn model from Section 1.5. An urn contains b blue balls and g green balls. At each time step, a ball is drawn uniformly at random from the urn, its color noted, and then together with c additional balls of the *same* color, put back into the urn. Note that the number of balls in the urn increases by c with each step.

In the following, for simplicity, we will take $c = 1$. Let X_n denote the number of blue balls in the urn at time n, and consider the process $M_n = X_n/(b + g + n)$ which gives the fraction of blue balls in the urn after n steps. The process $(M_n)_{n\geq 0}$ is a martingale with

respect to $(X_n)_{n\geq 0}$. Indeed, the random variables M_n are uniformly bounded by 1, and

$$\begin{aligned}
\mathbb{E}(M_{n+1} \,|\, X_0, ..., X_n) &= (b+g+n+1)^{-1}\mathbb{E}(X_{n+1} \,|\, X_n) \\
&= (b+g+n+1)^{-1}\left[X_n(1-\frac{X_n}{b+g+n}) + (X_n+1)\frac{X_n}{b+g+n}\right] \\
&= (b+g+n+1)^{-1}\left[X_n + \frac{X_n}{b+g+n}\right] \\
&= \frac{X_n}{b+g+n} = M_n\,.
\end{aligned}$$

We can apply the Martingale Convergence Theorem to $(M_n)_{n\geq 0}$. Thus there exists a random variable $M_\infty \leq 1$ such that with probability 1,

$$\lim_{n\to\infty} M_n = M_\infty\,.$$

By the Bounded Convergence Theorem (see Appendix C), we have

$$\lim_{n\to\infty} \mathbb{E}(M_n) = \lim_{n\to\infty} \mathbb{E}(M_0) = \frac{b}{b+g} = \mathbb{E}(M_\infty)\,.$$

We will now show that the random variable M_∞ has a beta distribution Beta(b, g). Recall the density $f(x)$ for a Beta(b, g) distribution:

$$f(x) = \frac{\Gamma(b+g)}{\Gamma(b)\Gamma(g)}\, x^{b-1}(1-x)^{g-1} \quad \text{for } 0 < x < 1$$

and 0 elsewhere. Here $\Gamma(y)$ is the Gamma function defined by

$$\Gamma(y) = \int_0^\infty x^{y-1}e^{-x}\,dx\,.$$

Special values of the Gamma function are $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$. A special case of the beta distribution is Beta$(1, 1)$ = Uniform$(0, 1)$.

We introduce a sequence of Bernoulli random variables $(Y_n)_{n\geq 1}$ defined by $Y_n = 1$ if the nth ball drawn is blue and $Y_n = 0$ if the nth ball drawn is green. In Section 1.5, we showed that the random variables Y_n are *identically* distributed (but not independent!) with $\mathbb{P}(Y_n = 1) = \frac{b}{b+g}$. By de Finetti's Theorem, if we choose a success probability p according to a beta distribution Beta(b, g) on $[0, 1]$, then conditional on p, the sequence $(Y_n)_{n\geq 1}$ is i.i.d Bernoulli(p). Thus, conditionally on p, by the Strong Law of Large Numbers,

$$\bar{Y}_n = \frac{1}{n}(Y_1 + \cdots + Y_n) \xrightarrow{n\to\infty} p \quad \text{a.s.}\,,$$

and therefore,

$$\bar{Y}_n = \frac{1}{n}(Y_1 + \cdots + Y_n) \xrightarrow{n\to\infty} Y \quad \text{a.s.},$$

where $Y \sim \text{Beta}(b, g)$. Note that

$$M_n = \frac{b + (Y_1 + \cdots + Y_n)}{b + g + n} = \frac{b + \bar{Y}_n n}{b + g + n},$$

and thus

$$\lim_{n\to\infty} M_n = \lim_{n\to\infty} \frac{b + \bar{Y}_n n}{b + g + n} = \lim_{n\to\infty} \bar{Y}_n = Y \quad \text{a.s.}.$$

This proves that

$$M_\infty \sim \text{Beta}(b, g).$$

Since $\text{Beta}(1, 1) \sim \text{Uniform}(0, 1)$, we have the following: If the process starts with one bue ball and one green ball, in the long run, the fraction of blue balls in Pólya's urn settles down to a fraction that is uniformly distributed on the unit interval.

Exercises

Exercise 6.1. Consider a martingale $(M_n)_{n\geq 0}$ with respect to $(X_n)_{n\geq 0}$ and nonnegative integers

$$n_1 < n_2 < \cdots < n_k < n_{k+1}.$$

Show that

$$\mathbb{E}(M_{n_{k+1}} \,|\, X_{n_1}, ..., X_{n_k}) = M_{n_k}.$$

Exercise 6.2. Consider a submartingale $(M_n)_{n\geq 0}$ with respect to $(X_n)_{n\geq 0}$. Show that for all $n, k \geq 0$,

$$\mathbb{E}(M_{n+k} \,|\, X_0, ..., X_n) \geq M_n$$

and

$$\mathbb{E}(M_{n+k}) \geq \mathbb{E}(M_n).$$

[If $(M_n)_{n\geq 0}$ is a supermartingale with respect to $(X_n)_{n\geq 0}$, then each of the two inequalities is reversed.]

Exercise 6.3. A sequence of i.i.d. Bernoulli flips with a fair coin. Slightly modify the game from Example 6.6.1 by considering the pattern HTT. What is the expected waiting time $\mathbb{E}(D)$ until we see the pattern HTT for the first time? Does it differ from the result in Example 6.6.1?

Exercise 6.4. Consider a sequence of i.i.d. coin tosses with a fair coin, i.e. for each coin toss we have $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. What is the expected number of tosses until we see the pattern HHHH for the first time? Use martingale theory to answer this question.

Exercise 6.5. Consider a Markov chain $(X_n)_{n\geq 0}$ with finite state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. Show that if f is a right eigenvector of $\mathbf{P}$ corresponding to eigenvalue $\lambda \neq 0$, then $(f(X_n)/\lambda^n)_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$.

Exercise 6.6. Consider a Markov chain $(X_n)_{n\geq 0}$ on discrete state space $\mathcal{S}$ and assume $a \in \mathcal{S}$ is an absorbing state. Consider the hitting (or first return) time $T^a \geq 1$, and define

$$f(x) = \mathbb{P}_x(T^a < \infty)$$

for $x \in \mathcal{S}$. Show that $(f(X_n))_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$.

Exercise 6.7. Consider a random variable Z and a stochastic process $(X_n)_{n\geq 0}$, both defined on the same probability space Ω. Furthermore, assume $\mathbb{E}(|Z|) < \infty$. Show that the process $(Y_n)_{n\geq 0}$ defined by

$$Y_n = \mathbb{E}(Z \,|\, X_0, X_1, ..., X_n)$$

for $n \geq 0$ is a martingale with respect to $(X_n)_{n\geq 0}$.

Exercise 6.8. Let $X_1, X_2, ...$ be i.i.d. random variables with $\mathbb{P}(X_1 = 1) = \frac{3}{5}$, $\mathbb{P}(X_1 = 0) = \frac{1}{5}$, $\mathbb{P}(X_1 = -2) = \frac{1}{10}$, and $\mathbb{P}(X_1 = -3) = \frac{1}{10}$. Consider random walk $(S_n)_{n\geq 0}$ with $S_n = \sum_{i=1}^n X_i$ for $n \geq 1$ and $S_0 = 0$. Furthermore, consider the hitting time

$$T = \min\{n \,:\, S_n \geq 20\}.$$

Use the Optional Stopping Theorem (justify that the conditions are satisfied!) to compute $\mathbb{E}(T)$. (*Hint*: Theorem 4.3.2 may be useful.)

Exercise 6.9. Let $X_1, X_2, ...$ be i.i.d. random variables with $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \mathbb{P}(X_1 = 0) = \frac{1}{3}$. Consider random walk $(S_n)_{n\geq 0}$ with $S_0 = 0$ and $S_n = \sum_{i=1}^n X_i$. Show that the process $(Y_n)_{n\geq 0}$ defined by

$$Y_n = 3\sin\left(\frac{\pi}{2} S_n\right)$$

for $n \geq 0$ is a martingale with respect to $(X_n)_{n\geq 0}$.

Exercise 6.10. Consider simple (symmetric or biased) random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$ and fix $n > 0$. Show that the process $(M_k)_{0\leq k\leq n}$ defined by

$$M_k = \frac{S_{n-k}}{n-k}$$

is a martingale with respect to itself. (*Hint*: Make use of the result from Exercise 4.9.)

Exercise 6.11 (The ballot problem, revisited)**.** Let $c > d > 0$. Consider an election between two candidates, Candidate 1 and Candidate 2, in which Candidate 1 receives c votes and Candidate 2 receives d votes. Corollary 4.5.7 states that the probability that, throughout the election, Candidate 1 was always ahead of Candidate 2 is

$$\frac{c-d}{c+d}.$$

Use the martingale from Exercise 6.10 (for a suitably chosen n) and the stopping time $T = \min\{T^0, n-1\}$ to re-prove this result.

Exercise 6.12. Let $x \in \mathbb{N}$. A gambler starts with a fortune of \$x and makes a sequence of \$1 bets against the house which has unlimited funds. Let $0 < p < 1$ and $p \neq \frac{1}{2}$. Assume that with each bet, the gambler either wins \$1 with probability p or loses \$1 with probability $1-p$. The game ends at time T^0 when the gambler's fortune has reached \$0 (bankruptcy). Use the Martingale Convergence theorem to compute the probability $\mathbb{P}(T^0 < \infty)$, i.e. the probability that the gambler will eventually go bankrupt. (*Hint:* Work with the martingale from Example 6.1.4 and distinguish cases.)

Exercise 6.13. Consider a Galton-Watson branching process $(X_n)_{n\geq 0}$ with offspring distribution μ for which we assume $\mu(0) > 0$ and $\mu(0)+\mu(1) < 1$. Let m denote the expectation of μ. We have shown in Section 6.6.3 that the process $(M_n)_{n\geq 0}$ with $M_n = X_n/m^n$ is a martingale with respect to $(X_n)_{n\geq 0}$. Use the Martingale Convergence theorem to prove that if $m \leq 1$, then with probability 1, the population will ultimately become extinct, that is, with probability 1,

$$\lim_{n\to\infty} X_n = 0.$$

Exercise 6.14. Consider a Galton-Watson branching process $(X_n)_{n\geq 0}$ with offspring distribution μ for which we assume $\mu(0) > 0$ and $\mu(0) + \mu(1) < 1$. Let m denote the expectation of μ, and assume $m > 1$ (the process is supercritical).

(a) The probability generating function f of μ has a unique fixed point e_0 in the open interval $(0,1)$ (recall Figure 5.2). Consider the process $(M_n)_{n\geq 0}$ defined by $M_n = e_0^{X_n}$. Show that $(M_n)_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$.

(b) Use the Martingale Convergence theorem to re-prove that e_0 is the probability of ultimate extinction for the branching process $(X_n)_{n\geq 0}$. (*Hint*: Here, taking $\lim_{n\to\infty} \mathbb{E}(M_n) = \mathbb{E}(M_0) = \mathbb{E}(M_\infty)$ is justified by the Dominated Convergence theorem; see Theorem C.3.3.)

Exercise 6.15. Recall Pólya's urn model discussed in Section 6.6.4. We starts with b blue balls and g green balls. Let X_n denote the number of blue balls in the urn after n steps.

(a) If, at each step, instead of one ball we add $c > 1$ balls of the same color that was drawn, is the fraction of red balls in the urn after n steps still a martingale with respect to $(X_n)_{n\geq 0}$?

(b) Further modify the urn process by not only adding $c \geq 1$ balls of the same color that was drawn, but by also adding $d \geq 1$ balls of the opposite color at each step. Is now the fraction of blue balls in the urn after each step a martingale with respect to $(X_n)_{n\geq 0}$?

Exercise 6.16. Consider an urn that initially contains a number b of blue balls and a number g of green balls, so in total $n = b + g$ balls. We perform sampling without replacement from this urn. Clearly, the probability of drawing a blue ball with the first draw is $\frac{b}{b+g}$. You play a game by which you choose a time $T \in \{0, 1, ..., n-1\}$ and predict that the $(T+1)^{\text{th}}$ draw will yield a blue ball. If the $(T+1)^{\text{th}}$ ball is indeed blue, you win the game. The time T has to be a stopping time. So in making your choice of T, you are allowed to use any information gained from observing the process up to that time. Is it possible to devise a strategy for choosing T that increases your initial probability of $\frac{b}{b+g}$ of correctly predicting "blue" for the $(T+1)^{\text{th}}$ draw?

Exercise 6.17. Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is called *convex* if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

for all $x, y \in \mathbb{R}$ and for all $\alpha \in [0, 1]$. If "$\leq$" is replaced by "$\geq$" in the above inequality, the function f is called *concave*. Consider a martingale $(X_n)_{n\geq 0}$ and a function f on $\mathbb{R}$ with $\mathbb{E}|f(X_n)| < \infty$ for $n \geq 0$. Show that if f is convex, then the process $(f(X_n))_{n\geq 0}$ is a submartingale, and if f is concave, then $(f(X_n))_{n\geq 0}$ is a supermartingale.

Chapter 7

Reversibility

7.1 Time reversal of a Markov chain

Consider a positive recurrent (not necessarily irreducible) Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S}$ and a strictly positive stationary distribution π for the chain. Assume that the Markov chain starts in π. Let P_{xy} for $x, y \in \mathcal{S}$ be the one-step transition probabilities of the chain. Fix a time $N > 0$. The **time reversed process** $(\tilde{X}_n)_{0\leq n\leq N}$ is defined by

$$\tilde{X}_n = X_{N-n}$$

for $0 \leq n \leq N$. The time reversed process is also a Markov chain. Indeed, we have

$$\begin{aligned}
&\mathbb{P}(\tilde{X}_{k+1} = y \,|\, \tilde{X}_0 = x_0, \tilde{X}_1 = x_1, ..., \tilde{X}_{k-1} = x_{k-1}, \tilde{X}_k = x) \qquad (7.1)\\
&= \mathbb{P}(X_{N-k-1} = y \,|\, X_N = x_0, X_{n-1} = x_1, ..., X_{N-k+1} = x_{k-1}, X_{N-k} = x)\\
&= \frac{\mathbb{P}(X_{N-k-1} = y,\, X_{N-k} = x,\, X_{N-k+1} = x_{k-1}, ..., X_{n-1} = x_1,\, X_N = x_0)}{\mathbb{P}(X_{N-k} = x,\, X_{N-k+1} = x_{k-1}, ..., X_{n-1} = x_1,\, X_N = x_0)}\\
&= \frac{\pi(y)P_{yx}\,\mathbb{P}(X_N = x_0, X_{n-1} = x_1, ..., X_{N-k+1} = x_{k-1} \,|\, X_{N-k} = x)}{\pi(x)\mathbb{P}(X_N = x_0, X_{n-1} = x_1, ..., X_{N-k+1} = x_{k-1} \,|\, X_{N-k} = x)}\\
&= \frac{\pi(y)}{\pi(x)}P_{yx}\,.
\end{aligned}$$

The above shows that the conditional probabilities (7.1) for the time reversed process are independent of the past of the trajectory (if we consider time k the present), that is, independent of $\tilde{X}_n$ for $n < k$. Thus the time reversed process $(\tilde{X}_n)_{0\leq n\leq N}$ is a Markov chain, and its one-step and, by induction, its n-step transition probabilities are

$$\tilde{P}_{xy} = \frac{\pi(y)}{\pi(x)}P_{yx}\,, \quad \text{and } \tilde{P}^n_{xy} = \frac{\pi(y)}{\pi(x)}P^n_{yx} \quad \text{ for } x, y \in \mathcal{S}\,, n \geq 1\,.$$

Note that in case $(X_n)_{n\geq 0}$ is reducible, the transition probabilities $\tilde{P}_{xy}$, $x, y \in \mathcal{S}$, do not depend on the particular choice of π (see Exercise 7.3). And if the original transition probabilities P_{xy}, $x, y \in \mathcal{S}$, define an irreducible Markov chain on $\mathcal{S}$, the transition probabilities $\tilde{P}_{xy}$ also define an irreducible Markov chain on $\mathcal{S}$. Indeed, take any $x, z \in \mathcal{S}$. Then either $P_{zx} > 0$ or there exists $n \geq 2$ and $y_1, ..., y_{n-1} \in \mathcal{S}$ such that $P_{z,y_1} P_{y_1,y_2} \cdots P_{y_{n-1},x} > 0$. Hence we have

$$0 < \frac{\pi(z)}{\pi(y_1)} P_{z,y_1} \frac{\pi(y_1)}{\pi(y_2)} P_{y_1,y_2} \cdots \frac{\pi(y_{n-1})}{\pi(x)} P_{y_{n-1},x} = \tilde{P}_{y_1,z} \tilde{P}_{y_2,y_1} \cdots \tilde{P}_{x,y_{n-1}} \leq \tilde{P}^n_{x,z}$$

which shows irreducibility of the Markov chain $(\tilde{X}_n)_{n\geq 0}$, which we refer to as the **time reversal** of $(X_n)_{n\geq 0}$. Furthermore $(X_n)_{n\geq 0}$ and $(\tilde{X}_n)_{n\geq 0}$ have the **same stationary distribution** π:

$$\sum_{x\in\mathcal{S}} \pi(x) \tilde{P}_{xy} = \sum_{x\in\mathcal{S}} \pi(x) \frac{\pi(y)}{\pi(x)} P_{yx} = \pi(y) \sum_{x\in\mathcal{S}} P_{yx} = \pi(y)\,.$$

Note that the transition matrix for the time reversal chain is

$$\tilde{\mathbf{P}} = \mathbf{D}^{-1} \mathbf{P}^t \mathbf{D}$$

where $\mathbf{P}^t$ denotes the transpose of $\mathbf{P}$ and $\mathbf{D}$ is the diagonal matrix $\mathbf{D} = \mathrm{diag}(\pi(1), ..., \pi(n))$.

Proposition 7.1.1. *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain with stationary distribution π, and assume $X_0 \sim \pi$. Consider its time reversal $(\tilde{X}_n)_{n\geq 0}$ with $\tilde{X}_0 \sim \pi$. Then for all $n \geq 1$,*

$$(X_0, X_1, ..., X_n) \sim (\tilde{X}_n, ..., \tilde{X}_1, \tilde{X}_0)\,,$$

that is, for all $n \geq 1$ and for all $x_i \in \mathcal{S}$,

$$\mathbb{P}(X_0 = x_0,\, X_1 = x_1,\, ...,\, X_n = x_n) = \mathbb{P}(\tilde{X}_0 = x_n,\, \tilde{X}_1 = x_{n-1},\, ...,\, \tilde{X}_n = x_0)$$

The proof is straightforward verification.

Example 7.1.1 (Biased random walk on the discrete cycle)**.** Consider random walk on the discrete cycle $\mathbb{Z}_d = \{0, 1, ..., d-1\}$. Let $0 < p < 1$ and $p \neq \frac{1}{2}$. At each time, the walk takes one step clockwise with probability p or one step counter-clockwise with probability $1 - p$. Thus the transition probabilities are

$$\begin{array}{lll} P_{k,k+1} = & p & = P_{d-1,0} \quad \text{for } k = 0, 1..., d-2, \quad \text{and} \\ P_{0,d-1} = & 1-p & = P_{k,k-1} \quad \text{for } k = 1, 2, ..., d-1 \end{array},$$

and zero otherwise. The stationary distribution is uniform distribution on $\mathbb{Z}_d$. The time reversed chain has transition probabilities

$$\begin{array}{rrll} \tilde{P}_{k,k+1} = & 1-p & = \tilde{P}_{d-1,0} & \text{for } k = 0, 1..., d-2, \quad \text{and} \\ \tilde{P}_{0,d-1} = & p & = \tilde{P}_{k,k-1} & \text{for } k = 1, 2, ..., d-1 \end{array} ,$$

and zero otherwise. We see that here the time reversal is random walk on $\mathbb{Z}_d$ that takes steps with "opposite bias" compared to the original chain. □

7.2 Reversible Markov chains

Markov chains for which the time reversal has the same transition probabilities as the original chain often are of special interest. We call such Markov chains **reversible**.

Definition 7.2.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain on (finite or countably infinite) state space $\mathcal{S}$. If there exists a positive probability measure π on $\mathcal{S}$ for which*

$$\pi(x)P_{xy} = \pi(y)P_{yx} \quad \text{for all } x, y, \in \mathcal{S}\,, \tag{7.2}$$

we call the Markov chain **reversible** *with respect to π. Equations (7.2) are called the* **detailed balance equations**.

Note that the detailed balance equations (7.2) are equivalent to the conditions

$$P_{xy} = \tilde{P}_{xy} \quad \text{for all } x, y \in \mathcal{S}\,,$$

and so, for a reversible Markov chain, we have

$$\mathbf{P} = \tilde{\mathbf{P}} = \mathbf{D}^{-1}\mathbf{P}^t\mathbf{D}$$

where $\mathbf{P}^t$ denotes the transpose of $\mathbf{P}$ and $\mathbf{D} = \text{diag}(\pi(1), ..., \pi(n))$.

Lemma 7.2.1. *If a Markov chain $(X_n)_{n\geq 0}$ is reversible with respect to a positive probability measure π, then π is a stationary distribution for the Markov chain. Hence the Markov chain is positive recurrent.*

Proof. Summing over x on both sider of Equation (7.2) yields

$$\sum_{x\in\mathcal{S}} \pi(x)P_{xy} = \sum_{x\in\mathcal{S}} \pi(y)P_{yx} = \pi(y)$$

which shows that π is a stationary distribution. The existence of a strictly positive stationary distribution implies positive recurrence of the chain. □

By Proposition 7.1.1, for a reversible chain $(X_n)_{n\geq 0}$ that starts in stationary distribution π, for any $n \geq 1$ and $x_0, x_1, ..., x_n \in \mathcal{S}$, we have

$$\mathbb{P}(X_0 = x_0,\, X_1 = x_1,\, ...,\, X_n = x_n) = \mathbb{P}(X_0 = x_n,\, X_1 = x_{n-1},\, ...,\, X_n = x_0)\,,$$

that is, we have equality of the joint distributions

$$(X_0, X_1, ..., X_n) \sim (X_n, ..., X_1, X_0)\,.$$

Observing a reversible chain as it unfolds, we cannot tell, probabilistically speaking, whether time runs forward or time runs backward.

Remark 7.2.2. *Equations (7.2) often lead to a manageable recursion for the computation of the stationary distribution π of an irreducible, positive recurrent, and reversible Markov chain. However, if the Markov chain is not reversible (and of course this applies to many Markov chains) and/or not positive recurrent, then the system of equations (7.2) will not have a positive solution that can be normalized to a probability distribution π on $\mathcal{S}$.*
Vice versa, equations (7.2) are often used to construct a Markov chain (i.e., its transition probabilities) that has a desired stationary distribution π. This is relevant for the construction of Markov chain Monte Carlo algorithms which are the subject of Chapter 9.

Example 7.2.1. Biased random walk (with $p \neq \frac{1}{2}$) on the discrete cycle $\mathbb{Z}_d$ is not reversible (see Example 7.1.1). Symmetric random walk (with $p = \frac{1}{2}$) on $\mathbb{Z}_d$ is reversible. □

Symmetric random walk on $\mathbb{Z}_d$ is an example of a Markov chain that has a symmetric transition matrix $\mathbf{P}$. Any finite-state Markov chain with symmetric transition matrix $\mathbf{P}$ is reversible with respect to uniform measure on its state space $\mathcal{S}$.

Example 7.2.2 (Random walk on a weighted graph is reversible)**.** Let $G(V, E)$ be a connected, finite graph and $C : E \to \mathbb{R}^+$ a weight function defined on the edge set E. We will write $C(v, w)$ instead of $C(\{v, w\})$. We have introduced random walk on $G(V, E)$ with weight function C in Section 1.5. Let $C(v) = \sum_{w:w\sim v} C(v, w)$. Recall that for two vertices $v, w \in V$, the one-step transition probability is defined by

$$P_{v,w} = \frac{C(v, w)}{C(v)} \quad \text{if } w \sim v\,,$$

and is zero otherwise. Let $C_G = \sum_{w \in V} C(w)$. Consider the distribution π on V defined by

$$\pi(v) = \frac{C(v)}{C_G} \qquad \text{for } v \in V\,.$$

Since $\sum_{v \in V} \frac{C(v)}{C_G} = 1$, the distribution π is a probability distribution. We also verify that the detailed balance equations hold for all $v, w \in V$:

$$\frac{C(v)}{C_G}\frac{C(v,w)}{C(v)} = \frac{C(w)}{C_G}\frac{C(w,v)}{C(w)}\,.$$

This shows that π is the unique stationary distribution, and that random walk on the weighted graph $G(V, E)$ with weight function C is reversible. Since simple random walk on a graph $G(V, E)$ is a special case (corresponding to weight function $C \equiv 1$), simple random walk on a graph is reversible. □

The following theorem gives an alternate criterion for reversibility of an irreducible Markov chain. Note that it does not require knowledge of the stationary distribution.

Theorem 7.2.3 (Kolmogorov's loop criterion)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain with state space $\mathcal{S}$. Then $(X_n)_{n\geq 0}$ is reversible if and only if the product of the one-step transition probabilities along any finite loop is the same as the product of the one-step transition probabilities along the reversed loop, that is, if for any $n \geq 2$ and any states $i_1, i_2, ..., i_n \in \mathcal{S}$,*

$$P_{i_1,i_2} P_{i_2,i_3} \cdots P_{i_{n-1},i_n} P_{i_n,i_1} = P_{i_1,i_n} P_{i_n,i_{n-1}} \cdots P_{i_2,i_1} \tag{7.3}$$

holds.

Proof. Assume $(X_n)_{n\geq 0}$ is reversible. Then $P_{i,j} = \frac{\pi(j)}{\pi(i)} P_{j,i}$ for all $i, j \in \mathcal{S}$. Replacing each factor $P_{i,j}$ on the left hand side of (7.3) with $\frac{\pi(j)}{\pi(i)} P_{j,i}$ yields the right hand side of (7.3). Conversely, assume that (7.3) holds. Then

$$\sum_{i_2,\ldots,i_{n-1}} P_{i_1,i_2} P_{i_2,i_3} \cdots P_{i_{n-1},i_n} P_{i_n,i_1} = \sum_{i_2,\ldots,i_{n-1}} P_{i_1,i_n} P_{i_n,i_{n-1}} \cdots P_{i_2,i_1}$$

from which we get

$$P^{n-1}_{i_1,i_n} P_{i_n,i_1} = P_{i_1,i_n} P^{n-1}_{i_n,i_1} \quad .$$

It follows that for all $n \geq 1$ and $x, y \in \mathcal{S}$ we have

$$P^n_{x,y} P_{y,x} = P^n_{y,x} P_{x,y}\,. \tag{7.4}$$

Let π denote the stationary distribution of the Markov chain. Taking the limit of the Césaro averages on both sides of (7.4) yields

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}P_{x,y}^{k}P_{y,x}=\pi(y)P_{y,x}=\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}P_{y,x}^{k}P_{x,y}=\pi(x)P_{x,y}$$

which shows reversibility of the chain $(X_n)_{n\geq 0}$. □

Note: Kolmogorov's loop criterion (7.3) may also be satisfied for a null recurrent Markov chain which has (necessarily) infinite state space $\mathcal{S}$. And the detailed balance equations (7.2) may have a positive solution, but this solution then cannot be normalized to a probability distribution π on $\mathcal{S}$. As an example, consider simple symmetric random walk on $\mathbb{Z}$ which is null recurrent. There is no stationary distribution. But the loop criterion (7.3) clearly holds. The detailed balance equations (7.2) have solution $\pi(x)\equiv c$ for any $c>0$.

Example 7.2.3 (Birth/death chains are reversible)**.** Consider an irreducible, positive recurrent birth/death chain. Without loss of generality we can take $\mathcal{S}=\{0,1,...,n\}$ in case $\mathcal{S}$ is finite, and $\mathcal{S}=\mathbb{N}_0$ in case $\mathcal{S}$ is infinite. Consider a finite loop $x_1\to x_2\to\cdots\to x_n\to x_1$. Since for an irreducible birth/death chain $P_{xy}\neq 0$ if and only if $|x-y|=1$, we may assume $|x_k-x_{k+1}|=1$ for $1\leq k\leq n-1$ and $|x_n-x_1|=1$. (Otherwise, both sides in (7.3) are 0.) But for any loop $x_1\to x_2\to\cdots\to x_n\to x_1$ with this property, the reversed loop $x_1\to x_n\to\cdots\to x_2\to x_1$ must make the *same* collection of one-step transitions as the original loop, possibly some of them at different times than the original loop. For this reason, equality (7.3) is satisfied, and it follows that any irreducible, positive recurrent birth/death chain is reversible. The detailed balance equations (7.2) give a recursion for computing the unique stationary distribution π. □

7.2.1 Linear-algebraic interpretation of reversibility

Let $\mathcal{S}$ be a finite state space which we identify with $\mathcal{S}=\{1,...,n\}$, and consider the real vector space $V=\mathbb{R}^n$ consisting of all functions

$$f:\mathcal{S}\to\mathbb{R}\,.$$

Let π be the stationary distribution of an irreducible Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{S}$ with transition matrix $\mathbf{P}$. The corresponding *Markov operator* $M:V\to V$ is the linear operator defined by

$$M(f)=\mathbf{P}f=\left(\sum_{j=1}^{n}P_{ij}f(j)\right)_{1\leq i\leq n}$$

where we identify $f \in V$ with a column vector. We can define an inner product $\langle \cdot, \cdot \rangle_\pi$ on V by

$$\langle f, g \rangle_\pi = \sum_{x \in \mathcal{S}} f(x) g(x) \pi(x) \quad \text{for } f, g \in V .$$

Lemma 7.2.4. *The Markov chain $(X_n)_{n \geq 0}$ is reversible if and only if the corresponding Markov operator M is self-adjoint, that is, if and only if*

$$\langle \mathbf{P} f, g \rangle_\pi = \langle f, \mathbf{P} g \rangle_\pi \tag{7.5}$$

for all $f, g \in V$.

Proof. Assume (7.5) holds and apply (7.5) to the functions $f = \mathbb{1}_{\{i\}}$ and $g = \mathbb{1}_{\{j\}}$ for $i \neq j$. Since $\mathbf{P} f = (P_{ki})_{1 \leq k \leq n}$ and $\mathbf{P} g = (P_{kj})_{1 \leq k \leq n}$, we get

$$\langle \mathbf{P} f, g \rangle_\pi = P_{ji} \pi(j) = \langle f, \mathbf{P} g \rangle_\pi = P_{ij} \pi(i)$$

which are the detailed balance equations (7.2), and so $(X_n)_{n \geq 0}$ is reversible.
Conversely, assume $(X_n)_{n \geq 0}$ is reversible. We have

$$\begin{aligned}
\langle \mathbf{P} f, g \rangle_\pi &= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} P_{ij} f(j) \right) g(i) \pi(i) \\
&= \sum_{i,j \in \mathcal{S}} P_{ij} f(j) g(i) \pi(i) \\
&= \sum_{i,j \in \mathcal{S}} P_{ji} f(j) g(i) \pi(j) \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} P_{ji} g(i) \right) f(j) \pi(j) = \langle f, \mathbf{P} g \rangle_\pi
\end{aligned}$$

for all $f, g \in V$, and so M is self-adjoint. □

We often need to understand the eigenvalues of the transition matrix $\mathbf{P}$. The importance of the reversibility condition stems from the *Spectral Theorem* for self-adjoint operators on a finite-dimensional inner product space. We quote a version of this classical theorem for real vector spaces.

Theorem 7.2.5 (Spectral Theorem)**.** *Let V be a finite-dimensional real vector space with inner product $\langle \cdot, \cdot \rangle$ and $T : V \to V$ a linear operator. Then T is self-adjoint if an only if V has an orthonormal basis consisting of eigenvectors of T. In particular, for a self-adjoint operator T, all eigenvalues are real valued.*

If a given Markov chain $(X_n)_{n\geq 0}$ is not reversible, it is often helpful to work with a modified version of $(X_n)_{n\geq 0}$ that *is* reversible. See Exercise 7.4 for possible "reversiblizations" of a Markov chain.

Exercises

Exercise 7.1. Consider an irreducible, positive recurrent birth/death chain $(X_n)_{n\geq 0}$ and its stationary distribution π. Verify that the detailed balance equations hold for π.

Exercise 7.2. Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ that is reversible with respect to a positive probability measure π on $\mathcal{S}$. Fix $k_0 \in \mathbb{N}$. Show that the Markov chain $(Y_n)_{n\geq 0}$ defined by $Y_n = X_{nk_0}$ is also reversible with respect to π.

Exercise 7.3. Consider a positive recurrent Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S}$ and a strictly positive stationary distribution π on $\mathcal{S}$. Assume the chain is reducible.

(a) Show that the transition probabilities $\tilde{P}_{xy}$, $x, y \in \mathcal{S}$, for the time-reversal chain $(\tilde{X}_n)_{n\geq 0}$ do not depend on the choice of π.

(b) Show that $(\tilde{X}_n)_{n\geq 0}$ has the same class structure, i.e. the same irreducible closed classes, as the original chain $(X_n)_{n\geq 0}$.

Exercise 7.4 (Additive and multiplicative reversiblizations)**.** Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on finite state space $\mathcal{S}$ with stationary distribution π and transition matrix $\mathbf{P}$. We assume that $(X_n)_{n\geq 0}$ is not reversible. Let $\tilde{\mathbf{P}}$ be the transition matrix of its time reversal.

(a) Show that the matrix $\mathbf{P}\tilde{\mathbf{P}}$ defines the transition matrix of a reversible Markov chain on $\mathcal{S}$. This chain is called the multiplicative reversiblization of $(X_n)_{n\geq 0}$.

(b) Show that the matrix $\frac{1}{2}(\mathbf{P} + \tilde{\mathbf{P}})$ defines the transition matrix of a reversible Markov chain on $\mathcal{S}$. This chain is called the additive reversiblization of $(X_n)_{n\geq 0}$.

Exercise 7.5. Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on finite state space $\mathcal{S}$ with stationary distribution π and transition matrix $\mathbf{P}$. Consider the diagonal matrix $\mathbf{D} = \mathrm{diag}(\pi(1), ..., \pi(n))$ and its square root $\mathbf{D}^{\frac{1}{2}} = \mathrm{diag}(\sqrt{\pi(1)}, ..., \sqrt{\pi(n)})$. Define the matrix

$$\hat{\mathbf{P}} = \mathbf{D}^{\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}}\,.$$

Show that $(X_n)_{n\geq 0}$ is reversible if and only if $\hat{\mathbf{P}}$ is a symmetric matrix.

Chapter 8

Markov Chains and Electric Networks

8.1 Reversible chains and graph networks

We have introduced random walk on a weighted graph in Section 1.5. In Example 7.2.2, we showed that random walk on a weighted graph is reversible, and we computed the stationary distribution. Conversely, any reversible Markov chain can be viewed as a random walk on a weighted graph: Assume $(X_n)_{n\geq 0}$ is an irreducible and reversible Markov chain with state space $\mathcal{S}$ and stationary distribution π. In order to frame this Markov chain as a random walk on a weighted graph, we view $\mathcal{S}$ as the vertex set V of a graph $G(V, E)$. Let $x, y \in \mathcal{S}$. As a consequence of reversibility, if $P_{xy} > 0$, then $P_{yx} > 0$, and

$$\pi(x)P_{xy} = \pi(y)P_{yx}\,.$$

We take $\{x, y\} \in E \iff P_{xy} > 0$, and we assign weight

$$C(x, y) = \pi(x)P_{xy}$$

to the edge $\{x, y\}$. We can directly verify that random walk on the so constructed weighted graph yields the same transition probabilities $\mathbf{P}$ as the ones for the Markov chain $(X_n)_{n\geq 0}$.

Viewing a reversible Markov chain as a random walk on a weighted graph allows us to interpret the process as an electrical network. In taking this analogy further, we can apply familiar laws from physics about voltage, current, resistance, etc. to the study of random walks on graphs. This often turns out very useful for computations. The electric network interpretation of random walks on graphs is the central focus of this chapter.

Definition 8.1.1. *Let $G(V, E)$ be a connected graph and $C : E \to \mathbb{R}^+$ a positive function on the edge set E.*

- *We call $G(V, E)$ together with the function C (i.e., the weighted graph) a* **network** *and denote this network by (G, C).*
- *We refer to the value $C(x, y)$ as the* **conductance** *of the edge $\{x, y\}$, and to the value $R(x, y) = C(x, y)^{-1}$ as the* **resistance** *of the edge $\{x, y\}$.*
- *Let $G^*(V^*, E^*)$ be a subgraph of $G(V, E)$ and $C^* = C|_{E^*}$ the restriction of C to $E^* \subseteq E$. We say (G^*, C^*) is a* **subnetwork** *of (G, C).*

Notation. We will use the following notation: We will often write $x \sim y$ for $\{x, y\} \in E$, and set

$$C(x) = \sum_{y \sim x} C(x, y) \quad \text{for } x \in V\,,$$

and

$$C_G = \sum_{x \in V} C(x)\,.$$

Note that for the special case of $C \equiv 1$, we have $C(x) = \deg(x)$ and $C_G = 2|E|$.

Example 8.1.1. Consider an irreducible Markov chain $(X_n)_{n \geq 0}$ on state space $\mathcal{S} = \{a, b, c, d, e, f\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & \frac{1}{5} & \frac{4}{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{6} & 0 & 0 & \frac{1}{6} & 0 & \frac{2}{3} \\ \frac{4}{7} & 0 & \frac{1}{7} & 0 & \frac{2}{7} & 0 \\ 0 & \frac{1}{5} & 0 & \frac{2}{5} & 0 & \frac{2}{5} \\ 0 & \frac{1}{7} & \frac{4}{7} & 0 & \frac{2}{7} & 0 \end{pmatrix}.$$

A straightforward computation yields the stationary distribution $\pi = (\frac{5}{32}, \frac{1}{16}, \frac{3}{16}, \frac{7}{32}, \frac{5}{32}, \frac{7}{32})$. We also verify that the detailed balance equations

$$\pi(x)P_{xy} = \pi(y)P_{yx} \quad \text{for all } x, y \in \mathcal{S}$$

hold, so $(X_n)_{n \geq 0}$ is reversible. Setting

$$C(x, y) = \pi(x)P_{xy}\,,$$

we get the corresponding network (or weighted graph) as shown in Figure 8.1 with edge conductances $C(x, y)$ marked in blue. The given Markov chain $(X_n)_{n \geq 0}$ is random walk on this graph network. □

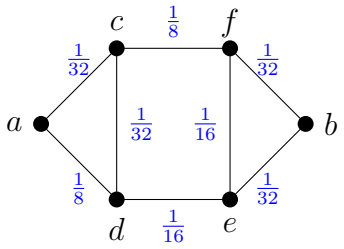


Figure 8.1

8.2 Harmonic functions

Let $\mathcal{S}$ be a discrete space, $(X_n)_{n\geq 0}$ a Markov chain with state space $\mathcal{S}$, and f a function on $\mathcal{S}$. Recall Definition 6.5.1 for discrete harmonic functions: We say $f : \mathcal{S} \to \mathbb{R}$ is *harmonic at* $x \in \mathcal{S}$ if

$$f(x) = \sum_{y\in\mathcal{S}} P_{xy}\, f(y)\,,$$

that is,we have $f(x) = \mathbb{E}(f(X_{n+1}) \,|\, X_n = x)$. If $A \subseteq \mathcal{S}$ and f is harmonic at each $z \in A$, we say f is harmonic on A.

Proposition 8.2.1 (Superposition principle)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and $A \subseteq \mathcal{S}$. Let f and g be two functions on $\mathcal{S}$ which are harmonic on A. Then for any constants $\alpha, \beta \in \mathbb{R}$, the function $\alpha f + \beta g$ is also harmonic on A.*

The proof of Proposition 8.2.1 is straightforward and left as an exercise. The notion of harmonicity of a function plays a major role in this chapter. We start by collecting a few basic facts about harmonic functions. First, recall Theorem 6.5.2. It states that for an irreducible and *recurrent* Markov chain, any function f that is harmonic *everywhere* on $\mathcal{S}$ must be a constant function.

Definition 8.2.1. *Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ and $W \subseteq \mathcal{S}$ a non-empty subset. Let $f_W : W \to \mathbb{R}$. A function $f : \mathcal{S} \to \mathbb{R}$ is called a* **harmonic extension** *of f_W to $\mathcal{S}$ if f_W and f agree on W and f is harmonic on $\mathcal{S} \setminus W$.*

Proposition 8.2.2 (Maximum principle)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on finite state space $\mathcal{S}$. Let $W \subset \mathcal{S}$ be a non-empty subset and f a function on $\mathcal{S}$ that is harmonic on $\mathcal{S} \setminus W$. Then there exists $w \in W$ such that $f(w) = \max_{x\in\mathcal{S}} f(x)$.*

Proof. Let $b = \max_{x\in\mathcal{S}} f(x)$ and consider the set $B = \{z \in \mathcal{S} : f(z) = b\}$. Let $z_0 \in B$. If $z_0 \in W$, there is nothing to prove. Assume $z_0 \in \mathcal{S} \setminus W$. Fix $w_0 \in W$. Since $(X_n)_{n\geq 0}$ is irreducible, z_0 leads to w_0, so there exists a sequence of states $z_1, ..., z_n$ with $z_n = w_0$ for which $P_{z_0 z_1} > 0$, $P_{z_1 z_2} > 0, ..., P_{z_{n-1} z_n} > 0$. Let $k_0 = \min_{1\leq k\leq n}\{k : z_k \in W\}$.
Since f is harmonic at z_0, we have

$$f(z_0) = b = \sum_{y\in\mathcal{S}} P_{z_0 y} f(y)\,. \tag{8.1}$$

Assume that for at least one $y \in \mathcal{S}$ for which $P_{z_0 y} > 0$ we have $f(y) < b$. But then (8.1) leads to the contradiction $b < b$. Hence we dismiss the assumption and, in particular, get $f(z_1) = b$. Applying the same argument to $f(z_1)$, ..., $f(z_{k_0-1})$ (which is valid since $z_1, ..., z_{k_0-1} \in \mathcal{S} \setminus W$), we arrive at $f(z_{k_0}) = b$. Since we assumed $z_{k_0} \in W$, the claim is proved. □

Proposition 8.2.3 (Uniqueness principle)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on finite state space $\mathcal{S}$ and let $W \subset \mathcal{S}$ be a non-empty subset. If f_1 and f_2 are two functions on $\mathcal{S}$ that are both harmonic on $\mathcal{S} \setminus W$ and agree on W (so $f_1(w) = f_2(w)$ for all $w \in W$), then*

$$f_1 = f_2\,.$$

Proof. Consider the function $f_1 - f_2$ which is 0 on W and harmonic on $\mathcal{S} \setminus W$. By the Maximum principle, $(f_1 - f_2) \leq 0$ on $\mathcal{S} \setminus W$. Reversing the roles of f_1 and f_2, we get $(f_2 - f_1) \leq 0$. It follows that $f_1 = f_2$ on $\mathcal{S}$. □

Proposition 8.2.4 (Existence)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on finite state space $\mathcal{S}$ and let $W \subset \mathcal{S}$ be a non-empty subset. Let $f_W : W \to \mathbb{R}$ be a function on W, and consider the hitting time $T^W = \min\{n \geq 0 : X_n \in W\}$. Then the function $f : \mathcal{S} \to \mathbb{R}$ defined by*

$$f(x) = \mathbb{E}_x[f_W(X_{T^W})] \qquad \text{for } x \in \mathcal{S}$$

is the unique harmonic extension of f_W to $\mathcal{S}$.

Proof. Clearly, for $w \in W$, $f(w) = \mathbb{E}_w[f_W(X_0)] = f_W(w)$, so f and f_W agree on W. We need to show that f is harmonic on $\mathcal{S} \setminus W$. Let $x \in \mathcal{S} \setminus W$ and assume $X_0 = x$. Conditioning on the state the Markov chain is in at time $n = 1$, we get

$$f(x) = \mathbb{E}_x[f_W(X_{T^W})] = \sum_{y\in\mathcal{S}} P_{xy}\, \mathbb{E}_x[f_W(X_{T^W}) \,|\, X_1 = y]\,.$$

By the Markov property, $\mathbb{E}_x[f_W(X_{T^W}) \,|\, X_1 = y] = \mathbb{E}_y[f_W(X_{T^W})] = f(y)$ for all $y \in \mathcal{S}$, and so we have

$$f(x) = \sum_{y\in\mathcal{S}} P_{xy}\, f(y) \qquad \text{for all } x \in \mathcal{S} \setminus W$$

which shows that f is a harmonic extension of f_W to $\mathcal{S}$. Uniqueness of f follows from the uniqueness principle. □

Example 8.2.1. Consider simple random walk on $\mathcal{S} = \{0, 1, ..., N\}$ with reflecting boundaries at 0 and at N whose transition graph is shown in Figure 8.2.

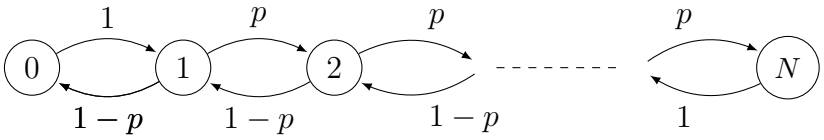


Figure 8.2

Let $W = \{0, N\}$ and define $f_W(0) = 1$ and $f_W(N) = 0$. Then the unique harmonic extension of f_W to $\mathcal{S}$ is

$$f(k) = \mathbb{E}_k[f_W(X_{T^W})] = \mathbb{P}_k(T^0 < T^N) \qquad \text{for } 1 \leq k \leq N - 1\,.$$

The function values $f(k)$ are the gambler's ruin probabilities (see Section 4.4). □

8.3 Voltage and Current

Throughout this section we assume that $G(V, E)$ is a finite, connected graph that has no self loops.

Definition 8.3.1. *Let (G, C) be a finite, irreducible network with vertex set V. Consider two vertices $a, b \in V$. A function Φ on V that is harmonic on $V \setminus \{a, b\}$ is called a* **voltage** *on (G, C). If $\Phi(a) \geq \Phi(b)$, we call a the* **source** *and b the* **sink** *of the network.*

Note that by the results from Section 8.2, for given boundary values $\Phi(a)$ and $\Phi(b)$, a voltage exists and is unique. We can interpret (G, C) together with a voltage on it as an electric circuit where electricity flows along its weighted edges (i.e., conductors between vertices with given resistances) according to known laws of physics.

Definition 8.3.2. *Let (G, C) be a finite, irreducible network and $a, b \in V$. A* **flow from** *a* **to** *b is an anti-symmetric function I on the* oriented *edges of (G, C), that is,*

$$I(x, y) = -I(y, x) \quad \text{for } \{x, y\} \in E\,, \tag{8.2}$$

for which **Kirchhoff's node** *(or* **Kirchhoff's current***)* **law** *holds for all $x \in V \setminus \{a, b\}$, that is, we have*

$$\sum_{y : y \sim x} I(x, y) = 0 \quad \text{for all } x \in V \setminus \{a, b\}\,, \tag{8.3}$$

and

$$\sum_{y : y \sim a} I(a, y) \geq 0\,.$$

Definition 8.3.3. *Given a voltage Φ on (G, C), its* **current flow** *I_Φ is the flow defined by*

$$I_\Phi(x, y) = \frac{\Phi(x) - \Phi(y)}{R(x, y)} \quad \text{for } \{x, y\} \in E\,. \tag{8.4}$$

Ohm's law: The definition of current flow in (8.4),

$$I_\Phi(x, y) = \frac{\Phi(x) - \Phi(y)}{R(x, y)}$$

matches *Ohm's law* for electrical circuits which states that the electrical current between two nodes equals their potential difference divided by the resistance of the conductor that connects them.

Notes: (1) For any flow I from a to b on a network (G, C), the antisymmetry property on oriented edges implies

$$\sum_{x\in V}\left(\sum_{y:y\sim x} I(x,y)\right) = \sum_{\{x,y\}\in E}[I(x,y)+I(y,x)] = 0\,. \tag{8.5}$$

(2) Kirchhoff's node law indeed holds for the current flow I_Φ associated with a voltage Φ on a network: For all vertices $x \in V \setminus \{a, b\}$, we have

$$\begin{aligned}\sum_{y:y\sim x} I_\Phi(x,y) &= \sum_{y:y\sim x}\frac{\Phi(x)-\Phi(y)}{R(x,y)} = C(x)\sum_{y\sim x}\frac{C(x,y)}{C(x)}[\Phi(x)-\Phi(y)]\\ &= C(x)\left[\sum_{y:y\sim x} P_{xy}\Phi(x) - \sum_{y:y\sim x} P_{xy}\Phi(y)\right]\\ &= C(x)\,[\Phi(x)-\Phi(x)] = 0\,.\end{aligned}$$

(3) A current flow does not change if we add the same constant c to the boundary values $\Phi(a)$ and $\Phi(b)$ (which then adds c to the resulting voltage on V). For this reason, we can take one of the boundary values to be 0 and the other to be positive. Without loss of generality, we will assume $\Phi(a) > 0$ (which makes a the source) and $\Phi(b) = 0$ (which makes b the sink). Since the network is irreducible, by Proposition 8.2.4, at least one neighboring vertex of a has voltage strictly smaller than $\Phi(a)$. Therefore, at the source a,

$$\sum_{y:y\sim a} I_\Phi(a,y) > 0\,. \tag{8.6}$$

Definition 8.3.4. *Let I be a flow from a to b. We call $\|I\| = \sum_{y:y\sim a} I(a,y)$ the* **strength** *of the flow I. If $\|I\| = 1$, we call the flow I a* **unit flow**.

