Finite Markov chains and Monte-Carlo methods

By Soumik Pal and Tim Mesikepp
with diagrams and student input from Jamie Forschmiedt and Celeste Zeng

January 21, 2023

Contents

Chapter 1

Markov Chains

1.1 Random walks on graphs

In this chapter we introduce *Markov chains* and their fundamental properties. Before we give the general definition, however, it will be helpful to begin with a very concrete and important class of examples, *random walks on graphs*. To understand a random walk on a *graph*, though, we need to understand what a graph is, and so we start by reviewing the very basics of graphs, after which we describe two kinds of random walks on their vertices. Both of these are Markov chains, and they will provide helpful intuition for what Markov chains are all about.

1.1.1 What are graphs?

A graph $G = (V, E)$ is just a collection of vertices V and edges E. Both of these are sets, and E is a subset of unordered pairs of vertices. An edge $\{u, v\} \in E$ iff (if and only if) the vertices u and v are "connected" in the graph G. When $\{u, v\} \in E$, say that the vertices u and v are *adjacent* or that they are *neighbors*, and write $u \sim v$. If there is an edge $\{u, u\}$ in the graph, it is said to be a loop. Unless otherwise specified, our graphs will not contain any loops.

This is all best seen in examples. For instance, for the graph in Figure 1.1, $V = \{1, 2, 3, 4, 5\}$ and

$$E = \{\{1, 2\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}.$$

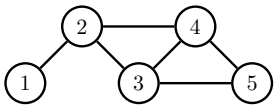

Figure 1.1: A graph with six vertices

Note that, as *sets*, $\{2,3\} = \{3,2\}$, and so we don't have to list them twice. Another example is the **6-cycle** in Figure 1.2. Here $V = \{1,2,3,4,5,6\}$ and

$$E = \{\{1,2\},\{2,3\},\{3,4\},\{4,5\},\{5,6\},\{6,1\}\}.$$

In Figure 1.2 we have arranged the vertices in a hexagonal shape, but keep in mind that the layout, or *embedding*, of the graph in two dimensions (or any dimension) doesn't matter. Our definition $G = (V, E)$, after all, does not care how we draw the graph; all that matters is what the vertices are and how they are connected. So, for instance, the picture of a trip across the US in Figure 1.3 gives exactly the same graph as Figure 1.2, even though the embeddings look very different.

Sometimes our vertex set will be infinite, as with the integer graph (V, E) in Figure 1.4. Here the vertices are $V = \{\ldots, -1, 0, 1, 2, \ldots\}$ and edges connect adjacent integers, $E = \{\,\{j, j+1\} \,:\, j \in \mathbb{Z}\,\}$. Often we will restrict ourselves to a finite portion of this graph, like $V' = \{0, 1, \ldots, n\}$ and the associated edges.

Graphs can even be generated randomly. In Figure 1.5 we have a sample of a *Erdős-Rényi* random graph, which is generated via two parameters: the number of vertices n, and a probability $0 \le p \le 1$ for edge inclusion. For

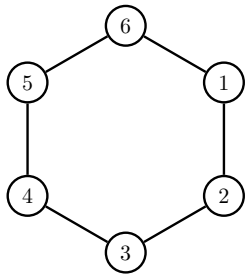

Figure 1.2: The 6-cycle graph. In general, the n**-cycle** has n vertices, n edges, and each vertex is connected to two neighbors, forming one "loop" around the entire graph.

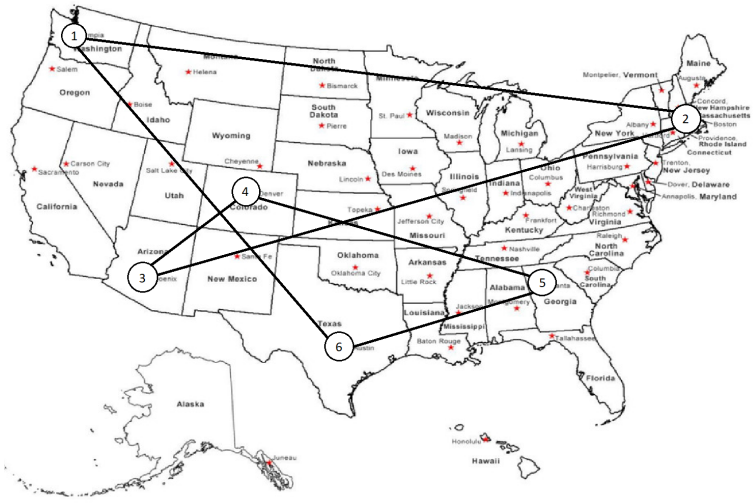

Figure 1.3: Another embedding of the 6-cycle. As a *graph*, this is identical to the hexagonal embedding of the 6-cycle in Figure 1.2.

each pair $1 \leq i \neq j \leq n$ of the vertices, the edge $\{i, j\}$ is included with probability p, independent of all the other edges. In Figure 1.5, there are $n = 50$ vertices and $p = 0.2$.