By (8.6), for a current flow I_Φ, we have

$$\|I_\Phi\| > 0\,,$$

and by (8.3), (8.5), and (8.6), we have at the sink b,

$$\sum_{y:y\sim b} I_\Phi(b,y) = -\|I_\Phi\| < 0\,.$$

Proposition 8.3.1 (Probabilistic interpretation of voltage I)**.** *Let (G, C) be a finite, irreducible network, and consider random walk on the network. Let $a, b \in V$. Then for all $x \in V$ and for the given boundary values $\Phi(a) = 1$ and $\Phi(b) = 0$, the voltage $\Phi(x)$ at x is*

$$\Phi(x) = \mathbb{P}_x(T^a < T^b)$$

for the random walk on (G, C).

Proof. Let $W = \{a, b\}$. By Proposition 8.2.4, the unique voltage Φ at x is

$$\Phi(x) = 1 \cdot \mathbb{P}_x(T^W = a) + 0 \cdot \mathbb{P}_x(T^W = b) = \mathbb{P}_x(T^a < T^b)\,.$$

(Note: Example 8.2.1 shows a special case.) □

Next, we are interested in the probabilistic meaning of current flow. For this, we will consider the random variable

$$V^x_{T^b} = \sum_{k=0}^{T^b-1} \mathbb{1}_{\{x\}}\,, \tag{8.7}$$

which is the number of visits to state x before the chain visits b for the first time. We will assume that the chain starts in state a and will use the notation

$$v(x) = \mathbb{E}_a(V^x_{T^b})\,.$$

Proposition 8.3.2 (Probabilistic interpretation of current)**.** *Let (G, C) be a finite, irreducible network, let $a, b \in V$, and consider random walk $(X_n)_{n\geq 0}$ on the network that starts at vertex a. Let Φ be a voltage on (G, C) with source a and sink b. For any edge $\{y, z\}$ with $y, z \neq b$, the current flow $I_\Phi(y, z)$ is proportional to the* **expected number of directed edge crossings** *between y and z, that is, the expected number of crossings from y to z minus the expected number of crossings from z to y before the chain visits state b for the first time. More precisely, for a given voltage Φ, there exists a constant c_Φ such that for every $y, z \in V \setminus \{b\}$ with $y \sim z$,*

$$I_\Phi(y, z) = c_\Phi \left[v(y)P_{yz} - v(z)P_{zy}\right]\,. \tag{8.8}$$

Proof. Note that $v(b) = \mathbb{E}_a(V_{T^b}^b) = 0$. Let $x \neq b$. If we modify the random walk by changing b into an absorbing state, then in the notation of Section 2.3, we have

$$v(x) = (\mathbf{V})_{a,x}\,,$$

that is, $v(x)$ is the (a,x)-entry in the fundamental matrix $\mathbf{V}$ for the transition matrix $\mathbf{P}$ of the modified chain. Recall that $\mathbf{V}(\mathbf{I}-\mathbf{Q}) = \mathbf{I}$, and so

$$\mathbf{VQ} = \mathbf{V} - \mathbf{I}\,. \tag{8.9}$$

For $x \neq a$, we get for the (a,x)-entry of the matrix defined by either side of (8.9),

$$\sum_{y\in V}(\mathbf{V})_{a,y}P_{yx} = (\mathbf{V})_{a,x}\,, \qquad \text{and so } \sum_{y\in V} v(y)P_{yx} = v(x)\,. \tag{8.10}$$

Because of the reversibility of the random walk on (G,C), we have $C(x)P_{xy} = C(y)P_{yx}$. Substituting $P_{yx} = P_{xy}C(x)/C(y)$ in (8.10) yields

$$\frac{v(x)}{C(x)} = \sum_{y\in V} P_{xy}\frac{v(y)}{C(y)}\,,$$

which shows that the function

$$\tilde{\Phi}(z) = \frac{v(z)}{C(z)} \qquad \text{for } z \in V \tag{8.11}$$

is harmonic on $V \setminus \{a,b\}$. It is the unique voltage corresponding to the boundary values $\tilde{\Phi}(a) = \dfrac{v(a)}{C(a)}$ and $\tilde{\Phi}(b) = 0$. The current flow $I_{\tilde{\Phi}}$ for this voltage is

$$\begin{aligned} I_{\tilde{\Phi}}(y,z) &= \left(\tilde{\Phi}(y) - \tilde{\Phi}(z)\right) C(y,z) \\ &= v(y)\frac{C(y,z)}{C(y)} - v(z)\frac{C(z,y)}{C(z)} = v(y)\,P_{yz} - v(z)\,P_{zy}\,. \end{aligned}$$

This shows that for the specific voltage (8.11) and any edge $\{y,z\}$ with $y,z \neq b$, the flow $I_{\tilde{\Phi}}(y,z)$ (which takes direction into account) is equal to the expected number of *directed* edge crossings $[v(y)\,P_{yz} - v(z)\,P_{zy}]$ before time T^b.

Now consider any voltage Φ with given boundary values $\Phi(a) > 0$ and $\Phi(b) = 0$. Define $c_\Phi = \Phi(a)/\tilde{\Phi}(a)$. By the Superposition principle (Proposition 8.2), we have $\Phi = c_\Phi\,\tilde{\Phi}$ on V, and hence $I_\Phi = c_\Phi\,I_{\tilde{\Phi}}$ on the directed edges of the network. This proves (8.8). □

Corollary 8.3.3. *For a given finite, irreducible network (G,C), the current flow corresponding to the voltage $\tilde{\Phi}$ defined by (8.11) is the* unit current flow. *That is,*

$$\|I_{\tilde{\Phi}}\| = 1.$$

Proof. By Proposition 8.3.2, $\|I_{\tilde{\Phi}}\| = \sum_{z:z\sim a} I_{\tilde{\Phi}}(a,z)$ is the expected number of times the random walk leaves state a minus the expected number of times the random walk enters state a, before hitting state b for the first time. This number must be 1, since the walk starts in a, and afterwards, for each time the walk returns to a, it will leave a. □

We now turn to the computation of $\tilde{\Phi}(a)$. We will denote by Φ_1 the voltage resulting from the boundary condition $\Phi_1(a) = 1$ and $\Phi_1(b) = 0$, and we will denote by I_1 the current flow resulting from Φ_1.

Lemma 8.3.4. *Let (G, C) be a finite, irreducible network and $a, b \in V$. With the above notation,*

$$\tilde{\Phi}(a) = \frac{1}{\|I_1\|}\,. \tag{8.12}$$

Proof. Note that multiplication of a voltage results in multiplication of the resulting current by the same factor. Thus

$$\frac{\tilde{\Phi}(a)}{\|I_{\tilde{\Phi}}\|} = \frac{\Phi_1(a)}{\|I_1\|}\,.$$

Since $\|I_{\tilde{\Phi}}\| = 1$ and $\Phi_1(a) = 1$, we have (8.12). □

The previous discussion yields the following, second interpretation of voltage.

Proposition 8.3.5 (Probabilistic interpretation of voltage II)**.** *Let (G, C) be a finite, irreducible network. Consider two distinct vertices $a, b \in V$ and random walk on the network starting at a. Let $\tilde{\Phi}$ denote the voltage on V defined by the boundary values $\tilde{\Phi}(b) = 0$ and $\tilde{\Phi}(a) > 0$ chosen such that the corresponding current flow is the unit current flow. Then*

$$\tilde{\Phi}(x) = \frac{\mathbb{E}_a(V^x_{T^b})}{C(x)} \quad \text{for all } x \in V \tag{8.13}$$

where $V^x_{T^b}$ is the number of visits to state x between (including) times 0 and $T^b - 1$.

As an immediate corollary of Proposition 8.3.5, we get the following identity for the *expected hitting time* of a state b:

Corollary 8.3.6. *Consider random walk on a finite network (G, C) and two distinct vertices $a, b \in V$. Then*

$$\mathbb{E}_a(T^b) = \sum_{x \in V} \tilde{\Phi}(x) C(x) \tag{8.14}$$

Example 8.3.1. Consider random walk on the network shown in Figure 8.3. Notice that this network is a scaled version of the network in Figure 8.1. We have scaled the conductances, but the resulting Markov chain is still the same as in Example 8.1.1. Compute (a) $\mathbb{P}_d(T^a < T^b)$, (b) $\mathbb{E}_a(V^a_{T^b})$, and (c) $\mathbb{E}_a(V^d_{T^b})$.

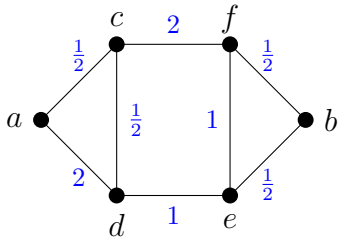


Figure 8.3

If we apply unit voltage at vertex a and zero voltage at vertex b, we get the voltages at the rest of the vertices as marked in Figure 8.4.

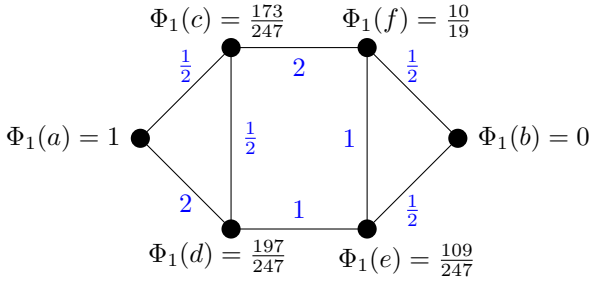


Figure 8.4: Voltages associated with the network in Figure 8.3

(a) By directly reading off the voltage at vertex d in Figure 8.4, we have

$$\mathbb{P}_d(T^a < T^b) = \Phi_1(d) = \frac{197}{247} \approx 0.80\,.$$

(b) First, we compute the strength of the flow. It is

$$\|I_1\| = I_1(a,c) + I_1(a,d) = (1 - \frac{173}{247})\frac{1}{2} + (1 - \frac{197}{247})2 = \frac{137}{247}\,.$$

By (8.11) and (8.12), we have

$$\frac{\mathbb{E}_a(V^a_{T^b})}{C(a)} = \tilde{\Phi}(a) = \frac{1}{\|I_1\|},$$

from which we conclude that

$$\mathbb{E}_a(V^a_{T^b}) = C(a)/\|I_1\| = 3\frac{247}{137} \approx 5.41\,.$$

(c) If the walk starts at vertex a, then the expected number of visits to vertex d before time T^b is

$$\mathbb{E}_a(V^d_{T^b}) = C(d)\Phi_1(d)/\|I_1\| = (\frac{7}{2})(\frac{197}{247})(\frac{247}{137}) \approx 5.03\,.$$

□

8.4 Effective resistance

Let (G, C) be a finite, irreducible network and $a, b \in V$. Let Φ be a voltage on V with given boundary values $\Phi(a) > 0$ and $\Phi(b) = 0$. Recall that the total flow at vertex a is equal to $\|I_\Phi\| > 0$, so in continuing with the physics analogy, electricity is flowing out of vertex a and into the network. By Kirchhoff's law, for any vertex $x \in V \setminus \{a, b\}$, the total flow at x is 0. Thus at vertex b, the total flow is $-\|I_\Phi\|$. The same amount of electricity that flows into the network at vertex a flows out of the network at vertex b.

Definition 8.4.1. *Let (G, C) be a finite, irreducible network, and let Φ be a voltage on V with source a and sink b. We define the* **effective resistance** *$R_{\text{eff}}(a, b)$ between a and b by*

$$R_{\text{eff}}(a, b) = \frac{\Phi(a) - \Phi(b)}{\|I_\Phi\|}. \tag{8.15}$$

Analogously, the **effective conductance** *$C_{\text{eff}}(a, b)$ between a and b is defined by*

$$C_{\text{eff}}(a, b) = R_{\text{eff}}(a, b)^{-1}\,.$$

Notice that the quotient in (8.15) does not depend on the specific voltage Φ. Hence $R_{\text{eff}}(a, b)$ is determined by the properties of the network (G, C) alone, and it can be computed from any voltage, say from Φ_1, as in the following proposition.

Proposition 8.4.1. *Let (G, C) be a finite, irreducible network, and let $a, b \in V$. Consider the voltage Φ_1 on V with boundary values $\Phi_1(a) = 1$ and $\Phi_1(b) = 0$, and let I_1 be its flow. Then*

$$R_{\text{eff}}(a, b) = \frac{1}{\|I_1\|} \qquad \text{and} \qquad C_{\text{eff}}(a, b) = \|I_1\| .$$

Interpretation of effective resistance: Imagine a second, small network consisting of only the two vertices a and b and a single edge $\{a, b\}$ with conductance $C_{\text{eff}}(a, b)$ (and therefore resistance $R_{\text{eff}}(a, b)$). If we apply the same voltage to a and b in both networks, the resulting current out of a and into b will be the same in both networks. Equivalently, we can say that $R_{\text{eff}}(a, b)$ is the difference in voltage at a and at b that is needed in order to create a unit current flow from a to b in the network.

We can now establish a connection between effective resistance and **escape probabilities**. Assume the random walk on the network starts at vertex x, and y is another vertex, distinct from x. We call the probability $\mathbb{P}_x(T^y < T^x)$, i.e., the probability that the walk will visit y *before* it returns back to x for the first time, an escape probability.

Proposition 8.4.2 (Escape probability)**.** *Let (G, C) be a finite, irreducible network, let $x, y \in V$ be two distinct vertices, and consider random walk on the network. Then*

$$\mathbb{P}_x(T^y < T^x) = \frac{1}{C(x)\, R_{\text{eff}}(x, y)} = \frac{C_{\text{eff}}(x, y)}{C(x)} .$$

Proof. We define a voltage Φ_1 on V with $\Phi_1(x) = 1$ and $\Phi_1(y) = 0$. The strength of the associated current flow is

$$\begin{aligned} \|I_1\| &= \sum_{z:z\sim x} (\Phi_1(x) - \Phi_1(z))C(x, z) = \sum_{z:z\sim x} (1 - \Phi_1(z))C(x, z) \\ &= C(x) - \sum_{z:z\sim x} \Phi_1(z)C(x, z) = C(x) \left(1 - \sum_{z:z\sim x} \Phi_1(z)\frac{C(x, z)}{C(x)}\right) \\ &= C(x) \left(1 - \sum_{z:z\sim x} P_{xz}\Phi_1(z)\right) = C(x)\mathbb{P}_x(T^y < T^x) . \end{aligned}$$

Since $R_{\text{eff}}(x, y) = 1/\|I_1\|$, it follows that

$$\mathbb{P}_x(T^y < T^x) = \frac{1}{C(x)R_{\text{eff}}(x, y)} .$$

□

Example 8.4.1. We continue Example 8.3.1. It is random walk on the network shown in Figure 8.3, for which we have computed $\|I_1\| = \frac{137}{247}$. Hence the effective resistance between a and b is $R_{\text{eff}}(a,b) = \frac{247}{137}$, and we get for the escape probability,

$$\mathbb{P}_a(T^b < T^a) = (\frac{2}{5})\frac{137}{247} \approx 0.22\,.$$

□

When computing the effective resistance (or effective conductance) between two vertices, it is often helpful to **simplify the network, without changing voltages and currents in the network**. Towards this end, we can apply any of the following four simplifications:

(1) Series law (resistances in series add): Let $a \neq b$. We would like to replace the two edges $\{a,v\}$ and $\{v,b\}$ with common endpoint v and with respective resistances $R_1 = R(a,v)$ and $R_2 = R(v,b)$ by a single edge $\{a,b\}$ with resistance R, without changing the current flow.

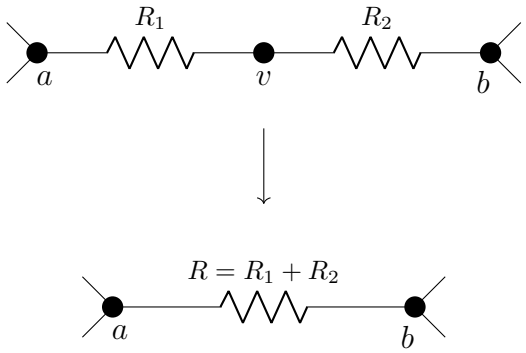


Figure 8.5: Series law

We assume the voltages $\Phi(a)$ and $\Phi(b)$ are the same for both networks. By Kirchhoff's law, $I = I(a,v) = I(v,b)$ in the above network. We would like the modified network to have the same flow I. By Ohm's law,

$$I = \frac{\Phi(a) - \Phi(v)}{R_1} = \frac{\Phi(v) - \Phi(b)}{R_2} = \frac{\Phi(a) - \Phi(b)}{R}$$

which yields

$$R_1\, I + R_2\, I = R\, I\,,$$

from which we get

$$R = R_1 + R_2\,.$$

(2) Parallel law (conductances in parallel add): Here we have two distinct edges with same endpoints a and b and with respective resistances R_1 and R_2 (and respective conductances C_1 and C_2). We would like to replace the two edges with a single edge $\{a, b\}$ with resistance R (and conductance C), without changing the current flow.
Again, we assume the voltages $\Phi(a)$ and $\Phi(b)$ are the same for both networks. Here the current flow at vertex a is

$$I = \frac{\Phi(a) - \Phi(b)}{R_1} + \frac{\Phi(a) - \Phi(b)}{R_2} = \frac{\Phi(a) - \Phi(b)}{R}\,.$$

From this we compute

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}\,,$$

and so

$$R = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \quad \text{and} \quad C = C_1 + C_2\,.$$

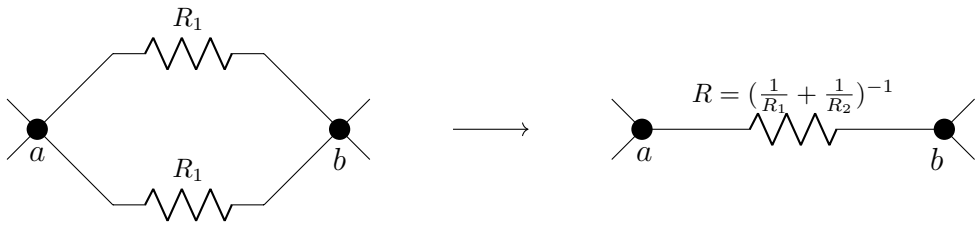


Figure 8.6: Parallel law

(3) Combining vertices that have the same voltage: If two or more vertices (none of them being vertices a or b) have the same voltage, they can be combined into one single vertex (while keeping all existing edges, with the exception of any resulting self-loops that may be deleted) without changing $R_{\text{eff}}(a, b)$. This follows from the fact that combining vertices of equal voltage will not change the voltages at the rest of the vertices, and therefore will not change the the strength of the current flow from a to b. Exercise 8.10 asks the reader to verify this.

(4) Deleting an edge whose end-vertices have the same voltage: If two vertices (neither of them being vertices a or b) have the same voltage, any edge between them can be deleted without changing $R_{\text{eff}}(a, b)$. This follows from the fact that deleting such an edge will not change the voltages in the network, and therefore will not change the strength of the current flow from a to b. Exercise 8.9 asks the reader to verify this.

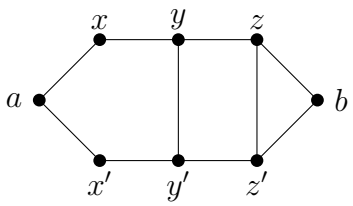
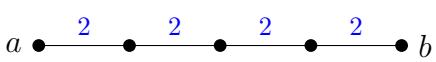


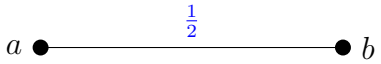
Figure 8.7

Figure 8.8

Example 8.4.2. Consider simple random walk (all edges have conductance 1) on the graph in Figure 8.7. Compute $R_{\text{eff}}(a, b)$.
We set $\Phi(a) = 1$ and $\Phi(b) = 0$. Because of the symmetry of the graph, it is clear that for the voltages at the rest of the vertices, we have $\Phi(x) = \Phi(x')$, $\Phi(y) = \Phi(y')$, and $\Phi(z) = \Phi(z')$. For the purpose of computing $R_{\text{eff}}(a, b)$, we can combine vertices that have equal voltage and delete any self-loops. This results in the simplified graph in Figure 8.8 which can be further simplified to



and



where the blue numbers are the edge conductances. Hence $R_{\text{eff}}(a, b) = 2$.

Alternate approach: We delete edges $\{y, y'\}$ and $\{z, z'\}$. Instead of combining x and x', y and y', z and z', we first apply the series law, then the parallel law. See Figure 8.9. This, again, yields $R_{\text{eff}}(a, b) = 2$.

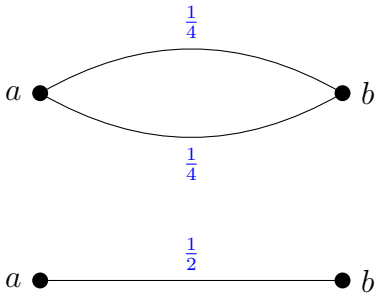


Figure 8.9

□

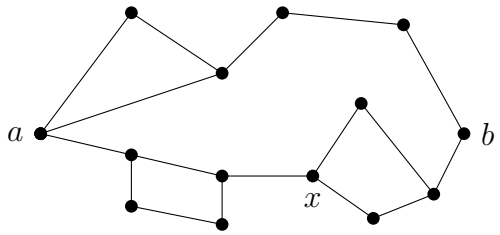


Figure 8.10

Example 8.4.3. Consider simple random walk on the graph in Figure 8.10. All edges are assumed to have conductance 1. Find the probabilities $\mathbb{P}_x(T^b < T^a)$ and $\mathbb{P}_a(T^b < T^a)$. In order to be able to compute the desired probabilities, we will simplify the network in a number of steps. See Figure 8.11. The numbers in blue are edge conductances:

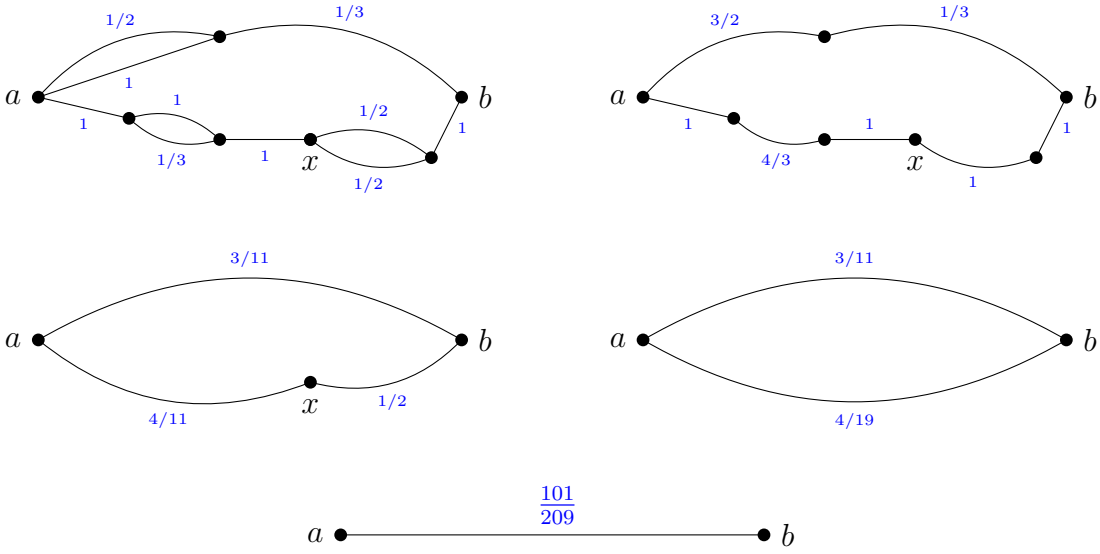


Figure 8.11: Simplifying the network in Figure 8.10

Hence $C_{\text{eff}}(a,b) = \frac{101}{209}$ and $R_{\text{eff}}(a,b) = \frac{209}{101}$. Here $C(a) = 3$, so we get for the escape probability for vertex a,

$$\mathbb{P}_a(T^b < T^a) = \frac{C_{\text{eff}}(a,b)}{C(a)} = \frac{101}{627}\,.$$

From the first network in the second row in Figure 8.10, i.e. the simplified network that contains exactly the three vertices x, a and b, we compute

$$\mathbb{P}_x(T^b > T^a) = \frac{1/2}{1/2 + 4/11} = \frac{11}{19}\,.$$

□

Example 8.4.4. Consider the gambler's ruin problem from Section 4.4. In order to phrase the process as a random walk on a network, we consider states 0 and state$N$ as reflecting boundaries, that is we set $P_{01} = P_{N,N-1} = 1$. For states $k = 1, ..., N-1$ we have $P_{k,k+1} = p$ and $P_{k,k-1} = 1-p$. Assume the gambler starts with fortune x at time 0. Because of the reflecting boundaries, the process is an irreducible birth/death chain, and it is also reversible. It is equivalent to random walk on network in Figure 8.12 (the blue numbers are conductances; we set $q = 1-p$).

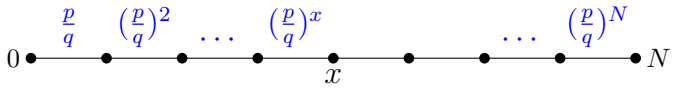


Figure 8.12

Note that the gambler's ruin probability $\mathbb{P}_x(T^0 < T^N)$ is equal to the voltage at vertex x, when setting the voltage to 1 at vertex 0, and to 0 at vertex N. In the following, we will set $r = \frac{q}{p}$ when $p \neq q$. After applying the series law (which does not change the voltage at x), the network simplifies to the network in Figure 8.13

Figure 8.13

For the case $p = \frac{1}{2}$, we get the network in Figure 8.14.

Figure 8.14

We can now derive the gambler's ruin probabilities from the above simplified networks. From

$$\mathbb{P}_x(T^0 < T^N) = \frac{C_{\text{eff}}(0,x)}{C_{\text{eff}}(0,x) + C_{\text{eff}}(x,N)}$$

we get

$$\mathbb{P}_x(T^0 < T^N) = \begin{cases} \dfrac{r^x - r^N}{1-r^N} = \dfrac{(\frac{q}{p})^x - (\frac{q}{p})^N}{1-(\frac{q}{p})^N} & \text{for } p \neq \frac{1}{2} \\ (N-x)/N & \text{for } p = \frac{1}{2}\,. \end{cases}$$

These formulas match the formulas we have derived in Section 4.4, using a different approach. □

Lemma 8.4.3. *Consider a finite network (G, C), two vertices $a, b \in V$, and a flow I from a to b. For any function $f : V \to \mathbb{R}$, we have*

$$(f(a) - f(b)) \|I\| = \frac{1}{2} \sum_{x,y \in V} (f(x) - f(y)) I(x, y) . \tag{8.16}$$

Proof.

$$\begin{aligned}
\frac{1}{2} \sum_{x,y \in V} (f(x) - f(y)) I(x, y) &= \frac{1}{2} \left(\sum_{x,y \in V} f(x) I(x, y) - \sum_{x,y \in V} f(y) I(x, y) \right) \\
&= \frac{1}{2} \left(\sum_{x,y \in V} f(x) I(x, y) + \sum_{x,y \in V} f(y) I(y, x) \right) \\
&= \sum_{x,y \in V} f(x) I(x, y) \\
&= \sum_{x \in V} f(x) \sum_{y \in V} I(x, y) \\
&= f(a)\|I\| - f(b)\|I\| = (f(a) - f(b)) \|I\|
\end{aligned}$$

□

Definition 8.4.2. *Consider a finite network (G, C), two vertices $a, b \in V$, and a flow I from a to b. The* **energy $\mathsf{E}(I)$ dissipated by the flow** *I is defined by*

$$\mathsf{E}(I) = \frac{1}{2} \sum_{x,y \in V} I(x, y)^2 R(x, y) . \tag{8.17}$$

Note that $\mathsf{E}(I) = \sum_{\{x,y\} \in E} I(x, y)^2 R(x, y)$.

If in particular the flow is a current flow I_ϕ, then by (8.16) and (8.17) with $f = \Phi$, the energy dissipated by the current flow is

$$\mathsf{E}(I_\Phi) = (\Phi(a) - \Phi(b)) \|I_\phi\| , \tag{8.18}$$

and using the definition of effective resistance in (8.18),

$$\mathsf{E}(I_\Phi) = R_{\text{eff}}(a, b) \|I_\phi\|^2 .$$

From this, we have the following proposition:

Proposition 8.4.4. *Let (G, C) be a finite network, $a, b \in V$, and I_Φ a current flow from a to b. The effective resistance $R_{\text{eff}}(a, b)$ is equal to $\mathsf{E}(I_\Phi)$ when I_Φ is the unit current flow.*

The following result is known as *Thomson's Principle* and will be useful.

Theorem 8.4.5 (Thomson's Principle)**.** *Let (G, C) be a finite network, $a, b \in V$, and I_Φ the* unit *current flow from a to b. Then*

$$\mathsf{E}(I_\Phi) = \min\{\mathsf{E}(I) : I \text{ is a unit flow from } a \text{ to } b\}$$

and $\mathsf{E}(I_\Phi) < \mathsf{E}(I)$ for all unit flows $I \neq I_\Phi$ from a to b.

Proof. Let I be a unit flow from a to b and I_Φ the unit current flow from a to b. We define a new flow $F = I - I_\Phi$ which is also a flow from a to b. Note that $\|F\| = 0$. We have

$$\begin{aligned}
2\mathsf{E}(I) &= 2\mathsf{E}(F + I_\Phi) \\
&= \sum_{x,y \in V} (F(x,y) + I_\Phi(x,y))^2 R(x,y) \\
&= \sum_{x,y \in V} F(x,y)^2 R(x,y) + 2 \sum_{x,y \in V} F(x,y) I_\Phi(x,y) R(x,y) + \sum_{x,y \in V} I_\Phi(x,y)^2 R(x,y) \\
&= \sum_{x,y \in V} F(x,y)^2 R(x,y) + 2 \sum_{x,y \in V} F(x,y)(\Phi(x) - \Phi(y)) + 2\mathsf{E}(I_\Phi)\,.
\end{aligned}$$

By (8.16), and since $\|F\| = 0$, the middle term in the sum on the right hand side of the last line is equal to zero. Indeed,

$$2 \sum_{x,y \in V} F(x,y)(\Phi(x) - \Phi(y)) = 4(\Phi(a) - \Phi(b))\|F\|\,.$$

Furthermore, since $R(x, y) > 0$ for any edge $\{x, y\}$ in the edge set E of the network, we have

$$\sum_{x,y \in V} F(x,y)^2 R(x,y) \geq 0$$

and

$$\sum_{x,y \in V} F(x,y)^2 R(x,y) = 0 \iff F \equiv 0 \iff I = I_\Phi\,.$$

Hence

$$\mathsf{E}(I_\Phi) \leq \mathsf{E}(I)$$

with equality holding if and only if $I = I_\Phi$. □

As a direct consequence of Thomson's principle, we now have the following theorem, known as *Raleigh's Monotonicity Principle*.

Theorem 8.4.6 (Raleigh's Monotonicity Principle)**.** *Let $G(V, E)$ be a connected graph and C and $C^\star$ two different assignments of conductances (and hence different corresponding resistances R and $R^\star$) to the edges of G. Then if $R(x, y) \leq R^\star(x, y)$ for all edges $\{x, y\} \in E$, we have*

$$R_{\text{eff}}(a, b) \leq R^\star_{\text{eff}}(a, b) \quad \text{for all } a, b \in V\,.$$

Proof. Consider two networks (G, C) and $(G, C^\star)$ on the same underlying graph $G(V, E)$ and $a, b \in V$. Assume

$$R(x, y) \leq R^\star(x, y) \quad \text{for all } \{x, y\} \in E\,.$$

Furthermore, consider the unit current flows I_Φ on (G, C) and $I^\star_{\Phi^\star}$ on $(G, C^\star)$, both from a to b. Note that $I^\star_{\Phi^\star}$ is also a unit flow on (G, C) from a to b. We have

$$\begin{aligned} R^\star_{\text{eff}}(a, b) &= \frac{1}{2} \sum_{x,y \in V} I^\star_{\Phi^\star}(x, y)^2 R^\star(x, y) \\ &\geq \frac{1}{2} \sum_{x,y \in V} I^\star_{\Phi^\star}(x, y)^2 R(x, y) \\ &= \mathsf{E}(I^\star_{\Phi^\star}) \quad \text{on the network } (G, C)\,. \end{aligned}$$

By Thomson's Principle (and Proposition 8.4.4),

$$\mathsf{E}(I^\star_{\Phi^\star}) \geq \mathsf{E}(I_\Phi) = R_{\text{eff}}(a, b)\,.$$

□

Remark 8.4.7. *Assume the underlying graph $G(V, E)$ in the network (G, C) is a simple graph, i.e. has no multiple edges or self-loops. Let $a, b \in V$. Theorem 8.4.6 implies that* **adding edges** *(that are not already in E) with arbitrary finite conductances to the underlying graph G will not increase (but may reduce) the effective resistance $R_{\text{eff}}(a, b)$. Indeed, we can think of any newly added edge $\{x, y\}$ as already being part of the original network but having conductance 0 and therefore resistance ∞, so no current will flow through $\{x, y\}$ in the original network. Thus "adding" the edge $\{x, y\}$ to the network means lowering its resistance from ∞ to a finite number. Intuitively, adding edges to the network may create more pathways for current to flow from a to b, hence may increase the flow. By Ohm's law, a larger current flow implies a smaller effective resistance between a and b.*
On the other hand, **cutting an existing edge** *in the network (i.e. making the resistance between its end vertices infinite) will never reduce but may increase $R_{\text{eff}}(a, b)$.*

Remark 8.4.8. *An extreme case of lowering resistance between two vertices x and y (each of which is neither a nor b) in the network is* **shorting** *them, i.e. combining x and y and deleting the edge between them, which can be viewed as making the resistance between them 0. If vertices x and y in the original network have distinct voltage, then shorting them will decrease $R_{\text{eff}}(a, b)$ in the new network. And as already discussed on page 234, if x and y have the same voltage, the shorting them will not change $R_{\text{eff}}(a, b)$ in the new network.*

We conclude this section with a lower bound for effective resistance. First, a definition:

Definition 8.4.3. *Let $G(V, E)$ be a finite connected graph and $a, b \in V$. A subset $S \subseteq E$ is called a* **cutset separating** *a* **from** *b if every path from a to b contains an edge in S.*

Theorem 8.4.9 (Nash–Williams Inequality)**.** *Consider a finite connected network (G, C) and two distinct vertices $a, b \in V$. Let $S_1, ..., S_k$ be a collection of pairwise disjoint cutsets that separate a from b. Then*

$$R_{\text{eff}}(a, b) \geq \sum_{i=1}^{k} \left(\sum_{\{x,y\} \in S_i} C(x, y) \right)^{-1}. \tag{8.19}$$

Proof. Let $S \subseteq E$ be a cutset that separates a from b. Consider the set $A \subseteq V$ defined by

$$A = \{x : a \text{ and } x \text{ are connected in the graph } \tilde{G}(V, E \setminus S)\}.$$

Let I be the unit current flow from a to b in $G(V, E)$. Since $\sum_{y:y\sim x} I(x, y) = 0$ for all $x \neq a, b$, we have

$$\sum_{x\in A} \sum_{y:y\sim x} I(x, y) = \|I\|.$$

Note that by antisymmetry of I,

$$\sum_{x\in A} \sum_{\substack{y:y\sim x \\ y\in A}} I(x, y) = 0$$

since for $x, y \in A$, both $I(x, y)$ and $I(y, x)$ appear in the sum. Furthermore, if $x \in A$, $y \notin A$, and $\{x, y\} \in E$, then $\{x, y\} \in S$. Thus

$$\|I\| = \sum_{x\in A} \sum_{\substack{y:y\sim x \\ y\notin A}} I(x, y) \leq \sum_{\{x,y\}\in S} |I(x, y)|. \tag{8.20}$$

Since we assume $\|I\| = 1$, we have

$$\begin{aligned} 1 &\leq \left(\sum_{\{x,y\}\in S} |I(x, y)| \right)^2 \\ &= \left(\sum_{\{x,y\}\in S} \sqrt{C(x, y)}\sqrt{R(x, y)}\, |I(x, y)| \right)^2 \\ &\leq \sum_{\{x,y\}\in S} C(x, y) \sum_{\{x,y\}\in S} R(x, y)I(x, y)^2 \end{aligned}$$

where the third line follows from the Cauchy-Schwarz inequality. This shows that

$$\left(\sum_{\{x,y\}\in S} C(x, y) \right)^{-1} \leq \sum_{\{x,y\}\in S} R(x, y)I(x, y)^2. \tag{8.21}$$

Setting $S = S_i$ in (8.21) and summing over i from 1 to k yields

$$\sum_{i=1}^{k} \left(\sum_{\{x,y\}\in S_i} C(x, y) \right)^{-1} \leq \sum_{i=1}^{k} \sum_{\{x,y\}\in S_i} R(x, y)I(x, y)^2 \leq \sum_{\{x,y\}\in E} R(x, y)I(x, y)^2.$$

The right-most sum is the energy dissipated by the unit current flow I. Hence, by Proposition 8.4.4,

$$R_{\text{eff}}(a, b) \geq \sum_{i=1}^{k} \left(\sum_{\{x,y\}\in S_i} C(x, y) \right)^{-1}.$$

□

8.5 Commute times and Cover times

Let (G, C) be a finite network and $x \neq y$ two vertices. Consider a random walk $(X_n)_{n\geq 0}$ on the network and assume $X_0 = x$. Recall (from Example 2.2.1) the definition of *commute time* $T^{x\leftrightarrow y}$ between x and y:

$$T^{x\leftrightarrow y} = \min\{n > T^y \,:\, X_n = x\}\,.$$

We will use the notation $t^{x\leftrightarrow y} = \mathbb{E}_x(T^{x\leftrightarrow y})$. By the Strong Markov property,

$$t^{x\leftrightarrow y} = \mathbb{E}_x(T^y) + \mathbb{E}_y(T^x)\,.$$

Remark 8.5.1. *Note that in general, $\mathbb{E}_x(T^y) \neq \mathbb{E}_y(T^x)$ (unless the network has special symmetry properties with respect to the two vertices x and y). However, the following* **cycle identity** *holds for random walk on any network: For all $x, y, z \in V$, we have*

$$\mathbb{E}_x(T^y) + \mathbb{E}_y(T^z) + \mathbb{E}_z(T^x) = \mathbb{E}_x(T^z) + \mathbb{E}_z(T^y) + \mathbb{E}_y(T^x). \tag{8.22}$$

For a proof of (8.22), see [35].

We first give a result for the commute times $t^{x\leftrightarrow y}$ for which x and y are neighboring vertices.

Proposition 8.5.2. *Consider random walk on a finite, irreducible network (G, C). Let V be the vertex set of G and E the edge set of G. Then*

$$\sum_{\{x,y\}\in E} C(x,y)\, t^{x\leftrightarrow y} = C_G\,(|V| - 1)\,. \tag{8.23}$$

Proof. Let π denote the stationary distribution. We have

$$\begin{aligned}
\sum_{\{x,y\}\in E} C(x,y)\, t^{x\leftrightarrow y} &= \frac{1}{2}\sum_{x\in V}\sum_{y\in V} C(x,y)(\mathbb{E}_x(T^y)+\mathbb{E}_y(T^x)) \\
&= \sum_{x\in V}\sum_{y\in V} C(x,y)\mathbb{E}_y(T^x) \\
&= \sum_{x\in V} C(x)\sum_{y\in V} P_{xy}\,\mathbb{E}_y(T^x) \\
&= \sum_{x\in V} C(x)(\mathbb{E}_x(T^x)-1) \\
&= C_G\sum_{x\in V}\pi(x)(\frac{1}{\pi(x)}-1) \\
&= C_G(|V|-1)\,.
\end{aligned}$$

□

Proposition 8.5.3 (Commute time identity)**.** *Consider random walk on a network (G,C) and $x,y\in V$ with $x\neq y$. Then the expected commute time between x and y is*

$$t^{x\leftrightarrow y} = C_G\, R_{\text{eff}}(x,y)\,. \tag{8.24}$$

Proof. Recall Example 2.2.1. There we have introduced the random variable $V^z_{T^y}$ (we reintroduced it in (8.7)) as the number of visits to state z strictly before time T^y. For $z=x$, we found

$$\mathbb{E}_x(V^x_{T^y}) = \frac{1}{\mathbb{P}_x(T^y<T^x)}\,.$$

By Remark 2.2.5, $\mu(z)=\mathbb{E}_x(V^z_{T^{x\leftrightarrow y}})$, $z\in V$, is an invariant measure for the random walk on the network (G,C) which, when normalized by the factor $1/t^{x\leftrightarrow y}$, becomes the unique stationary distribution π for the random walk. Also observe that, per definition of $T^{x\leftrightarrow y}$, for $z=x$, we have $\mathbb{E}_x(V^x_{T^y})=\mathbb{E}_x(V^x_{T^{x\leftrightarrow y}})$. Hence

$$\frac{\mu(x)}{t^{x\leftrightarrow y}} = \frac{1}{\mathbb{P}_x(T^y<T^x)\,t^{x\leftrightarrow y}} = \frac{C(x)}{C_G}$$

which, together with Proposition 8.4.2, yields the commute time identity

$$t^{x\leftrightarrow y} = C_G\, R_{\text{eff}}(x,y)\,.$$

□

Example 8.5.1. Consider simple random walk on the graph in Example 8.4.3. For the given graph $G(V, E)$, we have $|E| = 17$, and so $C_G = 34$. The expected commute time between vertices a and b is

$$t^{a\leftrightarrow b} = \mathbb{E}_a(T^b) + \mathbb{E}_b(T^a) = 34\frac{209}{101} = 70.36\,.$$

□

Example 8.5.2. Again we consider the gambler's ruin problem (see Example 8.4.4). Here we are interested in the expected duration $\mathbb{E}_x(T^{\{0,N\}})$ of the game when $p = \frac{1}{2}$, that is, the expected time until simple symmetric random walk visits either 0 or N for the first time. We solve this problem by combining vertices 0 and N into one single vertex. In doing so, we consider the boundary $\{0, N\}$ as one single state for the process. This transforms the process into simple symmetric random walk on the discrete N-cycle. See Figure 8.15.

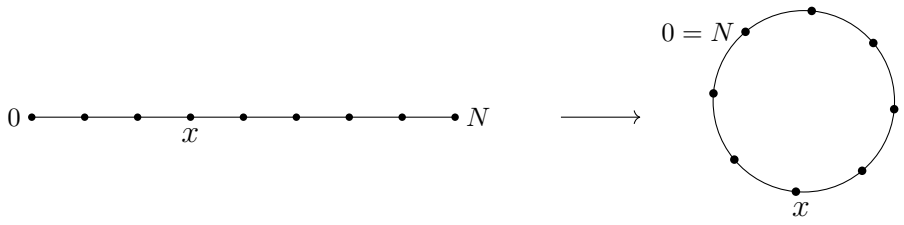


Figure 8.15

Because of symmetry, we have

$$\mathbb{E}_x(T^0) = \mathbb{E}_0(T^x) = \frac{1}{2}\,t^{x\leftrightarrow 0}$$

for simple symmetric random walk on the N-cycle. But $\mathbb{E}_x(T^0)$ for random walk on the N-cycle is equal to $\mathbb{E}_x(T^{\{0,N\}})$ for the gambler's ruin chain. We compute $R_{\text{eff}}(x, 0) = \frac{x(N-x)}{N}$ for the N-cycle via simplifying the network in a couple of steps. See Figure 8.16 (the blue numbers are conductances).

Figure 8.16

Here $C_G = 2N$, and thus by Proposition 8.5.3,

$$t^{x \leftrightarrow 0} = 2x(N-x)\,,$$

from which follows that the expected duration of the game is

$$\mathbb{E}_x(T^{\{0,N\}}) = x(N-x)\,,$$

which matches the formula we have computed in Section 4.4, using a different approach. □

Definition 8.5.1 (Cover time)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain on finite state space $\mathcal{S}$. The cover time random variable T^{cov} is the minimum number of steps it takes for the chain to visit all states in $\mathcal{S}$ at least once, that is,*

$$T^{\text{cov}} = \min\{n : \ \forall y \in \mathcal{S},\ \exists k \leq n,\ \text{s.t. } X_k = y\}\,.$$

We call its expectation

$$t_x^{\text{cov}} = \mathbb{E}_x(T^{\text{cov}})$$

the **cover time** *for the chain starting in state x.*

Example 8.5.3 (Cover time of the N-cycle)**.** Consider simple symmetric random walk on $\mathbb{Z}_N$. Because of symmetry, the distribution of the cover time random variable T^{cov} does not depend on the starting state x of the random walk, and we therefore simply write t^{cov} for its expectation. We can restate the cover time t^{cov} for random walk on $\mathbb{Z}_N$ as the expected time for simple symmetric unrestricted random walk on $\mathbb{Z}$ to visit a range of N distinct states (including the starting state) for the first time. By Proposition 4.7.1, the expected time until simple symmetric random walk on $\mathbb{Z}$ visits the Nth new state is $\mathbb{E}(T^{(N)}) = \frac{1}{2}N(N-1)$. It follows that the cover time t^{cov} for simple symmetric random walk on the discrete cycle $\mathbb{Z}_N$ is

$$t^{\text{cov}} = \frac{1}{2}N(N-1)\,.$$

□

Example 8.5.4 (Cover time for a star graph)**.** Consider simple random walk on a star graph $G(V,E)$ with n rays (so $|V| = n+1$ and $|E| = n$). Assume the walk starts at the center vertex c.

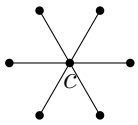


Figure 8.17: Star graph with 6 rays

Let us denote the time it takes for the walk to reach a kth *new* vertex and immediately afterwards step back to c by $T^{(k)}$ for $1 \leq k \leq n$. Then

$$T^{\text{cov}} + 1 = T^{(n)} = 2 + \sum_{k=2}^{n} (T^{(k)} - T^{(k-1)})$$

with $(T^{(k)} - T^{(k-1)}) = 2Y_k$ where

$$Y_k \sim \text{Geom}\left(\frac{n-k+1}{n}\right) \qquad \text{for } 2 \leq k \leq n\,.$$

Since $\mathbb{E}(Y_k) = \frac{n}{n-k+1}$, we get

$$\begin{aligned} t_c^{\text{cov}} + 1 &= 2 + 2\sum_{k=2}^{n} \frac{n}{n-k+1} \\ &= 2n \sum_{k=1}^{n} \frac{1}{k} \quad \approx \quad 2n \ln n \quad (\text{for large } n)\,. \end{aligned}$$

□

Example 8.5.5 (Cover time for a star with long rays)**.** We now consider a star with n *long* rays where each ray contains r vertices (without counting the center vertex c). For example, for the star graph on the left-hand side in Figure 8.18, $n = 6$ and $r = 4$. For a random walk starting at vertex c, find t_c^{cov}.

Figure 8.18

Here we denote by $T^{(k)}$, for $1 \leq k \leq n$, the commute time between c and a kth *new* extremal vertex (the vertex farthest away from c) on a ray. We also use the notation T^{return} for the time it takes the walk to move from an extremal vertex back to c. With this notation, we have

$$t_c^{\text{cov}} + \mathbb{E}(T^{\text{return}}) = \sum_{k=1}^{n} \mathbb{E}(T^{(k)}) \,.$$

To compute $\mathbb{E}(T^{\text{return}})$, note that because of symmetry, $\mathbb{E}(T^{\text{return}}) = \frac{1}{2} t^{0 \leftrightarrow r}$ for the chain graph of length r:

By the commute time identity, we have $t^{0 \leftrightarrow r} = 2r \cdot r$, and so $\mathbb{E}(T^{\text{return}}) = r^2$.
Next we compute $\mathbb{E}(T^{(k)})$. Assume the walk has already visited $k-1$ extremal vertices. In Figure 8.18, for the graph on the right-hand side, $k = 3$, and we mark the already visited vertices in black, the not-yet-visited vertices in white or blue. Since the order in which the walk visits all extremal vertices does not matter, we can combine (glue together) the $n-k+1$ extremal vertices that have not yet been visited. We'll call the combined vertex $G_{(n-k+1)}$.The commute time between c and $G_{(n-k+1)}$ is $\mathbb{E}(T^{(k)})$. The effective resistance between c and $G_{(n-k+1)}$ is

$$R_{\text{eff}}(c, G_{(n-k+1)}) = \frac{r}{n-k+1} \,,$$

and so, by the commute time identity,

$$\mathbb{E}(T^{(k)}) = 2nr \left(\frac{r}{n-k+1} \right) .$$

Altogether, we get for the cover time for a star with long rays,

$$t_c^{\text{cov}} = 2nr^2 \sum_{k=1}^{n} \frac{1}{k} \;-\; r^2 \,.$$

□

For graphs that exhibit a large amount of symmetries (we won't make this condition precise here), t_x^{cov} may not depend on the starting state x. Examples are the discrete cycle and the complete graph of n vertices. However, in general, the cover time t_x^{cov} *will* depend on the initial state x of the chain, and we are often interested in the worst-case scenario, that is, the largest cover time

$$t^{\text{cov}} = \max_{x \in \mathcal{S}} t_x^{\text{cov}}$$

among all starting states. The following theorem, which was first proved in [3], gives an upper bound for t^{cov} for simple random walks on graphs.

Theorem 8.5.4. *Let $G(V, E)$ be a finite, connected graph. The largest cover time t^{cov} for simple random walk on G satisfies*

$$t^{\text{cov}} \leq 2|E|\,(|V| - 1)\,.$$

Proof. Let T be a **spanning tree**[1] for the graph G. If $|V| = n$, any spanning tree for G will have $(n-1)$ edges. Let v_0 be any vertex in V. It is possible to traverse a tree T, starting at v_0, in a way that every vertex gets visited at least once, and every edge gets traversed exactly twice, once in each direction. Such a tour $v_0 \to v_1 \to \cdots \to v_{2n-2} = v_0$ is called a **depth-first tour** and is illustrated below with an example. The expected cover time (starting at v_0 is less than or equal to the expected time needed to move from v_0 to v_1, then from v_1 to v_2, and so on until the last move from v_{2n-1} to v_0. Thus

$$t^{\text{cov}}_{v_0} \leq \sum_{k=1}^{2n-2} \mathbb{E}_{v_{k-1}}(T^{v_k})\,.$$

Note that here the expected hitting times $\mathbb{E}_{v_{k-1}}(T^{v_k})$ are for simple random walk on the original graph $G(V, E)$, not on T. Since in the depth-first tour every edge gets traversed exactly once in each direction, it follows from the commute time identity that

$$t^{\text{cov}}_{v_0} \leq \sum_{\{v,w\}\in T} t^{v\leftrightarrow w} = 2|E| \sum_{\{v,w\}\in T} R_{\text{eff}}(v, w)\,.$$

By Raleigh's Monotonicity Principle (see Theorem 8.4.6 and the remarks that follow it), we have $R_{\text{eff}}(v, w) \leq 1$ for all $\{v, w\} \in T$, and hence

$$t^{\text{cov}}_{v_0} \leq 2|E|(n-1)\,. \tag{8.25}$$

Since the upper bound in (8.25) does not depend on the starting state v_0, the result of Theorem 8.5.4 follows. □

Example 8.5.6. Consider simple random walk on the graph $G(V, E)$ in Figure 8.19. The subset of blue edges in Figure 8.20 constitutes a spanning tree T for $G(V, E)$. A depth-first tour seeks to increase with each step, if possible, its edge distance to the starting point, but without duplicating directed edge crossings. If at a given time this is

[1] A spanning tree T for a connected graph $G(V, E)$ is a subgraph of G that contains all vertices V and a minimum number of edges from E. Every connected graph has at least one spanning tree.

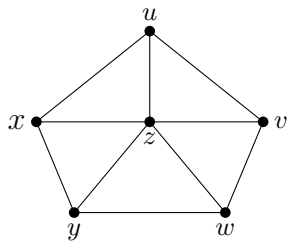


Figure 8.19

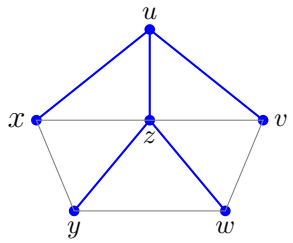


Figure 8.20: A spanning tree for the graph in Figure 8.19

not possible, the walk takes a step "backwards towards its starting point" which reduces its distance to the starting point by one edge. Along such a tour, each oriented edge in T will be traversed exactly once.
An example of a depth-first tour for the spanning tree in Figure 8.20 that starts at vertex u is

$$u \to x \to u \to z \to y \to z \to w \to z \to u \to v \to u\,.$$

Or, as another example, a depth-first tour for the same spanning tree that starts at vertex w, is

$$w \to z \to y \to z \to u \to x \to u \to v \to u \to z \to w\,.$$

□

Let $G(V, E)$ be a graph with $|V| = n$. Since, for any graph, we have $|E| \leq \binom{n}{2}$, Theorem 8.5.4 tells us that the cover time for any finite graph can be at most $O(n^3)$. For the star graph (see Example 8.5.4) we have shown by direct computation that t^{cov} is $O(n \ln n)$, which is of strictly lower order than the upper bound from Theorem 8.5.4 for this graph, namely $O(n^2)$. In most cases though, a direct computation of t^{cov} is not feasible, and Theorem 8.5.4 provides a useful upper bound.

Example 8.5.7 (Cover time for the "lollipop graph" is maximal)**.** Consider a so-called lollipop graph which consists of a complete graph K_m of m vertices (also called a clique)

with a chain graph of r vertices attached to it. Here we consider a lollipop graph G of n vertices (with n even) consisting of a clique $K_{n/2}$ of $n/2$ vertices and an attached chain graph of $n/2$ vertices. Figure 8.21 shows an example where $n = 10$.

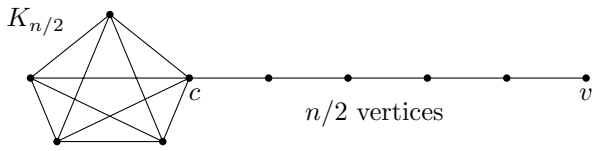


Figure 8.21: Lollipop graph

We claim that the cover time t_c^{cov} for simple random walk starting at c is $\Theta(n^3)$. By Theorem 8.5.4, the cover time t_c^{cov} is $O(n^3)$. Clearly,

$$t_c^{\text{cov}} \geq \mathbb{E}_c(T^v) = t^{c\leftrightarrow v} - \mathbb{E}_v(T^c).$$

Note that the number of edges of G is $|E| = \frac{n}{2}(\frac{n}{2} - 1)\frac{1}{2} + \frac{n}{2} = \Theta(n^2)$. The effective resistance between c and v is $R_{\text{eff}}(c, v) = \frac{n}{2}$. Thus, by the commute time identity,

$$t^{c\leftrightarrow v} = 2\left[\frac{n}{2}(\frac{n}{2} - 1)\frac{1}{2} + \frac{n}{2}\right]\frac{n}{2} = \Theta(n^3)\,.$$

And we have $\mathbb{E}_v(T^c) = (\frac{n}{2})^2 = \Theta(n^2)$ (recall a similar computation in Example 8.5.5). Altogether, it follows that $t^{c\leftrightarrow v}$ is $\Theta(n^3)$. □

Note: We can show that for simple random walk starting at the outer end v of the chain graph, the cover time t_v^{cov} for the lollipop graph from Example 8.5.7 is significantly smaller, namely $O(n^2)$.

Remark 8.5.5. *U. Feige[14] proved that for* **all** *graphs with n vertices,*

$$t^{\text{cov}} \leq \frac{4}{27}n^3 + o(n^3)\,.$$

For the lollipop graph consisting of a clique of $2n/3$ vertices and a chain graph of $n/3$ vertices, this upper bound is tight, that is, $t^{\text{cov}} = \frac{4}{27}n^3 + o(n^3)$. For a lower bound, Feige[15] proved that for all *graphs with n vertices, $t^{\text{cov}} = \Omega(n\ln(n))$. Note that for the star graph, we have shown $t^{\text{cov}} = \Theta(n\ln(n))$.*
Perhaps not surprisingly, graphs that possess a large amount of inherent symmetry (e.g., the star graph) tend to have a relatively small cover time. On the other hand, graphs with little or no symmetry properties (the lollipop graph is an extreme case) tend to have relatively large cover times.

8.6 Transience and Recurrence of Infinite Networks

We now turn to the study of *infinite* networks. We will consider networks (G, C) for which the underlying graph $G(V, E)$ is connected and the number of vertices is countably infinite. We will always assume that the graph $G(V, E)$ is *locally finite*, that is, each vertex $v \in V$ has finite degree. Our main focus are questions regarding transience or recurrence of random walks on such networks. We will study these questions with the use of analogous notions to effective resistance, flows, energy dissipated by a flow, etc. for infinite networks.

Let $G(V, E)$ be (finite or infinite) graph and (G, C) a network. Recall that $G^*(V^*, E^*)$ is called a **subgraph** of G if $V^* \subseteq V$ and $E^* \subseteq E$. A network (G^*, C^*) is called a **subnetwork** of (G, C) if G^* is a subgraph of G and C^* is the restriction of C from E to E^*.

Definition 8.6.1. *(a) Let $G(V, E)$ be a graph and $G^*(V^*, E^*)$ a subgraph of G. We say G^* is an* **induced subgraph** *of G if for all $x, y \in V^*$,*

$$\{x, y\} \in E \implies \{x, y\} \in E^* .$$

(b) Let $G(V, E)$ be an infinite graph and $\{G_n\}_{n\geq 1}$ with $G_n = (V_n, E_n)$ for $n \geq 1$ a sequence of finite induced subgraphs of G. We say $\{G_n\}_{n\geq 1}$ **exhausts** *G if*

$$\begin{aligned} V_n &\subseteq V_{n+1} \quad \forall n \geq 1, \text{ and} \\ V &= \bigcup_{n\geq 1} V_n . \end{aligned}$$

Now consider an infinite, connected, locally finite network (G, C) and a sequence $\{G_n\}_{n\geq 1}$ of finite induced subgraphs $G_n(V_n, E_n)$ that exhaust G. From this sequence of finite subgraphs, we construct the following sequence $\{(\tilde{G}_n, \tilde{C}_n)\}_{n\geq 1}$ with $\tilde{G}_n = (\tilde{V}_n, \tilde{E}_n)$ of finite networks: For each $n \geq 1$, we combine (short) the infinite set of vertices $V \setminus V_n$ so as to result in a single vertex b_n, and then delete any possibly resulting self-loops for b_n. We set $\tilde{V}_n = V_n \cup \{b_n\}$. Note that this construction may result in multiple (but always at most finitely many) edges between a vertex $x \in V_n$ and b_n as part of the edge set $\tilde{E}_n$. In order to distinguish such multiple edges, we can label them by z, i.e. write $\{x, b_n\}_z$ if $z \in V \setminus V_n$ and $\{x, z\} \in E$. The conductances $\tilde{C}_n$ are

$$\tilde{C}_n(x, y) = C(x, y) \text{ if } x, y \in V_n$$

and

$$\tilde{C}_n(x, b_n)_z = C(x, z) \text{ if } x \in V_n\,,\ z \in V \setminus V_n\,, \text{ and } \{x, z\} \in E\,.$$

Let $a \in V_1$. Since for $n \geq 1$, the finite network $(\tilde{G}_n, \tilde{C}_n)$ arises from $(\tilde{G}_{n+1}, \tilde{C}_{n+1})$ by shorting vertices, by Raleigh's Monotonicity Principle,

$$R_{\text{eff}}(a, b_n) \leq R_{\text{eff}}(a, b_{n+1}) \tag{8.26}$$

where the effective resistance on the left-hand side is for the network $(\tilde{G}_n, \tilde{C}_n)$, and the effective resistance on the right-hand side is for the network $(\tilde{G}_{n+1}, \tilde{C}_{n+1})$. By (8.26), $\lim_{n\to\infty} R_{\text{eff}}(a, b_n)$ exists. This limit may be finite or infinite.

Definition 8.6.2. *Consider an infinite, connected, locally finite network (G, C) and a corresponding sequence of finite networks $\{(\tilde{G}_n, \tilde{C}_n)\}_{n\geq 1}$ constructed as described above. Let $a \in V_1$. The effective resistance from vertex a to ∞ is defined by*

$$R_{\text{eff}}(a, \infty) = \lim_{n\to\infty} R_{\text{eff}}(a, b_n)\,.$$

Correspondingly, the effective conductance from a to ∞ is defined by

$$C_{\text{eff}}(a, \infty) = R_{\text{eff}}(a, \infty)^{-1}\,.$$

Recall Proposition 8.4.2. Using the same setup as in Definition 8.6.2, we have for $n \geq 1$,

$$\mathbb{P}_a(T^{b_n} < T^a) = \frac{1}{C(a)\, R_{\text{eff}}(a, b_n)} = \frac{C_{\text{eff}}(a, b_n)}{C(a)} \tag{8.27}$$

where all quantities refer to random walk on the network $(\tilde{G}_n, \tilde{C}_n)$. Furthermore,

$$\lim_{n\to\infty} \mathbb{P}_a(T^{b_n} < T^a) = \mathbb{P}_a(T^a = \infty) \tag{8.28}$$

where the probability on the right-hand side of (8.28) refers to random walk on the infinite network (G, C). Recall that random walk on an irreducible Markov chain is recurrent iff $\mathbb{P}_a(T^a = \infty) = 0$ for some state a and transient iff $\mathbb{P}_a(T^a = \infty) > 0$ for some state a. This yields the following result:

Proposition 8.6.1. *Consider an infinite, connected, locally finite network (G, C) and a corresponding sequence of finite networks $\{(\tilde{G}_n, \tilde{C}_n)\}_{n\geq 1}$ constructed as described above. Let $a \in V_1$. Random walk on (G, C) is recurrent iff*

$$R_{\text{eff}}(a, \infty) = \infty\,.$$

Equivalently, random walk on (G, C) is transient iff $R_{\text{eff}}(a, \infty) < \infty$.

Proof. By (8.27),

$$\lim_{n\to\infty} \mathbb{P}_a(T^{b_n} < T^a) = \frac{1}{C(a)R_{\text{eff}}(a,\infty)} \tag{8.29}$$

which, together with (8.28) and the paragraph following (8.28), proves the statement. □

Note: By (8.29) and (8.28), the effective resistance $R_{\text{eff}}(a,\infty)$ does not depend on the choice of sequence $\{G_n\}_{n\geq 1}$ of finite induced subgraphs of G that exhaust G.

Example 8.6.1 (Simple symmetric random walk on $\mathbb{N}_0$ with reflecting boundary at 0)**.** The sequence of chain graphs $G_n = [0,n]$, $n \geq 1$, exhausts $\mathbb{N}_0$. For each $n \geq 1$, the network $(\tilde{G}_n, \tilde{C}_n)$ (constructed in the manner described on page 252) is simply the chain graph $[0,(n+1)]$ with unit conductances. Since

$$R_{\text{eff}}(0,(n+1)) = n+1$$

for network $(\tilde{G}_n, \tilde{C}_n)$, $n \geq 1$, we have

$$R_{\text{eff}}(0,\infty) = \lim_{n\to\infty}(n+1) = \infty$$

which shows that simple symmetric random walk on $\mathbb{N}_0$ with reflecting boundary at 0 is recurrent. □

Example 8.6.2 (Recurrence of simple symmetric random walk on $\mathbb{Z}$)**.** Consider simple symmetric random walk on $G = \mathbb{Z}$. Here the sequence of chain graphs $G_n = [-n,n]$, $n \geq 1$, exhausts G. For each $n \geq 1$, the network $(\tilde{G}_n, \tilde{C}_n)$ as described on page 252 is the discrete cycle $\mathbb{Z}_{2n+2}$ with unit conductances. It arises by identifying the end-vertices $(n+1)$ and $-(n+1)$ in the chain graph $[-(n+1),(n+1)]$ and renaming part of the vertices. Applying the Series and Parallel laws, we compute

$$R_{\text{eff}}(0,(n+1)) = \frac{n+1}{2}$$

for network $(\tilde{G}_n, \tilde{C}_n)$, $n \geq 1$. It follows that

$$R_{\text{eff}}(0,\infty) = \lim_{n\to\infty} \frac{n+1}{2} = \infty,$$

and so simple symmetric random walk on $\mathbb{Z}$ is recurrent. □

We point out that by Definition 8.6.2, **Raleigh's Monotonicity principle** (Theorem 8.4.6 for finite networks) **carries over to $R_{\text{eff}}(a,\infty)$ for infinite networks**. For two conductances C and $C^\star$ on the same underlying infinite graph $G(V,E)$ for which $R(x,y) \le R^\star(x,y)$ for all $\{x,y\} \in E$, we have

$$R_{\text{eff}}(a,\infty) \le R^\star_{\text{eff}}(a,\infty) \tag{8.30}$$

where $R_{\text{eff}}(a,\infty)$ and $R^\star_{\text{eff}}(a,\infty)$ are the effective resistances from vertex a to ∞ for the networks (G,C) and $(G,C^\star)$, respectively.

Proposition 8.6.2 (Comparing networks)**.** *Consider two conductances C and $C^\star$ on the same infinite graph $G(V,E)$.*

(a) Assume $C^\star(x,y) \le C(x,y)$ for all $\{x,y\} \in E$. If random walk on (G,C) is recurrent, then random walk on $(G,C^\star)$ is also recurrent. In particular, if some vertices in (G,C) are shorted, the resulting self-loops deleted, and random walk on the resulting network is recurrent, then random walk on the original network (G,C) is also recurrent.

(b) Assume $C^\star(x,y) \le C(x,y)$ for all $\{x,y\} \in E$. If random walk on $(G,C^\star)$ is transient, then random walk on (G,C) is also transient. In particular, if some edges in (G,C) are deleted and random walk on the resulting network is transient, then random walk on the original network (G,C) is also transient.

(c) If there exist $0 < K_1 \le K_2 < \infty$ such that

$$K_1 C(x,y) \le C^\star(x,y) \le K_2 C(x,y)$$

for all $\{x,y\} \in E$, then random walk on (G,C) and random walk on $(G,C^\star)$ are either both recurrent or both transient.

Proof. Parts (a) and (b) directly follow from (8.30). For part (c), note that the three networks (G,C), (G,K_1C), and (G,K_2C) define the same random walk on G, hence are of the same type (transient or recurrent). The result follows from parts (a) and (b). □

Example 8.6.3. Consider biased random walk on $\mathbb{N}_0$ with reflecting boundary at 0. It is random walk on the network in Figure 8.22. Conductances are shown in blue.
If $p < q$, i.e., the walk is biased in the direction towards the boundary 0, we conclude from Example 8.6.2 and Proposition 8.6.2(a) that the walk is recurrent.

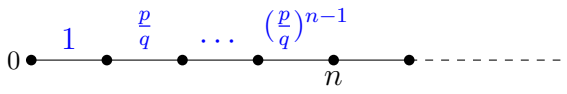


Figure 8.22

Alternatively, instead of using a comparison of networks, we could directly compute

$$R_{\text{eff}}(0,\infty) = \lim_{n\to\infty} R_{\text{eff}}(0,n) = \sum_{i=0}^{n-1} \left(\frac{q}{p}\right)^i = \infty$$

which shows that the walk is recurrent. □

Definition 8.6.3. *Let (G, C) be an infinite, connected, locally finite network.*

(a) A path in G is called a **simple** *path if it includes any edge in G at most once.*

(b) Let $a \in V$. A subset $S \subseteq E$ of edges in G is called a **cutset separating vertex a from ∞** *if every infinite, simple path that starts in vertex a includes at least one edge in S.*

The following theorem extends Theorem 8.4.9 to infinite networks. It gives a criterion for recurrence of a network.

Theorem 8.6.3 (Nash–Williams Inequality, infinite version)**.** *Let (G, C) be an infinite, connected, locally finite network and $a \in V$. If $\{S_i\}_{i\geq 1}$ is a sequence of finite, pairwise disjoint cutsets that separate a from ∞, then*

$$R_{\text{eff}}(a,\infty) \geq \sum_{i=1}^{\infty} \left(\sum_{\{x,y\}\in S_i} C(x,y) \right)^{-1}. \tag{8.31}$$

If the sum on the right-hand side of (8.31) is infinite, then random walk on (G, C) is recurrent.

Proof. Consider an infinite, connected, and locally finite network (G, C), a vertex $a \in V$, and a sequence $\{S_i\}_{i\geq 1}$ of finite, pairwise disjoint cutsets that separate vertex a from ∞. We construct a sequence of networks $\{(\tilde{G}_n, \tilde{C}_n)\}_{n\geq 1}$ as described (and with the same notation as) on page 253 for which $a \in \tilde{V}_1$ and

$$\bigcup_{i=1}^{n} S_i \subseteq \tilde{E}_n \quad \text{for all } n \geq 1.$$

Then for each of the finite networks $(\tilde{G}_n, \tilde{C}_n)$ in the sequence, the union of cutsets $\bigcup_{i=1}^{n} S_i$ separates vertex a from b_n. By Theorem 8.4.9,

$$R_{\text{eff}}(a, b_n) \geq \sum_{i=1}^{n} \left(\sum_{\{x,y\} \in S_i} C(x, y) \right)^{-1} . \tag{8.32}$$

Taking the limit as $n \to \infty$ on both sides of (8.32) yields (8.31).
If the the right-hand side of (8.31) is infinite, then $R_{\text{eff}}(a, \infty) = \infty$ which, by Proposition 8.6.1, implies that random walk on (G, C) is recurrent. □

Example 8.6.4 (Recurrence of simple symmetric random walk on $\mathbb{Z}^2$)**.** Consider simple symmetric random walk on $\mathbb{Z}^2$ starting at the origin a. All conductances in the network are considered to be 1.
In Figure 8.23 below, each set of points Y_i that have a fixed (graph) distance i from the origin is identified via connecting dashed lines that form a square. We have $Y_i = \{(x, y) : |x| + |y| = i\}$ for $i \geq 1$. Any simple infinite path that starts at the origin a must pass at some finite time from a vertex in set Y_i to a vertex in set Y_{i+1} for all $i \geq 1$. Therefore, each (finite) subset of edges S_i consisting of all edges that connect a vertex in Y_i with a vertex in Y_{i+1} is a cutset separating vertex a from ∞. By construction, the sets $\{S_i\}_{i \geq 0}$ are pairwise disjoint.

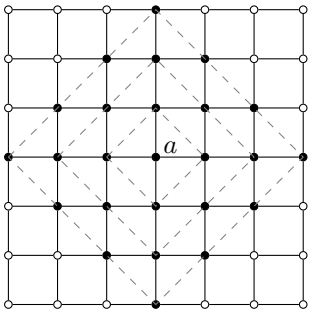


Figure 8.23

Applying Theorem 8.6.3, we get

$$R_{\text{eff}}(a, \infty) \geq \sum_{i=1}^{\infty} |S_i|^{-1}. \tag{8.33}$$

Since $|Y_i| = 4i$, with four of the vertices in Y_i contributing 3 edges and the rest of the vertices in Y_i contributing 2 edges to S_i, we have

$$|S_i| = 12 + 2(4i - 4) = 8i + 4 .$$

Thus, by (8.33),

$$R_{\text{eff}}(a,\infty) \geq \sum_{i=1}^{\infty} \frac{1}{8i+4} = \infty$$

which shows that $R_{\text{eff}}(a,\infty) = \infty$, and so simple symmetric random walk on $\mathbb{Z}^2$ is recurrent.

Note: Since $\mathbb{Z}$ can be viewed as a subnetwork of $\mathbb{Z}^2$, the result (once more) shows that simple symmetric random walk on $\mathbb{Z}$ is also recurrent. □

The following definitions for infinite networks are analogous to Definitions 8.3.2 and 8.4.2 for finite networks.

Definition 8.6.4. *Consider an infinite, connected, locally finite network (G,C) and $a \in V$.*

(a) A **flow from** *a* **to** *∞ is an anti-symmetric function I on the* oriented *edges of (G,C) for which* **Kirchhoff's node law** *holds, that is, for which*

$$\sum_{y:y\sim x} I(x,y) = 0 \quad \text{for all } x \in V \setminus \{a\}\,,$$

and for which

$$\|I\| := \sum_{y:y\sim a} I(a,y) \geq 0\,.$$

If $\|I\| = 1$, we call I a **unit flow** *from a to ∞.*

(b) Let I be a flow from a to ∞. The **energy $\mathsf{E}(I)$ dissipated by the flow** *I is defined by*

$$\mathsf{E}(I) = \sum_{\{x,y\}\in E} I(x,y)^2 R(x,y)\,.$$

Using the notion of *energy dissipated by a flow*, we arrive at the following criterion for transience:

Proposition 8.6.4. *Consider an infinite, connected, locally finite network (G,C). If there exists a vertex a and a unit flow I from a to ∞ with $\mathsf{E}(I) < \infty$, then random walk on (G,C) is transient.*

Proof. Let I be a unit flow on (G,C) from vertex a to ∞. Assume $\mathsf{E}(I) < \infty$ and consider a sequence $\{(\tilde{G}_n, \tilde{C}_n)\}_{n\geq 1}$ of finite networks constructed from (G,C) in a manner

described in Definition 8.6.1 and the paragraph following Definition 8.6.1. Without loss of generality, we assume $a \in \tilde{G}_1$. Then for all $n \geq 1$, the restriction of I to the oriented edges in $\tilde{G}_n$ defines a unit flow $\tilde{I}_n$ from a to b_n in $(\tilde{G}_n, \tilde{C}_n)$. We have

$$R_{\text{eff}}(a, b_n) \leq \mathsf{E}(\tilde{I}_n) \leq \mathsf{E}(I) < \infty \tag{8.34}$$

where the first inequality in (8.34) follows from Proposition 8.4.4 and Theorem 8.4.5. Hence

$$\lim_{n\to\infty} R_{\text{eff}}(a, b_n) = R_{\text{eff}}(a, \infty) \leq \mathsf{E}(I) < \infty\,.$$

By Proposition 8.6.1, random walk on (G, C) is transient. □

Example 8.6.5 (Transience of simple symmetric random walk on $\mathbb{Z}^d$ for $d \geq 3$)**.** Here we assume that all edge conductances are equal to 1. We start with simple random walk on $\mathbb{Z}^3$ and prove its transience by constructing a flow I on $\mathbb{Z}^3$ with finite energy $\mathsf{E}(I)$. The argument we present here is taken from [23].

Recall the process called *Pólya's urn* which was introduced in Section 1.5. Here we will consider a three-color Pólya's urn. The process starts with three balls of distinct color, one red, one green, and one blue ball. At each time step, a ball is drawn uniformly at random from the urn, its color noted, and then, together with a new ball of the same color, replaced into the urn. The random vector (R_n, G_n, B_n) gives the number of red, green, and blue balls at time $n \geq 0$. We have $(R_0, G_0, B_0) = (1, 1, 1)$ and $(R_n, G_n, B_n) \in (\mathbb{Z}^+)^3$ with $R_n + G_n + B_n = n + 3$.

By Proposition 1.5.2, for all $n \geq 1$, the random vector (R_n, G_n, B_n) is *uniformly distributed* over the set $V_n = \{(x, y, z) \in (\mathbb{Z}^+)^3 : x + y + z = n + 3\}$. Note that $|V_n|$ is equal to the number of ways in which n indistinguishable balls can be distributed over three distinguishable (colored) boxes, so

$$|V_n| = \binom{n+2}{2}\,. \tag{8.35}$$

We define a flow I on $\mathbb{Z}^3$ from $a = (1, 1, 1)$ to ∞ in the following way: For an *oriented* edge (u, v) in the first octant that points in the positive direction of either the x-axis, the y-axis, or the z-axis, we set

$$I(u, v) = \mathbb{P}(\text{the urn process moves from } u \text{ to } v)\,,$$

$$I(v, u) = -I(u, v)\,,$$

and define I to be 0 on all other oriented edges in $\mathbb{Z}^3$. This is indeed a flow since

$$\|I\| = \sum_{v:v\sim a} I(a, v) = 1$$

and for $u \in (\mathbb{Z}^+)^3$, $u \neq a$,

$$\sum_{\substack{w:w\sim u\\ I(w,u)>0}} I(w,u) = \mathbb{P}(\text{the urn process visits } u) = \sum_{\substack{v:v\sim u\\ I(u,v)>0}} I(u,v)\,, \tag{8.36}$$

and therefore

$$\sum_{v:v\sim u} I(u,v) = 0\,.$$

We now compute $\mathsf{E}(I)$. From (8.36) and (8.35) we have

$$\begin{aligned}
\sum_{\substack{w:w\sim u\\ I(w,u)>0}} I(w,u)^2 &\leq \left(\sum_{\substack{w:w\sim u\\ I(w,u)>0}} I(w,u)\right)^2 \\
&= \mathbb{P}(\text{the urn process visits } u)^2 \\
&= \binom{n+2}{2}^{-2}.
\end{aligned}$$

Thus we get

$$\begin{aligned}
\mathsf{E}(I) &= \sum_{\{w,u\}} I(w,u)^2 \\
&= \sum_{n=1}^{\infty} \sum_{u\in V_n} \sum_{\substack{w:w\sim u\\ I(w,u)>0}} I(w,u)^2 \\
&\leq \sum_{n=1}^{\infty} \binom{n+2}{n} \binom{n+2}{2}^{-2} \\
&= \sum_{n=1}^{\infty} \frac{2}{(n+2)(n+1)} < \infty\,.
\end{aligned}$$

By Proposition 8.6.4, simple random walk on $\mathbb{Z}^3$ is transient.

Finally, note that $\mathbb{Z}^3$ is a subgraph of $\mathbb{Z}^d$ for $d \geq 3$. Any flow I on a subgraph H of a graph G can be extended to a flow on G by simply defining the extended flow to be 0 on all oriented edges that are in G but not in H. This proves transience for simple random walk on $\mathbb{Z}^d$ for $d \geq 3$. □

Remark 8.6.5. *The converse of the statement of Proposition 8.6.4 is also true: If a connected, locally finite network (G, C) is transient, then there exists a unit flow I on the network for which $\mathsf{E}(I) < \infty$. We omit the proof (see [24] for a reference). Using this result, and therefore without the need to construct a specific flow of finite energy which may be difficult, we once again conclude (recall that we have already proved this earlier) that if random walk on a subnetwork (G^*, C^*) of a network (G, C) is transient, then random walk on the larger network (G, C) is also transient. And, equivalently, if random walk on (G, C) is recurrent, then random walk on any subnetwork (G^*, C^*) is also recurrent.*

Exercises

Exercise 8.1. Show that the Markov chain with state space $\mathcal{S} = \{a, b, c, d\}$ and transition matrix

$$\mathbf{P} = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1/6 & 1/6 & 0 & 2/3 \\ 1/5 & 2/5 & 2/5 & 0 \\ 0 & 1/3 & 1/6 & 1/2 \\ 4/9 & 0 & 1/3 & 2/9 \end{pmatrix} \end{array}$$

is reversible and therefore can be interpreted as a random walk on a weighted graph. Find a weighted graph for this Markov chain for which all weights are integers.

Exercise 8.2. Consider an irreducible birth/death chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S} = \{0, 1, ..., N\}$. The transition probabilities are

$$\begin{aligned} P_{x,x+1} &= p_x \quad \text{for } 0 \leq x \leq N-1 \\ P_{x,x-1} &= q_x \quad \text{for } 1 \leq x \leq N \end{aligned}$$

with $p_x + q_x = 1$ and $p_x, q_x > 0$ for all x. Figure 8.24 shows the transition graph.

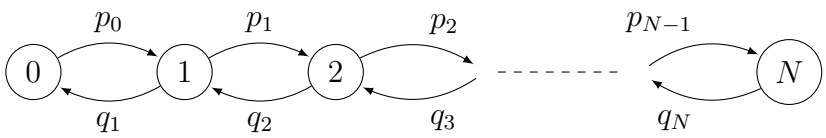


Figure 8.24: Birth/death chain

We have shown in Section 7.2 that $(X_n)_{n\geq 0}$ is reversible. Describe $(X_n)_{n\geq 0}$ as a random walk on a weighted graph. Describe such a weighted graph by giving an explicit formula for the weight of each edge.

Exercise 8.3. Consider the Ehrenfest chain with N particles (see Section 1.5).

(a) Describe the process as a random walk on a network (G, C).

(b) Let N be even. Use the network interpretation to compute a formula for

$$\mathbb{P}_{N/2}(T^0 < T^{N/2}) .$$

Exercise 8.4. Consider simple random walk on the graph in Figure 8.25.

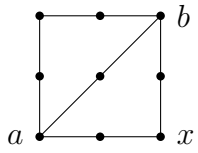


Figure 8.25

Compute the following:

(a) $\mathbb{E}_a(T^b)$,
(b) $\mathbb{P}_a(T^b < T^a)$,
(c) $\mathbb{E}_a(V^x_{T^b})$ (the expected number of visits to x before the first visit to b, given that the walk starts in a),
(d) $R_{\text{eff}}(x, b)$.

Exercise 8.5. A graph is called *d-regular* if every vertex has degree d. Let $G(V, E)$ be a finite, connected, d-regular graph with n vertices. Consider simple random walk on G. Find a positive integer N such that the probability that simple random walk on G of length N has not yet reached all vertices on the graph is at most $\frac{1}{10}$.

Exercise 8.6. Consider $G(V, E)$ and assume there exists an edge $e = \{x_0, y_0\}$ with $e \in E$ such that the removal of e results in two disjoint subgraphs G_{x_0} and G_{y_0} with respective edge sets E_{x_0} and E_{y_0}.

(a) Assuming all edge weights are equal to 1, show that $R_{\text{eff}}(x_0, y_0) = 1$.

(b) Consider simple random walk on $G(V, E)$. Use the result from part (a) to prove that

$$\mathbb{E}_{x_0}(T^{y_0}) = 2|E_{x_0}| + 1.$$

Exercise 8.7. Consider simple random walk on a finite, connected graph $G(V, E)$ with $|V| = n$. Let $x, y \in V$ and assume $\{x, y\} \in E$. Show that $\mathbb{E}_x(T^y)$ is at most $O(n^2)$. Is this necessarily true if x and y are not connected by an edge? If it is not true in general, give an example where $\mathbb{E}_x(T^y)$ is of strictly higher order than $\Theta(n^2)$.

Exercise 8.8. Consider a finite, irreducible network (G, C) with vertex set V and edge set E. **Foster's theorem** states that

$$\sum_{\{x,y\}\in E} R_{\text{eff}}(x, y)C(x, y) = |V| - 1\,.$$

Prove Foster's theorem.

Exercise 8.9. Let (G, C) be a finite, irreducible network, and let $a, b \in V$. Consider the unique voltage Φ_1 on V with boundary values $\Phi_1(a) = 1$ and $\Phi_1(b) = 0$. Assume there exist two vertices $x, y \in V \setminus \{a, b\}$ for which $\Phi_1(x) = \Phi_1(y)$ and that $e = \{x, y\}$ is in the edge set E of G.

(a) Show that removing the edge e from E does not change the voltage at any vertex. More precisely, consider the slightly altered network (G', C') that arises from (G, C) by deleting the edge e from the graph and keeping the weights on all remaining edges the same. Furthermore, consider the unique voltage Φ_1' on V' for the network (G', C') with boundary values $\Phi_1'(a) = 1$ and $\Phi_1'(b) = 0$. Show that $\Phi_1(x) = \Phi'(x)$ for all $x \in V$.

(b) Conclude from part (a) that $R_{\text{eff}}(a, b) = R'_{\text{eff}}(a, b)$, that is, deleting from the network an edge whose endpoints have equal voltage does not change the effective resistance.

Exercise 8.10. Let (G, C) be a finite, irreducible network, and let $a, b \in V$. Consider the unique voltage Φ_1 on V with boundary values $\Phi_1(a) = 1$ and $\Phi_1(b) = 0$. Assume there exist two vertices $x, y \in V \setminus \{a, b\}$ for which $\Phi_1(x) = \Phi_1(y)$.

(a) Show that combining the two vertices x and y into a single vertex z does not change the voltages in the system. More precisely, consider the slightly altered network (G', C') that arises from (G, C) by combining the two vertices x and y into a single vertex z and deleting any possibly resulting self-loops. All weights remain the same, except for weights of edges (from the original network) that have x or y as an endpoint, in which case we add their weights: We set $C'(z, v) = C(x, v) + C(y, v)$ for $v \neq x, y$. Furthermore, consider the unique voltage Φ_1' on V' for the network (G', C') with boundary values $\Phi_1'(a) = 1$ and $\Phi_1'(b) = 0$. Show that $\Phi_1(v) = \Phi_1'(v)$ for all $v \neq z$ and $\Phi_1'(z) = \Phi_1(x) = \Phi_1(y)$.

(b) Conclude from part (a) that $R_{\text{eff}}(a,b) = R'_{\text{eff}}(a,b)$, that is, combing vertices that have equal voltage (and deleting any possibly resulting self loops) does not change the effective resistance.

Exercise 8.11. Let (G, C) be a finite, irreducible network. Show that the effective resistance obeys the triangle inequality, that is, show that for all $a, b, c \in V$,

$$R_{\text{eff}}(a,c) \leq R_{\text{eff}}(a,b) + R_{\text{eff}}(b,c).$$

Exercise 8.12. Consider a *complete graph* K_n with n vertices and let x, y be two distinct vertices. Compute the effective resistance $R_{\text{eff}}(x,y)$.

Exercise 8.13. Consider simple random walk on the graph in Figure 8.26. Compute the commute time $t^{a \leftrightarrow b}$.

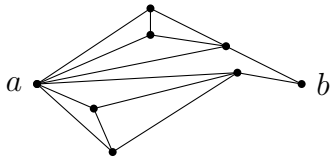


Figure 8.26

Exercise 8.14. Let K_n be a complete graph with n vertices. Show that the cover time t^{cov} for simple random walk on K_n is of order $O(n \ln n)$.

Exercise 8.15. Consider a chain graph G with N vertices as shown in Figure 8.27. Consider simple random walk starting at x for some $1 < x < N$.

(a) Compute $\mathbb{E}_x(T^{x+1})$. (b) Compute the cover time t^{cov}_x.

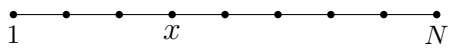


Figure 8.27

Exercise 8.16. Again, consider a chain graph G with N vertices as shown in Figure 8.27. Assume that simple random walk starts at vertex 1. Let $V^x_{T^N}$ be the number of visits to vertex x before the walk reaches vertex N for the first time. Show that

$$\mathbb{E}_1(V^1_{T^N}) = N - 1$$

and

$$\mathbb{E}_1(V^x_{T^N}) = 2(N-x) \qquad \text{for } 2 \leq x \leq N-1\,.$$

Exercise 8.17. Consider simple random walk on the 3-dimensional unit cube $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. See Figure 1.7. Let $\mathbf{a} = (0, 0, 0)$ and $\mathbf{c} = (1, 1, 1)$.

(a) Show that $R_{\text{eff}}(\mathbf{a}, \mathbf{c}) = 5/6$.

(b) Compute the expected hitting time $\mathbb{E}_{\mathbf{a}}(T^{\mathbf{c}})$.

Exercise 8.18. Consider simple random walk on the 3-dimensional unit cube $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ as in Exercise 8.17. In addition to $\mathbf{a} = (0, 0, 0)$ and $\mathbf{c} = (1, 1, 1)$, consider the vertex $\mathbf{b} = (1, 1, 0)$.

(a) Compute the probability $\mathbb{P}_{\mathbf{c}}(T^{\mathbf{a}} < T^{\mathbf{b}})$.

(b) Compute the expected hitting time $\mathbb{E}_{\mathbf{a}}(T^{\mathbf{b}})$.

Exercise 8.19. Recall the features of a *lollipop graph*. It consists of a complete graph K_m of m vertices and a chain graph of k vertices that is attached to K_m. Figure 8.28 shows an example with $m = 5$ and $k = 7$. Consider the vertices c and v as labeled in Figure 8.28 and any vertex a in K_m with $a \neq c$. For general m and k, find formulas for (a) $\mathbb{E}_a(T^v)$ and (b) $\mathbb{E}_c(T^a)$.

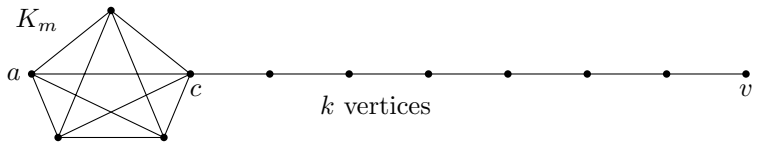


Figure 8.28: Lollipop graph

Exercise 8.20. Consider an infinite connected network (G, C) for which the edge set E is countably infinite. Show that random walk on the network is positive recurrent if and only if

$$\sum_{\{x,y\} \in E} C(x, y) < \infty \,.$$

Exercise 8.21. Consider an irreducible birth/death chain $(X_n)_{n \geq 0}$ on infinite state space $\mathcal{S} = \mathbb{N}_0$. For the transition probabilities, we assume $p_0 = 1$, and $p_x + q_x = 1$ and $p_x, q_x > 0$ for all $x \geq 1$. See the transition graph in Figure 8.29. Describe $(X_n)_{n \geq 0}$ as a random walk on an infinite network. Use this viewpoint to prove a **criterion for transience** of $(X_n)_{n \geq 0}$ in terms of the transition probabilities p_x, q_x, $x \geq 1$.

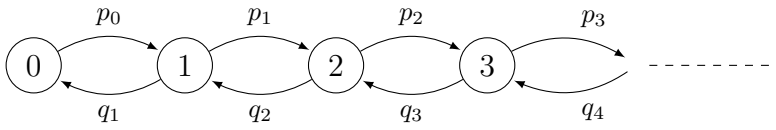


Figure 8.29: Infinite-state birth/death chain

Exercise 8.22. Fix a positive integer k. An infinite *rooted k-ary tree* is a tree that has a distinguished vertex 0 (the root) which has degree k and for which all other vertices have degree $k+1$. See Figure 8.30 for a picture of a binary (2-ary) tree. Let $R_{\text{eff}}^{(k)}(0,\infty)$ denote the effective resistance from 0 to infinity for a k-ary tree. Compute $R_{\text{eff}}^{(k)}(0,\infty)$. What can we conclude about transience or recurrence of simple random walk on a k-ary tree?

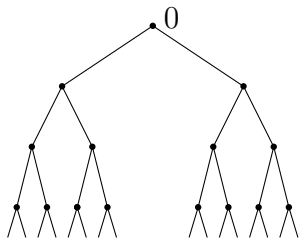


Figure 8.30: A binary tree. The graph continues to infinity in the same manner.

Exercise 8.23. Consider a network (G, C) whose underlying graph is an infinite rooted tree. Let $V_n \subset V$ denote the set of all vertices that have distance n from the root 0. For all $n \geq 1$, we assume that edges that have one endvertex in V_{n-1} and the other endvertex in V_n have the same conductance C_n.

(a) Find a sufficient condition (in terms of C_n and $|V_n|$, $n \geq 1$) for recurrence of random walk on (G, C).

(b) Fix $k \in \mathbb{N}$. Consider the special case where G is an infinite k-ary rooted tree and $C_n = \lambda^{-n}$ for some $\lambda > 1$. Find a necessary and sufficient condition on λ under which random walk on (G, C) is recurrent.

Chapter 9

Markov Chain Monte Carlo

The basic problem in Markov chain Monte Carlo can be described as follows. Given a probability distribution π on a large and often intractable state space $\mathcal{S}$, how can we draw samples from this distribution π? A Markov chain Monte Carlo algorithm *constructs* a Markov chain on $\mathcal{S}$ that is fairly easy to simulate and that converges to the desired distribution π. Running this Markov chain for a sufficiently long time and then taking samples will produce (almost) samples from π. We address the relevant and important topic of *convergence rates* of Markov chains (for how long should we run the algorithm?) in Chapter 11.

9.1 MCMC Algorithms

9.1.1 Metropolis-Hastings Algorithm

The goal is to construct an irreducible, aperiodic, positive recurrent Markov chain on a given state space $\mathcal{S}$ that has a desired stationary distribution π. One would like the transition probabilities to have the form

$$P_{xy} = a(x, y)\, T_{xy}$$

for $x, y \in \mathcal{S}$, where the T_{xy} are transition probabilities for an easy-to-implement **proposal** chain $\mathbf{T}$ on $\mathcal{S}$. The values $a(x, y)$ are **acceptance probabilities** according to which state y, if proposed by the proposal chain $\mathbf{T}$, will either be accepted or rejected. In the latter case, the chain remains in state x.

We now describe the algorithm in more detail. Assume the **target distribution** π is *strictly positive* on $\mathcal{S}$. (If not, simply delete from the state space all states x for which $\pi(x) = 0$.) Choose an irreducible proposal chain with transition matrix $\mathbf{T}$ that is easy to

simulate. From this, the **Metropolized chain** $(X_n)_{n\geq 0}$ with transition matrix $\mathbf{P}$ arises via the following algorithm.

Metropolis-Hastings Algorithm:

1. Start in any state.
2. Given that the chain is currently in state x, for the next step choose a state y according to the transition probabilities T_{xy} from the proposal chain.
3. Decide whether to accept y or to reject y and stay in x with the use of the acceptance probability
$$a(x,y) = \min(\frac{\pi(y)\,T_{yx}}{\pi(x)\,T_{xy}}, 1)\,.$$
That is, if a biased coin for which $\mathbb{P}(H) = a(x,y)$ lands on heads, accept the proposed state and move to y. If the biased coin lands on tails, reject the proposed state and remain in the current state x. Thus the transition probabilities P_{xy} for the Metropolized chain are:
$$\text{for } x \neq y: \qquad P_{xy} = \begin{cases} 0 & \text{if } T_{xy} = 0 \\ T_{xy}\,a(x,y) & \text{if } T_{xy} \neq 0 \end{cases}$$
$$\text{for } x = y: \qquad P_{xx} = 1 - \textstyle\sum_{y\neq x} P_{xy}\,.$$

Note that for this algorithm one only needs to know the stationary distribution π up to proportionality. The normalizing constant for π cancels out in the computation of $a(x,y)$. This is a great advantage of the algorithm. Often, in situations where the state space $\mathcal{S}$ is extremely large, its actual size may not be known. If, for example, we would like to sample from the uniform distribution on $\mathcal{S}$, we can simply work with any $\pi \propto (1,1,...,1)$ in the algorithm (the symbol "$\propto$" stands for "is proportional to").

Proposition 9.1.1. *Let $(X_n)_{n\geq 0}$ be constructed as in the Metropolis-Hastings algorithm. Then $(X_n)_{n\geq 0}$ is a positive recurrent and reversible Markov chain with stationary distribution π.*

Proof. We assume $\pi(x) > 0$ for all $x \in \mathcal{S}$. We will first show that the detailed balance equations (7.2) hold for the transition probabilities P_{xy} and the distribution π. For $x = y$, there is nothing to show.
Assume $x \neq y$ and $P_{xy} = 0$. Then it must be that either $T_{xy} = 0$, in which case $a(y,x) = 0$, and thus $P_{yx} = 0$ as well, or that $T_{xy} \neq 0$ but $a(x,y) = 0$, in which case $T_{yx} = 0$ and

therefore $P_{yx} = 0$ as well. This shows that for $P_{xy} = 0$, the detailed balance equations $\pi(x)P_{xy} = \pi(y)P_{yx}$ hold.
Assume $x \neq y$ and $P_{xy} \neq 0$. Then $T_{xy} \neq 0$ and $T_{yx} \neq 0$. We distinguish two cases.
Case 1: $a(x, y) \geq 1$. Then $P_{xy} = T_{xy}$ and $a(y, x) \leq 1$. Thus

$$\pi(x)\, P_{xy} = \pi(x)\, T_{xy} = \pi(x)\, T_{xy}\, \frac{\pi(y)T_{yx}}{\pi(y)T_{yx}} = \pi(y)\, T_{yx}\, a(y, x) = \pi(y)\, P_{yx}\,.$$

Case 2: $a(x, y) < 1$. Then $a(y, x) \geq 1$. We can apply the same argument as for Case 1, in reverse order, and get again

$$\pi(y)P_{yx} = \pi(x)P_{xy}\,.$$

Since the detailed balance equations hold for all $x, y \in \mathcal{S}$, $\mathbf{P}$ is reversible and π is a stationary distribution for $\mathbf{P}$. □

Notes: (1) The Metropolis-Hastings algorithm constructs a *reversible* Markov chain $\mathbf{P}$.
(2) Using an irreducible proposal chain $\mathbf{T}$ for which $T_{xy} > 0 \iff T_{yx} > 0$ will imply $T_{xy} > 0 \Rightarrow P_{xy} > 0$. In this case, irreducibility of the proposal chain implies irreducibility of the Metropolized chain.
(3) To show aperiodicity of the Metropolized chain, it suffices to show that at least one state x has positive holding probability P_{xx}. We can show this by showing that for at least one pair of states x, y, we we have $a(x, y) < 1$. Assume $a(x, y) \geq 1$ for all $x, y \in \mathcal{S}$. But then $P_{xy} = T_{xy}$ for all $x, y \in \mathcal{S}$, and so π is already stationary for the proposal chain, and there is no need for the construction of a Metropolized chain $\mathbf{P}$ in the first place.