Random graphs have many applications, such as in analyzing communication networks, groupings of friends on social media, and the spread of a respiratory illness through a population.

1.1.2 Simple, symmetric random walks on graphs

Now that we know what a graph $G = (V, E)$ is, we can describe how to move around its vertices on a random. The idea is simple: start at some vertex $X_0 \in V$, and at time $t = 1$ choose a new vertex $X_1 \in V$, and at $t = 2$ a new vertex $X_2 \in V$, and so on, generating a random sequence $X_0, X_1, X_2, \ldots \subset V$. The key question is how to choose the next vertex X_{n+1} given that our current vertex is $X_n = v$. We could just select X_{n+1} uniformly at random among V. That is not terribly interesting, however,

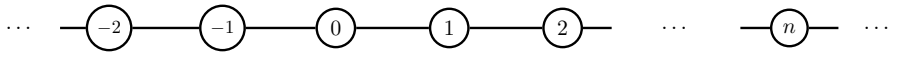

Figure 1.4: The integers $\mathbb{Z}$ as an infinite graph.

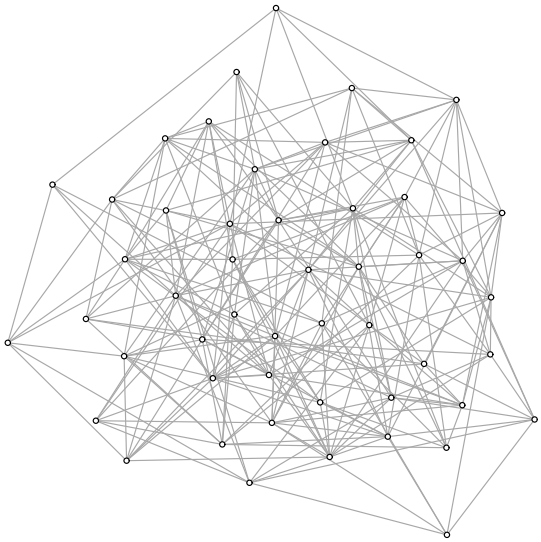

Figure 1.5: An Erdős-Rényi random graph with $n = 50$ vertices and edge inclusion probability $p = 0.2$.

because we are not respecting the geometry of the graph; we have totally ignored the edge identifications E. So the most natural way is to move to any of the *adjacent* vertices of X_n, each chosen with equal probability.

To make this precise we need some notation. We say the **degree** $\deg(v)$ of a vertex v is the number of **adjacent vertices** to v. This is the number of edges $\{v, w\} \in E$, or the number of edges coming out of v. In Figure 1.1, $\deg(1) = 1, \deg(2) = \deg(3) = \deg(4) = 3$ and $\deg(5) = 2$. So if our current position is $X_n = v$, then our next vertex X_{n+1} will be one of the $\deg(v)$ adjacent vertices to X_n, each chosen with equal probability. We thus have the transition probabilities

$$\mathbb{P}(X_{n+1} = w \,|\, X_n = v) = \frac{1}{\deg(v)} \tag{1.1}$$

for each $w \sim v$. Once we select a starting vertex X_0, the resulting sequence of random vertices $X_0, X_1, X_2, \ldots$ is a Markov chain which is called the **simple symmetric random walk** on the graph. Here *simple* means that we only jump to neighboring vertices, and *symmetric* means that each neighbor is

chosen with equal probability. Usually we will be a bit sloppy and just call such a walk the **simple random walk** as a shorthand.

If, at each step, we had chosen X_{n+1} uniformly at random in V, irrespective of the value of X_n, then all the X_n's would be independent of each other. Note well, however, that this is certainly not the case now: the next step X_{n+1} heavily depends on our current location X_n.

Going back to our examples, suppose $X_n = 2$ in Figure 1.1. Then X_{n+1} will be either 1,3 or 4, each chosen with probability 1/3. If we are currently at vertex 5, we next jump to either 3 or 4, each with probability 1/2. If $X_n = 1$, then $X_{n+1} = 2$ because that is the only option. For the 6-cycle in Figure 1.2, we flip a fair coin at each stage to determine if we move clockwise (CW) or counter-clockwise (CCW).