Example 9.1.1 (Simulating a Poisson distribution)**.** Here the state space is $\mathcal{S} = \mathbb{N}_0$ and the target distribution π is defined by $\pi(x) = \frac{e^{-\lambda}\lambda^n}{n!}$ for $n \in \mathbb{N}_0$. As an easy to simulate proposal chain, we can use simple symmetric random walk on $\mathbb{N}_0$ with reflecting boundary at 0. So

$$T_{n,n+1} = T_{n,n-1} = \frac{1}{2} \quad \text{for } n \geq 1 \quad \text{and } P_{01} = P_{00} = \frac{1}{2}\,.$$

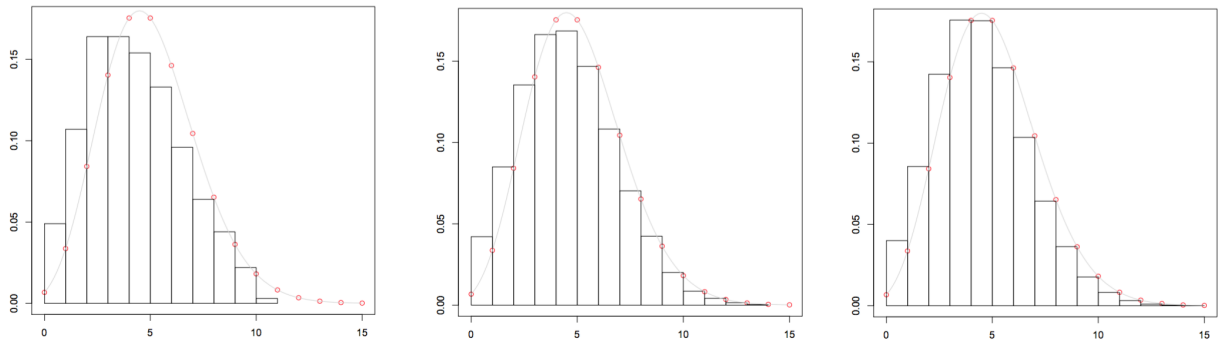
The acceptance probabilities are

$$\begin{aligned} a(n, n+1) &= \min(\tfrac{\lambda}{n+1}, 1) \ \text{ for } n \geq 0 \\ a(n, n-1) &= \min(\tfrac{n}{\lambda}, 1) \ \text{ for } n \geq 0\,. \end{aligned}$$

The transition probabilities of the Metropolized chain $\mathbf{P}$ on $\mathbb{N}_0$ are

$$P_{n,m} = \begin{cases} \frac{1}{2}\min(\frac{\lambda}{n+1}, 1) & \text{for } n \geq 0, m = n+1 \\ \frac{1}{2}\min(\frac{n}{\lambda}, 1) & \text{for } n \geq 1, m = n-1 \\ 1 - \frac{1}{2}[\min(\frac{\lambda}{n+1}, 1) + \min(\frac{n}{\lambda}, 1)] & \text{for } n \geq 1, n = m \\ 1 - \frac{1}{2}\min(\lambda, 1) & \text{for } n = m = 0 \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly, the chain $\mathbf{P}$ is irreducible and aperiodic. Figure 9.1 shows the result from three simulations of this Metropolis chain $\mathbf{P}$ for three different running times, namely for $k = 10^3$, $k = 10^4$, and $k = 10^5$ number of steps. The Poisson parameter for this simulation is $\lambda = 5$. (The simulation was done with the software package R.) The histogram for each simulation was scaled to a density histogram. The red overlaid dots represent the actual Poisson probabilities for Poisson(5). We can see from the three simulations that with increasing number of steps, the distribution of the Metropolis chain closer and closer approximates the target (stationary) distribution $\pi = \text{Poisson}(5)$. □



10^3 steps 10^4 steps 10^5 steps

Figure 9.1: Three simulations of a Metropolis chain whose target distribution is Poisson(5)

Example 9.1.2 (Random walk on a connected graph)**.** Consider a finite graph $G(V, E)$ which could model a social network, a computer network, etc. Suppose one does not know the overall size and global structure of the network, but for each vertex one knows its immediate nearest neighbors (for example, friends in a social network, linked websites to a given website). One can therefore choose one of its neighbors uniformly at random

and move there next. This is an example of simple random walk on a graph. If the graph is connected, the Markov chain is irreducible. The transition probabilities are

$$T_{vw} = \begin{cases} \frac{1}{\deg(v)} & \text{if } w \sim v \\ 0 & \text{otherwise}\,, \end{cases}$$

and the stationary distribution for simple random walk on the connected graph is

$$\nu(v) = \frac{\deg(v)}{2|E|} \qquad \text{for } v \in V$$

(recall Example 7.2.2). If the graph is not a regular graph, that is, if not every vertex v has the same degree, then in the long run, states with higher degree will be visited more often than states with lower degree.

Consider a function $f : V \to \mathbb{R}$ that models a certain property of each vertex v. We would like to assess the average value $f_{\text{avg}} = \frac{1}{|V|}\sum_{v\in V} f(v)$ of this property across the whole network. But $|V|$ may be extremely large (and not known), so a direct computation of f_{avg} may be impossible. A Monte Carlo approach to computing, or at least to closely approximating, f_{avg} proceeds as follows:

1. Construct an irreducible Metropolized chain $(X_n)_{n\geq 0}$ on $G(V, E)$ whose target distribution π is *uniform* distribution on V, so $\pi \propto (1, 1, ..., 1)$. In this case the acceptance probabilities are

   $$a(v, w) = \frac{\pi(w)T_{wv}}{\pi(v)T_{vw}} = \frac{\deg(v)}{\deg(w)}\,,$$

   and hence the transition probabilities are

   $$P_{vw} = \begin{cases} 0 & \text{if } w \nsim v \\ \dfrac{1}{\deg(v)} \min(\dfrac{\deg(v)}{\deg(w)}, 1) & \text{if } w \sim v\,. \end{cases}$$

   Notice that the acceptance probabilities bias the walk against moving towards higher-degree vertices (which would happen for unaltered simple random walk on the graph).

2. Simulate $(X_n)_{n\geq 0}$ and apply the Ergodic theorem for irreducible Markov chains (Theorem 3.1.1) to approximate f_{avg}. Recall that the Ergodic theorem states

   $$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) = \mathbb{E}(f(X)) \qquad \text{with probability } 1\,,$$

   where $X \sim \pi$ and π is the unique stationary distribution. For our case this yields

   $$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) = f_{\text{avg}} \qquad \text{with probability } 1\,.$$

As an additional question, we may want to find vertices at which f takes its maximum (or minimum) value. Such an optimization problem can be solved using a similar Monte Carlo approach. For this case we would choose a target distribution π that puts high probability on vertices v at which f attains the extreme value. We return to this question in Section 9.2. □

9.1.2 Gibbs Sampler

The Gibbs sampler, also called **Glauber dynamics**, is a Monte Carlo algorithm for sampling from probability distributions π on high-dimensional spaces.
Let $\mathcal{C}$ be a finite set. We consider the elements in $\mathcal{C}$ as labels or colors. We have an underlying graph $G(V, E)$ with large vertex set V. Each vertex $v \in V$ is assigned a label $c \in \mathcal{C}$ according to some rule. The collection of all such possible assignments makes up the state space $\mathcal{S} \subseteq \mathcal{C}^{|V|}$, which is also called the **configuration space**. A configuration (or state) $\mathbf{x} \in \mathcal{S}$ is a function

$$f_{\mathbf{x}} : V \to \mathcal{C}\,.$$

The Gibbs sampler is a Markov chain on the configuration space that, at each step, updates only one coordinate of the current state $\mathbf{x}$ in a prescribed way. The following describes the steps of the algorithm.

Gibbs Sampler algorithm:

Let $I = \{1, 2, ..., N\}$, and consider a strictly positive probability distribution π (the target distribution) on $\mathcal{S} \subseteq \mathcal{C}^N$. Assume the chain starts in state $\mathbf{x} = (x_1, x_2, ..., x_N)$ with $x_i \in \mathcal{C}$, $i \in I$.

1. Choose an index $k \in I$, independently of all previously chosen indices, according to a fixed, strictly positive probability distribution η on I (η is often uniform distribution on I).

2. Choose the new state $\mathbf{y}$ by updating the kth coordinate of $\mathbf{x}$ in the following way. Choose $\mathbf{y}$ from the probability distribution π on $\mathcal{S}$ conditioned on the set of configurations that agree with $\mathbf{x}$ at all indices $I \setminus \{k\}$. Let

   $$\mathcal{S}_{\mathbf{x},k} = \{\mathbf{y} \in \mathcal{S} \,:\, x_i = y_i \text{ for } i \in I \text{ and } i \neq k\}$$

   denote the set of all states $\mathbf{y}$ that agree with $\mathbf{x}$ at all indices, except possibly at index k. With this notation, the transition probabilities of the Gibbs sampler are

$$P_{\mathbf{xy}} = \begin{cases} \eta(k)\frac{\pi(\mathbf{y})}{\pi(\mathcal{S}_{\mathbf{x},k})} & \text{if } \mathbf{y} \in \mathcal{S}_{\mathbf{x},k} \text{ and } x_k \neq y_k \\ 0 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at more than one index}. \end{cases}$$

Clearly, for all $\mathbf{x} \in \mathcal{S}$, $P_{\mathbf{xx}} \neq 0$, so the chain is aperiodic. Let $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and assume that $\mathbf{y} \in \mathcal{S}_{\mathbf{x},k}$ for some $k \in I$. It follows that $\mathbf{x} \in \mathcal{S}_{\mathbf{y},k}$ and $\mathcal{S}_{\mathbf{x},k} = \mathcal{S}_{\mathbf{y},k}$. Using this equality, we verify that the detailed balance equations hold:

$$\pi(\mathbf{x})P_{\mathbf{xy}} = \pi(\mathbf{x})\eta(k)\frac{\pi(\mathbf{y})}{\pi(\mathcal{S}_{\mathbf{x},k})} = \pi(\mathbf{y})\eta(k)\frac{\pi(\mathbf{x})}{\pi(\mathcal{S}_{\mathbf{y},k})} = \pi(\mathbf{y})P_{\mathbf{yx}}\,.$$

If the Gibbs sampler with transition probabilities $P_{\mathbf{xy}}$ as defined above is also irreducible (which needs to be checked separately in each case), it will in fact converge to the target distribution π.

Example 9.1.3 (Disk packing)**.** This model is also known as the *hard-core model* in the literature. It is used in chemistry and statistical physics where it models a liquid or a gas whose particles are balls of positive radius who cannot overlap. The model starts with a graph $G(V, E)$ and the label set $\mathcal{C} = \{0, 1\}$. Label 1 stands for "occupied" and label 0 stands for "free". A label is assigned to each vertex in such a way that no two adjacent vertices, that is, vertices connected by an edge, are occupied. Such a configuration is called a *feasible* configuration. The state space $\mathcal{S}$ is the set of all feasible configurations. The below graph gives an example of a possible configuration where the graph is a rectangle in $\mathbb{Z}^2$. Black circles indicate occupied sites and, white circles indicate free sites. Figure 9.2 shows a possible configuration.

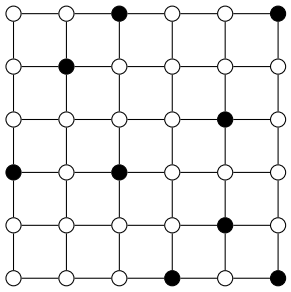


Figure 9.2: A configuration for disk packing on a square grid

In applications, the graph will have a large number of vertices which makes the state space $\mathcal{S} \subset \{0, 1\}^V$ extremely large. A natural quantity of interest is the average number of occupied sites. An exact computation would require an enumeration of all feasible

configurations $\mathbf{x}$ and counting their number $N(\mathbf{x})$ of occupied sites. Such a brute-force calculation may be very difficult and computationally costly, even prohibitive. A Monte Carlo approach provides a solution. Construct an irreducible, aperiodic Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{S}$ that converges to uniform distribution on $\mathcal{S}$. Sampling from the simulated (approximate) uniform distribution will allow an estimate for $\mathbb{E}(N(X))$ (where $X \sim \text{Unif}(\mathcal{S})$) with the use of the Ergodic theorem which states

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} N(X_k) = \mathbb{E}(N(X)) \,.$$

The Gibbs sampler applied to the hard-core model proceeds in the following way. Assuming the current state is $\mathbf{x}$,

1. choose a vertex v uniformly at random (and independently from all previous choices) from the vertex set v;
2. if at least one neighbor of v is occupied (which implies v has label 0), stay at $\mathbf{x}$;
3. if all neighbors of v are labeled 0, choose with equal probability a new label 1 or 0 and update vertex v with the new label. Leave the remaining vertices unchanged.

This algorithm results in the transition probabilities

$$P_{\mathbf{xy}} = \begin{cases} 0 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at more than one vertex} \\ \frac{1}{2|V|} & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at exactly one vertex} \\ 1 - \frac{n(\mathbf{x})}{2|V|} & \text{if } \mathbf{x} = \mathbf{y} \end{cases}$$

where $n(\mathbf{x})$ denotes the number of *nearest neighbors* of $\mathbf{x}$, that is, the number of feasible configurations $\mathbf{z}$ that differ from $\mathbf{x}$ at exactly one vertex. □

In Exercise 9.2, the reader is asked to prove irreducibility, aperiodicity, and reversibility (with respect to uniform distribution) of the Gibbs sampler for disk packing.

Example 9.1.4 (Proper graph coloring)**.** This problem has many applications in chemistry, scheduling problems, social networking, and other areas. Let $G(V, E)$ be a finite, connected graph and $\mathcal{C}$ a set of "colors" with $|\mathcal{C}| = k$. A proper k-coloring $\mathbf{x}$ of the graph G is an assignment of colors to the vertices of G with the property that no two adjacent vertices have the same color. See Figure 9.3. Let $\mathcal{S}$ be the set of all proper k-colorings of G. We have $\mathcal{S} \subset \mathcal{C}^V$. In applications, one would often like to sample from the uniform distribution on $\mathcal{S}$.

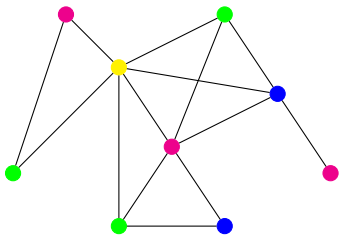


Figure 9.3: A proper graph coloring

A brute-force approach of enumerating all possible graph colorings for a given graph and a set of colors is likely not feasible due to the often enormous size of $\mathcal{S}$. Note that for a given graph G and k colors, a proper coloring may not exist. For k colors with $k \leq |V|$, proper graph colorings exist, and often a lot fewer colors are needed, depending on the graph structure. For the complete graph with $|V| = n$, we do need n different colors. For bipartite graphs, only two colors are needed.
The Gibbs sampler applied to proper graph coloring proceeds in the following way. Assuming the current state is the proper k-coloring (configuration) $\mathbf{x}$,

1. choose a vertex v uniformly at random (and independently from all previous choices) from the vertex set V;
2. choose a new color uniformly at random from the subset of colors that are "allowable" colors for v, that is, all colors not attained by any neighbor of v;
3. update vertex v with the new color (and leave all other vertices unchanged).

This algorithm results in the transition probabilities

$$P_{\mathbf{xy}} = \begin{cases} 0 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at more than one vertex} \\ \frac{1}{|V||\mathcal{S}_{\mathbf{x},v}|} & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at exactly vertex } v \end{cases}$$

where

$$\mathcal{S}_{\mathbf{x},v} = \{\mathbf{y} \in \mathcal{S} \,:\, \mathbf{x}(w) = \mathbf{y}(w) \text{ for } w \in V \text{ and } w \neq v\}$$

is the set of all proper k-colorings that agree with $\mathbf{x}$, except possibly at vertex v. □

In Exercise 9.6, the reader is asked to prove that the Gibbs sampler for proper graph coloring is aperiodic and reversible with respect to uniform distribution on $\mathcal{S}$. Irreducibility is often more difficult to determine and depends on the topological properties of the underlying graph G and the number k of available colors. However, for sufficiently large k, more specifically if

$$k \geq \max\{\deg(v) : v \in V\} + 2\,,$$

irreducibility is guaranteed.

Remark 9.1.2. *The* **chromatic number** $\chi(G)$ *of a graph* G *is the* minimal *number of colors needed to properly color the graph. It is often a difficult problem to determine* $\chi(G)$ *for a given graph.*
Four Color Theorem. *Any* **planar graph***, that is a graph that can be drawn in the plane without any of its edges crossing each other, has* **chromatic number at most four***. Francis Guthrie at University College London, England, first conjectured in 1852 that four colors suffice to color any map in a way that two countries sharing a boundary (that is a curve segment, not just a point) do not share a color. The theorem was proved much later, in 1976, by Kenneth Appel and Wolfgang Haken at U. Illinois.*

9.2 Stochastic Optimization and Simulated Annealing

We start with a large state space $\mathcal{S}$ and a function f on $\mathcal{S}$. The general problem is to find the global maximum or global minimum of f on $\mathcal{S}$. For large state spaces, an exhaustive search may be computationally prohibitive. And often problems of this type are combinatorial in nature for which a deterministic algorithm for enumerating the states may not even exist. An additional difficulty in the search of a global extremum could arise if f has a large number of local extrema on $\mathcal{S}$. While exact solutions may be elusive in many situations, Monte Carlo methods have often been very successful in producing close to optimal solutions to difficult, large-scale optimization problems.

Example 9.2.1 (The knapsack problem)**.** This is a famous problem in combinatorial optimization. We have n items labeled $1, 2, ..., n$. Each item i, for $1 \leq i \leq n$, has a weight w_i and a value v_i attached to it. Suppose we would like to put a selection of these items into a knapsack that allows a total upper weight limit of W. Find an optimal selection of items, that is, a selection that maximizes the total value of the items in the knapsack.
We identify any selection $A \subseteq \{1, 2, ..., n\}$ of items with a binary vector

$$\mathbf{z} = (z_1, ..., z_n) \in \{0, 1\}^n$$

where $z_i = 1$ iff item $i \in A$. The state space $\mathcal{S}$, which we call the set of **feasible solutions**, is the set

$$\mathcal{S} = \{\mathbf{z} \in \{0, 1\}^n \, : \, \sum_{i=1}^{n} w_i z_i \leq W\} \, .$$

We define the value function f on $\mathcal{S}$ by

$$f(\mathbf{z}) = \sum_{i=1}^{n} v_i z_i \,.$$

The optimization problem is to find

$$\max\{f(\mathbf{z}) \,:\, \mathbf{z} \in \mathcal{S}\} \,.$$

This problem has been widely studied in computer science. It is known to be NP-complete as a decision problem. No efficient algorithm is known (or is likely to exist) for an exact solution for the knapsack problem for large n. □

The basic idea for a Monte Carlo approach to solve an optimization problem of this kind is to simulate a Metropolis chain on $\mathcal{S}$ that converges to a stationary distribution π which puts high probability on states with extreme values of f. A standard distribution π that is used in this approach is the so-called *Boltzmann distribution*.

Definition 9.2.1. *Let $\mathcal{S}$ be a finite set, $f : \mathcal{S} \to \mathbb{R}$ a function on $\mathcal{S}$, and $T > 0$ a parameter. The* **Boltzmann distribution** $\pi_{f,T}$ *with* **energy function** f *and* **temperature parameter** T *is defined by*

$$\pi_{f,T}(s) = \frac{1}{Z_{f,T}} e^{-f(s)/T} \qquad \text{for } s \in \mathcal{S}$$

where $Z_{f,T} = \sum_{s\in\mathcal{S}} e^{-f(s)/T}$ is the normalizing constant that makes $\pi_{f,T}$ a probability distribution.

The energy function f in the Boltzmann distribution is the function f that arises from the given optimization problem. For a fixed temperature T, the Boltzmann distribution puts higher probability on states with relatively small values of f than on states with relatively large values of f. If the problem is to maximize f, then one would work with $-f$ as the energy function instead. Note that for high temperature T, the distribution $\pi_{f,T}$ is almost uniform on $\mathcal{S}$. The lower the temperature T, the more $\pi_{f,T}$ concentrates near states that minimize f.

Simulated Annealing

The idea of **simulated annealing** is to run a Metropolis chain for which one lets the temperature parameter T change with time. One starts with a high temperature T_1 and runs a Metropolis chain with target distribution π_{f,T_1} for N_1 time units. Since at high temperature T_1 the target distribution is almost uniform, the Markov chain will widely

explore the state space during that time (and not prematurely get trapped near a local extremum). After time N_1, one lowers the temperature to T_2 with $T_2 < T_1$ and allows the Metropolis chain to run with target distribution π_{f,T_2} until time N_2. Due to the lower temperature, π_{f,T_2} concentrates more on states that minimize f. After time N_2 one lowers the temperature to T_3 with $T_3 < T_2$, and so on. Theoretical results in this area confirm that under sufficiently slow "cooling", the probability that the Metropolis chain will be in an f-minimizing state at time n tends to 1 as $n \to \infty$.

A specific choice of a sequence of temperatures $T_1 > T_2 > \cdots$ with

$$\lim_{i\to\infty} T_i = 0$$

and corresponding sequence of times $N_1 > N_2 > \cdots$ is called a **cooling schedule**. Note that in simulated annealing, due to the change of the target distribution over time, the transition probabilities change over time as well. The result is a time-inhomogeneous Markov chain.

Aside: The term *annealing* is borrowed from metallurgy where it refers to a process of first heating and then slowly, and in a controlled way, cooling a metal to improve its physical properties.

Example 9.2.2 (The traveling salesman problem)**.** The traveling salesman problem is another famous optimization problem that has a long history of being studied. It has the same computational complexity as the knapsack problem. Suppose there is a list of n cities $\{1, 2, ..., n\}$, and for each (unordered) pair i, j of cities, a known distance d_{ij} between them. If $d_{ij} = 0$, then there is no possible direct connection between cities i and j. The problem is to find the shortest possible route for a salesman to visit each city exactly once, and in the end return to the city he started from.

The given connectivity/distance between the n cities is represented by a weighted, non-directed graph $G(V, E)$ with n vertices whose edges and edge weights represent the given distances between the cities. We assume $G(V, E)$ is connected. A tour that visits each vertex in a connected graph exactly once and returns to its starting point is called a *Hamiltonian cycle* for the graph. Figure 9.4 shows an example with five cities and two (not necessarily shortest) Hamiltonian cycles, one in purple and one in blue.

The traveling salesman problem is equivalent to finding a permutation $\sigma \in S_n$ (where S_n is the permutation group of n distinguishable objects) that minimizes the length $d(\sigma)$ of the tour

$$d(\sigma) = d_{\sigma(n),\sigma(1)} + \sum_{i=1}^{n-1} d_{\sigma(i),\sigma(i+1)} \tag{9.1}$$

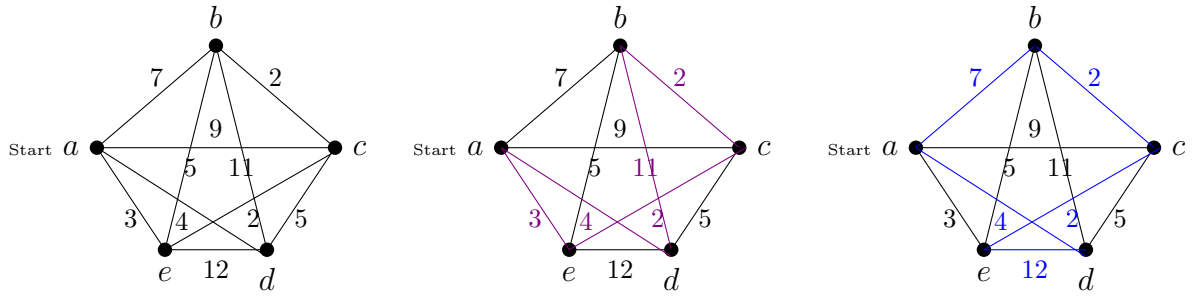


Figure 9.4: Two Hamiltonian cycles for a road system between 5 cities

and for which $d_{\sigma(i),\sigma(i+1)} > 0$, $d_{\sigma(n),\sigma(1)} > 0$. For large n, a brute force approach of checking all possible routes is computationally not feasible since $|S_n| = n!$ (which grows super-exponentially in n). To be precise, considering that the starting city stays fixed, and that the reverse tour has the same length, we need to minimize the length $d(\sigma)$ over $\frac{(n-1)!}{2}$ permutations σ.

We now describe a Metropolis chain that can be used in a simulated annealing algorithm for a solution to the traveling salesman problem. For simplicity we will assume that the underlying graph is complete, that is, $d_{ij} > 0$ for all $i, j \in V$. The state space is $\mathcal{S} = S_n$. We declare two permutations $\sigma, \sigma' \in \mathcal{S}$ to be neighbors, denoted by $\sigma \sim \sigma'$, if σ' arises from σ by reversing a substring of the tour. More precisely, if $\sigma = (\sigma_1, \sigma_2, ..., \sigma_n)$, then for $1 \leq i < j \leq n$, the corresponding substring is $(\sigma_i, \sigma_{i+1}, ..., \sigma_j)$. Reversing the order of the substring $(\sigma_i, \sigma_{i+1}, ..., \sigma_j)$ in σ yields the new permutation

$$\sigma' = (\sigma_1, ..., \sigma_{i-1}, \sigma_j, \sigma_{j-1}, ..., \sigma_i, \sigma_{j+1}, ..., \sigma_n)\,.$$

As an example, consider $\sigma = (1, 2, ..., n)$. The tour σ:

$$1 \to 2 \to \cdots \to \overbrace{i \to \cdots \to j}^{\text{reverse!}} \to \cdots \to n \to 1$$

becomes the tour σ':

$$1 \to 2 \to \cdots \to (i-1) \to \overbrace{j \to (j-1) \to \cdots \to (i+1) \to i}^{\text{reversed}} \to (j+1) \to \cdots \to n \to 1\,.$$

Figure 9.5 shows two tours that are considered neighbors (the example is for 9 cities and $i = 4$, $j = 7$). Notice that in Figure 9.4, the purple tour $a \to \overbrace{e \to c \to b}^{\text{reverse!}} \to d \to a$ and the blue tour $a \to \overbrace{b \to c \to e}^{\text{reversed}} \to d \to a$ are also neighbors.

The algorithm will only allow one-step transitions between neighbors. The target distribution for the Metropolis chain is the Boltzmann distribution from Definition 9.2.1 with

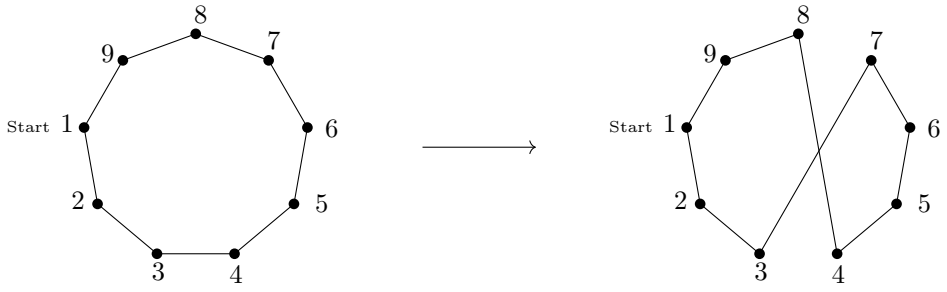


Figure 9.5: Two neighboring tours for 9 cities

energy function $f(\sigma) = d(\sigma)$ from (9.1). At each step, the proposal chain $\mathbf{T}$ chooses two indices i, j uniformly at random from $\{1, 2, ..., n\}$ with $i < j$. Given that the current state is σ, the proposal chain proposes the neighboring permutation σ' constructed as described above. In other words, the proposal chain $\mathbf{T}$ is simple random walk on a graph with vertex set $\mathcal{S} = S_n$ (each vertex is identified with a permutation $\sigma \in S_n$) and for which each vertex has $\binom{n}{2}$ neighbors.
For fixed temperature T, we get the following transition probabilities $P_{\sigma,\sigma'}$ for the Metropolis chain:

$$P_{\sigma,\sigma'} = \begin{cases} \dfrac{2}{n(n-1)} \min\left(e^{(d(\sigma)-d(\sigma'))/T}, 1\right) & \text{if } \sigma, \sigma' \text{ are neighbors} \\ 0 & \text{if } \sigma \neq \sigma' \text{ and } \sigma, \sigma' \text{ are not neighbors} \\ 1 - \displaystyle\sum_{\sigma":\sigma"\sim\sigma} \dfrac{2}{n(n-1)} \min\left(e^{(d(\sigma)-d(\sigma"))/T}, 1\right) & \text{if } \sigma = \sigma'. \end{cases}$$

Note that for this algorithm,

$$d(\sigma) - d(\sigma') = d_{i-1,i} + d_{j,j+1} - d_{i-1,j} - d_{i,j+1}$$

which shows that the computation of $d(\sigma) - d(\sigma')$ does not depend on n, and so there is low computational cost involved in this step of the algorithm.

The last step for implementing the algorithm is to decide on a cooling schedule $T_1, T_2, ...$ with corresponding $N_1, N_2, ...$ Finding a suitable cooling schedule for simulated annealing often requires some experimentation. □

Exercises

Exercise 9.1. Consider a standard 8×8 chessboard as shown in figure 9.6 below.

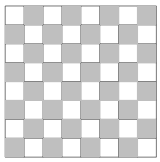


Figure 9.6: Standard chessboard

A king can move one square at a time in any direction (horizontally, vertically, or diagonally) that is available from his current location. Construct transition probabilities for the king's moves on the chessboard that guarantee that, in the long run, the king will spend equal percentage of time on each square of the board.

Exercise 9.2. Consider the disk-packing model introduced in Example 9.1.3. Prove that the Gibbs sampler described in this example defines an (a) irreducible, (b) aperiodic, and, (c) with respect to uniform distribution, reversible Markov chain $(X_n)_{n\geq 0}$ on the space of feasible configurations $\mathcal{S}$.

Exercise 9.3 (A generalization of the disk-packing model)**.** One can generalize Example 9.1.3 by introducing a parameter $\lambda > 0$ to the stationary distribution which will change the weight given to different "packing sizes". More precisely, one defines the target probability distribution π_λ on $\mathcal{S}$ by

$$\pi_\lambda(\mathbf{x}) = \frac{\lambda^{N(\mathbf{x})}}{Z_\lambda} \quad \text{for } \mathbf{x} \in \mathcal{S}$$

where $N(\mathbf{x})$ is the number of occupied sites of configuration $\mathbf{x}$, and Z_λ is the normalizing constant

$$Z_\lambda = \sum_{\mathbf{x}\in\mathcal{S}} \lambda^{N(\mathbf{x})}$$

for the distribution. Note that the case $\lambda = 1$ reduces to the standard disk-packing model.

(a) Describe the transition probabilities for the Gibbs sampler algorithm for this model.
(b) Verify that π_λ is the stationary distribution.

Exercise 9.4 (A 1-dimensional disk-packing model)**.** Consider a chain graph of n vertices and let $\mathcal{S}_n$ be the set of feasible configurations for disk packing as described in Example 9.1.3. For an example of a feasible configuration for a chain graph of 12 vertices see Figure 9.7.

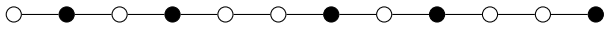


Figure 9.7

We would like to draw samples from the uniform distribution on $\mathcal{S}_n$. Construct the transition probabilities for a Markov chain on $\mathcal{S}_n$ that converges to Unif($\mathcal{S}_n$). How would you then use this Markov chain to estimate the average number m_n of occupied sites?

Exercise 9.5. Consider the disk-packing model for a chain graph of m vertices as in Exercise 9.4. Fix $n > 0$ even.

(a) Let $n_{j,m}$ be the number of feasible configurations with exactly j occupied sites. Show that

$$n_{j,m} = \binom{n+1-j}{j} \quad \text{for } j = 0, 1, ..., \frac{n}{2}\,.$$

(b) Recall the *Fibonacci sequence* $1, 1, 2, 3, 5, 8, 13, ...$ It is described by the recurrence

$$f_n = f_{n-1} + f_{n-2}$$

for $n \geq 3$, with $f_1 = f_2 = 1$. Show that $|\mathcal{S}_n| = f_{n+2}$, and as a consequence, we have

$$f_{n+2} = \sum_{j=0}^{n/2} \binom{n+1-j}{j}\,.$$

Exercise 9.6. Recall Example 9.1.4. Show that the Gibbs sampler for proper k-colorings defines a Markov chain that is (a) aperiodic, and (b) reversible with respect to uniform distribution on $\mathcal{S}$.

Exercise 9.7 (Sampling from a power law distribution)**.** Fix $s > 1$. Let $\mathcal{S} = \mathbb{N}$ and consider the distribution π defined by

$$\pi(n) \propto \frac{1}{n^s} \qquad \text{for } n \geq 1\,.$$

(a) Using the Metropolis algorithm, construct a Markov chain $(X_n)_{n\geq 0}$ whose limiting distribution is π. As a proposal chain, use simple symmetric random walk on $\mathbb{N}$ with reflecting boundary at 1. Describe how the algorithm proceeds. (Note that we do not need to know the normalizing constant $c = \left(\sum_{n=1}^{\infty} \frac{1}{n^s}\right)^{-1}$.)

(b) For $s = 2$, explicitly state the transition probabilities for $(X_n)_{n\geq 0}$ resulting from the algorithm in part (a).

Exercise 9.8 (The knapsack problem)**.** Recall Example 9.2.1. We have n items labeled $1, 2, ..., n$. Each item i, for $1 \leq i \leq n$, has a weight w_i and a value v_i attached to it. Suppose we would like to put a selection of these items into a knapsack that allows a total upper weight limit of W. The goal is to find an optimal selection of items, that is, a selection that maximizes the total value of the items in the knapsack. We represent a selection of items as a binary vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, where $z_j = 1$ if Item j is part of the selection, and zero otherwise. The constraint is then expressed as

$$\sum_{j=1}^{n} z_j w_j \leq W.$$

Let $\mathcal{S}$ be the set of binary vectors $\mathbf{z}$ that satisfy the above inequality.

(a) Consider the following Markov chain. Start at $(0, 0, \ldots, 0)$. At each step, choose an index j uniformly at random from $\{1, 2, \ldots, n\}$, replace z_j with $1 - z_j$ in the current selection vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, and call the resulting vector $\mathbf{y}$. If $\mathbf{y}$ is not in $\mathcal{S}$, remain at $\mathbf{z}$. If $\mathbf{y}$ is in $\mathcal{S}$, move to $\mathbf{y}$. Show that the uniform distribution on $\mathcal{S}$ is stationary for this chain.

(b) Show that the chain from part (a) is irreducible.

(c) The chain from part (a) is a useful way to draw samples from an approximately uniform distribution on $\mathcal{S}$. However, we are interested in optimizing (maximizing) the total value of our selection of items. We therefore construct a Markov chain with a stationary distribution π that puts much higher probability on any high-value solution than on any low-value solution. Specifically, we want to simulate from the distribution

$$\pi(x) \propto e^{\lambda f(\mathbf{z})}$$

where $f(\mathbf{z}) = \sum_{j=1}^{n} v_j z_j$ and λ is some positive constant. Using the Metropolis-Hastings algorithm, create a Markov chain whose stationary distribution is π.

Chapter 10

Random Walks on Groups

10.1 Basic notions

10.1.1 Generators, convolution powers

For many naturally occurring Markov chains, the state space $\mathcal{S}$ has the additional structure of that of a *group*, and the Markov chain is random walk on that group. Examples that we have already encountered include random walk on $\mathbb{Z}$ with group operation + and random walk on the discrete N-cycle $\{0, 1, ..., N-1\}$ with group operation + (mod N). In both examples, the Markov chain proceeds from state x to a new state by choosing an element y from $\mathcal{S}$, independently of x and according to a given *step distribution* μ on $\mathcal{S}$, and then by composing (adding) $x+y$, which is the resulting new state. For example, for simple symmetric random walk on $\mathbb{Z}$, at each time n, the Markov chain chooses its next step uniformly at random (and independently from all other steps) from $\{-1, 1\}$. This process of successively adding (or whatever the group operation is) i.i.d. steps can be generalized to any group. Section 10.2 introduces the larger class of examples of random walks on the *symmetric group* $\mathrm{S_n}$, i.e. the group of permutations of n distinguishable objects. We start with some basic definitions:

Definition 10.1.1. *A* **discrete group** $(G, \circ)$ *is a finite or countably infinite set* G *together with a* **group operation** $\circ : G \times G \to G$ *that has the following properties:*

(a) The operation $\circ$ *is associative.*

(b) There exists an identity element $id \in G$ *such that* $x \circ id = id \circ x = x$ *for all* $x \in G$*.*

(c) For all $x \in G$ *there exists an inverse* $x^{-1} \in G$ *such that* $x^{-1} \circ x = x \circ x^{-1} = id$*.*

If, in addition,

(d) $x \circ y = y \circ x$ *for all* $x, y \in G$*,*

then G *is called a* **commutative** *or* **abelian group**.

Definition 10.1.2. *Let* G *be a discrete group and* $U \subset G$*. We say*

(a) We say U **generates** G*, or* U *is a set of* **generators** *for* G*, if every* $g \in G$ *can be written as a finite product of elements in* U*.*

(b) If U *generates* G *and we have* $u \in U \iff u^{-1} \in U$*, the set* U *is called a* **symmetric set of generators** *for* G*.*

Note that the set of generators U may be a finite set or an infinite set. In the following, we will suppress writing $\circ$. Instead of writing $x \circ y$, we will simply write xy. Let G be a discrete group. Any probability distribution μ on G defines a *random walk on* G:

Definition 10.1.3. *Let* G *be a discrete group and* μ *a probability distribution on* G*. A* **right random walk** $(X_n)_{n \geq 0}$ *on* G *with* **step distribution** μ *is a Markov chain on* G *whose transition probabilities are*

$$P_{xy} = \mu(x^{-1}y) \quad \text{for all } x, y \in G\,.$$

Similarly, we define a **left random walk** *with step distribution* μ *to be the Markov chain whose transition probabilities are* $P_{xy} = \mu(yx^{-1})$ *for all* $x, y \in G$*.*

Recall the notion of *support*, denoted by $\mathrm{supp}(\mu)$, of a distribution μ on G:

$$\mathrm{supp}(\mu) = \{g \in G : \mu(g) > 0\}\,.$$

Example 10.1.1. In Example 1.5.6 we have introduced simple random walk on the k-dimensional hypercube. Here the state space is $\mathbb{Z}_2^k = \{0, 1\}^k$. It is a finite Abelian group

whose group operation is component-wise addition modulo 2. The step distribution μ in this example is given by

$$\mu(1,0,...,0) = \mu(0,1,...,0) = \cdots = \mu(0,...,0,1) = \frac{1}{k}\,.$$

Note that $\mathrm{supp}(\mu)$ is a symmetric set of generators for the group $\mathbb{Z}_2^k$. □

Proposition 10.1.1. *Let G be a discrete group and μ a probability distribution on G. Both left and right random walks on G with step distribution μ are irreducible if and only if $\mathrm{supp}(\mu)$ generates G.*

Proof. We show this for right random walk. The proof for left random walk is analogous. Set $U = \mathrm{supp}(\mu)$. First, assume U generates G, and let $x, y \in G$. Then there exist k and $u_1, ..., u_k \in U$ such that $x^{-1}y = u_1 u_2 \cdots u_k$. Hence

$$P^k_{xy} \geq P_{x,xu_1} P_{xu_1,xu_1u_2} \cdots P_{xu_1\cdots u_{k-1},y} = \mu(u_1)\mu(u_2)\cdots\mu(u_k) > 0\,,$$

which shows that the random walk is irreducible. Conversely, let us assume that the random walk is irreducible. Let $x \in G$. Then there exists m such that $P^m_{id,x} > 0$. It follows that there exist $u_1, ..., u_{m-1} \in U$ such that

$$P_{id,u_1} P_{u_1,u_1u_2} \cdots P_{u_1\cdots u_{m-1},x} > 0\,,$$

and, consequently, there must also exist u_m such that $x = u_1 \cdots u_{m-1}u_m$. This shows that U generates G. □

Let $(X_n)_{n\geq 0}$ be (right) random walk on the discrete group G with step distribution μ. Assume the random walk starts at the identity id of G. Thus $X_0 = id$ and $X_1 \sim \mu$. The distribution of X_2 (after the walk has taken 2 steps) is the **convolution** $\mu \star \mu$. It is defined by

$$\mu \star \mu(x) = \sum_{y\in G} \mu(y)\mu(y^{-1}x) \qquad \text{for all } x \in G\,.$$

We write $\mu^{\star 2} = \mu \star \mu$. It is straightforward to show that $\mu^{\star 2}$ is again a probability distribution. From this, we get the distribution of X_3 (after the walk has taken 3 steps). It is $\mu^{\star 3} = \mu^{\star 2} \star \mu$ defined by

$$\mu^{\star 3}(x) = \sum_{y\in G} \mu^{\star 2}(y)\mu(y^{-1}x) \qquad \text{for all } x \in G\,.$$

By induction, $X_n \sim \mu^{\star n}$ where

$$\mu^{\star n}(x) = \sum_{y\in G} \mu^{\star(n-1)}(y)\mu(y^{-1}x) \qquad \text{for all } x \in G\,.$$

Thus the n-step transition probabilities for a random walk on a group G are given by the n**-fold convolution powers** of μ:

$$P_{xy}^n = \mu^{\star n}(x^{-1}y) \quad \text{for all } x, y \in G\,.$$

As a consequence, the n-step transition matrices $\mathbf{P}^n$ for random walk on a finite group are *doubly stochastic*:

Definition 10.1.4. *A stochastic matrix* $\mathbf{P}$ *is called* **doubly stochastic** *if each column sums to* 1*.*

Indeed, for any random walk on a finite group G with step distribution μ, the yth column of $\mathbf{P}$ sums to

$$\sum_{x\in G} P_{xy} = \sum_{x\in G} \mu(x^{-1}y) = \sum_{g\in G} \mu(g) = 1\,.$$

The second equality follows from the fact that for any $y \in G$, the map $f(x) = x^{-1}y$ is a one-to-one function on G. Hence $\mathbf{P}$ (in fact $\mathbf{P}^n$ for all $n \geq 1$; to see this simply replace P_{xy} with P_{xy}^n and μ with $\mu^{\star n}$ in the above equations) is doubly stochastic.

For the following, recall the notion of an *invariant measure* (which is not necessarily a probability measure) from Definition 2.2.2.

Proposition 10.1.2. *Let G be a discrete group and μ a step distribution for (right) random walk on G. Then for any $c > 0$, the constant measure $\pi \equiv c$ on G is an invariant measure for the random walk.*

Proof. For any $y \in G$,

$$\sum_{x\in G} \pi(x)P_{xy} = c\sum_{x\in G} \mu(x^{-1}y) = c\sum_{g\in G} \mu(g) = c = \pi(y)\,.$$

□

Corollary 10.1.3. *Let G be a discrete group and μ be a step distribution whose support* $\mathrm{supp}(\mu)$ *generates G. Then the following holds for the associated random walk:*

(a) If G is finite, then the uniform distribution $\mu \equiv \frac{1}{|G|}$ is the unique stationary distribution.

(b) If G is infinite, then the random walk is either null recurrent or transient. There is no stationary distribution.

Proof. By Proposition 10.1.1, the random walk is irreducible. If G is finite, then the walk is positive recurrent and has a unique stationary distribution π. By Proposition 10.1.2, π is constant on G, hence π is uniform distribution on G. If G is infinite and the walk is recurrent, by Theorem 2.2.6, the constant measures on G are the only invariant measures for the random walk. However, since G is infinite, a constant measure on G cannot be normalized to a probability measure on G. Hence the random walk must be null recurrent. If the random walk is transient, then no stationary distribution exists either. □

It follows that for finite G, if the random walk on G with step-distribution μ is irreducible and aperiodic, then

$$\lim_{n\to\infty} \mu^{\star n}(x) = \frac{1}{|G|} \quad \text{for all } x \in G\,.$$

Note that $\mu(id) > 0$ is a sufficient (but not necessary) condition for aperiodicity of the associated random walk.

If a random walk $(X_n)_{n\geq 0}$ is periodic (and hence $\mu(id) = 0$), we often work with a so-called *lazy version* $(\tilde{X}_n)_{n\geq 0}$ of the random walk instead. We construct a lazy version by introducing positive holding probability in the following way: Consider a sequence $(Z_n)_{n\geq 1}$ of i.i.d. Bernoulli trials (coin tosses) that are also independent of $(X_n)_{n\geq 0}$. Fix a positive holding probability p with $p < 1$ and let $\mathbb{P}(Z_1 = 1) = p = 1 - \mathbb{P}(Z_1 = 0)$. The lazy version $(\tilde{X}_n)_{n\geq 0}$ is defined by $\tilde{X}_0 = X_0$ and

$$\tilde{X}_n = X_{S_n} \qquad \text{with } S_n = \sum_{k=1}^{n} Z_k$$

for $n \geq 1$. Roughly speaking, at each step, a coin toss decides whether the random walk stays in place or progresses (according to the original transition probabilities). The resulting step distribution $\tilde{\mu}$ for the lazy walk is

$$\tilde{\mu} = p\mu + (1-p)\delta_{id}$$

where δ_{id} denotes unit mass at the identity id.

10.1.2 Time reversal of a random walk

For some examples of random walks on groups, it turns out to be easier to analyze their *time reversal* (see Section 7.1), rather than the original random walk. Since for random walk on a discrete group G the constant measures are the only invariant measures, the time reversed random walk has step distribution $\tilde{\mu}$ defined by

$$\tilde{\mu}(g) = \mu(g^{-1}) \quad \text{for } g \in G\,.$$

We have shown in Section 7.1 that if a Markov chain is irreducible, so is its time reversal. The following lemma restates this fact for the special case of random walks on groups.

Lemma 10.1.4. *Let G be a discrete group and $U \subset G$ a set of generators. The the set $\tilde{U}$ defined by*

$$u^{-1} \in \tilde{U} \iff u \in U$$

is a set of generators for G.

Proof. Let $g \in G$. Since U generates G, there exist $u_1, ..., u_m \in U$ such that $g^{-1} = u_1 \cdots u_m$. So

$$g = u_m^{-1} \cdots u_1^{-1} ,$$

and since $u_1^{-1}, ..., u_m^{-1} \in \tilde{U}$, the set $\tilde{U}$ generates G. □

In Section 10.2 we will introduce examples of random walks on the permutation group $\mathrm{S_n}$. We will be interested in studying their rates of convergence to stationarity (to uniform measure). For such a study, we first need to decide on a specific notion of distance between two probability measures ν, π on G. One commonly used notion of distance is the L_1-distance defined by

$$\|\nu - \pi\|_{L_1} = \sum_{g \in G} |\nu(g) - \pi(g)| ,$$

or rather $\|\nu - \pi\|_{TV} := \frac{1}{2}\|\nu - \pi\|_{L_1}$, which is called *total variation distance* and which has a nice probabilistic interpretation (see Section 11.1). The following lemma explains why we can work with the time reversed random walk when studying rates of convergence with respect to L_1- (or total variation) distance.

Lemma 10.1.5. *Let G be a finite group, μ a step distribution on G whose support generates G, and π uniform distribution on G. Let $\tilde{\mu}$ be the step distribution of the time reversed walk. Then*

$$\|\mu^{*n} - \pi\|_{L_1} = \|\tilde{\mu}^{*n} - \pi\|_{L_1} \qquad \text{for } n \geq 1 .$$

Proof. We assume that the random walk is right random walk and starts at *id*. Let $n \geq 1$. The random walk proceeds by successively and independently choosing $x_1, x_2, ...$ from G according to μ and right multiplying each new chosen element with the current product. Thus

$$\mathbb{P}(X_n = x) = \mu^{*n}(x) = \sum_{\substack{x_1,\ldots,x_n: \\ x_1 \cdots x_n = x}} \mu(x_1) \cdots \mu(x_n)$$

which is equal to

$$\sum_{\substack{x_1,\dots,x_n:\\ x_1\cdots x_n=x}} \tilde{\mu}(x_1^{-1})\cdots\tilde{\mu}(x_n^{-1}) = \sum_{\substack{x_1,\dots,x_n:\\ x_n^{-1}\cdots x_1^{-1}=x^{-1}}} \tilde{\mu}(x_n^{-1})\cdots\tilde{\mu}(x_1^{-1}) = \tilde{\mu}^{*n}(x^{-1}) = \mathbb{P}(\tilde{X}_n = x^{-1})\,.$$

We get

$$\|\mu^{*n} - \pi\|_{L_1} = \sum_{x\in G}\left|\mu^{*n}(x) - \frac{1}{|G|}\right| = \sum_{x\in G}\left|\tilde{\mu}^{*n}(x^{-1}) - \frac{1}{|G|}\right| = \|\tilde{\mu}^{*n} - \pi\|_{L_1}\,.$$

□

Note: In general it is not true that, for a given time n, the distance to stationarity for a Markov chain and the distance to stationarity for its time reversal are equal.

Definition 10.1.5. *Let G be a discrete group and μ a probability distribution on G. We say μ is a* **symmetric probability distribution** *if*

$$\mu(x) = \mu(x^{-1}) \quad \text{for all } x \in G\,.$$

Lemma 10.1.6. *An irreducible random walk on a finite group G is reversible if and only if its step distribution μ is symmetric.*

Proof. For any random walk on a finite group G, uniform distribution on G is a stationary distribution. The statement follows from the fact that $P_{xy} = \mu(x^{-1}y)$. □

It follows that for a reversible random walk on a finite group G, the corresponding one-step transition matrix $\mathbf{P}$ is a symmetric matrix (and all eigenvalues of $\mathbf{P}$ are real).

Lastly, we point out that certain examples of Markov chains can be "lifted" to a random walk on a group, so the original Markov chain can be viewed as a lumped version (recall Section 1.7) of the random walk on the group. This viewpoint is often helpful towards analyzing the Markov chain. We have already seen examples of lumping: The Ehrenfest chain is a lumped version of simple random walk on the hypercube (see Example 1.7.4 and Subsection 11.2.4). Random walk on the integers $\{0, 1, ..., N\}$ with reflecting boundary at 0 and N can be viewed as a lumped version of simple random walk on a discrete cycle (see Example 1.7.5).

10.2 Card shuffling

Shuffling a deck of n cards is a Markov chain Monte Carlo approach to producing a (uniform or near uniform) random permutation of n cards. The Markov chain $(X_n)_{n\geq 0}$ that models the process of shuffling is a random walk on the symmetric group $\mathrm{S_n}$ which is the group of all permutations of n distinguishable objects. Recall that a permutation $\sigma \in \mathrm{S_n}$ is a bijective map

$$\sigma : \{1, ..., n\} \to \{1, ..., n\}$$

and thus can be identified with a particular order of the deck. The group operation on $\mathrm{S_n}$ is composition of two such functions. Note that the symmetric group $\mathrm{S_n}$ is non-abelian for $n \geq 3$.

Although shuffling a deck of cards is an entertaining way to think about random walks on the symmetric group $\mathrm{S_n}$, it is by no means the only application. The study of "mixing up n distinguishable objects" is relevant in many areas, for example in genetics or in cryptography. Note that here the size of the state space grows super-exponentially in n, which makes the state space prohibitively large for any direct enumeration of the states in most applications. For example, for a standard deck of 52 cards, the state space $\mathrm{S_n}$ consists of $|\mathrm{S_n}| = 52! > 8 \cdot 10^{67}$ permutations.

How we shuffle a deck, that is, which *shuffling mechanism* is used, is determined by the step distribution μ of the random walk. In theory, any distribution μ that is supported on a set of generators of the group $\mathrm{S_n}$ could be considered. But in praxis, only a few distributions μ translate into a convenient algorithm and have "fast enough" convergence to uniformity (we discuss rates of convergence in Chapter 11). In the following we describe two shuffling mechanisms that have been studied in great detail over the past decades (see Exercise 10.8 for additional examples).

Top-to-random shuffle

Consider top-to-random card shuffling chain with n cards. Initially, the deck is in perfect order. At each time, the top card is taken off and randomly inserted into the deck. See Figure 10.1 for a possible shuffle. The state space consists of all possible orders of the deck, i.e., all possible permutations of n distinguishable cards.

This process is modeled as a random walk on $\mathrm{S_n}$. The step distribution μ is concentrated and uniformly distributed on the set of cyclic permutations $C = \{\sigma_1, \sigma_2, ..., \sigma_n\}$ of the form

$$\sigma_k = (1 \to 2 \to 3 \to \cdots \to k \to 1) \quad \text{for } k = 1, 2, ..., n\,. \tag{10.1}$$

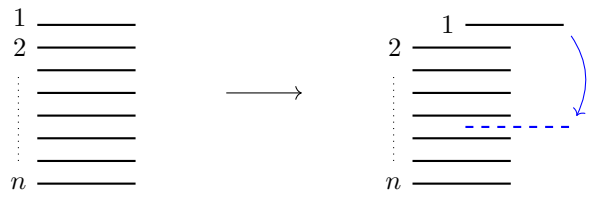


Figure 10.1: Top-to-random shuffle

Here $(1 \to 2 \to 3 \to \cdots \to k \to 1)$ in (10.1) stands for the cyclic permutation σ_k defined by

$$\sigma_k(i) = i+1 \quad \text{for } 1 \le i \le k-1\,, \qquad \sigma_k(k) = 1\,, \qquad \sigma_k(j) = j \quad \text{for } k+1 \le i \le n\,.$$

Note that $\sigma_1 = id$. It is well known that the set C is a set of generators for the symmetric group $\mathrm{S_n}$. Top-to-random shuffling proceeds as follows. At first, the deck of n cards is in perfect order (i.e. the chain starts at id). Perfect order means the cards are stacked according to descending label, so the card labelled 1 is the top card, and the card labelled n is the bottom card. For each shuffle, a cyclic permutation $\sigma_i \in C$ is chosen according to the step distribution

$$\mu(\sigma_i) = \frac{1}{n} \quad \text{for } 1 \le i \le n\,.$$

For the actual shuffle, this means that the top card is taken off and inserted into the deck at a uniformly randomly chosen location. This corresponds to multiplying (composing) the current permutation (which describes the current order of the deck) on the right by the chosen σ_i. Figure 10.2 shows an example (for $n = 5$) of a sample path (X_0, X_1, X_2, X_3) of three shuffles.

position	X_0	$\xrightarrow{\sigma_3}$	X_1	$\xrightarrow{\sigma_4}$	X_2	$\xrightarrow{\sigma_2}$	X_3
1	1	$\longrightarrow$	2	$\longrightarrow$	3	$\longrightarrow$	1
2	2	$\longrightarrow$	3	$\longrightarrow$	1	$\longrightarrow$	3
3	3	$\longrightarrow$	1	$\longrightarrow$	4	$\longrightarrow$	4
4	4	$\longrightarrow$	4	$\longrightarrow$	2	$\longrightarrow$	2
5	5	$\longrightarrow$	5	$\longrightarrow$	5	$\longrightarrow$	5

Figure 10.2: Sample path for three top-to-random shuffles

The last column gives the deck after three shuffles, where for the first shuffle the top card was inserted below the third card, for the second shuffle the top card was inserted below the fourth card, and for the third shuffle the top card was inserted below the second card.

This results, at time 3, in the permutation $\sigma_3 \circ \sigma_4 \circ \sigma_2$ which is given in the last row of Figure 10.3.

i	:	1	2	3	4	5
$\sigma_3(i)$	:	2	3	1	4	5
$\sigma_3(\sigma_4(i))$	:	3	1	4	2	5
$\sigma_3(\sigma_4(\sigma_2(i)))$	:	1	3	4	2	5

Figure 10.3: Resulting permutation of the deck after each shuffle

Note that when we write $\sigma(i) = k$ for a permutation σ, it means **"Card k is in position i"**. Thus the top-to-random shuffle with step distribution concentrated on C constitutes a **right random walk on** $\mathrm{S_n}$.

The **time reversal** of top-to-random shuffle is **random-to-top** shuffle. Its step distribution $\tilde{\mu}$ is uniform measure on the set $\tilde{C} = \{id, \sigma_2^{-1}, ..., \sigma_n^{-1}\}$ where

$$\sigma_k^{-1} = (k \to (k-1) \to (k-2) \to \cdots \to 1 \to k) \qquad \text{for } 2 \le k \le n\,.$$

A shuffle is performed by choosing a card uniformly at random, taking it out of the deck, and putting it on top of the pile. A possible random-to-top shuffle is shown in Figure 10.4. This random walk is also called the **move-to-front chain** or **Tsetlin library** (think of a librarian returning each randomly selected and used book back to the front of the shelf).

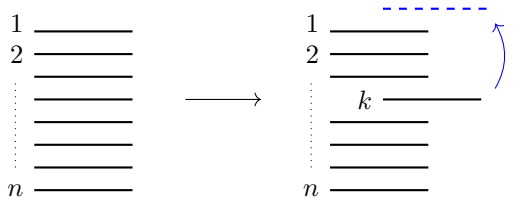


Figure 10.4: Random-to-top shuffle

Riffle shuffle

Riffle shuffling is a more realistic way of shuffling cards than, say, top-to-random shuffling. The model was proposed by Gilbert and Shannon (1955) and later, independently, by Reeds (1981). It has been closely studied ([1], [5]) and become famous for the result "7 shuffles mix a deck of cards" in [5].
The shuffling mechanism is as follows. Start with a perfectly ordered deck of n cards. At each step, cut the deck in half according to a binomial distribution $\mathrm{Bin}(n, \frac{1}{2})$. This gives

2 stacks, one of size m and one of size $n-m$, which happens with probability $\binom{n}{n}1/2^n$. Then "riffle" (interlace) the two stacks, but keep the relative order of the cards in each stack. There are $\binom{n}{m}$ ways to riffle the two stacks together. Assume each of these $\binom{n}{m}$ arrangements is equally likely to occur. Figure 10.5 shows a possible riffle shuffle with $m = 3$. It occurs with probability $1/2^n$.

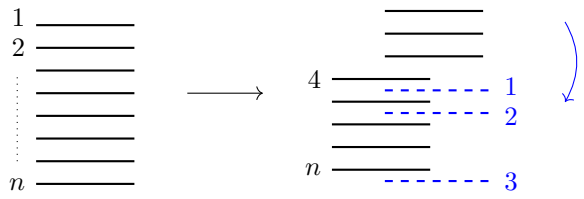


Figure 10.5: Riffle shuffle

We now define the concept of a **rising sequence** in a permutation: A rising sequence is a maximal consecutively increasing subsequence of the permutation. For example, the shuffle in the above picture results in a permutation with 2 rising sequences (viewed from top down): $(1, 2, 3)$ and $(4, 5, ..., n)$. For every permutation σ, the set $\{1, 2, ..., n\}$ decomposes into a disjoint union of rising sequences. A single riffle shuffle results in a permutation that has exactly two rising sequences or is the identity id (which has exactly one rising sequence). Two riffle shuffles performed in a row result in a permutation that has at most 4 rising sequences. With each shuffle, the number of rising sequences can at most double.

The step distribution μ for a riffle shuffle is given by

$$\mu(\sigma) = \begin{cases} 1/2^n & \text{if } \sigma \text{ has exactly two rising sequences} \\ (n+1)/2^n & \text{if } \sigma = id \\ 0 & \text{otherwise}. \end{cases}$$

Note that a single shuffle results in id if either $m = 0$ or $m = n$, or $m = j$ for $1 \leq j \leq n-1$ and the "riffling" puts the two stacks back on top of each other, resulting in the original order of the deck. Hence $\mu(id) = (n+1)/2^n$.

After k shuffles, the order of the deck has distribution μ^{*k} (the k-fold convolution of μ) on the permutation group $\mathrm{S_n}$. In [5], Bayer and Diaconis compute the exact formula for $\mu^{*k}(\sigma)$: Let $R(\sigma)$ denote the number of rising sequences of permutation σ. Then

$$\mu^{*k}(\sigma) = \frac{1}{2^{nk}} \binom{2^k + n - R(\sigma)}{n}.$$

The **time reversal** for the riffle shuffle proceeds as follows. At each step, we mark each card with either 0 or 1, according to i.i.d. Bernoulli random variables. We then sort the

cards according to this marking by bringing all cards marked with 0 to the top of the pile, leaving their relative order at the time of the marking intact. We will refer to one such step in the process as an inverse riffle shuffle. See Figure 10.6 for an illustration.

starting deck	mark cards
1	1
2	0
3	1
4	0
5	0

deck after 1 inv. riffle	sorted bits
2	0
4	0
5	0
1	1
3	1

Figure 10.6: Inverse riffle shuffle

For inverse riffle shuffling, the analogous notion to rising sequences for riffle shuffles is the notion of **descents of a permutation**. We say a permutation σ has a descent at k for $1 \leq k \leq n-1$ if

$$\sigma(k) > \sigma(k+1)\,.$$

Note that in the the example in Figure 10.6, the resulting permutation has a descent at $k = 3$. An inverse riffle shuffle results in a permutation that has exactly one descent or is the identity id (which has zero descents). The step distribution $\tilde{\mu}$ for the inverse shuffle is given by

$$\tilde{\mu}(\sigma) = \begin{cases} 1/2^n & \text{if } \sigma \text{ has exactly one descent} \\ (n+1)/2^n & \text{if } \sigma = id \\ 0 & \text{otherwise}\,. \end{cases}$$

We will study top-to-random shuffling and riffle shuffling in more detail in Chapter 11.

10.3 Random walks on finite abelian groups

In order to study the long-term behavior of finite state Markov chains with transition matrix $\mathbf{P}$, we need to understand the matrix powers $\mathbf{P}^n$ which in turn requires understanding the eigenvalues of $\mathbf{P}$ (see Section 11.2). For random walks on finite abelian (commutative) groups, the eigenvalues and eigenvectors of the transition matrix $\mathbf{P}$ are often fairly easy to compute. We have already encountered examples of finite abelian groups: the discrete circle $\mathbb{Z}_n$ (i.e., the cyclic group of order n) with addition modulo n, and the hypercube $\mathbb{Z}_2^k$ (i.e., the direct product of k copies of $\mathbb{Z}_2$) with component-wise addition modulo 2. In fact, any finite abelian group is of similar form:

Definition 10.3.1. *Consider two discrete groups $(G, \circ)$ and $(H, \diamond)$ and a function $f : G \to H$. We say f is a* **group homomorphism** *from G to H if*

$$f(x \circ y) = f(x) \diamond f(y) \quad \text{for all } x, y \in G\,. \tag{10.2}$$

If a function $f : G \to H$ with property (10.2) is also one-to-one, it is called a **group isomorphism** *from G to H, and the groups G and H are called* **isomorphic**.

Isomorphic groups have the same group structure and therefore can be identified as groups.

Theorem 10.3.1 (Fundamental Theorem of Finite Abelian Groups)**.** *Any finite, abelian group G is isomorphic to a direct product of cyclic groups whose order is a power of a prime number, that is,*

$$G \simeq \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k} \tag{10.3}$$

where $n_j = p_j^{k_j}$ with p_j prime (not necessarily distinct), $k_j \in \mathbb{N}$, for all $j = 1, ..., k$.

Under this identification, the group operation on G is component-wise addition modulo n_j in each slot. The order (cardinality) of G is $|G| = n_1 n_2 \cdots n_k$. Note that the symmetric group $\mathrm{S_n}$, i.e., the group of all permutations of n distinguishable objects, which we have encountered in modeling card shuffling, is not abelian.

10.3.1 Characters and eigenvalues

Since any finite abelian group can be identified with a group of the form (10.3), we will write the group operation on G as $+$. We start with the definition of a group character.

Definition 10.3.2. *Let $(G, +)$ be a finite abelian group. Consider the multiplicative group $(\mathbb{U}, \cdot)$ of complex numbers of modulus 1. A* **character** *χ of G is a group homomorphism $\chi : G \to \mathbb{U}$. That is,*

$$\chi(x + y) = \chi(x) \cdot \chi(y) \qquad \text{for all } x, y \in G\,. \tag{10.4}$$

Consider a finite abelian group $(G, +)$ of order $|G| = n$. Let id be its identity element and χ a character of G. It follows from property (10.4) that $\chi(id) = 1$ and $\chi(-x) = \overline{\chi(x)}$ for all $x \in G$. Furthermore, any function value $\chi(x)$ for $x \in G$ must be an nth root of unity. Indeed, it is know that for any $x \in G$, we have $x + x \cdots + x = id$ (n summands; see Exercise 10.3). And so by property (10.4), $(\chi(x))^n = 1$. In particular, the constant

function 1 is a character. We call it the trivial character $\chi_{triv} \equiv 1$. In the following, i will always denote the complex number $\sqrt{-1}$.

Proposition 10.3.2. *Let $G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$. For any $\mathbf{m} = (m_1, ..., m_k)$ with $m_j \in \mathbb{Z}_{n_j}$ (for $j = 1, ..., k$), the function $\chi_{\mathbf{m}}$ on G defined by*

$$\chi_{\mathbf{m}}(x) = \exp\left[2\pi i((x_1 m_1/n_1) + \cdots + (x_k m_k/n_k))\right] \qquad \text{for } \mathbf{x} = (x_1, ..., x_k) \in G$$

is a character of G.

There are $n = n_1 n_2 \cdots n_k = |G|$ distinct k-tuples $\mathbf{m}$ of the kind described in Proposition 10.3.2. Each such k-tuple defines a character for G, and two distinct $\mathbf{m}_1$, $\mathbf{m}_2$ define distinct characters for G. We will show in Proposition 10.3.3 that, in fact, these characters account for *all* characters for G.

For a given abelian group G, denote its set of characters by $\widehat{G}$. The set $\widehat{G}$ forms an abelian group with respect to pointwise multiplication:

(a) If $\chi_1, \chi_2 \in \widehat{G}$, then $\chi_1\chi_2 \in \widehat{G}$ and $\chi_1\chi_2 = \chi_2\chi_1$.
(b) The trivial character is $\chi_{triv} \equiv 1$ the identity element for pointwise multiplication.
(c) If $\chi \in \widehat{G}$, then $\bar{\chi}$ (the complex conjugate of χ) is also a character and $\chi\bar{\chi} = 1$. We have $\bar{\chi}(x) = \chi(-x)$.

Verification of properties (a)-(c) is straightforward. Moreover, the groups G and $\widehat{G}$ are isomorphic under the bijection $f : G \to \widehat{G}$ with $f(\mathbf{m}) = \chi_{\mathbf{m}}$.

Proposition 10.3.3. *Let $G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_l}$. Consider the vector space V_G of complex valued functions $f : G \to \mathbb{C}$, and on this vector space V_G the inner product $\langle \cdot, \cdot \rangle_G$ defined by $\langle f, g \rangle_G = \frac{1}{|G|} \sum_{x \in G} f(x)\bar{g}(x)$ for $f, g \in V_G$. We then have the following:*

(a) For any non-trivial character χ we have $\sum_{x \in G} \chi(x) = 0$.
(b) The set of characters $\widehat{G}$ forms an orthonormal system with respect to $\langle \cdot, \cdot \rangle_G$.
(c) $\widehat{G}$ forms a basis for the vector space V_G.

Proof. (a) Let $\chi \neq \chi_{triv}$ and take $x_0 \in G$. Then

$$\sum_{x \in G} \chi(x) = \sum_{x \in G} \chi(x + x_0) = \chi(x_0) \sum_{x \in G} \chi(x)\,. \tag{10.5}$$

Since χ is non-trivial, we can find $x_0 \in G$ such that $\chi(x_0) \neq 1$. Therefore it follows from (10.5) that

$$\sum_{x\in G} \chi(x) = 0\,.$$

(b) Let $n = |G|$. By part (a),

$$\langle \chi_{triv}, \chi\rangle_G = \frac{1}{n}\sum_{x\in G}\chi(x) = 0\,.$$

We also have $\langle \chi_{triv}, \chi_{triv}\rangle_G = \frac{1}{n}\sum_{x\in G} 1 = 1$. For χ_1 and χ_2 non-trivial characters, we have

$$\langle \chi_1, \chi_2\rangle_G = \frac{1}{n}\sum_{x\in G}\chi_1(x)\bar{\chi}_2(x) = \frac{1}{n}\sum_{x\in G}\chi(x) = 0\,,$$

where we have taken $\chi = \chi_1\chi_2$ (which is also a character since $\widehat{G}$ is a group). Lastly, for any non-trivial character χ we have

$$\langle \chi, \chi\rangle_G = \frac{1}{n}\sum_{x\in G}\chi(x)\bar{\chi}(x) = \frac{1}{n}\sum_{x\in G}\chi(x)\chi(-x) = \frac{1}{n}\sum_{x\in G}\chi(id) = 1\,.$$

(c) Since, by part (b), the set of characters forms and orthonormal system in V_G, the characters are linearly independent. The vector space V_G has dimension $n = |G|$. By Proposition 10.3.2, any finite abelian group has $n = |G|$ distinct characters. Hence the characters as described in Proposition 10.3.2 are all characters for G, and they form an orthonormal basis of V_G. □

Consider a random walk on a finite abelian group G with transition matrix $\mathbf{P}$. We next show that the characters of G are eigenvectors of $\mathbf{P}$. This fact will allow us to compute the eigenvalues of $\mathbf{P}$.

Proposition 10.3.4. *Let G be a finite abelian group and μ a probability distribution on G. Consider random walk on G with step distribution μ. Let χ be a character of G. Then χ is a right eigenvector for the transition matrix $\mathbf{P}$ with corresponding eigenvalue $\lambda = \mathbb{E}_\mu(\chi) = \sum_{z\in G}\mu(z)\chi(z)$.*

Proof. We have

$$\begin{aligned}(\mathbf{P}\chi)(y) &= \sum_{x\in G} P_{yx}\chi(x) = \sum_{x\in G}\mu(x-y)\chi(x)\\ &= \sum_{z\in G}\mu(z)\chi(z+y) = \chi(y)\sum_{z\in G}\mu(z)\chi(z) = \lambda\,\chi(y)\,.\end{aligned}$$

□

Proposition 10.3.4 gives an easy way for computing all eigenvalues of $\mathbf{P}$. Note that if μ is uniform distribution on G, the corresponding transition matrix $\mathbf{P}$ has eigenvalue 1 (with multiplicity 1), and the rest of the eigenvalues are 0. We can of course see this directly, since $\mathbf{P}$ is the matrix all of whose entries are equal to $\frac{1}{|G|}$. But it also follows from Proposition 10.3.3(a) and Proposition 10.3.4.

Definition 10.3.3. *Let G be a finite abelian group and μ a probability distribution on G. We call the map $\widehat{\mu} : \widehat{G} \to \mathbb{C}$ defined by*

$$\widehat{\mu}(\chi) = \mathbb{E}_\mu(\chi) = \sum_{z \in G} \mu(z)\chi(z)$$

the **Fourier transform** *of the probability distribution μ.*

Thus the image $\mathrm{Im}(\widehat{\mu})$ of the Fourier transform $\widehat{\mu}$ is the set of eigenvalues of the transition matrix $\mathbf{P}$ for a random walk with step distribution μ. Since $\mathbf{P}^k$ is the transition matrix for a random walk with step distribution μ^{*k}, we get the following immediate relationship between the Fourier transform of μ and the Fourier transform of its kth convolution power μ^{*k}:

Proposition 10.3.5. *Let μ be a probability distribution on a finite abelian group G and μ^{*k} its kth convolution power. Then*

$$\widehat{\mu^{*k}}(\chi) = \mathbb{E}_{\mu^{*k}}(\chi) = (\mathbb{E}_\mu(\chi))^k$$

for all $\chi \in \widehat{G}$.

In the following, we compute the eigenvalues for two of our running examples of random walks on abelian groups, namely simple random walk on $\mathbb{Z}_n$ and simple random walk on the hypercube $\mathbb{Z}_2^k$.

Example 10.3.1. Let G be the discrete circle $\mathbb{Z}_n$, i.e., the numbers $\{0, 1, ..., n-1\}$ together with the group operation $+ \pmod n$. We consider simple, symmetric random walk on $\mathbb{Z}_n$. The characters of $\mathbb{Z}_n$ can be labeled by k for $0 \le k \le n-1$ and are defined by

$$\chi_k(x) = e^{2\pi i x k/n} \qquad for\ x \in \mathbb{Z}_n\,.$$

Since the step distribution μ is uniform on $\{1, -1\}$, the corresponding eigenvalues are

$$\lambda_k = \mathbb{E}_\mu(\chi_k) = \frac{1}{2}e^{2\pi i k/n} + \frac{1}{2}e^{-2\pi i k/n} = \cos(2\pi k/n)$$

for $k = 0, 1, ..., n-1$. All eigenvalues are real, hence for each k, the real part $Re(\chi_k)$ of the eigenvector χ_k is also an eigenvector corresponding to λ_k. That is, for each k, the function $f_k(x) = \cos(2\pi kx/n)$ for $x \in \mathbb{Z}_n$ is a right eigenvector of the transition matrix $\mathbf{P}$ corresponding to eigenvalue $\lambda_k = \cos(2\pi k/n)$.

If instead we consider lazy simple random walk with $\tilde{\mu}(0) = \frac{1}{2}$ and $\tilde{\mu}(1) = \tilde{\mu}(-1) = \frac{1}{4}$, the eigenvectors remain the same, and we get the corresponding eigenvalues

$$\tilde{\lambda}_k = \frac{1}{2} + \frac{1}{4}e^{2\pi ik/n} + \frac{1}{4}e^{-2\pi ik/n} = \frac{1}{2} + \frac{1}{2}\cos(2\pi k/n)$$

for $k = 0, 1, ..., n-1$. Again, all eigenvalues of $\mathbf{P}$ are real since the random walk is reversible (has symmetric step distribution). □

Example 10.3.2. Consider $G = \mathbb{Z}_2^k$, the hypercube of dimension k. Note that $|\mathbb{Z}_2^k| = 2^k$. For simple random walk with holding probability $\frac{1}{2}$, the step distribution μ is defined by $\mu(0, ..., 0) = \frac{1}{2}$ and

$$\mu(1, 0, ..., 0) = \mu(0, 1, 0, ..., 0) = \cdots = \mu(0, ..., 0, 1) = \frac{1}{2k}\,.$$

Here the characters can be labeled by the binary k-tuples $\mathbf{k} = (m_1, ..., m_k)$ with $m_j \in \{0, 1\}$. So the character labeled by $\mathbf{k}$ is defined by

$$\chi_{\mathbf{k}}(\mathbf{x}) = e^{2\pi i(x_1m_1 + \cdots + x_km_k)/2} \qquad \text{for } \mathbf{x} = (x_1, ..., x_k) \in \mathbb{Z}_2^k\,.$$

For any non-trivial character $\chi_{\mathbf{k}}$, we can identify the label $\mathbf{k}$ with a subset $J \subseteq \{1, 2, ..., k\}$ (that is, the set J of indices j for which $m_j = 1$). Thus, equivalently, we can write for the character

$$\chi_J(\mathbf{x}) = \prod_{j \in J}(-1)^{x_j} \qquad \text{for } \mathbf{x} = (x_1, ..., x_k) \in \mathbb{Z}_2^k\,.$$

The corresponding eigenvalues are

$$\lambda_J = \frac{1}{2} + \frac{1}{2k}\left((k - |J|) - |J|\right) = \frac{k - |J|}{k}\,.$$

Since there are $\binom{k}{|J|}$ distinct subsets of size $|J|$ of $\{1, ..., k\}$, the multiplicity of eigenvalue $\frac{k-|J|}{k}$ is $\binom{k}{|J|}$. To summarize, the eigenvalues of lazy random walk on the hypercube $\mathbb{Z}_2^k$ are

$$\frac{k-j}{k} \text{ with multiplicity } \binom{k}{j} \qquad \text{for } j = 0, 1, ..., k\,.$$

Here again, all eigenvalues are real since the random walk is reversible. □

Exercises

Exercise 10.1. Consider two probability measures μ and ν on a discrete group G. Show that the convolution product $\mu \star \nu$ is also a probability measure.

Exercise 10.2. Show that the transition matrix $\mathbf{P}$ of a finite-state Markov chain is doubly stochastic if and only if uniform measure is a stationary distribution for the chain. If $\mathbf{P}$ is doubly stochastic, does it follow that each n-step transition matrix $\mathbf{P}^n$ is also doubly stochastic?

Exercise 10.3. Consider a finite abelian group $(G, +)$ with $|G| = n$. Show that for all $x \in G$,

$$x + x + \cdots + x = id \quad (n \text{ summands}) .$$

(*Hint*: Consider a function $f : G \to G$ defined by $f(g_j) = x + g_j$.)

Exercise 10.4. Let $(X_n)_{n\geq 0}$ be random walk on a discrete group G with step distribution μ for which $\mu(id) = 0$. Consider a lazy version $(\tilde{X}_n)_{n\geq 0}$ of this random walk where, at each step, an independent and biased coin toss determines whether or not the process holds in its current state or moves to a next state, according to the transition probabilities for $(X_n)_{n\geq 0}$. We assume the coin tosses form an i.i.d. sequence of Bernoulli(p) random variables. Find a formula for the n-step transition probabilities for $(\tilde{X}_n)_{n\geq 0}$ in terms of the k-step convolution powers of μ.

Exercise 10.5. The following is a generalization of a certain "translation invariance" feature that random walks on groups possess: Let $(X_n)_{n\geq 0}$ be a Markov chain on state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. We say the Markov chain is **transitive** if for any two states $r, s \in \mathcal{S}$ there exists a bijection $f = f_{(r,s)} : \mathcal{S} \to \mathcal{S}$ with $f(r) = s$ that preserves the transition probabilities, that is, for which $P_{u,v} = P_{f(u),f(v)}$ for all $u, v \in \mathcal{S}$. Intuitively, if a Markov chain is transitive, then, for any two states in the state space, there is always a suitable re-labeling of the states, so that the chain "looks probabilistically the same" if we use either one of the two states as the starting state for the chain.

(a) Show that a random walk on a group is a transitive Markov chain.

(b) Show that for any transitive Markov chain with finite state space $\mathcal{S}$, the uniform distribution on $\mathcal{S}$ is a stationary distribution.

Exercise 10.6. Consider a transitive Markov chain $(X_n)_{n\geq 0}$ on a finite state space $\mathcal{S}$ and let π be the stationary distribution. Assume the chain starts in state $x_0 \in \mathcal{S}$. Let $X_n \sim \mu_n$ for $n \geq 0$. Show that the L_1-distance to stationarity $\|\mu_n - \pi\|_{L_1}$ does not depend on the starting state x_0.

Exercise 10.7. Consider a probability measure μ on a discrete group G. Show that if $\text{supp}(\mu)$ is a symmetric set of generators for G, then the associated random walk on G is either aperiodic or has period 2. Give an example of μ and G for which $\mu(id) = 0$ and for which the random walk is aperiodic.

Exercise 10.8. For each of the following shuffling mechanisms of a deck of n cards, i.e. random walks on $\mathrm{S_n}$, describe the probability measure μ on $\mathrm{S_n}$ that is the step distribution for the random walk.

(a) *Random transpositions.* At each time step, choose uniformly at random two cards (with replacement) from the deck. Then interchange the location of the two chosen cards in the deck.

(b) *Random-to-random insertions.* At each time step, choose uniformly at random a card from the deck, remove it from the deck, and reinsert it into the deck in a uniformly at random chosen location.

Exercise 10.9. Consider a card shuffling process of n cards, i.e. an irreducible random walk $(X_n)_{n\geq 0}$ on the symmetric group $\mathrm{S_n}$ for a given step distribution μ on $\mathrm{S_n}$, as described in Section 10.2. Recall that for $\sigma \in \mathrm{S_n}$, for right random walk, $\sigma(i) = k$ means "Card k is in position i in the deck". Now fix a k, and consider the function $f : \mathrm{S_n} \to \{1, ..., \mathrm{n}\}$ defined by $f(\sigma) = \sigma^{-1}(k)$. The function f gives the position of Card k in the deck.

(a) Show that $(X_n)_{n\geq 0}$ is lumpable with respect of f, that is, the process $(f(X_n))_{n\geq 0}$ is a Markov chain on state space $\{1, ..., n\}$.

(b) Show that uniform distribution on $\{1, ..., n\}$ is stationary for the lumped Markov chain from part (a).

Exercise 10.10. Consider top-to-random shuffling for $n = 5$ cards. Explicitly give the transition matrix for the Markov chain from Exercise 10.9(a) which tracks the location of Card 1 in the deck.

Exercise 10.11. Consider a finite abelian group G and a probability measure μ on G. Let $\pi \sim \text{Unif}(G)$ be uniform probability measure on G. Prove that if there exists $k_0 \geq 1$ such that $\mu^{\star k_0} \sim \pi$, then $\mu \sim \pi$.

Exercise 10.12. Let $(G, \circ)$ and $(H, \diamond)$ be two finite abelian groups. Their direct product $(G \times H, \star)$ is also an abelian group with group operation $\star$ being defined component-wise. That is, for all $(g_1, h_1), (g_2, h_2) \in G \times H$, we define

$$(g_1, h_1) \star (g_2, h_2) = (g_1 \circ g_2, h_1 \diamond h_2)\,.$$

(a) Show that if χ_G is a character of G and χ_H is a character of H, then (χ_G, χ_H) defined by

$$(\chi_G, \chi_H)(g, h) = \chi_G(g)\chi_H(h) \quad \text{for all } g \in G,\ h \in H \tag{10.6}$$

is a character of $G \times H$.

(b) Recall that $\widehat{G}$ denotes the set of characters of G. Show that (10.6) defines a bijection between $\widehat{G} \times \widehat{H}$ and $\widehat{G \times H}$.

Chapter 11

Rates of Convergence

11.1 Basic set-up

Quantitative results about the *speed of convergence* of a given Markov chain are of major practical interest in applications of Markov chain Monte Carlo methods. For how long do we need to run the Markov chain so that samples drawn are "sufficiently good approximations" to samples from its stationary distribution π? Precise statements about the related *mixing time* of the Markov chain will be based on an appropriate choice of distance between probability measures. The following is a commonly used notion of distance between probability measures in this context:

Definition 11.1.1. *Let $\mathcal{S}$ be a discrete set and π and μ two probability distributions on $\mathcal{S}$. The* **total variation distance** *between π and μ is defined by*

$$\|\mu - \pi\|_{TV} = \sup_{E \subseteq \mathcal{S}} |\mu(E) - \pi(E)| . \tag{11.1}$$

It gives the maximum error made when we use μ to approximate π.

The definition implies $0 \leq \|\mu - \pi\|_{TV} \leq 1$. The following proposition shows that total variation distance is equal to one-half of the L_1-distance of the two distributions. It is sometimes easier to work with (11.2) rather than with the probabilistic definition (11.1) of total variation distance.

Lemma 11.1.1. *Let $\mathcal{S}$ be a discrete set and μ and π two probability distributions on $\mathcal{S}$. Then*

$$\|\mu - \pi\|_{TV} = \frac{1}{2} \sum_{x \in \mathcal{S}} |\mu(x) - \pi(x)| . \tag{11.2}$$

Proof. Consider the set $A = \{x \in \mathcal{S} : \mu(x) \geq \pi(x)\}$. See Figure **??** below. We have

$$\sup_{E \subseteq \mathcal{S}} (\mu(E) - \pi(E)) = \mu(A) - \pi(A)\,.$$

We then consider the set $A^c = \{x \in \mathcal{S} : \mu(x) < \pi(x)\}$ and reverse the roles of μ and π. This yields

$$\sup_{E \subseteq \mathcal{S}} (\pi(E) - \mu(E)) = \pi(A^c) - \mu(A^c)\,.$$

Note that $\mu(A) - \pi(A) = \pi(A^c) - \mu(A^c)$. We get

$$\begin{aligned} \|\mu - \pi\|_{TV} &= \sup_{E \subseteq \mathcal{S}} |\pi(E) - \mu(E)| = \mu(A) - \pi(A) \\ &= \frac{1}{2}(\mu(A) - \pi(A) + \pi(A^c) - \mu(A^c) = \frac{1}{2}\sum_{x \in \mathcal{S}} |\mu(x) - \pi(x)|\,. \end{aligned}$$

□

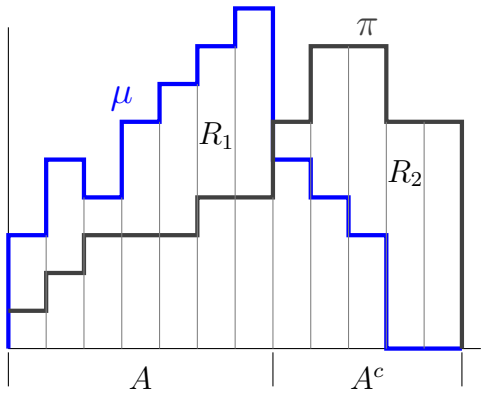


Figure 11.1: The areas of regions R_1 and R_2 are the same.

In the above Figure **??**, $\text{Area}(R_1) = \mu(A) - \pi(A)$ and $\text{Area}(R_2) = \pi(A^c) - \mu(A^c)$. The two areas are equal. We have

$$\|\mu - \pi\|_{TV} = \text{Area}(R_1) = \text{Area}(R_2)\,.$$

Note that $\|\mu - \pi\|_{TV} = 0$ if and only if $\mu(x) = \pi(x)$ for all $x \in \mathcal{S}$, and furthermore $\|\mu - \pi\|_{TV} = 1$ if and only if μ and π are supported on disjoint subsets of $\mathcal{S}$. For a sequence μ_n, $n \geq 0$, of probability measures on $\mathcal{S}$ we have

$$\lim_{n \to \infty} \|\mu_n - \pi\|_{TV} = 0 \quad \Longleftrightarrow \quad \lim_{n \to \infty} \mu_n(x) = \pi(x) \;\; \forall x \in \mathcal{S}\,. \tag{11.3}$$

If $\mathcal{S}$ is finite, statement (11.3) is immediate. The proof of (11.3) for countably infinite $\mathcal{S}$ is the content of Exercise 11.5.

Often times in applications, for a given Markov chain $(X_n)_{n\geq 0}$ with $X_n \sim \mu_n$, **quantitative results** about the **rate of convergence** of $\|\mu_n - \pi\|_{TV}$ to 0 are needed. In this chapter we will discuss methods for studying such rates of convergence. We first show that as the Markov chain progresses in time, its total variation distance to stationarity can only decrease with time.

Lemma 11.1.2. *Let $(X_n)_{n\geq 0}$ be a Markov chain and π a stationary distribution for the chain. Let $X_n \sim \mu_n$ for $n \geq 0$. Then*

$$\|\mu_{n+1} - \pi\|_{TV} \leq \|\mu_n - \pi\|_{TV} \qquad \text{for all } n \geq 0\,.$$

Proof. Using $\mu_{n+1}(x) = \sum_{y\in\mathcal{S}} \mu_n(y)\, P_{yx}$, we have

$$\begin{aligned}
\|\mu_{n+1} - \pi\|_{TV} &= \frac{1}{2}\sum_{x\in\mathcal{S}} \left| \sum_{y\in\mathcal{S}} \mu_n(y)\, P_{yx} - \sum_{y\in\mathcal{S}} \pi(y)\, P_{yx} \right| \\
&= \frac{1}{2}\sum_{x\in\mathcal{S}} \left| \sum_{y\in\mathcal{S}} (\mu_n(y) - \pi(y))\, P_{yx} \right| \\
&\leq \frac{1}{2}\sum_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} |\mu_n(y) - \pi(y)|\, P_{yx} \\
&= \frac{1}{2}\sum_{y\in\mathcal{S}} |\mu_n(y) - \pi(y)| \sum_{x\in\mathcal{S}} P_{yx} \\
&= \frac{1}{2}\sum_{y\in\mathcal{S}} |\mu_n(y) - \pi(y)| \\
&= \|\mu_n - \pi\|_{TV}
\end{aligned}$$

where the rearrangement of the sum in line 4 is justified because of absolute convergence. □

Example 11.1.1. Recall the 2-state chain on state space $\mathcal{S} = \{0, 1\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

for fixed $a, b \in (0, 1)$. Let us assume the chain starts in state 0. Thus $\mu_n = (P^n_{00}, P^n_{01})$. Recall Example 1.3.2 where we have computed the n-step transition probabilities for this

chain. We have

$$\mu_n = \left(\frac{b}{a+b} + \frac{a}{a+b}(1-a-b)^n, \quad \frac{a}{a+b} - \frac{a}{a+b}(1-a-b)^n \right) .$$

The unique stationary (limiting) distribution is

$$\pi = \left(\frac{b}{a+b}, \quad \frac{a}{a+b} \right) .$$

Thus we get the following explicit formula for total variation distance to stationarity as a function of time n:

$$\begin{aligned} \|\mu_n - \pi\|_{TV} &= \tfrac{1}{2} \left(|P_{00}^n - \pi(0)| + |P_{01}^n - \pi(1)| \right) \\ &= \frac{a}{a+b} |1-a-b|^n . \end{aligned} \tag{11.4}$$

Since we assumed $a, b \in (0,1)$, we have $|1-a-b| < 1$. Thus we see from (11.4) that $\|\mu_n - \pi\|_{TV}$ decays exponentially fast. Note that $(1-a-b)$ is the second largest eigenvalue of the transition matrix $\mathbf{P}$. In the next section we will see that the eigenvalues of $\mathbf{P}$, in particular the (in modulus) second largest eigenvalue of $\mathbf{P}$, play an important role in the rate of convergence to stationarity. □

In studying rates of convergence for Markov chain, one of the obvious questions of interest is: *How many steps are needed (or suffice) for the particular chain to be "ϵ-close" to stationarity*? This question is about the *mixing time* of a Markov chain which we will define next.

Definition 11.1.2. *Consider an irreducible, positive recurrent, and aperiodic Markov chain $(X_n)_{n \geq 0}$ with stationary distribution π. Let $X_n \sim \mu_n$ for $n \geq 0$. For a given $\epsilon > 0$, we define the* **mixing time** *t_ϵ^{mix} by*

$$t_\epsilon^{\text{mix}} = \min\{n : \|\mu_n - \pi\|_{TV} \leq \epsilon\} .$$

Note that the notion of *mixing time* makes sense in light of the monotonicity property of total variation distance (Lemma 11.1.2).

11.2 Spectral bounds

The distribution of a Markov chain at time n is $\mu_n = \mu_0 \mathbf{P}^n$. Clearly, when studying convergence rates of μ_n, we need to understand the convergence behavior of the matrix powers $\mathbf{P}^n$. For finite state space, this behavior will be determined by the eigenvalues

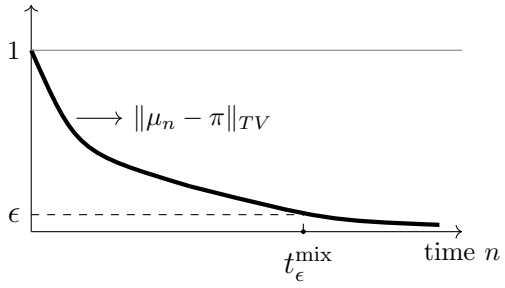


Figure 11.2: Mixing time

(the spectrum) of the finite transition matrix $\mathbf{P}$. Recall that for the special case of random walks on a finite abelian group, we have computed the spectrum of the transition matrix in Section 10.3. The Perron–Frobenius theorem (see Appendix A.6) provides useful information for the general case. The following proposition summarizes some important properties of the eigenvalues of a stochastic matrix $\mathbf{P}$.

Proposition 11.2.1. *Let $\mathbf{P}$ be a stochastic $(n \times n)$-matrix. We have the following:*

(a) $\lambda_1 = 1$ is an eigenvalue of $\mathbf{P}$, and there exists a nonnegative left eigenvector $\mathbf{v}$ corresponding to $\lambda_1 = 1$. If λ is an eigenvalue (possibly a complex eigenvalue) of $\mathbf{P}$, then $|\lambda| \leq 1$.

(b) If $\mathbf{P}$ is irreducible and aperiodic (in particular, if $\mathbf{P}$ is strictly positive), then eigenvalue $\lambda_1 = 1$ has algebraic and geometric multiplicity 1, and there exists a strictly positive left eigenvector $\mathbf{v}$ corresponding to eigenvalue 1. For all other eigenvalues $\lambda \neq 1$, we have $|\lambda| < 1$.

(c) If $\mathbf{P}$ is irreducible and periodic with period $c > 1$, then $\mathbf{P}$ has exactly c eigenvalues $\lambda_1, \lambda_2, ..., \lambda_c$ of modulus 1. The λ_i are the cth roots of unity. Each λ_i has algebraic and geometric multiplicity one.

Proof. (a) Consider the constant column n-vector $\mathbf{w}^{\mathrm{t}} = (1, 1, ..., 1)^{\mathrm{t}}$. Since $\mathbf{P}$ is stochastic, we have $\mathbf{P}\mathbf{w}^{\mathrm{t}} = \mathbf{w}^{\mathrm{t}}$. So 1 is a right and therefore also left eigenvalue of $\mathbf{P}$. Let $\mathbf{v} = (v_1, ..., v_n)$ be a left eigenvector corresponding to eigenvalue 1. We will show that $|\mathbf{v}| = (|v_1|, ..., |v_n|)$ is also a left eigenvector corresponding to eigenvalue 1, so $|\mathbf{v}|\mathbf{P} = |\mathbf{v}|$. Note that by the triangle inequality,

$$|v_j| \leq \sum_{i=1}^{n} |v_i| P_{ij}$$

for all $j = 1, ..., n$. If $|\mathbf{v}|\mathbf{P} \neq |\mathbf{v}|$, then there exists at least one index j_0 such that $|v_{j_0}| < \sum_{i=1}^{n} |v_i| P_{ij_0}$. Thus

$$\sum_{j=1}^{n} |v_j| < \sum_{j=1}^{n} \sum_{i=1}^{n} |v_i| P_{ij} = \sum_{i=1}^{n} |v_i|$$

which is a contradiction. It follows that the nonnegative vector $|\mathbf{v}|$ is in fact a left eigenvector corresponding to eigenvalue 1.

Let λ be an eigenvalue of $\mathbf{P}$ and the column vector $\mathbf{s}^{\mathrm{t}} = (s_1, ..., s_n)^{\mathrm{t}}$ a corresponding right eigenvector, so $\mathbf{P}\mathbf{s}^{\mathrm{t}} = \lambda \mathbf{s}^{\mathrm{t}}$. Let m be an index such that $|s_m| = \max_{1 \leq i \leq n} |s_i|$. Then

$$\begin{aligned} |\lambda|\,|s_m| &= |\lambda s_m| = |(\mathbf{P}\mathbf{s}^{\mathrm{t}})_m| \\ &= \left| \sum_{i=1}^{n} P_{mi} s_i \right| \leq \sum_{i=1}^{n} P_{mi} |s_i| \qquad (11.5) \\ &\leq |s_m| \sum_{i=1}^{n} P_{mi} = |s_m| \qquad (11.6) \end{aligned}$$

from which it follows that $|\lambda| \leq 1$.

(b) We first prove the statement for a strictly positive stochastic matrix $\mathbf{P}$. Assume $P_{ij} > 0$ for all $i, j \in \mathcal{S}$. Let $\mathbf{v}$ be a left eigenvector corresponding to eigenvalue 1. As proved in part (a), the vector $|\mathbf{v}|$ is also an eigenvector for eigenvalue 1, and hence so is the vector $\mathbf{u} = \frac{1}{2}(\mathbf{v} + |\mathbf{v}|)$. Note that by construction, $\mathbf{u}$ has nonnegative entries. Since $\mathbf{P}$ is a strictly positive matrix, it follows from $\mathbf{u}\mathbf{P} = \mathbf{u}$ that either $\mathbf{u} = \mathbf{0}$ (the zero vector) or $\mathbf{u}$ is a strictly positive vector. This implies that either all entries of $\mathbf{v}$ are strictly negative (in which case $-\mathbf{v}$ is a strictly positive eigenvector) or all entries of $\mathbf{v}$ are strictly positive. In order to show that eigenvalue 1 has geometric multiplicity 1, consider two left eigenvectors $\mathbf{v}'$ and $\mathbf{v}''$ for eigenvalue 1. By the above, we can assume that each of them is a strictly positive vector (if not, take its negative). Furthermore, we will assume that $\mathbf{v}'$ and $\mathbf{v}''$ have been normalized, so for both vectors the entries sum to one. Then the vector $\mathbf{u} = \mathbf{v}' - \mathbf{v}''$ is also a left eigenvector corresponding to eigenvalue 1, and its entries sum to 0. This means that either $\mathbf{u}$ has entries of mixed sign (which we have shown to be impossible), or all entries of $\mathbf{u}$ are 0. We must have the latter, and so we have shown that $\mathbf{v}' = \mathbf{v}''$. This proves that the eigenspace corresponding to eigenvalue 1 has dimension 1, and so eigenvalue 1 has geometric multiplicity 1.

We will now show that eigenvalue $\lambda_1 = 1$ has algebraic multiplicity 1. Assume its algebraic multiplicity is greater than one. Then, for some suitable basis of $\mathbb{C}^n$ (a Jordan basis), the matrix $\mathbf{P}$ contains a $(k \times k)$ Jordan block (with $k \geq 2$) for eigenvalue 1. And there exists a (possibly complex) column vector $\mathbf{u}^{\mathrm{t}}$ such that $\mathbf{P}\mathbf{u}^{\mathrm{t}} = \mathbf{u}^{\mathrm{t}} + \mathbf{w}^{\mathrm{t}}$ where $\mathbf{w}^{\mathrm{t}} = (1, 1, ..., 1)^{\mathrm{t}}$.

In the following, $Re\,(x)$ denotes the real part of a complex number x. Let m be an index such that $Re\,(u_m) = \max\limits_{1\le i\le n} Re\,(u_i)$. Then

$$\begin{aligned} Re\,(u_m) + 1 = Re\,(\mathbf{P}\mathbf{u}^{\mathrm{t}})_m &= Re\left(\sum_{i=1}^{n} P_{mi}u_i\right) \\ &\le Re\left(\sum_{i=1}^{n} P_{mi}u_m\right) = Re\,(u_m)\,. \end{aligned}$$

Since this is a contradiction, any Jordan block for eigenvalue 1 must have dimension 1. Since we have also shown that there exists exactly one Jordan block for eigenvalue 1, it follows that eigenvalue 1 has algebraic multiplicity 1.

By Corollary 2.4.5, the transition matrix $\mathbf{P}$ of an irreducible and aperiodic Markov chain is regular. So there exists $N > 0$ such that $\mathbf{P}^n$ is strictly positive for $n \ge N$. We apply the above results to $\mathbf{P}^N$. Also note that if λ is an eigenvalue of $\mathbf{P}$, then λ^N is an eigenvalue of $\mathbf{P}^N$, and that the corresponding eigenvectors are the same. It follows that $\lambda_1 = 1$ is an eigenvalue of $\mathbf{P}$. Since taking powers of a matrix preserves the algebraic multiplicities of the corresponding eigenvalues of its matrix powers, the algebraic multiplicity of $\lambda_1 = 1$ for $\mathbf{P}$ must be 1, and hence the geometric multiplicity of $\lambda_1 = 1$ for $\mathbf{P}$ is also 1.

(c) This is a direct consequence Part II of the Perron–Frobenius theorem (see Theorem A.6.2(e)). We omit the proof for this part. □

Note that, although we have given direct proofs for parts (a) and (b) of Proposition 11.2.1, these parts would also directly follow from Part I of the Perron–Frobenius theorem (Theorem A.6.1).

From Proposition 11.2.1 we can derive an alternate proof for convergence to stationarity for a finite-state, ergodic Markov chain: Let $\mathbf{P}$ be the transition matrix for an irreducible, aperiodic, finite-state Markov chain. Then there exists a (possibly complex) invertible matrix $\mathbf{S}$ (whose column vectors form a Jordan basis of $\mathbb{C}^n$ with respect to $\mathbf{P}$, and whose first column vector we choose to be $\mathbf{w}^{\mathrm{t}} = (1, 1, ..., 1)^{\mathrm{t}}$) such that

$$\mathbf{J} = \mathbf{S}^{-1}\mathbf{P}\mathbf{S}$$

where $\mathbf{J}$ is the block-diagonal matrix

$$\mathbf{J} = \mathrm{diag}(1, \mathbf{J}^{(\lambda_2,n_2)}, ..., \mathbf{J}^{(\lambda_k,n_k)})\,.$$

The $(n_i \times n_i)$ square matrices $\mathbf{J}^{(\lambda_i, n_i)}$ along the diagonal of $\mathbf{J}$ are called the Jordan blocks. For each index $i = 2, ..., k$, its corresponding Jordan block is of the form

$$\mathbf{J}^{(\lambda_i, n_i)} = \begin{pmatrix} \lambda_i & 1 & & \\ 0 & \lambda_i & \ddots & \\ & \ddots & \ddots & 1 \\ & & 0 & \lambda_i \end{pmatrix}$$

where λ_i is an eigenvalue of $\mathbf{P}$, $|\lambda_i| < 1$, and all matrix entries that are not marked are understood to be 0. We have $1 + n_2 + \cdots + n_k = n$.
Since for $|\lambda_i| < 1$ we have

$$\left[\mathbf{J}^{\lambda_i, n_i}\right]^m \xrightarrow{m \to \infty} \mathbf{0}_{(n_i \times n_i)} \quad \text{(the } (n_i \times n_i) \text{ zero matrix)},$$

it follows that

$$\mathbf{P}^m = \mathbf{S}\mathbf{J}^m\mathbf{S}^{-1} \xrightarrow{m \to \infty} \mathbf{S} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \mathbf{S}^{-1},$$

and so the limit $\lim_{m \to \infty} \mathbf{P}^m$ exists. Moreover, note that

$$\mathbf{S} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}.$$

It follows that the matrix limit $\mathbf{L} = \mathbf{S}[\text{diag}(1, 0, ..., 0)]\mathbf{S}^{-1}$ is a stochastic matrix all of whose rows are the same (they are equal to the first row of $\mathbf{S}^{-1}$). Let π denote this constant row vector. Altogether, we have shown that

$$\lim_{m \to \infty} \mathbf{P}^m = \begin{pmatrix} \leftarrow & \pi & \rightarrow \\ & \vdots & \\ \leftarrow & \pi & \rightarrow \end{pmatrix}.$$

The distribution π is the limiting distribution, hence the unique stationary distribution for the Markov chain that has transition matrix $\mathbf{P}$. This reproves the convergence theorem for finite-state ergodic Markov chains (which we have proved in a more general setting in Section 3.2). Alternatively, the result follows from the statement of Theorem A.6.1 part (e) (Perron–Frobenius Theorem, Part I).

In this chapter we present several approaches that allow us to obtain quantitative results for the rate of convergence to stationarity.

In the following subsection, we will show that for a finite-state ergodic Markov chain, there is always an exponential rate of convergence to stationarity, and we will find a way to estimate this rate of convergence in terms of the eigenvalues of the transition matrix $\mathbf{P}$. Towards this end, we need to understand the *spectral representation* of $\mathbf{P}$. In order to avoid too many technicalities, we will focus our discussion on ergodic Markov chains that are also reversible and therefore have a diagonalizable transition matrix $\mathbf{P}$ (over $\mathbb{R}$).

11.2.1 Spectral decomposition of the transition matrix

Throughout this section, we will assume that $\mathbf{P}$ is the transition matrix for an **irreducible, aperiodic, and reversible** Markov chain on state space $\mathcal{S} = \{1, ..., n\}$.

Let π be the stationary distribution for this Markov chain. Since the Markov chain is irreducible, π is strictly positive on $\mathcal{S}$. Consider the diagonal matrix $\mathbf{D} = \text{diag}(\pi(1), ..., \pi(n))$ and its square root

$$\mathbf{D}^{\frac{1}{2}} = \text{diag}(\sqrt{\pi(1)}, ..., \sqrt{\pi(n)}) \,.$$

Let $\mathbf{P}^* = \mathbf{D}^{\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}}$. Then the matrix entries P^*_{ij} of $\mathbf{P}^*$ are

$$P^*_{ij} = \frac{\sqrt{\pi(i)}}{\sqrt{\pi(j)}} P_{ij} \,.$$

Since we assume $\mathbf{P}$ is reversible with respect to π, we have

$$P^*_{ij} = P^*_{ji} \,,$$

and so $\mathbf{P}^*$ is a symmetric matrix. By the Spectral Theorem (see Theorem 7.2.5), $\mathbf{P}^*$ is orthogonally diagonalizable. Let $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$ be an orthonormal basis of $\mathbb{R}^n$ (with respect to the standard Euclidian inner product) of right eigenvectors of $\mathbf{P}^*$. We assume $\mathbf{P}^*\mathbf{s}_1 = \mathbf{s}_1$ and $\mathbf{P}^*\mathbf{s}_k = \lambda_k \mathbf{s}_k$ for $2 \leq k \leq n$. Since $\mathbf{P}$ and $\mathbf{P}^*$ are similar matrices, they have the same set of eigenvalues $\{1, \lambda_2, ..., \lambda_n\} \subset (-1, 1]$. Recall that, since the Markov chain is irreducible and aperiodic, the multiplicity of eigenvalue $\lambda_1 = 1$ is one, and therefore -1 is not an eigenvalue. We can write

$$\mathbf{P}^* = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{\text{t}} \tag{11.7}$$

where $\mathbf{S}$ is the matrix whose column vectors are $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$ (and so its matrix entries are $S_{ik} = \mathbf{s}_k(i)$), and furthermore $\mathbf{S}^{\text{t}}$ denotes the transpose of $\mathbf{S}$, and

$$\mathbf{\Lambda} = \text{diag}(1, \lambda_2, ..., \lambda_n) \,.$$

From (11.7) we get

$$P^*_{ij} = \sum_{k=1}^{n} S_{ik}\lambda_k S_{jk}$$

and thus

$$P_{ij} = \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} \sum_{k=1}^{n} S_{ik}\lambda_k S_{jk}\,.$$

Moreover, since $\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{\mathrm{t}}\mathbf{D}^{\frac{1}{2}}$, we have

$$\mathbf{P}^m = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Lambda}^m\mathbf{S}^{\mathrm{t}}\mathbf{D}^{\frac{1}{2}}\,,$$

and so

$$P^m_{ij} = \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} \sum_{k=1}^{n} S_{ik}\lambda^m_k S_{jk}\,. \tag{11.8}$$

We can rewrite (11.8) as

$$P^m_{ij} = \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} S_{i1}S_{j1} + \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} \sum_{k=2}^{n} S_{ik}\lambda^m_k S_{jk}\,. \tag{11.9}$$

Since $P^m_{ij} \xrightarrow{m\to\infty} \pi(j)$ and $\lim_{m\to\infty} \lambda^m_k = 0$ for $2 \le k \le n$, equation (11.9) implies

$$\boxed{P^m_{ij} = \pi(j) + \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} \sum_{k=2}^{n} S_{ik}\lambda^m_k S_{jk}\,.} \tag{11.10}$$

Alternatively, we can see that $\frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} S_{i1}S_{j1} = \pi(j)$, by noting that a normalized right eigenvector $\mathbf{s}_1$ of $\mathbf{P}^*$ corresponding to $\lambda_1 = 1$ is the column vector

$$\mathbf{s}_1 = (\sqrt{\pi(1)}, ..., \sqrt{\pi(n)})^{\mathrm{t}}\,. \tag{11.11}$$

So $S_{i1} = \sqrt{\pi(i)}$ and $S_{j1} = \sqrt{\pi(j)}$, which gives the result.

Expression (11.10) is called the **spectral representation** of the transition probabilities. From it we see that the absolute values of the non-trivial eigenvalues $\lambda_2, ..., \lambda_n$ play a crucial role in the rate of convergence to stationarity. Define

$$\lambda_* = \max\{|\lambda_2|, ..., |\lambda_n|\}\,.$$

The spectral representation (11.10) shows that, for all $i, j \in \mathcal{S}$, there exists a constant C_{ij} such that

$$|P_{ij}^m - \pi(j)| \leq C_{ij}\lambda_*^m \,.$$

Since the state space is finite, the maximum $\max_{i,j\in\mathcal{S}} C_{i,j}$ exists. We conclude that for any initial distribution μ_0, the distributions $\mu_m = \mu_0\mathbf{P}^m$ converge to π in total variation distance **exponentially fast**. That is, there exists a constant $\tilde{C}$ such that

$$\|\mu_m - \pi\|_{TV} \leq \tilde{C}\,\lambda_*^m \quad \text{for all } m \geq 1\,.$$

For specific examples, it may however be difficult to compute the constants C_{ij}, since they involve knowing the eigenvectors, and therefore difficult to compute $\tilde{C}$. The next section gives an upper bound for total variation distance in terms of the eigenvalues that does not require computation of these constants.

When a Markov chain with transition matrix $\mathbf{P}$ is periodic, there is no convergence. To avoid this issue, we often consider a *lazy version* of the Markov chain by adding positive holding probability to each state. In such a case, it is fairly standard to add holding probability of $\frac{1}{2}$. We then work with the modified transition matrix

$$\tilde{\mathbf{P}} = \frac{1}{2}\mathbf{P} + \frac{1}{2}\mathbf{I}\,.$$

where $\mathbf{I}$ is the identity matrix of order $|\mathcal{S}|$. Note that $\mathbf{P}$ and $\tilde{\mathbf{P}}$ have the same stationary distribution, and that $\tilde{\mathbf{P}}$ is reversible if $\mathbf{P}$ is reversible. A additional advantage of working with $\tilde{\mathbf{P}}$ is that all eigenvalues of $\tilde{\mathbf{P}}$ are nonnegative:

Lemma 11.2.2. *Let $\mathbf{P}$ be the transition matrix of a reversible, irreducible Markov chain. Consider its lazy version $\tilde{\mathbf{P}} = \frac{1}{2}\mathbf{P} + \frac{1}{2}\mathbf{I}$. Then all eigenvalues of $\tilde{\mathbf{P}}$ are nonnegative.*

Proof. Let $\mathbf{s}$ be an eigenvector of $\tilde{\mathbf{P}}$ corresponding to eigenvalue $\tilde{\lambda}$. Thus

$$\tilde{\mathbf{P}}\mathbf{s} = (\frac{1}{2}\mathbf{P} + \frac{1}{2}\mathbf{I})\mathbf{s} = \frac{1}{2}\mathbf{P}\mathbf{s} + \frac{1}{2}\mathbf{s} = \tilde{\lambda}\mathbf{s}\,.$$

It follows that

$$\mathbf{P}\mathbf{s} = (2\tilde{\lambda} - 1)\mathbf{s}\,,$$

and therefore $(2\tilde{\lambda} - 1)$ is an eigenvalue of $\mathbf{P}$. But $(2\tilde{\lambda} - 1) \geq -1$, and therefore $\tilde{\lambda} \geq 0$.

□

Adding holding probability $\frac{1}{2}$ to each state does slow down the convergence rate of the process, but not in a significant way. Very roughly, it will double the mixing time.

11.2.2 Spectral bounds on total variation distance

Consider a Markov chain $(X_n)_{n\geq 0}$ on finite state space $\mathcal{S}$. Assuming the Markov chain starts in state x, we will denote the distribution of X_m by $P^m_{x,\cdot}$.

Theorem 11.2.3. *Let* $\mathbf{P}$ *be the transition matrix for an irreducible, aperiodic, and reversible Markov chain on state space* $\mathcal{S} = \{1, ..., n\}$ *with stationary distribution* π. *Let* $\{1, \lambda_2, ..., \lambda_n\}$ *be the set of (necessarily real) eigenvalues of* $\mathbf{P}$, *and let* $\lambda_* = \max\{|\lambda_2|, ..., |\lambda_n|\}$. *Then*

$$4\|P^m_{x,\cdot} - \pi\|^2_{TV} \leq \frac{1-\pi(x)}{\pi(x)}\lambda_*^{2m}\,. \tag{11.12}$$

If the Markov chain arises from a random walk on a group G *with symmetric step distribution* μ, *then*

$$4\|P^m_{x,\cdot} - \pi\|^2_{TV} \leq \sum_{k=2}^{n} \lambda_k^{2m} \qquad \forall x \in G\,. \tag{11.13}$$

Recall from Exercise 10.6 that for a random walk on a group G with step distribution μ, we have $\|P^m_{x,\cdot} - \pi\|_{TV} = \|\mu^{\star m} - \pi\|_{TV}$ for all $x \in G$. In this case, total variation distance to stationarity does not depend on the initial distribution.

Proof of Theorem 11.2.3.

$$\begin{aligned}
4\|P^m_{x,\cdot} - \pi\|^2_{TV} &= \left(\sum_{y\in\mathcal{S}} |P^m_{xy} - \pi(y)|\right)^2 \\
&= \left(\sum_{y\in\mathcal{S}} \frac{|P^m_{xy} - \pi(y)|}{\sqrt{\pi(y)}}\sqrt{\pi(y)}\right)^2 \\
&\leq \sum_{y\in\mathcal{S}} \frac{(P^m_{xy} - \pi(y))^2}{\pi(y)} \\
&= \sum_{y\in\mathcal{S}} \frac{1}{\pi(y)}\left((P^m_{xy})^2 - 2P^m_{xy}\pi(y) + (\pi(y))^2\right) \\
&= \sum_{y\in\mathcal{S}} \frac{1}{\pi(y)}(P^m_{xy})^2 - 1 \\
&= \frac{1}{\pi(x)}P^{2m}_{xx} - 1
\end{aligned} \tag{11.14}$$

where the inequality in the third line is an application of the Cauchy–Schwartz inequality,

and (11.14) follows from reversibility. Indeed, since $\pi(y)P^m_{yx} = \pi(x)P^m_{xy}$, we have

$$\sum_{y\in\mathcal{S}} \frac{1}{\pi(y)}(P^m_{xy})^2 = \sum_{y\in\mathcal{S}} \frac{1}{\pi(x)} P^m_{xy}P^m_{yx} = \frac{1}{\pi(x)} P^{2m}_{xx}\,.$$

From (11.10) we have

$$P^{2m}_{xx} = \pi(x) + \sum_{k=2}^{n} S_{xk}\lambda_k^{2m} S_{xk}$$

an thus

$$\begin{aligned} P^{2m}_{xx} &\le \pi(x) + \lambda_*^{2m} \sum_{k=2}^{n} S_{xk}S_{xk} \\ &= \pi(x) + \lambda_*^{2m}(1-\pi(x))\,. \end{aligned} \tag{11.15}$$

Note that equality (11.15) follows from (11.11) and the orthogonality of the matrix $\mathbf{S}$. Inequality (11.12) follows from combining (11.14) and (11.15).

Now assume the Markov chain arises from random walk on a group G with step distribution μ. The stationary distribution π is uniform on G, that is, $\pi(x) = \frac{1}{|G|}$ for all $x \in G$. Furthermore, the diagonal elements P^{2m}_{xx} of $\mathbf{P}^{2m}$ do not depend on x since $P^{2m}_{xx} = \mu^{*2m}(id)$. Thus

$$\begin{aligned} \frac{1}{\pi(x)} P^{2m}_{xx} - 1 &= |G|P^{2m}_{xx} - 1 \\ &= \text{trace}(\mathbf{P}^{2m}) - 1 = \sum_{k=2}^{n} \lambda_k^{2m}\,, \end{aligned}$$

and (11.13) follows from (11.14). □

11.2.3 Random walk on the discrete circle

Let $(Y_n)_{n\ge 0}$ be symmetric random walk on the discrete N-cycle $\mathbb{Z}_N$. The group operation is + (mod N). The step distribution is $\mu(1) = \mu(-1) = \frac{1}{2}$. To avoid periodicity, we assume N is odd. See Figure 11.3. Assume the walk starts in state 0. Since this is random walk on a group, the stationary distribution is uniform distribution Unif($\mathbb{Z}_N$).

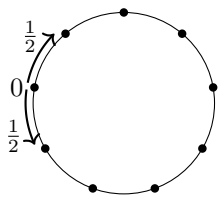


Figure 11.3

We computed the eigenvalues for this random walk in Example 10.3.1. They are

$$\lambda_k = \cos(2\pi k/N) \quad \text{for } k \in \{0, 1, ..., (N-1)\}\,.$$

Note that each eigenvalue other than eigenvalue 1 has multiplicity 2. Therefore we can write the sum in (11.13) as

$$2\sum_{k=1}^{(N-1)/2} [\cos(2\pi k/N)]^{2m}\,.$$

Using the Taylor expansion for $\cos x$, we get $\cos x \le 1 - (\frac{x^2}{2} - \frac{x^4}{4!})$. Since $\frac{1}{2} - \frac{x^2}{4!} > \frac{1}{12}$ for for $0 \le x \le \pi$, we have

$$\cos x \le 1 - \frac{x^2}{12} \le e^{-\frac{x^2}{12}} \qquad \text{for } 0 \le x \le \pi.$$

Thus we get from (11.13),

$$\begin{aligned}
\|\mu^{*m} - \text{Unif}(\mathbb{Z}_N)\,\|_{TV}^2 &\le \frac{1}{2}\sum_{k=1}^{(N-1)/2} [\cos(2\pi k/N)]^{2m} \\
&\le \frac{1}{2}\sum_{k=1}^{(N-1)/2} e^{-(2\pi^2 k^2 m)/(3N^2)} \\
&= \frac{1}{2} e^{-(2\pi^2 m)/(3N^2)} \sum_{k=1}^{(N-1)/2} e^{-(2\pi^2 (k^2-1) m)/(3N^2)} \\
&\le \frac{1}{2} e^{-(2\pi^2 m)/(3N^2)} \sum_{j=0}^{\infty} e^{-(2\pi^2 j m)/(3N^2)} \\
&= e^{-(2\pi^2 m)/(3N^2)} \frac{1}{2(1 - e^{-(2\pi^2 m)/(3N^2)})}\,.
\end{aligned}$$

For $m \ge N^2$, the denominator in the last expression is

$$2(1 - e^{-(2\pi^2 m)/(3N^2)}) \ge 1\,.$$

Thus, altogether, we get

$$\|\mu^{*m} - \text{Unif}(\mathbb{Z}_N)\,\|_{TV} \le e^{-(\pi^2 m)/(3N^2)}$$

for $m \ge N^2$. This shows that total variation distance is guaranteed to be less than 0.037 if the number m of steps is $m \ge N^2$. We conclude that the mixing time as a function of the size N of the state space is of order $O(N^2)$.

Cover time: Recall the definition of cover time t^{cov} (Definition 8.5.1) for a Markov chain $(Y_n)_{n\ge 0}$: It is defined as $t^{\text{cov}} = \max_{x\in\mathcal{S}} \mathbb{E}_x(T^{\text{cov}})$ for the random variable

$$T^{\text{cov}} = \min\{n : \ \forall y \in \mathcal{S},\ \exists k \le n,\ \text{s.t. } Y_k = y\}\,.$$

It should be intuitively clear that the cover time t^{cov} is related to the mixing time t^{mix}, since the Markov chain must visit every state at least once before it can be "reasonably close" to stationarity. Recall Example 8.5.3: The cover time t^{cov} for simple symmetric random walk on the cyclic group $\mathbb{Z}_N$ is

$$t^{\text{cov}} = \frac{1}{2}N(N-1).$$

Thus, asymptotically as $N \to \infty$, the cover time is of order $\Theta(N^2)$. This matches the upper bound for mixing rate which we have computed using eigenvalues and found to be of order $O(N^2)$.

The last new state visited: We include a surprising result about simple symmetric random walk on $\mathbb{Z}_N$ that concerns the distribution of the *last new* state the walk visits. The following proposition is taken from [27]. See also [2].

Proposition 11.2.4. *Consider simple symmetric random walk on the discrete cycle $\mathbb{Z}_N$ with vertices $\mathcal{S} = \{0, 1, 2..., (N-1)\}$. Without loss of generality, assume that the walk starts at vertex 0. Let L be the random variable that is the last new vertex reached by the random walk. Then L has uniform distribution on the set of vertices $\{1, 2, ..., (N-1)\}$.*

Proof. Let $k \in \{1, 2, ..., (N-1)\}$. The walk starts at vertex 0. For the adjacent vertices $k = 1$ or $k = N-1$, cutting open the circle (either at vertex $k = 1$ or at vertex $k = N-1$), flattening the circle into a discrete line segment, and using the gambler's ruin formulas (4.18) yields

$$\mathbb{P}(L = 1) = \mathbb{P}(L = N-1) = \frac{1}{N-1}\,.$$

For $k \in \{2, ..., N-2\}$, we compute $\mathbb{P}(L = k)$ by conditioning on the second-to-last new vertex visited (which is either $k-1$ or $k+1$). Doing so, we get

$$\mathbb{P}_0(L = k) = \mathbb{P}_0(T^{k-1} < T^{k+1})\,\mathbb{P}_{k-1}(T^{k+1} < T^k) + \mathbb{P}_0(T^{k+1} < T^{k-1})\,\mathbb{P}_{k+1}(T^{k-1} < T^k)\,.$$

Again, using formulas (4.18), we compute

$$\mathbb{P}_0(L = k) = \frac{(N-k-1)}{(N-2)}\frac{1}{(N-1)} + \frac{(k-1)}{(N-2)}\frac{1}{(N-1)} = \frac{1}{N-1} \qquad \text{for } k \in \{2, ..., N-2\}\,.$$

□

Remark 11.2.5. *By symmetry, a complete graph with N vertices has the same property that Proposition 11.2.4 states for the N-cycle: For simple random walk on a complete graph, for any starting vertex k, the last vertex that is reached is uniformly distributed over all vertices (not including k). Lovász and Winkler prove in [27] that the N-cycle and the complete graph with N vertices are in fact the only graphs that have this property.*

11.2.4 The Ehrenfest chain

We have introduced the Ehrenfest chain for N particles in Section 1.5 as an urn model for gas diffusion through a porous membrane. The state of the system at time n is the number of particles in Box 1 at time n. The state space is $\mathcal{S} = \{0, 1, ..., N\}$, and the stationary distribution is $\text{Bin}(N, \frac{1}{2})$.

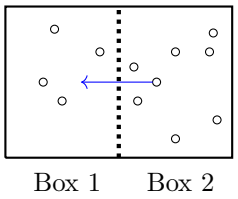


Figure 11.4: Ehrenfest chain

The Ehrenfest chain can be "lifted" to simple random walk on the hypercube $\mathbb{Z}_2^N$ (or, equivalently stated, the Ehrenfest chain is a lumped version of simple random walk on $\mathbb{Z}_2^N$). How many steps suffice for the system to be in or very near equilibrium, that is, for the number of balls in Box 1 to be approximately distributed $\text{Bin}(N, \frac{1}{2})$? To answer this question, we will apply the upper bound (11.13) to random walk on to $\mathbb{Z}_2^N$. Assume the random walk starts in state $(0, ..., 0)$ (i.e., the Ehrenfest chain starts in state 0). To avoid periodicity, we consider lazy simple random walk. The step distribution μ is defined by $\mu(0, ..., 0) = \frac{1}{2}$ and

$$\mu(1, 0, ..., 0) = \mu(0, 1, 0, ..., 0) = \cdots = \mu(0, ..., 0, 1) = \frac{1}{2N}\,.$$

We have computed the eigenvalues for (lazy) random walk on the hypercube in Example 10.3.2. They are

$$\frac{N-j}{N} \quad \text{with multiplicity } \binom{N}{j} \qquad \text{for } j = 0, 1, ..., N\,,$$

and so, by (11.13),

$$4\|\mu^{*m} - \mathrm{Unif}(\mathbb{Z}_2^N)\,\|_{TV}^2 \leq \sum_{j=1}^{N}(1-\frac{j}{N})^{2m}\binom{N}{j}\,.$$

Using the estimate $1-\frac{j}{N} \leq e^{-\frac{j}{N}}$, we get

$$4\|\mu^{*m} - \mathrm{Unif}(\mathbb{Z}_2^N)\,\|_{TV}^2 \leq \sum_{j=1}^{N} e^{-\frac{j}{N}2m}\binom{N}{j} = (1+e^{-\frac{2m}{N}})^N - 1\,.$$

Take $c>0$ and let $m = \frac{1}{2}N\ln N + cN$. This yields

$$(1+e^{-\frac{2(\frac{1}{2}N\ln N+cN)}{N}})^N - 1 = (1+\frac{1}{N}e^{-2c})^N - 1\,,$$

which, since $\lim\limits_{n\to\infty} \uparrow (1+\frac{1}{n})^n = e$, yields the estimate

$$4\|\mu^{*m} - \mathrm{Unif}(\mathbb{Z}_2^N)\,\|_{TV}^2 \leq e^{e^{-2c}} - 1\,.$$

The expression on the right hand side can be made arbitrarily small for a suitable $c>1$. Hence $\|\mu^{*m} - \mathrm{Unif}(\mathbb{Z}_2^N)\,\|_{TV}$ will be small for $m = \frac{1}{2}N\ln N + cN$ and for suitable $c>1$. Running the random walk a number of $m = O(N\ln N)$ steps suffices to be close to stationarity.
Since the Ehrenfest chain for N particles is a lumped version of random walk on $\mathbb{Z}_2^N$, the same upper bound for distance to stationarity is valid (see Exercises 11.2 and 11.3). We conclude that for the Ehrenfest chain with N particles, a number of $m = O(N\ln N)$ steps suffice for the chain to be close to its stationary distribution $\mathrm{Bin}(N,\frac{1}{2})$. □

11.3 Coupling

11.3.1 Definition of Coupling

In probability, *coupling* refers to a method by which two or more random variables (or sequences of random variables) are constructed on a common probability space Ω. A coupling is in effect a **construction of a joint probability distribution**, that is, a dependency structure for given random variables, that preserves their given marginal distributions. With the use of such a coupling one can derive information about each of the random variables by exploiting certain properties of their joint distribution. There are many ways in which one can couple given random variables. The usefulness of the method will depend on a judicious choice among all possible couplings, given the specific problem at hand.

Definition 11.3.1. *Let $\mathcal{S}$ be a discrete space and μ and ν two probability distributions on $\mathcal{S}$. A* **coupling** *of μ and ν is a pair of random variables $(X, Y) : \Omega \to \mathcal{S} \times \mathcal{S}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the marginal distribution of X is μ and the marginal distribution of Y is ν, that is,*

$$\mathbb{P}(X \in A) = \mu(A) \quad \text{and} \quad \mathbb{P}(Y \in B) = \nu(B) \qquad \text{for all } A, B \subseteq \mathcal{S}\,.$$

The underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ provides the **common source of randomness for the two random variables X and Y.**

Note that a specific coupling (X, Y) of μ and ν induces a specific joint probability distribution ω (i.e., the distribution of the random vector (X, Y)) on the product space $\mathcal{S} \times \mathcal{S}$ whose marginals are μ and ν:

$$\mu(x) = \sum_{y \in \mathcal{S}} \omega(x, y) \quad \text{and} \quad \nu(z) = \sum_{y \in \mathcal{S}} \omega(y, z) \qquad \text{for all } x, z \in \mathcal{S}\,.$$

Conversely, any bivariate distribution ω on $\mathcal{S} \times \mathcal{S}$ that has marginals μ and ν defines random variables $X \sim \mu$ and $Y \sim \nu$ with joint distribution ω and thus defines a coupling (X, Y) of μ and ν.

The common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ underlying a coupling is not unique. We can always consider $\Omega = \mathcal{S} \times \mathcal{S}$, in which case the random vector (X, Y) of the coupling becomes the identity map on $\mathcal{S} \times \mathcal{S}$, and the common probability space is

$$(\mathcal{S} \times \mathcal{S},\ \mathcal{P}(\mathcal{S} \times \mathcal{S}),\ \omega)\,.$$

Example 11.3.1 (Two coin tosses)**.** Consider two biased coins. Let the state space for each coin toss be $\mathcal{S} = \{0, 1\}$, and consider $X \sim \mu$ with $\mu(1) = p,\ \mu(0) = 1 - p$ and $Y \sim \nu$ with $\nu(1) = q,\ \nu(0) = 1 - q$. Assume $p < q$.

(a) *Independent coupling.* The joint distribution is given by

$X \setminus Y$	0	1
0	$(1-q)(1-p)$	$q(1-p)$
1	$p(1-q)$	pq

.

(b) Consider a uniform random variable $U \sim \text{Unif}([0, 1])$. We define

$$X = \begin{cases} 1 & \text{if } 0 < U \le p \\ 0 & \text{if } p < U \le 1 \end{cases}, \quad Y = \begin{cases} 1 & \text{if } 0 < U \le q \\ 0 & \text{if } q < U \le 1 \end{cases}.$$

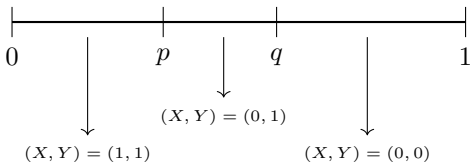


Figure 11.5: A possible coupling for two coin tosses with distinct biases

Here the random variable U is the common source of randomness for the coupling, and we can take $\Omega = [0, 1]$. See Figure 11.5. The joint distribution is given by

$X \setminus Y$	0	1
0	$1-q$	$q-p$
1	0	p

.

□

Example 11.3.2 (Two identically distributed coin tosses)**.** The state space is $\mathcal{S} = \{0, 1\}$. Consider $X \sim \mu_1$, $Y \sim \mu_2$ with $\mu_i(1) = p,\ \mu_i(0) = 1 - p$ for $i = 1, 2$. For each s with $0 \le s \le \min\{p, 1-p\}$ we have a coupling ω_s of μ_1 and μ_2 given by

$X \setminus Y$	0	1
0	$1-p-s$	s
1	s	$p-s$

.

For $s = p(1-p)$, the two coin tosses are independent. For $s = 0$, we have $X = Y$ which constitutes maximum dependence. □

Proposition 11.3.1. *Let μ and ν be two probability distributions on $\mathcal{S}$ and (X, Y) a coupling of μ and ν. Then*

$$\|\mu - \nu\|_{TV} \le \mathbb{P}(X \ne Y)\,. \tag{11.16}$$

Hence

$$\|\mu - \nu\|_{TV} \le \inf_{\text{couplings } (X,Y)} \mathbb{P}(X \ne Y) \tag{11.17}$$

where the infimum in (11.17) is taken over all couplings (X, Y) of μ and ν.

Proof. Let $A \subseteq \mathcal{S}$ and (X, Y) a coupling of μ and ν. We have

$$\begin{aligned}
\mu(A) - \nu(A) &= \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\
&= \mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X = Y) - \mathbb{P}(Y \in A, X \neq Y) \\
&= \mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X \neq Y) \\
&\leq \mathbb{P}(X \in A, X \neq Y) \\
&\leq \mathbb{P}(X \neq Y)\,.
\end{aligned}$$

Reversing the roles of μ and ν yields

$$|\mu(A) - \nu(A)| \leq \mathbb{P}(X \neq Y)$$

from which we get

$$\max_{A \subseteq \mathcal{S}} |\mu(A) - \nu(A)| = \|\mu - \nu\|_{TV} \leq \mathbb{P}(X \neq Y)\,.$$

This proves (11.16). Taking the infimum over all couplings on both sides of (11.16) yields (11.17). □

Example 11.3.3. We return to Example 11.3.1 of two biased coin tosses. The total variation distance of μ and ν is

$$\|\mu - \nu\|_{TV} = \frac{1}{2}(|p - q| + |(1 - p) - (1 - q)|) = |q - p|\,.$$

(a) For the independent coupling, we get

$$\mathbb{P}(X \neq Y) = q(1 - p) + p(1 - q) = q - p(2q - 1)\,.$$

Recall that we are assuming $q > p$, and so

$$\|\mu - \nu\|_{TV} = q - p \leq q - p(2q - 1) = \mathbb{P}(X \neq Y)$$

which confirms (11.16).

(b) For the coupling in Example 11.3.1(b), we get

$$\mathbb{P}(X \neq Y) = q - p\,.$$

So here we have

$$\|\mu - \nu\|_{TV} = \mathbb{P}(X \neq Y)\,.$$

This coupling is an example of an *optimal coupling* (see the following definition). □

Definition 11.3.2. *A coupling (X, Y) of μ and ν is called an* **optimal coupling** *if*

$$\|\mu - \nu\|_{TV} = \mathbb{P}(X \neq Y)\,.$$

For an optimal coupling, the two random variables are most strongly coupled in the sense that $\mathbb{P}(X = Y)$ is as large as possible.

Proposition 11.3.2. *Let μ and ν be two probability distributions on $\mathcal{S}$. There always exists an optimal coupling (X, Y) for μ and ν. As a consequence, we get the following improvement to (11.17):*

$$\|\mu - \nu\|_{TV} = \min_{\text{couplings } (X,Y)} \mathbb{P}(X \neq Y)\,. \tag{11.18}$$

Exercise 11.7 guides the reader through the construction of an optimal coupling.

11.3.2 Coupling of Markov chains

Recall that above, in order to construct a coupling of two random variables X and Y, we have constructed a common underlying probability space (i.e. a common source of randomness) on which the two random variables are defined. Here we will extend the idea of coupling to two Markov chains, that is to two entire sequences of random variables.

Let $\mathcal{S}$ be a (finite or infinite) discrete state space and $\mathcal{S}$ and $\mathbf{P}$ a stochastic matrix indexed by elements in $\mathcal{S}$. We will call a Markov chain with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$ a $\mathbf{P}$-Markov chain.

Definition 11.3.3. *Let $\mathcal{S}$ be a discrete state space and $\mathbf{P}$ a stochastic matrix indexed by $\mathcal{S}$. A coupling of a $\mathbf{P}$-Markov chain with initial distribution μ_0 and a $\mathbf{P}$-Markov chain with initial distribution ν_0 is a stochastic process $(X_n, Y_n)_{n\geq 0}$ with state space $\mathcal{S} \times \mathcal{S}$ such that*

(a) all random variables X_n and Y_n are defined on the same probability space,
(b) $(X_n)_{n\geq 0}$ is a $\mathbf{P}$-Markov chain with initial distribution μ_0,
(c) $(Y_n)_{n\geq 0}$ is a $\mathbf{P}$-Markov chain with initial distribution ν_0.

Note that Definition 11.3.3 does not require the stochastic process $(X_n, Y_n)_{n\geq 0}$ that defines a coupling to be a Markov chain. In all examples that we will present, the coupling will be a Markov chain.

Example 11.3.4. Consider simple random walk on the N-cycle $\mathbb{Z}_N$ (the integers mod N). Here we include holding probability of $h = \frac{1}{2}$ for each state. There are two reasons for adding holding probability: Adding holding probability eliminates periodicity if N is even, and it allows us to construct a fairly simple coupling which we describe below. Note that modifying a chain by adding a fixed holding probability to each state and proportionally changing all other transition probabilities does not significantly change the mixing time. For example, adding holding probability of $\frac{1}{2}$ to simple random walk on the N-cycle will slow down convergence by a factor of 2. With holding probability $\frac{1}{2}$, on average, the walk will move to one of its neighbors only half of the time. See also Remark 4.4.2.
Fix $p \in (0, \frac{1}{2})$. The transition probabilities for the $\mathbf{P}$-Markov chain on $\mathbb{Z}_N$ are $P_{zz} = \frac{1}{2}$, $P_{z,z+1} = p$, and $P_{z,z-1} = q$ for $z \in \mathbb{Z}_N$. We assume $p+q+\frac{1}{2} = 1$. We let the chain $(X_n)_{n\geq 0}$ start in state x (so $\mu_0 \sim \delta_x$), and we let the chain $(Y_n)_{n\geq 0}$ start in state y (so $\nu_0 \sim \delta_y$). See Figure 11.6.

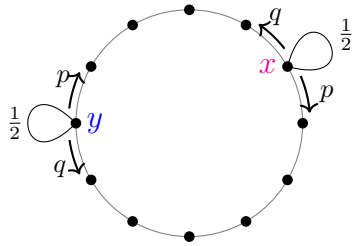


Figure 11.6: Simple random walk on $\mathbb{Z}_N$ with holding probability $\frac{1}{2}$

Consider a sequence $(U_n)_{n\geq 1}$ of i.i.d random variables with $U_1 \sim \text{Unif}([0,1])$. A possible coupling $(X_n, Y_n)_{n\geq 0}$ of the two chains proceeds as follows. The process starts in state $(x, y) \in \mathbb{Z}_N \times \mathbb{Z}_N$. For $n \geq 1$, if $(X_{n-1}, Y_{n-1}) = (x', y')$, then

$$\begin{cases} (X_n, Y_n) = (x'+1, y') & \text{if } 0 \leq U_n \leq p\,, \\ (X_n, Y_n) = (x'-1, y') & \text{if } p < U_n \leq \frac{1}{2}\,, \\ (X_n, Y_n) = (x', y'+1) & \text{if } \frac{1}{2} < U_n \leq \frac{1}{2}+p\,, \\ (X_n, Y_n) = (x', y'-1) & \text{if } \frac{1}{2}+p < U_n \leq 1\,. \end{cases}$$

See Figure 11.7. Here the common underlying source of randomness for the two chains is the sequence $(U_n)_{n\geq 1}$. An alternate description of this coupling would be to say that $(X_n, Y_n)_{n\geq 0}$ is a random walk on the *discrete torus* $\mathbb{Z}_N \times \mathbb{Z}_N$ that starts at state (x, y) and whose *step random variable* (ξ^X, ξ^Y) has distribution

$$\mathbb{P}\left((\xi^X, \xi^Y) = (1,0)\right) = \mathbb{P}\left((\xi^X, \xi^Y) = (0,1)\right) = p\,, \text{and}$$

$$\mathbb{P}\left((\xi^X, \xi^Y) = (-1,0)\right) = \mathbb{P}\left((\xi^X, \xi^Y) = (0,-1)\right) = q\,.$$

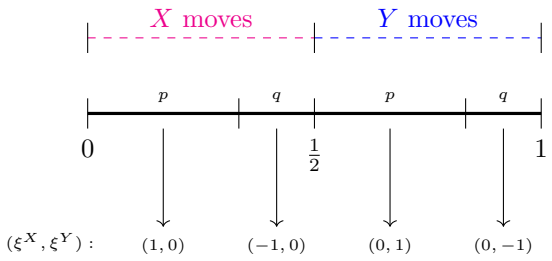


Figure 11.7: A coupling of two simple random walks on $\mathbb{Z}_N$

So

$$(X_n, Y_n) = (x, y) + \sum_{k=1}^{n} (\xi_k^X, \xi_k^Y)$$

where $(\xi_k^X, \xi_k^Y)_{k\geq 1}$ is an i.i.d. sequence with $(\xi_1^X, \xi_1^Y) \sim (\xi^X, \xi^Y)$. For this description of the coupling, we can view the common underlying probability space Ω for all random variables X_n and Y_n as the space of all infinite sequences $\omega = (\omega_1, \omega_2, ...)$ with entries $\omega_i \in \{(\pm 1, 0), (0, \pm 1)\}$ for $i \geq 1$. □

For a given coupling $(X_n, Y_n)_{n\geq 0}$ of two Markov chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$, the so-called **coupling or coalescence time** T_{couple} defined by

$$T_{\text{couple}} = \min\{n \geq 0 : X_n = Y_n\}$$

will play an important role. It is a stopping time for the process $(X_n, Y_n)_{n\geq 0}$. We can (and in fact, we usually will) define any coupling $(X_n, Y_n)_{n\geq 0}$ in such a way that once the two chains meet, that is from time T_{couple} onwards, they will move in lockstep. We do so by stipulating

$$X_n = Y_n \quad \text{for } n \geq T_{\text{couple}}\,.$$

See Figure 11.8.

Example 11.3.5 (Independent coupling until time T_{couple})**.** Consider a (finite or infinite) stochastic matrix $\mathbf{P}$ indexed by a state space $\mathcal{S}$ and two $\mathbf{P}$-chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ on $\mathcal{S}$. Here we allow the two chains to move independently until $X_n = Y_n$ for the first time, which happens at time T_{couple}. From then onwards, the two chains make the same jumps. More precisely, we define a Markov chain $(X_n, Y_n)_{n\geq 0}$ on $\mathcal{S} \times \mathcal{S}$ with transition probabilities

$$\begin{aligned} P_{(x,y),(x',y')} &= P_{x,x'} P_{y,y'} \text{ if } x \neq y, \text{and} \\ P_{(x,y),(x',y')} &= P_{x,x'} \text{ if } x = y \text{ and } x' = y' \\ P_{(x,y),(x',y')} &= 0 \text{ otherwise}\,. \end{aligned}$$

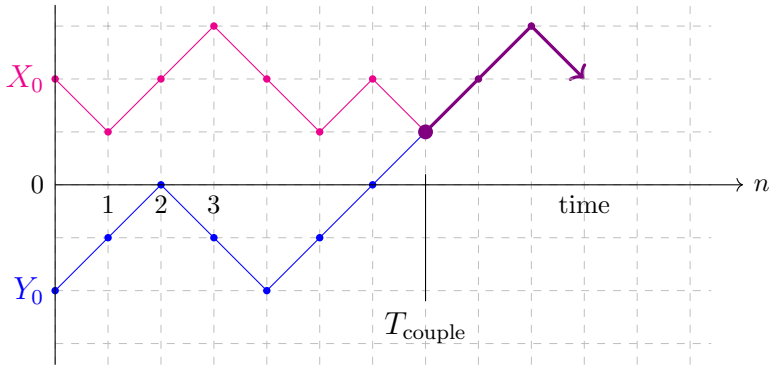


Figure 11.8: Coalescence time for a sample path for a coupling $(X_n, Y_n)_{n\geq 0}$

Exercise 11.10 asks the reader to verify that the above probabilities define a coupling for $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$. Note that T_{couple} is not necessarily a finite random variable. In this context, also recall Example 3.2.1. It is an example of a coupling that is a reducible Markov chain. □

Theorem 11.3.3 (Coupling inequality)**.** *Let $(X_n, Y_n)_{n\geq 0}$ be a coupling of a $\mathbf{P}$-Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{S}$ and a $\mathbf{P}$-Markov chain $(Y_n)_{n\geq 0}$ on $\mathcal{S}$. We denote the distribution of X_n by μ_n and the distribution of Y_n by ν_n for $n \geq 0$. Then*

$$\|\mu_n - \nu_n\|_{TV} \leq \mathbb{P}(T_{\text{couple}} > n)\,. \tag{11.19}$$

Proof. For any fixed time n, (X_n, Y_n) is a coupling of the two distributions μ_n and ν_n on $\mathcal{S}$. By Proposition 11.3.1,

$$\|\mu_n - \nu_n\|_{TV} \leq \mathbb{P}(X_n \neq Y_n)\,.$$

Since, per our assumption, the two chains run in lockstep after coupling, we have

$$\mathbb{P}(X_n \neq Y_n) = \mathbb{P}(T_{\text{couple}} > n)$$

which establishes (11.19). □

Consider an irreducible, positive recurrent, and aperiodic Markov chain $(X_n)_{n\geq 0}$ with initial distribution μ_0 and stationary distribution π. Let μ_n denote the distribution of the chain at time n. The Coupling inequality (11.19) re-establishes (we have seen this before) convergence to stationarity with respect to total variation distance, that is, it implies

$$\|\mu_n - \pi\|_{TV} \xrightarrow{n\to\infty} 0\,. \tag{11.20}$$

Proving (11.20), based on (11.19), is the object of Exercise 11.12.

Example 11.3.6 (Random-to-top shuffle)**.** We have introduced random-to-top shuffling in Section 10.2 as the time reversed random walk to top-to-random shuffling. Both random walks have the same rate of convergence to stationarity. We will construct a coupling for random-to-top shuffling.

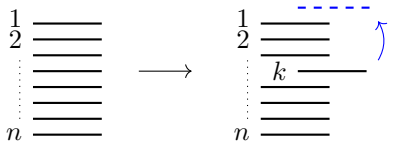


Figure 11.9: Random-to-top shuffle

Consider two **P**-Markov chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ on the permutation group $\mathrm{S_n}$. Each of the two **P**-chains is a random walk on $\mathrm{S_n}$ with step distribution $\tilde{\mu} \sim \mathrm{Unif}(\tilde{C})$ for $\tilde{C} = \{id, \sigma_2, ..., \sigma_n\}$, where σ_k denotes the cyclic permutation

$$\sigma_k = (k \to (k-1) \to (k-2) \to \cdots \to 1 \to k) \qquad \text{for } 2 \leq k \leq n\,.$$

Assume Markov chain $(X_n)_{n\geq 0}$ starts at the identity id (perfect order), and Markov chain $(Y_n)_{n\geq 0}$ starts in stationary (i.e., here uniform) distribution π. The coupling $(X_n, Y_n)_{n\geq 0}$ is constructed as follows. At each step, draw $k \in \{1, ..., n\}$ uniformly at random. In each of the two decks, take Card k and put it on the top of the pile. Both chains are evolving according to the transition probabilities for a random-to-top shuffle. Since Markov chain $(Y_n)_{n\geq 0}$ starts in stationary distribution, it remains in stationary distribution. Note that once Card k has been selected an put on top of both piles, it will be in the same location in both piles at all future times. Consider the stopping time

$$T = \min\{n : \text{all cards have been selected at least once}\}\,.$$

Clearly,

$$T_{\text{couple}} \leq T\,,$$

and therefore,

$$\mathbb{P}(T_{\text{couple}} > n) \leq \mathbb{P}(T > n)\,.$$

The random time T has the same distribution as the waiting time for the coupon collectors problem (see Section B.6). Using Lemma B.6.1, in combination with the coupling inequality (11.19), we get for $k = n \ln n + cn$ random-to-top shuffles of n cards,

$$\|\mu_k - \pi\|_{TV} \leq e^{-c}\,.$$

After k random-to-top shuffles where $k = O(n \ln n)$, the deck is near "completely mixed up". The same is true for top-to-random shuffle. □

In applications, it is often easiest to construct a coupling $(X_n, Y_n)_{n\geq 0}$ of two $\mathbf{P}$-Markov chains for which $(X_n)_{n\geq 0}$ starts in a given state x and $(Y_n)_{n\geq 0}$ starts in a given state y (as in Example 11.3.4). We can use such couplings to study the rate of convergence to stationarity for a $\mathbf{P}$-Markov chain that starts in a more general initial distribution μ_0. Theorem 11.3.5 below is a related result.
We will use the following notation: $(X_n^{(x)})_{n\geq 0}$ will denote a $\mathbf{P}$-Markov chain starting in state x and $\mu_n^{(x)}$ will denote the distribution of $X_n^{(x)}$. Note that with this notation, we have $\mu_n^{(x)}(z) = P_{xz}^n$ for $z \in \mathcal{S}$. Furthermore, $\mathbb{P}^{(x,y)}$ will denote the probability that is associated with a coupling $(X_n^{(x)}, X_n^{(y)})_{n\geq 0}$.

Lemma 11.3.4. *Let $(X_n)_{n\geq 0}$ be an irreducible, aperiodic, and positive recurrent $\mathbf{P}$-Markov chain on finite state space $\mathcal{S}$, and let π be the stationary distribution. We denote the distribution of X_n by μ_n for $n \geq 0$. Then*

$$\|\mu_n - \pi\|_{TV} \leq \max_{x,y\in\mathcal{S}} \|\mu_n^{(x)} - \mu_n^{(y)}\|_{TV} \,. \tag{11.21}$$

Proof. Fix $x \in \mathcal{S}$. We will first show $\|\mu_n^{(x)} - \pi\|_{TV} \leq \max_{y\in\mathcal{S}} \|\mu_n^{(x)} - \mu_n^{(y)}\|_{TV}$. Note that since π is the stationary distribution, we have

$$\pi(E) = \sum_{y\in\mathcal{S}} \pi(y)\mu_n^{(y)}(E) \quad \text{for all } E \subseteq \mathcal{S} \,.$$

We have

$$\begin{aligned}
\|\mu_n^{(x)} - \pi\|_{TV} &= \sup_{E\subseteq\mathcal{S}} |\mu_n^{(x)}(E) - \pi(E)| \\
&= \sup_{E\subseteq\mathcal{S}} \left| \sum_{y\in\mathcal{S}} \pi(y) \left[\mu_n^{(x)}(E) - \mu_n^{(y)}(E)\right] \right| \\
&\leq \sup_{E\subseteq\mathcal{S}} \sum_{y\in\mathcal{S}} \pi(y) \left|\mu_n^{(x)}(E) - \mu_n^{(y)}(E)\right| \\
&\leq \max_{y\in\mathcal{S}} \|\mu_n^{(x)} - \mu_n^{(y)}\|_{TV} \sum_{y\in\mathcal{S}} \pi(y) = \max_{y\in\mathcal{S}} \|\mu_n^{(x)} - \mu_n^{(y)}\|_{TV} \,.
\end{aligned}$$

This proves

$$\|\mu_n^{(x)} - \pi\|_{TV} \leq \max_{y\in\mathcal{S}} \|\mu_n^{(x)} - \mu_n^{(y)}\|_{TV} \,. \tag{11.22}$$

Proving the inequality

$$\|\mu_n - \pi\|_{TV} \leq \max_{x \in \mathcal{S}} \|\mu_n^{(x)} - \pi\|_{TV} \tag{11.23}$$

is the content of Exercise 11.6. Combining (11.22) and (11.23) yields

$$\|\mu_n - \pi\|_{TV} \leq \max_{x \in \mathcal{S}} \left[\max_{y \in \mathcal{S}} \|\mu_n^{(x)} - \mu_n^{(y)}\|_{TV} \right]$$

which proves (11.21). □

Theorem 11.3.5. *Let $(X_n)_{n \geq 0}$ be an irreducible, aperiodic, and positive recurrent $\mathbf{P}$-Markov chain on finite state space $\mathcal{S}$, and let π be the stationary distribution. We denote the distribution of X_n by μ_n for $n \geq 0$. Assume for each pair of states $x, y \in \mathcal{S}$ there exists a coupling $(X_n^{(x)}, X_n^{(y)})_{n \geq 0}$. Then*

$$\|\mu_n - \pi\|_{TV} \leq \max_{x,y \in \mathcal{S}} \mathbb{P}^{(x,y)}(T_{\text{couple}} > n) \tag{11.24}$$

and, as a consequence,

$$\|\mu_n - \pi\|_{TV} \leq \max_{x,y \in \mathcal{S}} \frac{\mathbb{E}^{(x,y)}(T_{\text{couple}})}{n+1}. \tag{11.25}$$

Proof. Inequality (11.24) is a combination of (11.21) and (11.19). By Markov's inequality (see Appendix B),

$$\mathbb{P}^{(x,y)}(T_{\text{couple}} > n) \leq \frac{\mathbb{E}^{(x,y)}(T_{\text{couple}})}{n+1}$$

which implies (11.25). □

Example 11.3.7 (Rate of convergence for simple random walk on $\mathbb{Z}_N$)**.** We return to Example 11.3.4 and its same set-up. See Figure 11.10.

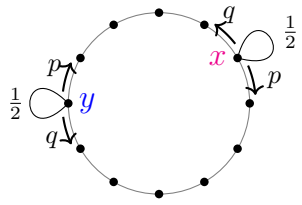


Figure 11.10: Simple random walk on $\mathbb{Z}_N$ with holding probability $\frac{1}{2}$

Fix $x, y \in \mathbb{Z}_N$. For a coupling $(X_n^{(x)}, X_n^{(y)})_{n\geq 0}$ as defined in Example 11.3.4, consider the process $(D_n)_{n\geq 0}$ that tracks the *clockwise distance* from $X_n^{(x)}$ to $X_n^{(y)}$. Note that in the above picture, $D_0 = 7$. We have

$$T_{\text{couple}} = \min\{n : D_n = 0 \text{ or } D_n = N\}\,.$$

The process $(D_n)_{0\leq n\leq T_{\text{couple}}}$ is simple symmetric random walk (without holding probability) on the integers $\{0, 1, ..., N\}$. See Figure 11.11.

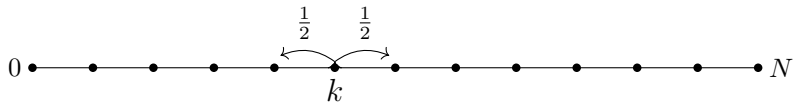


Figure 11.11: Tracking the clockwise distance of the two random walks on $\mathbb{Z}_N$

Recall formula (4.19). Let T be the time until simple symmetric walk $(D_n)_{n\geq 0}$ on $\{0, 1, ..., N\}$ hits the boundary $\{0, N\}$. Given $D_0 = k$, the expected time $\mathbb{E}(T \,|\, D_0 = k)$ is

$$\mathbb{E}(T \,|\, D_0 = k) = k(N - k) = \mathbb{E}(T_{\text{couple}})\,.$$

Thus we have

$$\max_{x,y\in\mathcal{S}} \mathbb{E}^{(x,y)}(T_{\text{couple}}) = \max_{1\leq k\leq N-1} \mathbb{E}(T \,|\, D_0 = k) \leq \frac{N^2}{4}\,. \tag{11.26}$$

Combining (11.26) with (11.25), we get for the total variation distance to stationarity at time n,

$$\|\mu_n - \pi\|_{TV} \leq \frac{N^2}{4(n+1)} < \frac{N^2}{4n}\,. \tag{11.27}$$

The right hand side of (11.27) shows that for $n \geq N^2$, the distance to stationarity is at most $\frac{1}{4}$. Equivalently, we can state this result in terms of the mixing time for $\epsilon = \frac{1}{4}$:

$$t_{1/4}^{\text{mix}} \leq N^2\,.$$

For simple random walk on $\mathbb{Z}_N$ with holding probability $\frac{1}{2}$, a number of $n = N^2$ steps suffice for the random walk to be $\frac{1}{4}$-close to stationarity. For simple random walk without holding probability, $n = \frac{1}{2}N^2$ steps suffice. Notice that this result matches out result for simple symmetric random walk on $\mathbb{Z}_N$ that we have obtained in Section 11.2.3 with the use of eigenvalue techniques. □

11.4 Strong Stationary Times

Consider one of our running examples, simple random walk on the hypercube $\mathbb{Z}_2^n$, i.e., the set of binary n-tuples. To avoid periodicity, we consider the lazy version of the walk that puts holding probability $\frac{1}{2}$ on each state. We can think of this random walk as selecting a spot j uniformly at random from $\{1, 2, ..., n\}$, followed by independently selecting a bit b uniformly at random from $\{0, 1\}$. Then the current state is being updated in location j with bit b. Note that once every location j has been chosen at least once and updated ("refreshed" with a random bit), the chain is in uniform distribution. The random time T_{ref} it takes to refresh every location at least ones is a stopping time for the random walk on $\mathbb{Z}_2^n$. Note that at time T_{ref}, the chain is **in exact stationary distribution**. It should be intuitively clear that the distribution of this random time T_{ref} is related to the rate of convergence to stationarity for the chain. The goal of this section is to make this relationship precise.

Definition 11.4.1. *Consider a Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S}$ and transition matrix $\mathbf{P}$. Let $(Y_n)_{n\geq 1}$ be a sequence of i.i.d. random variables taking values in a state space $\mathcal{R}$ and let $f : \mathcal{S} \times \mathcal{R} \to \mathcal{S}$ be a function. If*

$$\mathbb{P}(f(x, Y_1) = y) = P_{xy} \quad \text{for all } x, y \in \mathcal{S}\,,$$

we call f together with $(Y_k)_{k\geq 1}$ a **random mapping representation** *of $(X_n)_{n\geq 0}$.*

We can directly verify that a random mapping representation in fact constructs the Markov chain: Given a random variable X_0 taking values in a state space $\mathcal{S}$, and given f and $(Y_n)_{n\geq 1}$ as defined in Definition 11.4.1, with X_0 independent of $(Y_n)_{n\geq 1}$, the process $(X_n)_{n\geq 0}$ defined by the recurrence relation

$$X_n = f(X_{n-1}, Y_n) \quad \text{for } n \geq 1 \tag{11.28}$$

is a Markov chain on state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. The auxiliary random variables $(Y_n)_{n\geq 1}$, together with X_0 which determines the initial distribution, are the underlying source of randomness that determine the evolution of the Markov chain $(X_n)_{n\geq 0}$.

Example 11.4.1. Recall the 2-state chain on state space $\mathcal{S} = \{0, 1\}$. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

for some fixed $a, b \in (0, 1)$. Consider a sequence $(U_n)_{n\geq 1}$ of i.i.d. random variables with uniform distribution on the interval $(0, 1)$. We define a function $f : \{0, 1\} \times (0, 1) \to \{0, 1\}$

by

$$\begin{aligned} f(0,u) &= 0 \quad \text{if } u \in (0, 1-a] \\ f(0,u) &= 1 \quad \text{if } u \in (1-a, 1) \\ f(1,u) &= 0 \quad \text{if } u \in (0, b] \\ f(1,u) &= 1 \quad \text{if } u \in (b, 1)\,. \end{aligned}$$

The function f and the sequence $(U_n)_{n\geq 1}$ define a random mapping representation for the 2-state chain. Using a similar approach, we can construct a random mapping representation for *any* finite state Markov chain. □

Example 11.4.2. Consider random walk $(X_n)_{n\geq 0}$ on a group G with step distribution μ. Let $(Y_n)_{n\geq 1}$ be an i.i.d. sequence of random variable taking values in G with $Y_1 \sim \mu$, and let f be a function $f : G \times G \to G$ defined by

$$f(x, g) = xg \quad \text{for } x, g \in G\,.$$

Then f and $(Y_n)_{n\geq 1}$ define a random mapping representation for the random walk. The i.i.d. sequence $(Y_n)_{n\geq 1}$ tracks the steps the random walk takes along its trajectory. □

Random mapping representations are not unique. For any given finite-state Markov chain, one can always construct more than one random mapping representation. The following example is an illustration.

Example 11.4.3. Recall 'lazy' random walk on the hypercube $\mathbb{Z}_2^k$ from Example 10.3.2. The group $(\mathbb{Z}_2^k, +)$ is the set of binary vectors $\mathbf{x} = (x_1, ..., x_k)$ of length k, together with component-wise addition mod 2. The step distribution μ for this random walk is defined by $\mu(0, ..., 0) = \frac{1}{2}$ and

$$\mu(1, 0, ..., 0) = \mu(0, 1, 0, ..., 0) = \cdots = \mu(0, ..., 0, 1) = \frac{1}{2k}\,.$$

The following are two random mapping representations.

(a) Consider an i.i.d. sequence of random variables $(Y_n)_{n\geq 1}$ that take values in the set $\mathcal{R} = \{0, 1, ..., k\}$ and have distribution $\mathbb{P}(Y_1 = 0) = \frac{1}{2}$, and $\mathbb{P}(Y_1 = i) = \frac{1}{2k}$ for $i = 1, ..., k$. We define a function $f : \mathbb{Z}_2^k \times \mathcal{R} \to \mathbb{Z}_2^k$ by

$$f(\mathbf{x}, 0) = \mathbf{x} \quad \text{and } f(\mathbf{x}, i) = (x_1, ..., x_i + 1, ..., x_k)$$

for $\mathbf{x} \in \mathbb{Z}_2^k$ and $i = 1, ..., k$.

(b) Consider two independent sequences of random variables $(U_n)_{n\geq 1}$ and $(V_n)_{n\geq 1}$. The sequence $(U_n)_{n\geq 1}$ is an i.i.d. sequence of Bernoulli random variables with uniform distribution on the set $\{0, 1\}$ (fair coin tosses). The sequence $(V_n)_{n\geq 1}$ is an i.i.d. sequence of random variables with uniform distribution on the set $\mathcal{V} = \{1, ..., k\}$. We define a function

$$f : \mathbb{Z}_2^k \times (\{0, 1\} \times \mathcal{V}) \to \mathbb{Z}_2^k$$

by

$$f(\mathbf{x},(0,i))=\mathbf{x} \quad \text{and } f(\mathbf{x},(1,i))=(x_1,...,x_i+1,...,x_k)$$

for $\mathbf{x}\in\mathbb{Z}_2^k$ and $i\in\mathcal{V}$. □

Definition 11.4.2. *Consider a Markov chain $(X_n)_{n\geq 0}$ and a stochastic process $(Y_n)_{n\geq 0}$. We say $(X_n)_{n\geq 0}$* **is adapted to** *$(Y_n)_{n\geq 0}$ if for all $m\geq 0$, X_m is a function of $Y_0, Y_1, ..., Y_m$.*

Note that a random mapping representation f and $(Y_n)_{n\geq 1}$ of a Markov chain $(X_n)_{n\geq 0}$ represents a special scenario in which $(X_n)_{n\geq 0}$ is adapted to another process. In this case, the Markov chain $(X_n)_{n\geq 0}$ is adapted to the process $(X_0, Y_1, Y_2, ...)$.

Definition 11.4.3. *Consider a Markov chain $(X_n)_{n\geq 0}$ that is adapted to a process $(Y_n)_{n\geq 0}$. A* **randomized stopping time** *T for $(X_n)_{n\geq 0}$ is a random variable taking values in $\mathbb{N}_0\cup\{\infty\}$ such that for all $m\in\mathbb{N}_0$, the event $\{T=m\}$ can be determined from the values of $Y_0, Y_1, ..., Y_m$.*

Notice that a randomized stopping time is a more general notion than a stopping time (recall Definition 1.4.1). Indeed, assume T is a stopping time for a Markov chain $(X_n)_{n\geq 0}$ which is adapted to a process $(Y_n)_{n\geq 0}$. Since the indicator random variable $\mathbb{1}_{\{T=m\}}$ is a function of $X_0, X_1, ..., X_m$, and in turn each random variable X_i is a function of $Y_0, Y_1, ..., Y_i$, the event $\{T=m\}$ is determined by $Y_0, Y_1, ..., Y_m$. Example 11.4.4 below gives an example of a randomized stopping time that is *not* a stopping time for the Markov chain.

Definition 11.4.4. *Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain with state space $\mathcal{S}$ and stationary distribution π. Let T be a randomized stopping time for $(X_n)_{n\geq 0}$. We say T is a* **strong stationary time** *if the distribution of the Markov chain at time T is π, that is, if*

$$\mathbb{P}(X_T=x)=\pi(x) \quad \text{for all } x\in\mathcal{S},$$

and if, in addition, X_T **is independent of** *T, that is,*

$$\mathbb{P}(T=n, X_T=x)=\mathbb{P}(T=n)\pi(x) \quad \text{for all } x\in\mathcal{S}, \text{ for all } n\geq 0.$$

The condition of independence in the above definition will be important (see the proof of Lemma 11.4.1 below). A randomized stopping time T for which $X_T\sim\pi$, but for which X_T and T are not independent, is called a **stationary time**.

Example 11.4.4. We return to random walk $(X_n)_{n\geq 0}$ on $\mathbb{Z}_2^k$. The stationary distribution π is uniform distribution on $\mathbb{Z}_2^k$. Consider the random mapping representation for the walk described in Example 11.4.3 (b). The randomized stopping time T_{ref} defined by

$$T_{\text{ref}} = \min\{m : \forall i \in \mathcal{V}, \exists j \leq m,\ s.t.\ V_j = i\}$$

is the first time all k coordinates have been "refreshed" (i.e. updated by a uniformly and independently chosen bit from $\{0, 1\}$) at least once. This is the cover time (recall Definition 8.5.1) for the process $(V_n)_{n\geq 1}$. Because of the independence of the $U_1, U_2, ..., V_1, V_2...$ as well as the uniformity of their distributions on their respective state spaces, it should be clear that at time T_{ref}, all binary k-vectors are equally likely to occur, independently of the value of T_{ref}. Hence T_{ref} **is a strong stationary time for** $(X_n)_{n\geq 0}$.
We point out that T_{ref} is however *not* a stopping time for the random walk $(X_n)_{n\geq 0}$. Since the walk has positive holding probability at each step, it is not possible to determine "from the surface", i.e. directly from the trajectories of $(X_n)_{n\geq 0}$, whether or not T_{ref} has ocurred.

Notice also that for the random mapping representation described in Example 11.4.3 (a), the (similarly defined) randomized stopping time T defined by

$$T = \min\{m : \forall i \in \{1, ..., k\}, \exists j \leq m,\ s.t.\ Y_j = i\}$$

is *not a strong stationary time* for the random walk on $\mathbb{Z}_2^k$. □

Lemma 11.4.1. *Let Let $(X_n)_{n\geq 0}$ be an irreducible, positive recurrent Markov chain with state space $\mathcal{S}$ and stationary distribution π. Let T be a strong stationary time. Then*

$$\mathbb{P}(T \leq n, X_n = x) = \mathbb{P}(T \leq n)\pi(x)\,.$$

Proof. We have

$$\begin{aligned}
\mathbb{P}(T \leq n, X_n = x) &= \sum_{k\leq n} \mathbb{P}(T = k, X_n = x) \\
&= \sum_{k\leq n}\sum_{y\in\mathcal{S}} \mathbb{P}(T = k, X_k, = y, X_n = x) \\
&= \sum_{k\leq n}\sum_{y\in\mathcal{S}} \mathbb{P}(T = k)\pi(y)P_{yx}^{n-k} \\
&= \sum_{k\leq n} \mathbb{P}(T = k)\pi(x) = \mathbb{P}(T \leq n)\pi(x)\,.
\end{aligned}$$

□

Note that we have used the independence of T and X_T in the third line in the above proof. We are now ready to state an upper bound in terms of the tail distribution of a strong stationary time T.

Proposition 11.4.2. *Consider an irreducible, positive recurrent Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{S}$ and stationary distribution π. Let $X_n \sim \mu_n$. If T is a strong stationary time, then*

$$\|\mu_n - \pi\|_{TV} \leq \mathbb{P}(T > n)\,.$$

Proof. If $\mu_n \neq \pi$, then there exists $A \subseteq \mathcal{S}$ such that $\mu_n(A) > \pi(A)$. Let A be such a set. We have

$$\begin{aligned}
\mathbb{P}(X_n \in A) &= \mathbb{P}(X_n \in A, T \leq n) + \mathbb{P}(X_n \in A, T > n) \\
&= \mathbb{P}(T \leq n)\pi(A) + \mathbb{P}(X_n \in A, T > n) \\
&\leq \pi(A) + \mathbb{P}(T > n)\,.
\end{aligned}$$

It follows that

$$\|\mu_n - \pi\|_{TV} = \max_{A\subseteq\mathcal{S}} |\mathbb{P}(X_n \in A) - \pi(A)| \leq \mathbb{P}(T > n)\,.$$

□

Example 11.4.5 (Top-to-random shuffle). Recall top-to-random shuffle from Section 10.2. We start with a perfectly ordered deck of n cards. At each step, the top card is taken off and inserted in a uniformly chosen random position. See Figure 11.12. This process $(X_n)_{n\geq 0}$ is a random walk on the symmetric group $\mathrm{S_n}$. The step distribution μ is uniform distribution on the set of cyclic permutations $C = \{id, \sigma_2, ..., \sigma_n\}$ as defined in (10.1).

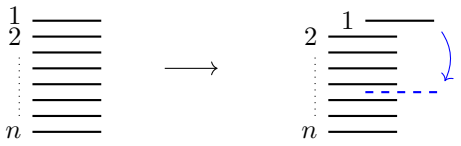


Figure 11.12: Top-to-random shuffle

The random time T= "the first time at which Card n has risen from the bottom to the top of the deck" is a stopping time for the random walk, since whether or not T has occurred at time n can be determined directly from the evolution of the random walk up

to time n. Thus $\tilde{T} = T + 1$ is a strong stationary time for the random walk: At time T, Card n is in top position and any arrangement of Cards 1 through $(n-1)$ below Card n is equally likely to occur. After one more time step, Card n has been inserted into the deck in a uniformly random position, and so at time $\tilde{T}$ any arrangement of the deck is equally likely to occur, and $X_{\tilde{T}}$ and $\tilde{T}$ are independent.
Note that if Card n is currently in location k of the deck, then the probability that Card n rises one spot upwards is $\frac{n-k+1}{n}$ for the next shuffle. So the waiting time for Card n to rise to location $(k-1)$ has a geometric distribution with success probability $\frac{n-k+1}{n}$. Clearly,

$$\tilde{T} = Y_n + Y_{n-1} + \cdots + Y_2 + 1$$

where $Y_k \sim \text{Geom}(\frac{n-k+1}{n})$ for $2 \leq k \leq n$. So $\tilde{T}$ has the same distribution as the waiting time in the Coupon collector's problem (see Section B.6). Using Proposition 11.4.2 and Lemma B.6.1, we get the same result as in Example 11.3.6: After $k = n \ln n + cn$ top-to-random shuffles of a deck of n cards, we have

$$\|\mu_k - \pi\|_{TV} \leq e^{-c}$$

which shows that we have for the mixing time t_ϵ^{mix},

$$t_\epsilon^{\text{mix}} \leq n \ln n + \ln(\epsilon^{-1})n \,.$$

□

Example 11.4.6 (Riffle shuffle). We have introduced riffle shuffling (and its time reversal) in Section 10.2. The below Figure 11.13 illustrates one possible step in this process.

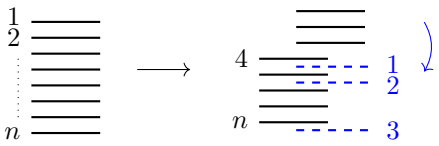


Figure 11.13: Riffle shuffle

Recall that by Lemma 10.1.5, the rate of convergence to stationarity for a random walk and the rate of convergence for its time reversal are the same with respect to total variation distance. We will make use of this result here, since working with the time reversal of riffle shuffling turns out to be easier than working with the original random walk. The time reversal of riffle shuffling (we will simply call it inverse shuffling) can be viewed as a type of sorting process. It proceeds in the following way: At each step, we mark each card with either 0 or 1, according to i.i.d. Bernoulli random variables. We then sort the

cards according to this marking by bringing all cards marked with 0 to the top of the pile, leaving their relative order at the time of the marking intact.
Keeping track of the 0 or 1 markings on each card over time produces a binary sequence on each card. After a finite number of k steps, each card is marked with a binary k-vector, and these binary k-vectors appear ordered in the deck from top to bottom with respect to right-to-left lexicographic order. Figure 11.14 gives an illustration for 5 cards and $k = 3$.

starting deck	step 1	step 2	step 3
1	0	1	0
2	1	1	1
3	1	0	1
4	0	1	1
5	0	0	0

deck after 3 inv. shuffles	sorted bits
5	$(0,0,0)$
1	$(0,1,0)$
3	$(1,0,1)$
4	$(0,1,1)$
2	$(1,1,1)$

Figure 11.14: Three inverse shuffles for a deck of 5 cards

The random walk $(X_n)_{n\geq 0}$ on $\mathrm{S_n}$ that is the so-described inverse shuffling is adapted to an i.i.d. sequence $(Y_n)_{n\geq 1}$ of random variables that have uniform distribution on the set of all binary vectors of length n (since the deck has n cards). The time T at which all n binary vectors of length T are distinct for the first time is a randomized stopping time for inverse shuffling. In fact, it is a strong stationary time for inverse shuffling: At time T, because of the independence of the 0–1 Bernoulli random variables, any arrangement of the deck is equally likely to occur, and X_T and T are independent.
Note that once T has occurred, from then onwards all (growing in length) binary vectors will be distinct at any future time as well.
In order to be able to apply Proposition 11.4.2, we need to estimate $\mathbb{P}(T > k)$, that is, the probability that after k inverse shuffles not all binary k-vectors are distinct. We have 2^k possible distinct rows. Each is equally likely to have occurred. So

$$\mathbb{P}(T > k) = 1 - \prod_{j=0}^{n-1} \frac{2^k - j}{2^k}\,. \tag{11.29}$$

We will now estimate for which values k (as a function of n) the right-hand side of (11.29) becomes small. Write

$$\prod_{j=0}^{n-1} \frac{2^k - j}{2^k} = \exp\left(\sum_{j=0}^{n-1} \ln\left(1 - \frac{j}{2^k}\right)\right)\,.$$

For small x, we can use $1 - x \approx e^{-x}$. From this we have

$$\prod_{j=0}^{n-1} \frac{2^k - j}{2^k} \approx \exp\left(-\sum_{j=0}^{n-1} \frac{j}{2^k}\right) \approx \exp(-\frac{n^2}{2 \cdot 2^k}) \,. \tag{11.30}$$

If we take $k = 2\log_2 n$ on the right-hand side of (11.30), we get

$$\exp(-\frac{n^2}{2 \cdot 2^{2\log_2 n}}) = e^{-1/2} \approx 0.607 \,.$$

We can improve things by taking $k = 2\log_2 n + c$ for some positive constant c instead. This yields

$$\exp(-\frac{n^2}{2 \cdot 2^{2\log_2 n + c}}) = e^{-1/2^{1+c}} \,,$$

which for $c = 3$ gives

$$e^{-1/16} \approx 0.94 \,.$$

Hence

$$\mathbb{P}(T > k) \approx 0.06 \,,$$

for $k = 2\log_2 n + 3$ (and of course an even smaller value for $k = 2\log_2 n + c$ with $c > 3$). Applying this estimate for the tail probability of T to Proposition 11.4.2, we conclude that after $k = 2\log_2 n + 3$ inverse riffle shuffles (and hence also riffle shuffles) the deck is reasonably close to random.

We will quote a result that improves on this estimate in the following section. □

11.5 The Cut-off phenomenon

In 1981, Diaconis and Shashahani [9] observed an interesting phenomenon while studying *random transpositions*, a type of card shuffling modeled as random walk on the symmetric group S_n: For this random walk, convergence to stationarity does not happen gradually, as one might expect, but rather abruptly around a certain *cut-off point*. They proved that, within a relatively small time interval around that cut-off point, total variation distance to uniformity drops from near 1 to near 0. Random transpositions shuffling proceeds in the following way: At at each step, two cards are chosen uniformly at random from the deck and their position is swapped. Diaconis and Shahshahani [9] proved that for random transpositions with a deck on n cards, the sharp drop in total variation distance happens within a relatively small time interval around time $k = \frac{1}{2} n \ln n$. Figure 11.15 below illustrates this type of convergence behavior.

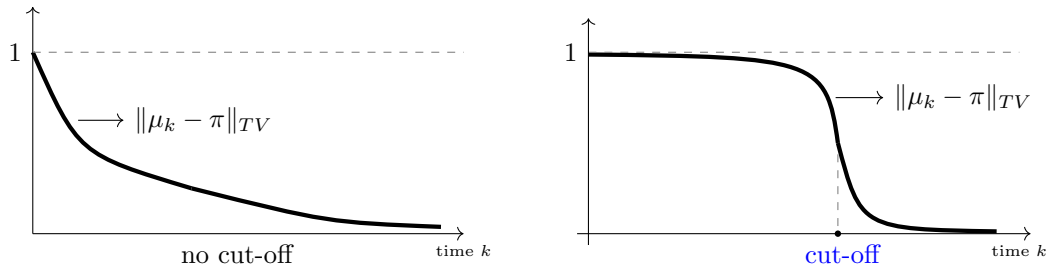


Figure 11.15: No cut-off versus a cut-off

Since this result was first published, the cut-off phenomenon has been proved (or disproved) for a number of Markov chains. This is an active area of research. The following example describes another random walk for which the cut-off phenomenon occurs.

Example 11.5.1 (Cut-off for riffle shuffle). We introduced riffle shuffling in Section 10.2 (for an illustration, see Figure 11.16) and found an upper bound for its rate of convergence to stationarity in Example 11.4.6. Here we quote a precise result that demonstrates the cut-off phenomenon for this random walk.

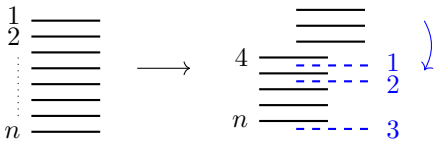


Figure 11.16: Riffle shuffling

This model was first analyzed by Aldous in [1] who found the asymptotic mixing time, and later by Bayer and Diaconis in [5] who found an exact formula for its distance to stationary which sharpened the result from [1]. The below table in Figure 11.17 gives the exact values of total variation distance for a deck of $n = 52$ cards. It is taken from [5].

k	1	$\cdots$	4	5	6	7	8	9	10
$\|\mu_k - \pi\|_{TV}$	1	$\cdots$	1	0.924	0.614	0.334	0.167	0.085	0.043

Figure 11.17: Total variation distance for riffle shuffling for a deck of 52 cards

A plot of this data in Figure 11.18 more clearly displays the sharp drop-off.
In [5], Bayer and Diaconis give a precise formula for for total variation distance to stationarity for a deck of n cards. They show that, for large n, for a number of

$$k = \frac{3}{2}\log_2 n + c$$

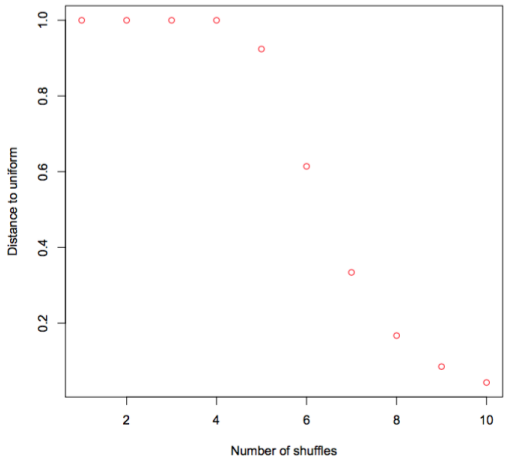


Figure 11.18: Plot of the data in Figure 11.17

shuffles, where c is a positive or negative constant,

$$\|\mu_k - \pi\|_{TV} = 1 - 2\Phi\left(\frac{-2^{-c}}{4\sqrt{3}}\right) + O\left(\frac{1}{n^{1/4}}\right) . \tag{11.31}$$

Here

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx$$

is the cumulative distribution function of a standard normal random variable. We can verify from (11.31) that for large n, if we choose a sufficiently large c, then the expression in (11.31) will be near 0. And if we choose a negative c that is sufficiently large in absolute value, the expression in (11.31) will be close to 1. This shows that a number of $\frac{3}{2}\log_2 n + c$ shuffles are **necessary and sufficient to mix the deck** of n cards. See Figure 11.19.

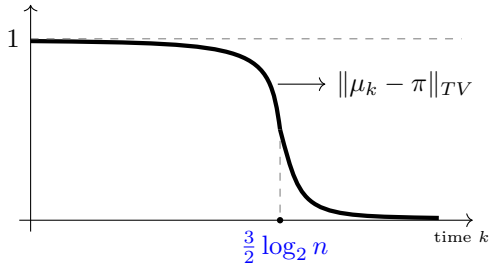


Figure 11.19: Cut-off at $k = \frac{3}{2}\log_2 n$ steps for riffle shuffling a deck of n cards

We point out that for most examples, a precise formula for total variation distance, such as in (11.31), is not available. □

We now give a precise mathematical definition of the cut-off phenomenon. For this, consider a natural sequence of Markov chains $(X_k^{(n)})_{k\geq 0}$ with $n \geq 1$ for which the size of the state space increases in a natural way, and for which the transition probabilities for each of the growing state spaces can be defined in an analogous way. For example, consider riffle shuffling a deck of n cards. For a given n, the size of the state space is $|\mathrm{S_n}| = \mathrm{n!}$. With growing n, the size of the state space grows accordingly. We regard n as the size parameter for the state space $\mathrm{S_n}$. The transition probabilities for riffle shuffles are defined in an analogous way for *any* n. Our viewpoint is a simultaneous time and space asymptotic: We want to understand the **mixing time** for a Markov chain, such as for riffle shuffling, **as a function of the size parameter n of the state space**.

Notation: For such a sequence of Markov chains $(X_k^{(n)})_{k\geq 0}$, we denote the stationary distribution for the Markov chain with size parameter n by $\pi^{(n)}$. And we denote the distribution at time k of the Markov chain with size parameter n by $\mu_k^{(n)}$.

Definition 11.5.1. *Consider a sequence of Markov chains $(X_k^{(n)})_{k\geq 0}$ whose state spaces have size parameter n. Let $t(n)$ and $w(n)$ be two nonnegative functions with*

$$\lim_{n\to\infty} t(n) = \infty \qquad \text{and} \qquad \lim_{n\to\infty} \frac{w(n)}{t(n)} = 0\,.$$

We say the Markov chain has a **cut-off at** *$t(n)$ if*

$$\lim_{n\to\infty} \|\mu^{(n)}_{t(n)+cw(n)} - \pi^{(n)}\|_{TV} = f(c)$$

where $f(c)$ is a function with

$$\lim_{c\to-\infty} f(c) = 1 \qquad \text{and} \qquad \lim_{c\to\infty} f(c) = 0\,.$$

The function $w(n)$ is called the **window** for the cut-off. In "little-oh" notation, we have $w(n) = o(t(n))$. For the above example of riffle shuffling of a deck of n cards, the cut-off is at $t(n) = \frac{3}{2}\log_2 n$, and the window for the cut-off is $w(n) = 1$.
Phenomenon reveals itself in a joint time and space asymptotic.

The following example proves a cut-off phenomenon for top-to-random shuffling.

Example 11.5.2 (Cut-off for the top-to-random shuffle). Recall Examples 11.3.6 and 11.4.5. Both show that for $k = n\ln n + cn$, for top-to-random shuffles of n cards, we have

$$\|\mu_k^{(n)} - \pi^{(n)}\|_{TV} \leq e^{-c}\,. \tag{11.32}$$

For sufficiently large c, the value e^{-c} will be close to 0. In order to prove a cut-off phenomenon, we also need to prove a matching lower bound for $t(n) = n \ln n$ and $w(n) = n$.

A matching lower bound for top-to-random shuffle.

In order to find a lower bound for total variation distance at time k, a successful approach can be to find a "bad" set A, in the sense that $|\mathbb{P}(X_k^{(n)} \in A) - \pi^{(n)}(A)|$ is still fairly large, and so total variation distance will also be still large. We will construct such a set. Let

$$A = \{\sigma \in \mathrm{S_n} : \text{Card n is above Card } (\mathrm{n}-1)\}\,.$$

Note that $\pi^{(n)}(A) = \frac{1}{2}$. We will show that for $k = n \ln n - 3n$,

$$\mathbb{P}(X_k^{(n)} \in A) < \frac{1}{4}\,,$$

and thus

$$\|\mu_k^{(n)} - \pi^{(n)}\|_{TV} > \frac{1}{4}\,.$$

Recall that initially the deck is in perfect order. So for event A to happen, Card $(n-1)$ must first rise to the top and then be inserted back into the deck somewhere below Card n. It follows that

$$\mathbb{P}(X_k^{(n)} \in A) \leq \mathbb{P}(T < k)$$

where T is the stopping time "Card $(n-1)$ has risen to the top of the pile". We will show that $\mathbb{P}(T < k) < \frac{1}{4}$. Towards this end, we write

$$T = \sum_{j=2}^{n-1} T_j$$

where T_j is the random time it takes for Card $(n-1)$ to rise from spot $(n-j+1)$ to spot $(n-j)$ in the pile. Note that the random variable T_j has a geometric distribution with success parameter $p_j = \frac{j}{n}$ (since the current top card must be inserted below Card $(n-1)$ for Card $(n-1)$ to rise by one spot).

Recall that for a geometric random variable $X \sim Geom(p)$, we have $\mathbb{E}(X) = \frac{1}{p}$ and $\mathrm{Var}(X) = \frac{1-p}{p^2}$. So we have

$$\mathbb{E}(T_j) = \frac{n}{j} \text{ and } \mathrm{Var}(T_j) \leq \frac{n^2}{j^2} \qquad \text{for } j = 2, 3, \ldots n-1\,.$$

We get

$$\mathbb{E}(T) = \sum_{j=2}^{n-1} \mathbb{E}(T_j) = n \sum_{j=2}^{n-1} \frac{1}{j} \;\geq\; n(\ln n - 1)\,, \tag{11.33}$$

where for the last inequality we have used the fact $\ln n \leq \sum_{j=1}^{n-1} \frac{1}{j}$. For an estimate for $\mathrm{Var}(T)$, recall (B.4). From it we get

$$\mathrm{Var}(T) = \sum_{j=2}^{n-1} \mathrm{Var}(T_j) \leq n^2 \sum_{j=2}^{\infty} \frac{1}{j^2} \leq \frac{2}{3} n^2 \,.$$

Recall that our goal is to show that $\mathbb{P}(k > T) < \frac{1}{4}$. Let $k = n \ln n - 3n$. This yields

$$\begin{aligned} \mathbb{P}(n \ln n - 3n > T) &= \mathbb{P}((n \ln n - n) - T - 2n > 0) \\ &\leq \mathbb{P}(\mathbb{E}(T) - T - 2n > 0) = \mathbb{P}(\mathbb{E}(T) - T > 2n) \end{aligned}$$

where the above inequality is due to (11.33). Applying Chebychev's Inequality to the random variable T and the probability $\mathbb{P}(\mathbb{E}(T) - T > 2n)$, we get

$$\mathbb{P}(\mathbb{E}(T) - T > 2n) \leq \frac{\mathrm{Var}(T)}{4n^2} \leq \frac{\frac{2}{3}n^2}{4n^2} < \frac{1}{4} \,.$$

Hence, for $k = n \ln n - 3n$, we get

$$\mathbb{P}(X_k^{(n)} \in A) < \frac{1}{4} \,,$$

and consequently

$$|P(X_k^{(n)} \in A) - \pi^{(n)}(A)| > \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \,.$$

This implies that for $k = n \ln n - 3n$,

$$\|\mu_k^{(n)} - \pi^{(n)}\|_{TV} > \frac{1}{4} \,, \tag{11.34}$$

and so total variation distance to stationarity is still fairly large for $k = n \ln n - 3n$ steps. See Figure 11.20 for an illustration. Together, inequalities (11.32) and (11.34) prove a cut-off phenomenon at $t(n) = n \ln n$ with window $w(n) = n$. □

Example 11.5.3 (No cut-off for random walk on the discrete N-cycle). Recall Section 11.2.3 where we have shown that for simple random walk on $\mathbb{Z}_N$, we have

$$\|\mu_k^{(N)} - \pi^{(N)}\|_{TV} < 0.037$$

for $k \geq N^2$. It is known (see [7], [8]) that for large N,

$$\|\mu_k^{(N)} - \pi^{(N)}\|_{TV} \approx f\left(\frac{k}{N^2}\right)$$

where f is a positive, decreasing, and continuous function with $f(0) = 1$ and $\lim_{x \to \infty} f(x) = 0$. This shows that the decay of total variation distance for simple random walk on $\mathbb{Z}_N$ happens gradually at around $k = N^2$ steps, without a sharp cut-off. See Figure 11.21. □

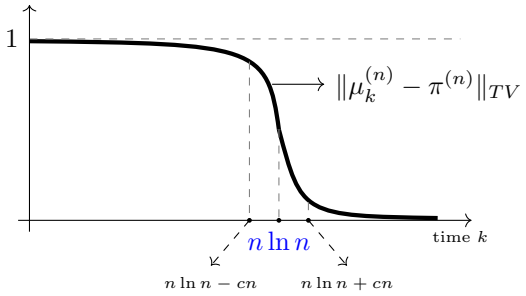


Figure 11.20: Cut-off at $k = n \ln n$ steps for top-to-random shuffling a deck of n cards

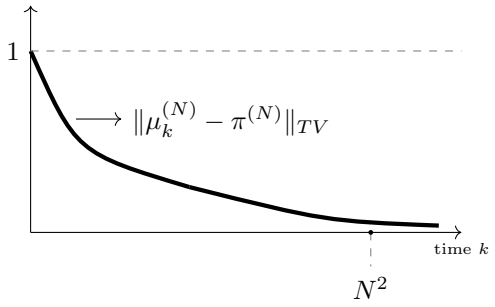


Figure 11.21: No sharp cut-off for total variation distance for random walk on $\mathbb{Z}_N$

Exercises

Exercise 11.1. Consider a state space $\mathcal{S}$ and the space $\mathcal{P}$ of all probability distributions on $\mathcal{S}$. Show that the function $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ defined by

$$d(\mu, \nu) = \|\mu - \nu\|_{TV}$$

for all $\mu, \nu \in \mathcal{P}$ defines a *metric* on $\mathcal{P}$.

Exercise 11.2. Consider an ergodic Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S}$ with stationary distribution π. Assume $(X_n)_{n\geq 0}$ is *lumpable* (see Section 1.7) with respect to a partition $\mathcal{A} = \{A_1, A_2, ...\}$ of $\mathcal{S}$. We denote the lumped chain by $(\langle X_n \rangle)_{n\geq 0}$ and its stationary distribution on $\mathcal{A}$ by $\widehat{\pi}$. Show that for all $n \geq 1$,

$$\|\widehat{\mu}_n - \widehat{\pi}\|_{TV} \leq \|\mu_n - \pi\|_{TV}$$

where μ_n and $\widehat{\mu}_n$ denote the distributions of the processes $(X_n)_{n\geq 0}$ and $(\langle X_n \rangle)_{n\geq 0}$ at time n, respectively.

Exercise 11.3. Consider the Ehrenfest chain $(Y_n)_{n\geq 0}$ for N particles and its *lift* $(X_n)_{n\geq 0}$ to simple random walk on the hypercube $\mathbb{Z}_2^N$. More precisely, $(Y_n)_{n\geq 0}$ is the lumped

version of $(X_n)_{n\geq 0}$ under Hamming weight (recall Example 1.7.4). Assume $Y_0 = 0$ and let $Y_n \sim \widehat{\mu}_n$ and $X_n \sim \mu_n$ for $n \geq 0$. Show that the total variation distance to stationarity is the same for the Ehrenfest chain and for its lift to the hypercube, that is, show that

$$\|\widehat{\mu}_n - \mathrm{Bin}(N, \frac{1}{2})\|_{TV} = \|\mu_n - \mathrm{Unif}(\mathbb{Z}_2^N)\|_{TV}.$$

Exercise 11.4. Consider a chain graph G with N vertices (see Figure 11.22) and simple random walk on this graph. Find the eigenvalues and their multiplicities for this random walk. (*Hint*: View the random walk as a lumped version of another Markov chain. Recall Example 1.7.5.)

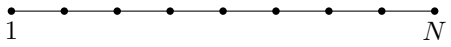


Figure 11.22

Exercise 11.5. Assume $\mathcal{S}$ is a countably infinite state space. Let μ_n, $n \geq 0$, be a sequence of probability measures on $\mathcal{S}$ as well as π be a probability measures on $\mathcal{S}$. Prove that

$$\lim_{n\to\infty} \|\mu_n - \pi\|_{TV} = 0 \quad \Longleftrightarrow \quad \lim_{n\to\infty} \mu_n(x) = \pi(x) \ \ \forall x \in \mathcal{S}\,.$$

Exercise 11.6. Let $\mathcal{S}$ be a finite state space and $\mathbf{P}$ the transition matrix for an irreducible, aperiodic Markov chain on $\mathcal{S}$. Let π be the stationary distribution for the chain. Consider the following distributions $\mu_n^{(x)}$ on $\mathcal{S}$: For $n \geq 0$ and $x \in \mathcal{S}$, let $\mu_n^{(x)}$ denote the distribution of $X_n^{(x)}$ where $(X_n^{(x)})_{n\geq 0}$ is the $\mathbf{P}$-Markov chain starting in state x. Furthermore, for $n \geq 0$, μ_n is the distribution of X_n where $(X_n)_{n\geq 0}$ is the $\mathbf{P}$-Markov chain starting in initial distribution μ_0. Prove that

$$\|\mu_n - \pi\|_{TV} \leq \max_{x\in\mathcal{S}} \|\mu_n^{(x)} - \pi\|_{TV}\,.$$

Exercise 11.7. Let $\mathcal{S}$ be a discrete space. Proposition 11.3.2 states that for any two distributions μ and ν on $\mathcal{S}$, there exists an *optimal coupling* (X, Y) in the sense that

$$\mathbb{P}(X \neq X) = \|\mu - \nu\|_{TV}\,.$$

Let $B = \{x \in \mathcal{S} : \mu(x) > \nu(x)\}$. Consider the following probabilities:

$$\begin{aligned}
\mathbb{P}((X,Y) = (x,x)) &= \min\{\pi(x), \nu(x)\} \quad \text{for } x \in \mathcal{S} \\
\mathbb{P}((X,Y) = (x,y)) &= \frac{(\mu(x) - \nu(x))(\nu(y) - \mu(y))}{\|\mu - \nu\|_{TV}} \quad \text{for } x \in B \text{ and } y \in B^c \\
\mathbb{P}((X,Y) = (x,y)) &= 0 \quad \text{otherwise}\,.
\end{aligned}$$

Show that the given probabilities define a coupling of μ and ν and that this coupling is optimal.

Exercise 11.8. Let $\mathcal{S} = \{0, 1, 2\}$. Consider the two distributions $\mu \sim \text{Unif}(\mathcal{S})$ and ν defined by $\nu(0) = \nu(1) = \frac{1}{2}$ and $\nu(2) = 0$ on $\mathcal{S}$. Give the joint distribution for an optimal coupling of μ and ν.

Exercise 11.9. Consider an ergodic Markov chain $(X_n)_{n\geq 0}$ with stationary distribution π. Fix $\epsilon > 0$, and recall the definition of the mixing time t_ϵ^{mix} (Definition 11.1.2). Using the same notation as in Theorem 11.3.5, show that

$$t_\epsilon^{\text{mix}} \leq \epsilon^{-1} \max_{x,y\in\mathcal{S}} \mathbb{E}^{(x,y)}(T_{\text{couple}}) \,.$$

Exercise 11.10. Verify that the probabilities given in Example 11.3.5 define a coupling for two $\mathbf{P}$-chains $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$.

Exercise 11.11. Give an example of a coupling of two Markov chains for which T_{couple} is not a finite random variable, that is, for which $\mathbb{P}(T_{\text{couple}} < \infty) < 1$.

Exercise 11.12. Consider an ergodic Markov chain $(X_n)_{n\geq 0}$ with transition matrix $\mathbf{P}$ on state space $\mathcal{S}$. Let π be its unique stationary distribution and let μ_n denote the distribution of X_n for $n \geq 1$. Use a coupling argument to prove convergence with respect to total variation distance. That is, prove that for any initial distribution μ_0,

$$\|\mu_n - \pi\|_{TV} \xrightarrow{n\to\infty} 0 \,.$$

(*Hint*: Couple $(X_n)_{n\geq 0}$ with a second $\mathbf{P}$-chain $(Y_n)_{n\geq 0}$ that starts in π. The key part is to prove that $\mathbb{P}(T_{\text{couple}} < \infty) = 1$.)

Exercise 11.13. Consider simple symmetric random walk on the N-cycle $\mathbb{Z}_N$ (the integers mod N) without holding probability. Let $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ be two such random walks with $(X_n)_{n\geq 0}$ starting in $x \in \mathbb{Z}_N$ and $(Y_n)_{n\geq 0}$ starting in $y \in \mathbb{Z}_N$. We assume $x \neq y$. Construct three distinct couplings $(X_n, Y_n)_{n\geq 0}$ for which, in each case, the source of randomness is a sequence of independent rolls of a 6-sided fair die.

Exercise 11.14. Consider simple biased random walk on $\mathbb{Z}_{20}$ (the integers mod 20) with transition probabilities $P_{x,x+1} = \frac{2}{3}$ and $P_{x,x-1} = \frac{1}{3}$. Let $(X_n)_{n\geq 0}$ be a copy of the random walk that starts in state 1, and let $(Y_n)_{n\geq 0}$ be a copy of the random walk that starts in state 9.

(a) Construct a bivariate process $(X_n, Y_n)_{n\geq 0}$ in the following way: At each time step, roll a fair 6-sided die. Assuming the current state is (x, y), if the die shows 1 or 2, the process moves to $(x + 1, y - 1)$. If the die shows 3 or 4, the process moves to $(x+1, y+1)$. And if the die shows 5 or 6, the process moves to $(x-1, y+1)$. Show that this is a coupling for $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$.

(b) For the coupling from part (a), compute $\mathbb{E}(T_{\text{couple}})$, that is, the expected time until the two random walks hit the same state for the first time.

Exercise 11.15. Consider a finite-state Markov chain $(X_n)_{n\geq 0}$ on state space $\mathcal{S}$ with transition matrix $\mathbf{P}$. Construct a random matrix representation for $(X_n)_{n\geq 0}$ using an i.i.d. sequence $(U_n)_{n\geq 1}$ of random variables with uniform distribution on the unit interval.

Exercise 11.16. Recall lazy random walk on the hypercube $\mathbb{Z}_2^k$ and its strong stationary time T_{ref} that we have introduced in Example 11.4.4. Let $c > 0$.

(a) Use the results from Section 11.4 to show that for this walk, after $n = k \ln k + ck$ steps, the distance to stationarity (with respect to total variation distance) is at most e^{-c}. (*Hint*: The result from Appendix B.6 may be useful.)

(b) Use the result from part (a) to show that for lazy random walk on the hypercube $\mathbb{Z}_2^k$,

$$t_\epsilon^{\text{mix}} \leq k \ln k + \ln(\epsilon^{-1})k\,.$$

Appendix A

A.1 Miscellaneous

Binomial Identities. We assume all integers are nonnegative.

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

Recursion:

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$$

Diagonal sums for $0 \leq k \leq n$:

$$\binom{k}{k} + \binom{k+1}{k} + \binom{k+2}{k} + \cdots + \binom{n}{k} = \binom{n+1}{k+1}$$

Vandermonde's Identity:

$$\sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j} = \binom{m+n}{k}$$

Stirling's approximation.

Stirlings approximation for factorials states

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1\,.$$

For large $n \in \mathbb{N}$, we can use the approximation

$$n! \approx n^n e^{-n}\sqrt{2\pi n}\,. \tag{A.1}$$

A.2 Bipartite graphs

Definition A.2.1. *Let $G(V, E)$ be an undirected graph. (a) We call the graph* **bipartite** *if there exists a partition $V = V_1 \cup V_2$ such that for any two vertices $v, w \in V$,*

$$\{v, w\} \in E \iff v \in V_1 \text{ and } w \in V_2 \quad (\text{or vice versa}) .$$

(b) We call a sequence of vertices $v_0\, v_1 \cdots v_{n-1}\, v_n$ a **path of length** *n if*

- *$v_i \neq v_j$ for $i \neq j$ and $0 \leq i, j \leq n$, and*
- *$\{v_i, v_{i+1}\} \in E$ for $0 \leq i \leq n-1$.*

(c) We call a sequence of vertices $v_0\, v_1 \cdots v_{n-1}\, v_0$ a **cycle of length** *n if $n \geq 3$ and*

- *$v_0\, v_1 \cdots v_{n-1}$ is a path of length $(n-1)$ and $\{v_{n-1}, v_0\} \in E$.*

Proposition A.2.1 (Bipartite graphs)**.** *Let $G(V, E)$ be a connected graph. $G(V, E)$ is bipartite if and only if it does not contain an odd-length cycle.*

Proof. Assume $G(V, E)$ is bipartite with $V = V_1 \cup V_2$. Assume $v_0 \in V_1$, and let $v_0\, v_1 \cdots v_{n-1}\, v_0$ be a cycle. Since the graph is bipartite, it must be that $v_0, v_2, v_4, ..., v_{n-2} \in V_1$ and that $n-2$ is even. It follows that n is also even, and hence the length of the cycle is even.
Conversely, assume that $G(V, E)$ does not contain an odd-length cycle. For any two vertices v and w, let $d(v, w)$ be the *distance* of v and w defined as the minimum path length among all possible paths from v to w. If there is no path from v to w, we set $d(v, w) = \infty$. Now fix a vertex $v \in V$. Consider the set $W = \{w \in V \,:\, d(v, w) \text{ is odd}\}$. Clearly, W is not empty since the graph is connected and so there is at least one edge emanating from v. If W contains more than one vertex, take two distinct $w_1, w_2 \in W$. We have two (minimal-length) paths $v\, s_1\, s_2 \cdots s_{k-1}\, w_1$ and $v\, u_1\, u_2 \cdots u_{l-1}\, w_2$. Let $\tilde{k}$ and $\tilde{l}$ be the largest indices for which $s_{\tilde{k}} = u_{\tilde{l}}$. Since both paths are of minimal length, it must follow that $\tilde{k} = \tilde{l}$. Now assume that $\{w_1, w_2\} \in E$. But then the resulting cycle

$$s_{\tilde{k}}\, s_{\tilde{k}+1} \cdots w_1\, w_2\, u_{l-1} \cdots u_{\tilde{l}}$$

is of odd length which contradicts our assumption for the graph. Hence we must dismiss the assumption that there exist two vertices $w_1, w_2 \in W$ that are joined by an edge.
The same argument can be applied to the set $W^c = V \setminus W$, and we conclude that no two vertices $r_1, r_2 \in W^c$ are joined by and edge. It follows that the graph $G(V, E)$ is bipartite for the partition $V = W \cup W^c$. □

A.3 Schur's theorem

Theorem A.3.1 (Schur's theorem)**.** *Let $C \subset \mathbb{N}$ such that $\gcd(C) = c$. Then there exists an integer N (depending on the set C) such that for all $n \geq N$ we can write cn as a linear combination of elements of C with nonnegative integer coefficients.*

Proof. Note that it suffices to prove the theorem for the case $\gcd(C) = 1$. For if $\gcd(C) = c > 1$, we can factor out c from each element in C to end up with a set C' whose elements are relative prime. The result for C' then implies the statement for C by multiplication of each linear combination by the factor c.

Assume $\gcd(C) = 1$. Let $L(C)$ denote the set of all values that can be computed as a linear combination of elements in C with nonnegative integer coefficients. We first show that there exist two consecutive positive integers $m, m+1 \in L(C)$. Assume this is not true. So there exists $d \geq 2$ such that $|m_1 - m_2| \geq d$ for all $m_1, m_2 \in L(C)$, and furthermore, there exist $m, m+d \in L(C)$. Since $\gcd(C) = 1$, there exists at least one integer $n \in C$ for which d is not a divisor. Hence we can write

$$n = k\,d + r$$

for some $k \geq 0$ and for a remainder $1 \leq r < d$. Clearly, both $(k+1)(m+d)$ and $n + (k+1)m$ are elements of $L(C)$. However their difference is

$$(k+1)(m+d) - (n + (k+1)m) = d - r < d\,,$$

which contradicts the assumption that d is the smallest difference in absolute value between two elements in $L(C)$. It follows that $L(C)$ must contain two consecutive positive integers m and $m+1$.

We now claim that the statement of the theorem follows for $N = m^2$. Indeed, assume $n \geq m^2$. We can write

$$n - m^2 = lm + s$$

for some $l \geq 0$ and for a remainder $0 \leq s < m$. Clearly, both $s(m+1)$ and $(m - s + l)m$ are elements of $L(C)$, and therefore their sum

$$s(m+1) + (m - s + l)m = n$$

is also in $L(C)$. □

A.4 Iterated double series

Definition A.4.1.

- *A double sequence $(a_{ij})_{i,j\geq 1}$ of real or complex numbers is a function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ (or $\mathbb{C}$) where we set $f(i,j) = a_{ij}$.*
- *Let $(a_{ij})_{i,j\geq 1}$ be a double sequence. An iterated series is an expression of the form*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} \quad \text{or} \quad \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}\,.$$

- *We say the iterated series $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij}$ converges to $A \in \mathbb{R}$ (or $\mathbb{C}$) if for each $j \geq 1$,*

$$\sum_{i=1}^{\infty} a_{ij} = A_j$$

for some $A_j \in \mathbb{R}$ (or $\mathbb{C}$), and

$$\sum_{j=1}^{\infty} A_j = A\,.$$

Theorem A.4.1 (Fubini's theorem for series)**.** *Consider a (real or complex) double sequence $(a_{ij})_{i,j\geq 1}$. If the iterated series*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |a_{ij}|$$

converges, then both $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij}$ and $\sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}$ converge and

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}\,.$$

For a proof see [31] (Theorem 8.3).

Corollary A.4.2. *Let $(a_{ij})_{i,j\geq 1}$ be a double sequence of nonnegative real numbers. Then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}$$

(both sides may be infinite).

A (finite or infinite) stochastic matrix $\mathbf{P}$ is a square matrix

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & \cdots & \cdots \\ \vdots & \vdots & & \\ \vdots & \vdots & & \end{pmatrix}$$

for which $P_{ij} \geq 0$ for all $i, j \geq 0$ and $\sum_{j \geq 0} P_{ij} = 1$ for all $i \geq 0$.

Corollary A.4.3. *The product of two (finite and of equal dimension, or infinite) stochastic matrices* $\mathbf{P}$ *and* $\tilde{\mathbf{P}}$ *is a stochastic matrix.*

Proof. The (i, j)th entry of the matrix product $\mathbf{P}\tilde{\mathbf{P}}$ is

$$(\mathbf{P}\tilde{\mathbf{P}})_{ij} = \sum_{k \geq 0} P_{ik} \tilde{P}_{kj} \geq 0 \,.$$

For finite matrices $\mathbf{P}$ and $\tilde{\mathbf{P}}$, it is clear that $\sum_{j \geq 0} \sum_{k \geq 0} P_{ik} \tilde{P}_{kj} = 1$ since changing the order of summation is not an issue. Assume both matrices are infinite matrices. Since we have

$$\sum_{k \geq 0} \sum_{j \geq 0} P_{ik} \tilde{P}_{kj} = 1 \,,$$

we can apply Theorem A.4.1 and get for the row sum of the ith row

$$\sum_{j \geq 0} (\mathbf{P}\tilde{\mathbf{P}})_{ij} = \sum_{j \geq 0} \sum_{k \geq 0} P_{ik} \tilde{P}_{kj} = 1 \,.$$

□

Corollary A.4.4. *Matrix multiplication for infinite stochastic matrices is associative.*

Proof. Let $\mathbf{P}$, $\tilde{\mathbf{P}}$, and $\mathbf{P}'$ be three infinite stochastic matrices. We need to show that $[(\mathbf{P}\tilde{\mathbf{P}})\mathbf{P}']_{ij} = [\mathbf{P}(\tilde{\mathbf{P}}\mathbf{P}')]_{ij}$. Since

$$[(\mathbf{P}\tilde{\mathbf{P}})\mathbf{P}']_{ij} = \sum_{n \geq 0} \left(\sum_{k \geq 0} P_{ik} \tilde{P}_{kn} \right) P'_{nj} = \sum_{n \geq 0} \sum_{k \geq 0} P_{ik} \tilde{P}_{kn} P'_{nj}$$

is a convergent double sum (it is the (i, j)th entry of a stochastic matrix) and all terms in the sum are nonnegative, we can apply Theorem A.4.1 and conclude that

$$\sum_{n \geq 0} \sum_{k \geq 0} P_{ik} \tilde{P}_{kn} P'_{nj} = \sum_{k \geq 0} \sum_{n \geq 0} P_{ik} \tilde{P}_{kn} P'_{nj} = \sum_{k \geq 0} P_{ik} \left(\sum_{n \geq 0} \tilde{P}_{kn} P'_{nj} \right) = [\mathbf{P}(\tilde{\mathbf{P}}\mathbf{P}')]_{ij}$$

from which we get

$$[(\mathbf{P}\tilde{\mathbf{P}})\mathbf{P}']_{ij} = [\mathbf{P}(\tilde{\mathbf{P}}\mathbf{P}')]_{ij}\,.$$

□

A.5 Infinite products

Definition A.5.1. *Let $(c_j)_{j\geq 1}$ be a sequence of* positive *constants. Consider the sequence of successive products $(\prod_{j=1}^{n} c_j)_{n\geq 1}$. We say the infinite product $\prod_{j=1}^{\infty} c_j$ exists if the sequence of successive products converges to a finite* positive *number:*

$$0 < \lim_{n\to\infty} \prod_{j=1}^{n} c_j = \prod_{j=1}^{\infty} c_j < \infty\,.$$

Lemma A.5.1. *Let $c_j = 1+\epsilon_j$ (resp. $c_j = 1-\epsilon_j$) with $0 \leq \epsilon_j < 1$ for all $j \geq 1$. Then $\prod_{j=1}^{\infty} c_j = \prod_{j=1}^{\infty}(1+\epsilon_j)$ (resp. $\prod_{j=1}^{\infty}(1-\epsilon_j)$) exists if an only if $\sum_{j=1}^{\infty} \epsilon_j$ converges.*

Proof. First, recall the Limit Comparison Test for positive infinite series: Let $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ be two infinite series with $a_n, b_n > 0$ for all $n \geq 1$. If

$$\lim_{n\to\infty} \frac{a_n}{b_n} = c$$

with $0 < c < \infty$, then either both series diverge or both series converge.

Note that the infinite product $\prod_{j=1}^{\infty}(1+\epsilon_j)$ converges if and only if $\sum_{j=1}^{\infty} \ln(1+\epsilon_j)$ converges to a finite number. Furthermore, $\epsilon_j = 0 \iff \ln(1+\epsilon_j) = 0$. Assume $\lim_{j\to\infty} \epsilon_j = 0$ and consider the subsequence ϵ'_k of strictly positive values. Since

$$\lim_{k\to\infty} \frac{\ln(1+\epsilon'_k)}{\epsilon'_k} = 1\,,$$

by the Limit Comparison Test, $\prod_{j=1}^{\infty}(1+\epsilon_j)$ converges if and only if $\sum_{j=1}^{\infty} \epsilon_j$ converges.

For the case $\prod_{j=1}^{\infty}(1-\epsilon_j)$ a similar argument applies. We can apply the Limit Comparison Test to the series $(-1)\sum_{j=1}^{\infty} \ln(1-\epsilon_j)$ and $\sum_{j=1}^{\infty} \epsilon_j$.

□

A.6 The Perron–Frobenius Theorem

This is a classical theorem from Linear Algebra. It has several parts. Here we quote the parts that are most relevant for our purposes. For a reference see [32] or [18].

Theorem A.6.1 (Perron–Frobenius, Part I)**.** *Let* $\mathbf{P}$ *be strictly positive* $(n \times n)$*-matrix. Then the following statements hold.*

(a) There exists a positive real number λ^**, called the Perron–Frobenius eigenvalue, such that* λ^* *is an eigenvalue of* $\mathbf{P}$*, and for all other eigenvalues* λ *of* $\mathbf{P}$ *we have*

$$|\lambda| < \lambda^* .$$

(b) The Perron–Frobenius eigenvalue λ^* *satisfies*

$$\min_i \sum_j P_{i,j} \leq \lambda^* \leq \max_i \sum_j P_{i,j} .$$

(c) The algebraic multiplicity, and therefore also the geometric multiplicity, of λ^* *is one. In particular, the eigenspace corresponding to the Perron–Frobenius eigenvalue* λ^* *is one-dimensional.*

(d) There exists a left eigenvector $\mathbf{v}$ *corresponding to eigenvalue* λ^* *for which all entries* v_j*,* $1 \leq j \leq n$*, are strictly positive. There also exists a right eigenvector* $\mathbf{w}^{\mathrm{t}}$ *corresponding to* λ^* *for which all entries* w_j*,* $1 \leq j \leq n$*, are strictly positive.*

(e) Let $\mathbf{v}$ *and* $\mathbf{w}^{\mathrm{t}}$ *be the positive left and right eigenvectors from part (d). Under the normalization* $\sum_{j=1}^n w_j = 1$ *and* $\sum_{j=1}^n v_j w_j = 1$*, we have*

$$\frac{1}{(\lambda^*)^k} \mathbf{P}^k \xrightarrow{k\to\infty} \mathbf{w}^{\mathrm{t}} \mathbf{v} .$$

(f) The Perron–Frobenius eigenvalue λ^* *is the only eigenvalue for which there exist strictly positive right and left eigenvectors.*

Definition A.6.1. *A real $(n\times n)$-matrix $\mathbf{P}$ is called* **irreducible** *if $\mathbf{P}$ is nonnegative and for any pair of indices i, j, there exists $k > 0$ such that*

$$(\mathbf{P}^k)_{i,j} > 0\,.$$

Definition A.6.2. *Let $\mathbf{P}$ be a (nonnegative) irreducible $(n\times n)$-matrix. The matrix $\mathbf{P}$ is called periodic, or cyclic, with period $c > 1$ if for a $k \in \{1, ..., n\}$ (and hence for all $k \in \{1, ..., n\}$)*

$$c = \gcd\{m : (\mathbf{P}^m)_{k,k} > 0\}\,.$$

If $c = \gcd\{m : (\mathbf{P}^m)_{k,k} > 0\} = 1$, the matrix $\mathbf{P}$ is called aperiodic, or acyclic.

Theorem A.6.2 (Perron–Frobenius, Part II)**.** *Let $\mathbf{P}$ be a (nonnegative) irreducible $(n \times n)$-matrix. Then the following statements hold.*

(a) There exists a positive real number λ^*, called the Perron–Frobenius eigenvalue, such that λ^* is an eigenvalue of $\mathbf{P}$, and for all other eigenvalues λ of $\mathbf{P}$ we have

$$|\lambda| \le \lambda^*\,.$$

(b) The Perron–Frobenius eigenvalue λ^* satisfies

$$\min_i \sum_j P_{i,j} \le \lambda^* \le \max_i \sum_j P_{i,j}\,.$$

(c) The algebraic, and therefore also the geometric multiplicity of λ^* is one.

(d) There exists a left eigenvector $\mathbf{v}$ corresponding to eigenvalue λ^* for which all entries v_j, $1 \le j \le n$, are strictly positive. There also exists a right eigenvector $\mathbf{w}^{\mathrm{t}}$ corresponding to λ^* for which all entries w_j, $1 \le j \le n$, are strictly positive.

(e) If $\mathbf{P}$ is periodic with period $c > 1$, then $\mathbf{P}$ has precisely c distinct eigenvalues λ of modulus $|\lambda| = \lambda^*$. These c eigenvalues are

$$e^{2\pi ik/c}\lambda^* \quad \text{for } k = 0, 1, ..., c-1\,. \tag{A.2}$$

Each of the eigenvalues in (A.2) has algebraic (and hence also geometric) multiplicity one.

Appendix B

B.1 Sigma algebras, Probability spaces

Definition B.1.1. *Let Ω be a set. A collection $\mathcal{F}$ of subsets of Ω is called a σ-algebra (or σ-field) if the following three properties hold:*

(a) $\Omega \in \mathcal{F}$
(b) If $E \in \mathcal{F}$*, then* $E^c \in \mathcal{F}$*.*
(c) $\mathcal{F}$ *is closed under countable union, that is, if* $E_k \in \mathcal{F}$ *for* $k \geq 1$*, then*

$$\bigcup_{k\geq 1} E_k \in \mathcal{F}\,.$$

The fundamental notion is that of a probability space:

Definition B.1.2 (Probability space)**.** *A triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space if*

(a) Ω *is a set (the set of all possible outcomes),*
(b) $\mathcal{F}$ *is a sigma-algebra of* Ω*,*
(c) $\mathbb{P}$ *is a probability measure, that is, a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *for which*

- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}$ *has the σ-additivity property: If E_k, $k \geq 1$, are pairwise disjoint sets of $\mathcal{F}$, then*

$$\mathbb{P}\left(\bigcup_{k\geq 1} E_k\right) = \sum_{k\geq 1} \mathbb{P}(E_k)\,.$$

The elements of $\mathcal{F}$ are called *events*. They are exactly those collections of outcomes (i.e., subsets of Ω) for which a probability is defined. If Ω is discrete, we usually take $\mathcal{F}$ to be the power set (i.e., the set of all subsets) of Ω. In this case, *any* subset of Ω has a well

defined probability assigned to it. If Ω is an *uncountable* set, then $\mathcal{F}$ usually is a strict subset of the power set of Ω, and so for some subsets of Ω there is no defined probability.

Definition B.1.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say an event $B \in \mathcal{F}$ happens* **almost surely (a.s.)** *if $\mathbb{P}(B) = 1$. Equivalently, B happens almost surely if $\mathbb{P}(B^c) = 0$.*

The following continuity property of $\mathbb{P}$ is a consequence of σ-additivity:

Lemma B.1.1 (Continuity of probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and E_k, $k \geq 1$, a sequence of events.*

(a) If $E_1 \subseteq E_2 \subseteq \cdots$, then

$$\lim_{k\to\infty} \uparrow \mathbb{P}(E_k) = \mathbb{P}\left(\bigcup_{k\geq 1} E_k\right).$$

(b) If $E_1 \supseteq E_2 \supseteq \cdots$, then

$$\lim_{k\to\infty} \downarrow \mathbb{P}(E_k) = \mathbb{P}\left(\bigcap_{k\geq 1} E_k\right).$$

Proof. (a) Let $E_0 = \emptyset$ and consider the events $A_k = E_k - E_{k-1}$ for $k \geq 1$. The events A_k are pairwise disjoint. Clearly,

$$\bigcup_{k\geq 1} E_k = \bigcup_{k\geq 1} A_k\,,$$

and so,

$$\mathbb{P}\left(\bigcup_{k\geq 1} E_k\right) = \sum_{k\geq 1} \mathbb{P}(A_k) = \sum_{k\geq 1} (\mathbb{P}(E_k) - \mathbb{P}(E_{k-1}))\,.$$

But

$$\sum_{k\geq 1} (\mathbb{P}(E_k) - \mathbb{P}(E_{k-1})) = \lim_{n\to\infty} \sum_{k=1}^{n} (\mathbb{P}(E_k) - \mathbb{P}(E_{k-1})) = \lim_{n\to\infty} \uparrow \mathbb{P}(E_n)\,,$$

and thus

$$\mathbb{P}\left(\bigcup_{k\geq 1} E_k\right) = \lim_{n\to\infty} \uparrow \mathbb{P}(E_n)\,.$$

(b) To prove (b), we apply the result from part (a) to the sequence of events E_k^c, $k \geq 1$, for which then $E_1^c \subseteq E_2^c \subseteq \cdots$ holds. We omit the details. □

Let us now assume that we have n probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$, ..., $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$. We can form the direct product $\Omega = \Omega_1 \times \cdots \times \Omega_n$ which consists of all points $\omega = (\omega_1, ..., \omega_n)$ with $\omega_i \in \Omega_i$ for $1 \leq i \leq n$. We then consider the *direct product σ-algebra* $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$ on Ω which is the σ-algebra that consists of all sets of the form

$$F = F_1 \times \cdots \times F_n \qquad \text{with } F_i \in \mathcal{F}_i \text{ for } 1 \leq i \leq n\,.$$

In order to make $(\Omega, \mathcal{F})$ into a probability space, we need to define a probability $\mathbb{P}$ on $\mathcal{F}$. How we do this, will depend on the particular dependency structure we would like $\mathbb{P}$ (the joint probability distribution on the product space for which, a priori, we only know the marginal probability distributions for each component) to represent. The simplest probability measure $\mathbb{P}$ on product space $\Omega = \Omega_1 \times \cdots \times \Omega_n$ is **product measure** defined by

$$\mathbb{P}(F) = \mathbb{P}_1(F_1)\mathbb{P}_2(F_2)\cdots\mathbb{P}_n(F_n) \qquad \text{for all } F \in \mathcal{F}$$

where $F = F_1 \times F_2 \times \cdots \times F_n$. Product measure models **independence**. If Ω_i for $1 \leq i \leq n$ are discrete sets and we write $p_i(x_i) = \mathbb{P}_i(x_i)$, we have

$$\mathbb{P}(F) = \sum_{\substack{x_i \in F_i \\ 1 \leq i \leq n}} p_1(x_1)\, p_2(x_2) \cdots p_n(x_n)\,.$$

Example B.1.1. Consider n i.i.d. Bernoulli trials (coin flips) $(X_1, ..., X_n)$ with $\mathbb{P}(X_1 = 1) = p$ and $\mathbb{P}(X_1 = 0) = 1 - p$. We can associate with each coin flip X_i a probability space $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ where $\Omega_i = \{0, 1\}$ and $\mathcal{F}_i = \{\emptyset, \{1\}, \{0\}, \{0, 1\}\}$. The direct product space $\Omega = \Omega_1 \times \cdots \times \Omega_n$ consists of all binary vectors (with entries 0 or 1) of length n. We have $|\Omega| = 2^n$. For a specific outcome $\omega = (\omega_1, ..., \omega_n)$ with $\omega_i \in \{0, 1\}$ we have

$$\mathbb{P}(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n - \sum_{i=1}^n \omega_i}\,.$$

□

Definition B.1.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and a discrete set $\mathcal{S}$. A discrete* **random variable** *X is a map*

$$X : \Omega \to \mathcal{S}$$

such that $X^{-1}(\{s\}) \in \mathcal{F}$ for all $s \in \mathcal{S}$. The **distribution (law)** *of X is a probability measure μ on $\mathcal{S}$. It is the so-called* **pushforward** *measure to $\mathcal{S}$ of the probability measure $\mathbb{P}$ on Ω under the map X, that is,*

$$\mu = \mathbb{P} \circ X^{-1}\,.$$

B.2 Expectation, Basic inequalities

Expectation of a nonnegative integer-valued random variable:

Let X be a nonnegative integer-valued random variable. Then

$$\mathbb{E}(X) = \sum_{n \geq 1}^{\infty} \mathbb{P}(X \geq n)\,. \tag{B.1}$$

Markov's Inequality:

Let X be a random variable and $c > 0$. Then

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}(|X|)}{c}\,. \quad \text{(Markov's inequality)}$$

As a corollary, for higher moments,

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}(|X|^n)}{c^n} \quad \text{for } n \geq 1\,.$$

Assume $\mathbb{E}(|X|) < \infty$. Applying Markov's Inequality to the random variable $(X - \mathbb{E}(X))^2$, we get **Chebychev's Inequality:**

Let X be a random variable with $\mathbb{E}(|X|) < \infty$. Then

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}\,. \quad \text{(Chebychev's Inequality)}$$

Similar inequalities hold for higher central moments.

B.3 Properties of Conditional Expectation

Here we collect some main properties of conditional expectation. Assume all random variables are defined on the same probability space, and all expectations exist. Let $a, b, c \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$.

The following properties hold for conditional expectation. Note that statements (b)-(j) are meant to hold **with probability** 1.

(a) $\mathbb{E}(Y \,|\, X_1, ..., X_k)$ is a function of $X_1, ..., X_k$.

(b) $\mathbb{E}(c \,|\, X) = c$.

(c) *(Linearity property)* $\mathbb{E}(aX_1 + bX_2 \,|\, Y) = a\mathbb{E}(X_1 \,|\, Y) + b\mathbb{E}(X_2 \,|\, Y)$.

(d) *(Positivity)* If $Y \leq 0$, then $\mathbb{E}(Y \,|\, X) \leq 0$.

(e) If X and Y are independent, then $\mathbb{E}(Y \,|\, X) = \mathbb{E}(Y)$.

(f) If $Y = g(X)$, then $\mathbb{E}(Y \,|\, X) = \mathbb{E}(g(X) \,|\, X) = g(X)$.

(g) *(Pull through property)* $\mathbb{E}(Yg(X) \,|\, X) = g(X)\mathbb{E}(Y \,|\, X)$.

(h) *(Total expectation)* $\mathbb{E}(\mathbb{E}(Y \,|\, X)) = \mathbb{E}(Y)$.

(i) *(Tower property)* Let $k < n$. Then

$$\mathbb{E}\left(\mathbb{E}(Y \,|\, X_1, ..., X_k) \,|\, X_1, ..., X_n\right) = \mathbb{E}(Y \,|\, X_1, ..., X_k)\,,$$

and

$$\mathbb{E}\left(\mathbb{E}(Y \,|\, X_1, ..., X_n) \,|\, X_1, ..., X_k\right) = \mathbb{E}(Y \,|\, X_1, ..., X_k)\,.$$

(j) *(Jensen's Inequality)* If $f : \mathbb{R} \to \mathbb{R}$ is a convex function and $\mathbb{E}(|X|) < \infty$, then

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))\,,$$

and

$$f(\mathbb{E}(X \,|\, Y)) \leq \mathbb{E}(f(X) \,|\, Y)\,.$$

B.4 Modes of Convergence of Random Variables

In the following four definitions, let $(X_n)_{n\geq 0}$ be a sequence of random variables and X a random variable defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition B.4.1 (Almost sure convergence)**.** *We say the sequence $(X_n)_{n\geq 0}$ converges to X* **almost surely** *or* **with probability 1**, *if*

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1\,.$$

We write $X_n \xrightarrow{a.s.} X$.

Note that almost sure convergence means **pointwise convergence almost everywhere**: There exists a subset $B \subset \Omega$ with $B \in \mathcal{F}$ and $\mathbb{P}(B) = 1$ such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in B$.

Definition B.4.2 (Convergence in distribution)**.** *Let F and F_n, $n \geq 0$, be the cumulative distribution functions of X and X_n, $n \geq 0$, respectively. We say the sequence $(X_n)_{n\geq 0}$ converges to X* **in distribution** *if*

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

for all $x \in \mathbb{R}$ at which $F(x)$ is continuous. We write $X_n \xrightarrow{\mathcal{D}} X$.

Almost sure convergence is often referred to as *strong convergence*. Convergence in distribution is often referred to as *weak convergence*.

Definition B.4.3 (Convergence in probability)**.** *We say the sequence $(X_n)_{n\geq 0}$ converges to X* **in probability** *if for any $\epsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0\,.$$

We write $X_n \xrightarrow{p} X$.

Definition B.4.4 (L^p convergence or convergence in p^{th} mean)**.** *Let $p \geq 1$. We say the sequence $(X_n)_{n\geq 0}$ converges to X* **in L^p or in the p^{th} mean** *if*

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0\,.$$

We write $X_n \xrightarrow{L^p} X$.

Theorem B.4.1. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X a random variable defined on the same probability space. Then*

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{\mathcal{D}} X\,,$$

and furthermore,

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X\,.$$

Let c be a constant*. Then*

$$X_n \xrightarrow{\mathcal{D}} c \implies X_n \xrightarrow{p} c\,.$$

Example B.4.1 (Convergence in distribution does not imply convergence in probability)**.** Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and $P(\omega_i) = 1/4$ for $1 \leq i \leq 4$. Define the following random variables:

$$X_n(\omega_1) = X_n(\omega_2) = 1, \;\; X_n(\omega_3) = X_n(\omega_4) = 0 \quad \text{for all } n \geq \mathbb{N}\,.$$

$$X(\omega_1) = X(\omega_2) = 0, \;\; X(\omega_3) = X(\omega_4) = 1\,.$$

Clearly, $F_{X_n} = F_X$ for all $n \geq \mathbb{N}$ with

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1\,. \end{cases}$$

Since $F_{X_n}(x) = F_X(x)$ for all n, it is obvious that $X_n \xrightarrow{\mathcal{D}} X$. Observe that $|X_n(\omega_i) - X(\omega_i)| = 1$ for all $n \in \mathbb{N}$ and $1 \leq i \leq 4$. Hence

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \frac{1}{2}) = 1\,,$$

and the sequence $(X_n)_{n\geq 0}$ does not converge to X in probability. □

Example B.4.2 (Almost sure convergence does not imply convergence in L^1 (in mean))**.** Consider $\Omega = [0, 1]$ together with standard uniform (Lebesgue) measure $\mathbb{P}$ on $[0, 1]$. Let the sequence $(X_n)_{n\geq 1}$ of random variables $X_n : [0, 1] \to \mathbb{R}$ be defined by

$$X_n(x) = \begin{cases} n & \text{for } 0 \leq x \leq \frac{1}{n} \\ 0 & \text{for } \frac{1}{n} < x \leq 1 \end{cases}$$

for all $n \geq 1$. Then for all $x \in \Omega$ with $0 < x \leq 1$ we have $\lim_{n\to\infty} X_n(x) = 0$, and so

$$X_n \xrightarrow{a.s.} X \equiv 0\,.$$

We have $\mathbb{E}(|X_n|) = 1$ for all $n \geq 1$, and so

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 1 \neq 0$$

from which it follows that $(X_n)_{n\geq 1}$ does not converge to X in L^1 (in mean). □

B.5 Classical Limit Theorems

Theorem B.5.1 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables with finite first moment. Let $\mathbb{E}(X_1) = \mu$ and $S_n = \sum_{k=1}^n X_k$. Then*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu\,.$$

If the random variables $X_1, X_2, ...$ are nonnegative and $\mathbb{E}(X_1) = \infty$, then

$$\frac{S_n}{n} \xrightarrow{a.s.} \infty\,.$$

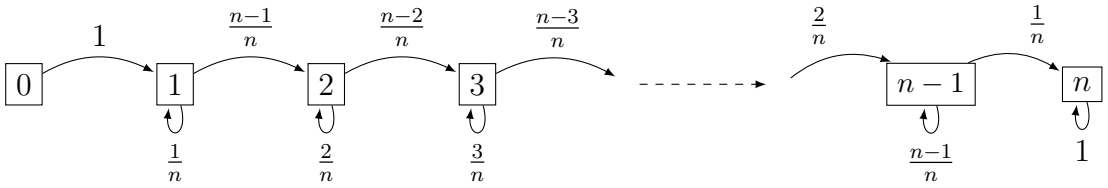
Theorem B.5.2 (Central Limit Theorem)**.** *Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables with $\mathrm{Var}(X_1) = \sigma^2 < \infty$ and $\mathbb{E}(X_1) = \mu$. Set $S_n = \sum_{k=1}^n X_k$. Then*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{\mathcal{D}} N(0,1)$$

where $N(0,1)$ is a standard normal random variable.

B.6 Coupon Collector's Problem.

This is a familiar problem in basic probability. Since applications of this problem appear in various places in our analysis of rates of convergence of Markov chains, we give a brief review: A cereal company issues n distinct types of coupons which a collector collects one-by-one. Each cereal box the collector purchases contains exactly one coupon, and the probability of finding a certain type of coupon in a purchased box is uniform over all types of coupons. Let X_k denote the number of distinct types the collector has after having purchased k boxes. The sequence $(X_n)_{n\geq 0}$ is a Markov chain on state space $\mathcal{S} = \{0, 1, ..., n\}$. It is a pure birth chain, and state n is absorbing. The following the the transition diagram for the chain:



Of particular interest for this chain is the expected time T until absorption in n. As can be seen from the transition diagram, for $i = 0, 1, ..., n-1$, the waiting time T^j for going from state j to state $j+1$ has a geometric distribution with success parameter $p_j = \frac{n-j}{n}$. By the linearity of expectation, we get

$$\mathbb{E}(T) = \sum_{j=0}^{n-1} \frac{n}{n-j} = n \sum_{j=1}^{n} \frac{1}{j} \,. \tag{B.2}$$

From the inequality

$$\ln(n+1) \leq \sum_{j=1}^{n} \frac{1}{j} \leq 1 + \ln n$$

we conclude

$$\lim_{n\to\infty} \left(\sum_{j=1}^{n} \frac{1}{j} \right) / \ln n = 1 \,,$$

and hence for large $n \in \mathbb{N}$,

$$\mathbb{E}(T) \approx n \ln n \,. \tag{B.3}$$

The variance of T^j is

$$\mathrm{Var}(T^j) = \left(\frac{n}{n-j} \right)^2 \left(1 - \frac{n-j}{n} \right) = \frac{j\, n}{(n-j)^2} \leq \frac{n^2}{(n-j)^2} \,,$$

and so

$$\mathrm{Var}(T) \leq \; n^2 \sum_{j=1}^{n} \frac{1}{j^2} \leq \; \frac{5}{3} n^2 \,, \tag{B.4}$$

where we have used $\sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$.

We can say more about the distribution of T and will compute the probabilities $\mathbb{P}(T \leq k)$ for $k \geq n$. We do so by rephrasing coupon collecting as an occupancy problem: Place k indistinguishable balls into n distinct boxes, one at a time and independently of each other, where each time a ball is placed, a box is chosen uniformly at random. The event of having collected n distinct coupons by time k has the same probability as the event that

each of the n boxes contains at least one ball. Its computation involves two combinatorial questions. The first is: In how many ways can one distribute k balls into n distinct boxes? It is the number of all possible linear arrangements of $n-1$ dividers and k balls. This number is $\binom{(n-1)+k}{n-1}$. Here is an example for 7 boxes and 10 balls:

$$| \circ\circ\circ \; | \; | \; \circ \; | \; \circ\circ\circ\circ\circ \; | \; | \circ$$

Boxes $1, 3$ and 6 are empty. Box 2 contains three balls, Boxes 4 and 7 contain one ball each, and Box 5 contains five balls. The second combinatorial question is: In how many ways can one arrange k balls into n boxes so that each box contains at least one ball? To answer, proceed as follows. First put exactly one ball into each box, which leaves $k-n$ balls unassigned. Distribute these $k-n$ balls in any way among the n boxes. The number of ways in which one can do this is $\binom{(n-1)+(k-n)}{n-1}$. Thus, altogether, we get

$$\mathbb{P}(T \le k) = \binom{k-1}{n-1} \Big/ \binom{n+k-1}{n-1} \quad \text{for } k \ge n\,.$$

In several places in our analysis of rates of convergence, we need an estimate for the tail probability of T for large times k. The following proposition gives such an estimate.

Lemma B.6.1. *Let T be the waiting time for the coupon collector's problem. For sufficiently large n and any $c > 0$, we have*

$$\mathbb{P}(T > n\ln n + cn) \le e^{-c}\,. \tag{B.5}$$

Proof. Note that the event $\{T > n\ln n + cn\}$ is the same as the event $\bigcup_{j=1}^{n} C_j$ where C_j is the even that Coupon j has not been collected by time $k = n\ln n + cn$. The probability of event C_j is $\mathbb{P}(C_j) = (1-\frac{1}{n})^{\lfloor n\ln n + cn\rfloor}$. Since $\mathbb{P}(\bigcup_{j=1}^{n} C_j) \le \sum_{j=1}^{n} \mathbb{P}(C_j)$, we get

$$\begin{aligned}
\mathbb{P}(T > n\ln n + cn) &\le n\left(1-\frac{1}{n}\right)^{\lfloor n\ln n + cn\rfloor} \\
&\le n e^{-\frac{n\ln n + cn - 1}{n}} \\
&\approx e^{-c}\,.
\end{aligned}$$

□

Appendix C

C.1 Growth Rates of Functions

Definition C.1.1 (Big-Oh)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. We say that f* **is asymptotically of at most the order of** *g, or f* **is big-Oh of** *g, if there exist positive constants x_0 and M such that*

$$|f(x)| \leq M\,|g(x)| \quad \text{for all } x \geq x_0\,.$$

In this case we write $f(x) = \mathrm{O}(g(x))$.

Example C.1.1. (a) Let $f(n) = n \ln n + 3n^2 - 5n$ for $n \geq 1$. Then $f(n) = \mathrm{O}(n^2)$.
(b) Let $f(x) = 9x^3 + \sqrt{x^4 + x^2}$. Then $f(x) = \mathrm{O}(x^3)$. It would also be correct to state $f(x) = \mathrm{O}(x^5)$, for example. □

Definition C.1.2 (Big-Omega)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. We say that f* **is asymptotically of at least the order of** *g, or f* **is big-Omega of** *g, if there exist positive constants x_0 and M such that*

$$|f(x)| \geq M\,|g(x)| \quad \text{for all } x \geq x_0\,.$$

In this case we write $f(x) = \Omega(g(x))$.

Notice that $f(x) = \Omega(g(x))$ if and only if $g(x) = \mathrm{O}(f(x))$.

Definition C.1.3 (Big-Theta)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. We say that f* **is asymptotically of the same order as** *g, or f* **is big-Theta of** *g, if*

$$f(x) = \mathrm{O}(g(x)) \text{ and } g(x) = \mathrm{O}(f(x)) .$$

In this case we write $f(x) = \Theta(g(x))$.

Example C.1.2. Let $f(n) = n \ln n + 3n^2 - 5n$ for $n \geq 1$. Then $f(n) = \Theta(n^2)$. □

Definition C.1.4 (Little-oh)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. We say that f* **is asymptotically of smaller order than** *g, or f* **is little-oh of** *g, if*

$$f(x) = \mathrm{O}(g(x)) \text{ and } f(x) \neq \Theta(g(x)) .$$

In this case we write $f(x) = \mathrm{o}(g(x))$.

Example C.1.3. Let $f(n) = n \ln n + 3n^2 - 5n$ for $n \geq 1$. Then $f(n) = o(n^2 \ln n)$. It would also be correct to state $f(n) = o(n^3)$, for example. □

C.2 Lim sup, Lim inf

Definition C.2.1 (Supremum, Infimum)**.** *Let E be a non-empty subset of $\mathbb{R}$. We say $b \in \mathbb{R}$ is the* **supremum of** *E and write $b = \sup(E)$ if b is the smallest upper bound of E. That is, $x \leq b$ for all $x \in E$, and for any other upper bound M of E we have $b \leq M$.*
We say $a \in \mathbb{R}$ is the **infimum of** *E and write $a = \inf(E)$ if a is the largest lower bound of E. That is, $a \leq x$ for all $x \in E$, and for any other lower bound m of E we have $m \leq a$.*

Definition C.2.2 (Lim sup, Lim inf)**.** *Let $(a_n)_{n\geq 0}$ be a sequence of real numbers.*

(a) The **limit supremum** *of $(a_n)_{n\geq 0}$, denoted by $\limsup_{n\to\infty} a_n$, is defined by*

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \left(\sup_{i\geq n} a_i \right) .$$

(b) The **limit infimum** *of $(a_n)_{n\geq 0}$, denoted by $\liminf_{n\to\infty} a_n$, is defined by*

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \left(\inf_{i\geq n} a_i \right) .$$

Unlike the limit, which may not exist for a given sequence $(a_n)_{n\geq 0}$, the limit supremum and the limit infimum always exist for any sequence. If $\lim_{n\to\infty} a_n$ does exist, then

$$\lim_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n .$$

The limit superior $\limsup_{n\to\infty} a_n$ is the *largest accumulation point* for the sequence $(a_n)_{n\geq 0}$. Thus for any $\epsilon > 0$, for *only finitely many* $k \in \mathbb{N}$,

$$a_k > \epsilon + \limsup_{n\to\infty} a_n$$

and for *infinitely many* $j \in \mathbb{N}$,

$$a_j > -\epsilon + \limsup_{n\to\infty} a_n .$$

Similarly, the limit inferior is the *smallest accumulation point* for the sequence $(a_n)_{n\geq 0}$. For only finitely many $k \in \mathbb{N}$,

$$a_k < -\epsilon + \liminf_{n\to\infty} a_n$$

and for infinitely many $j \in \mathbb{N}$,

$$a_j < \epsilon + \liminf_{n\to\infty} a_n .$$

C.3 Interchanging Limit and Integration

Here we state several classical limit theorems a for integration. We phrase the theorems in terms of expectations of random variables. For a reference see [31].

Theorem C.3.1 (Monotone Convergence Theorem)**.** *Let $(X_n)_{n\geq 0}$ be a sequence of nonnegative random variables and X a (not necessarily finite) random variable with*

$$\lim_{n\to\infty} X_n = X \qquad a.s.$$

If

$$0 \leq X_0 \leq X_1 \leq X_2 \leq \cdots \qquad a.s.\,,$$

then

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)\,.$$

Recall: We assume that the random variables $(X_n)_{n\geq 0}$ and X are defined on the *same* probability space $(\Omega, \mathcal{F}, P)$. When we write

$$\lim_{n\to\infty} X_n = X \quad \textbf{a.s.}, \tag{C.1}$$

the "a.s." stands for **almost sure convergence** or (alternatively stated) **convergence with probability** 1 (recall Definition B.4.1). More precisely, (C.1) means that we have

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

for all $\omega \in \Omega$, except possibly for $\omega \in A$ where $A \subset \Omega$ is an event of probability 0. Similarly, $0 \leq X_0 \leq X_1 \leq X_2 \leq \cdots$ a.s. means

$$0 \leq X_0(\omega) \leq X_1(\omega) \leq X_2(\omega) \leq \cdots$$

for all $\omega \in \Omega$, except possibly for $\omega \in B$ where $B \subset \Omega$ is an event of probability 0.
As a consequence of Theorem C.3.1, we have the following corollary for the interchange of $\mathbb{E}$ and $\sum$ for nonnegative random variables:

Corollary C.3.2. *Let $(Y_n)_{n\geq 0}$ be a sequence of nonnegative random variables and X a (not necessarily finite) random variable with*

$$\sum_{n=0}^{\infty} Y_n = X \qquad a.s.$$

Then

$$\sum_{n=0}^{\infty} \mathbb{E}(Y_n) = \mathbb{E}(X)\,.$$

The corollary follows from Theorem C.3.1 by setting

$$X_n = \sum_{k=0}^{n} Y_k \qquad \text{for } n \geq 0\,.$$

Note that any one of the expectations in Corollary C.3.2 may be ∞.

Theorem C.3.3 (Dominated Convergence Theorem)**.** *Let $(X_n)_{n\geq 0}$ be a sequence of random variables and X a random variable with*

$$\lim_{n\to\infty} X_n = X \qquad a.s.$$

If there exists a random variable Y with $\mathbb{E}(|Y|) < \infty$ such that

$$|X_n| \leq Y \qquad a.s. \text{ for all } n \geq 0\,,$$

then

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)\,.$$

Corollary C.3.4 (Bounded Convergence)**.** *Let $(X_n)_{n\geq 0}$ be a sequence of random variables and X a random variable with*

$$\lim_{n\to\infty} X_n = X \qquad a.s.$$

If there exists a constant K_0 such that

$$|X_n| \leq K_0 \qquad a.s. \text{ for all } n \geq 0\,,$$

then

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)\,.$$

Combining Corollary C.3.2 and Theorem C.3.3 yields the following proposition.

Proposition C.3.5. *If a sequence of random variables $(Z_n)_{n\geq 0}$ satisfies $\sum_{n=0}^{\infty} \mathbb{E}(|Z_n|) < \infty$, then*

$$\sum_{k=0}^{\infty} \mathbb{E}(Z_k) = \mathbb{E}\left(\sum_{k=0}^{\infty} Z_k\right)\,.$$

Proof. Set $Y_n = |Z_n|$ and apply Corollary C.3.2. This yields

$$\sum_{n=0}^{\infty} \mathbb{E}(Y_n) = \sum_{n=0}^{\infty} \mathbb{E}(|Z_n|) = \mathbb{E}\left(\sum_{n=0}^{\infty} |Z_n|\right) < \infty\,.$$

Since the random variable $Y = \sum_{n=0}^{\infty} |Z_n|$ has finite expectation, it is finite almost surely. As a consequence, $\sum_{n=0}^{\infty} Z_n$ converges almost surely. Set $X_n = \sum_{n=0}^{n} Z_k$ and $X = \sum_{k=0}^{\infty} Z_k$. Then

$$\lim_{n\to\infty} X_n = X \qquad a.s.$$

Since we have $\mathbb{E}(|Y|) = \mathbb{E}(Y) < \infty$ and

$$|X_n| \leq \sum_{k=0}^{n} |Z_k| \leq Y \qquad a.s. \text{ for all } n \geq 0\,,$$

by Theorem C.3.3, we conclude

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \lim_{n\to\infty} \sum_{k=0}^{n} \mathbb{E}(Z_n) = \sum_{k=0}^{\infty} \mathbb{E}(Z_k) = \mathbb{E}(X) = \mathbb{E}\left(\sum_{k=0}^{\infty} Z_k\right)\,.$$

□

In general, a sequence of random variables $(X_n)_{n\geq 0}$ may not converge a.s. However, for all $\omega \in \Omega$,

$$\liminf_{n\to\infty} X_n(\omega) \quad \text{and} \quad \limsup_{n\to\infty} X_n(w)$$

do always exist, and

$$\liminf_{n\to\infty} X_n \quad \text{and} \quad \limsup_{n\to\infty} X_n$$

are random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Fatou's Lemma tells us what happens when we interchange $\mathbb{E}$ and lim sup (or lim inf).

Lemma C.3.6 (Fatou's Lemma)**.** *Let $(X_n)_{n\geq 0}$ be a sequence of random variables.*

(a) If there exists a random variable Y with $\mathbb{E}(|Y|) < \infty$ such that

$$Y \leq X_n \quad a.s. \text{ for all } n \geq 0\,,$$

then

$$\mathbb{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)\,.$$

(b) Similarly, if there exists a random variable Y with $\mathbb{E}(|Y|) < \infty$ such that

$$X_n \leq Y \quad a.s. \text{ for all } n \geq 0\,,$$

then

$$\mathbb{E}\left(\limsup_{n\to\infty} X_n\right) \geq \limsup_{n\to\infty} \mathbb{E}(X_n)\,.$$

Note that the random variable Y may be a constant. Clearly, (b) follows from (a) by replacing X_n with $-X_n$ and noting that

$$\liminf_{n\to\infty} X_n = \limsup_{n\to\infty}(-X_n)\,.$$

□

Bibliography

[1] D. Aldous (1983), *Random walks on finite groups and rapidly mixing Markov chains,* Séminaire de probabilités XVII, Lecture Notes in Math. 986, 243-297, Springer, New York.

[2] D. Aldous and J. Fill (2002), *Reversible Markov Chains and Random Walks on Graphs,* available at `http://www.stat.berkeley.edu/~aldous/RWG/book.html`.

[3] R. Aleliunas, R. M. Karp, R. J. Lipton, L. Lovász, C. Rackoff (1979), *Random walks, universal traversal sequences, and the complexity of maze problems,* 20th Annual Symposium on Foundations of Computer Science (San Juan, Puerto Rico), IEEE, 218-223.

[4] K. B. Athreya and P. E. Ney (1972). *Branching Processes,* Dover Publications, Inc.

[5] D. Bayer and P. Diaconis (1992), *Trailing the dovetail shuffle to its lair,* Annals of Applied Prob. **2**, No. 2, 294-313.

[6] P. Brémaud (1999), *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues,* Texts in Applied Mathematics **31**, Springer Verlag.

[7] P. Diaconis (1988), *Group Representations in Probability and Statistics,* Lecture Notes – Monograph Series, **11**, Institute of Mathematical Statistics, Hayward, California.

[8] P. Diaconis (1996), *The cut-off phenomenon in finite Markov chains,* Proc. Natl. Acad. Sci. USA, **93**, 1659-1664.

[9] P. Diaconis and M. Shahshahani (1981), *Generating a random permutation with random transpositions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, **57**, 159-179.

[10] P. Diaconis and M. Shahshahani (1987), *Time to reach stationarity in the Bernoulli-Laplace diffusion model,* SIAM J. Math. Anal., Vol. **18** No. 1.

[11] P. G. Doyle and J. L. Snell (1984), *Random walks and electric networks,* Carus Math. Monographs **22**, Mathematical Association of America.

[12] R. Durrett (2010), *Probability, Theory and Examples,* 4th Edition, Cambridge University Press.

[13] F. Eggenberger and G. Pólya (1923), *Über die Statistik verketteter Vorgänge,* Zeitschrift für angewandte Mathematik und Mechanik **3**, Heft 4, 279-289.

[14] U. Feige (1995), *A tight upper bound on the cover time for random walks on graphs,* Random Structures & Algorithms **6**, 51-54.

[15] U. Feige (1995), *A tight lower bound on the cover time for random walks on graphs,* Random Structures & Algorithms **6**, 433-438.

[16] W. Feller (1970), *An Introduction to Probability Theory and Its Applications, Volume II,* 2nd Edition, John Wiley & Sons, Inc.

[17] G. Grimmett and D. Stirzaker (2001), *Probability and Random Processes,* 3rd Edition, Oxford University Press.

[18] R. Horn and Ch. Johnson (2013), *Matrix Analysis,* 2nd Edition, Cambridge University Press.

[19] S. Kakutani (1945), *Markov processes and the Dirichlet problem,* Proc. Jap. Acad., **21**, 227-233.

[20] A. N. Kolmogorov (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung,* Julius Springer, Berlin.

[21] J. Lamperti (1996), *Probability - A Survey of the Mathematical Theory,* 2nd Edition, Wiley Series in Probability and Statistics.

[22] G. Lawler (2006), *Introduction to Stochastic Processes,* 2nd Edition, Chapman & Hall/CRC.

[23] D. Levin and Y. Peres (2010), *Pólya's Theorem on Random Walks via Pólya's Urn,* The American Mathematical Monthly **117**, Issue 3, 220-231.

[24] R. Lyons and Y. Peres (2016), *Probability on Trees and Networks,* Cambridge University Press.

[25] D. Levin and Y. Peres (2017), *Markov Chains and Mixing Times,* 2nd Edition, AMS, Providence, Rhode Island.

[26] L. Lovász (1993), *Random walks on graphs: A survey,* Combinatorics, Paul Erdös is Eighty (Volume 2), Bolyai Society (Hungary), 1-46.

[27] L. Lovász and P. Winkler (1993), *On the last new vertex visited by a random walk,* J. Graph Theory **17**, 593-596.

[28] J. R. Norris (2009), *Markov Chains,* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

[29] S. Orey (1962), *An Ergodic Theorem for Markov Chains,* Z. Wahrscheinlichkeitstheorie **1**, 174-176.

[30] S. I. Resnick (2002), *Adventures in Stochastic Processes,* Birkhäuser, Boston.

[31] W. Rudin (1976), *Principles of Mathematical Analysis,* 3rd Edition, McGraw-Hill.

[32] E. Seneta (1981), *Non-negative Matrices and Markov Chains,* 2nd Edition, Springer Verlag.

[33] A. N. Shiryaev (1995), *Probability,* 2nd Edition, Graduate Texts in Mathematics **95**, Springer Verlag.

[34] F. Spitzer (1964), *Principles of Random Walk,* Graduate Texts in Mathematics **34**, Springer Verlag.

[35] P. Tetali (1991), *Random walks and the effective resistance of networks,* Journal of Theoretical Probability, Vol. 4, No. 1, 101-109.

[36] D. Williams (1991), *Probability with Martingales,* Cambridge University Press.

Index