Note that if we sum the transition probability (1.1) over *all* $w \sim v$, we have

$$\sum_{w \sim v} \mathbb{P}(X_{n+1} = w \,|\, X_n = v) = \sum_{w \sim v} \frac{1}{\deg(v)} = \deg(v) \cdot \frac{1}{\deg(v)} = 1, \qquad (1.2)$$

as we would expect: our probabilities for how we take the next step should add up to 1. Note also that how we take the next step is determined *entirely* from our current position; the earlier history has no influence. If we are currently at $v = 5$ on the 6-cycle, our position two steps ago, or twenty steps ago, has no say in how we determine the next step: regardless of the past we move to either 4 or 6, with equal property. This "forgetfullness" is called the **Markov property** and is why the simple random walk on a graph is a particular case of a Markov chain.*

Exercise 1.1. *For an Erdős-Rényi random graph with n vertices and edge probability p, find the expected degree $\mathbb{E}(\deg(v))$ of a vertex.*

1.1.3 Weighted walks on graphs

Often our random walks on graphs will be the simple, as described above, but not always; we may wish to bias moving one direction over another, and hence not chose the among the adjacent vertices to v with equal probability. In general we can assign a probability $p_{vw} \geq 0$ when we wish to move from

*We'll formally define the Markov property below in Definition 1.3.

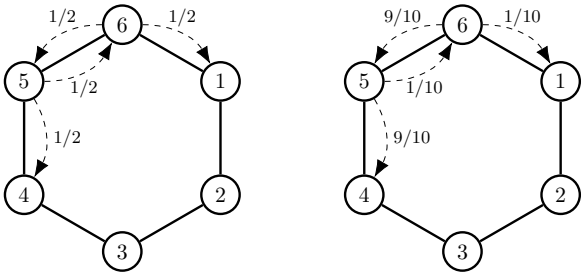

Figure 1.6: A simple and biased walk on the 6-cycle.

v to w along the edge $\{v, w\}$, where

$$\sum_{w \sim v} \mathbb{P}(X_{n+1} = w \,|\, X_n = v) = \sum_{w \sim v} p_{vw} = 1, \tag{1.3}$$

as for the simple walk (1.2).

In the 6-cycle, for example, we may want to move clockwise with probability p, and counterclockwise with probability $1 - p$, for some $0 \leq p \leq 1$. Similarly with the number line: perhaps we want to go right with probability p, and left with probability $1 - p$. Such walks are called **weighted** or **biased random walks** (as opposed to symmetric) on the underlying graph. See Figure 1.6. Note that we need not have $p_{vw} = p_{wv}$, as in the 6-cycle, $p_{65} = \frac{9}{10} \neq \frac{1}{10} = p_{56}$. And, of course, we don't have to have some fixed "rule" to assign the probabilities, as in the probability of going CW is always p. We merely need to have (1.3) for each vertex v.

1.2 Transition matrices

Whether we do the simple or biased walk on our graph, we need some way of keeping track of the probabilities of the possible paths of the random walk. We do this through a matrix P, where P is $n \times n$ if our graph has n vertices. Entry P_{ij} of P is the probability $\mathbb{P}(X_{k+1} = j \,|\, X_k = i)$, for any time point k. This *transition matrix* will be a key tool in our analysis of Markov chains.

Definition 1.1. A **transition matrix**, or **transition probability matrix**, is an $n \times n$ matrix P where

(*i*) All the entries are non-negative, $P_{jk} \geq 0$ for all $1 \leq j, k \leq n$.

(*ii*) Each row sums to one, $\sum_{k=1}^{n} P_{jk} = 1$ for each j.

Every random walk (simple, or with weights) on a graph corresponds to a transition matrix P, and each transition matrix defines a random walk. For example, the matrix for the simple walk on the graph in Figure 1.1 is

$$P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}. \tag{1.4}$$

Entry P_{23} is $1/3$ because the probability of moving to 3, starting from vertex 2, is $1/3$. $P_{14} = 0$ because there is no edge from vertex 1 to 4. And so on. Note that, in general, we need not have $p_{ij} = p_{ji}$, as here $p_{45} \neq p_{54}$. The transition matrix for the 6-cycle is

$$P_2 = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 \end{bmatrix}. \tag{1.5}$$

This is a good time for you to review your old class notes on matrices; in time we will see that surprisingly deep and rich properties of our Markov chains are encoded in the transition matrix P.

Exercise 1.2. *Consider the transition matrix P_1 in* (1.4).

(*i*) *What do the rows add up to? Explain what this means intuitively.*

(*ii*) *What do the columns add up to?*

(*iii*) *Given an eigenvector for P_1 corresponding to eigenvalue $\lambda = 1$.*

(*iv*) *Answer the same questions for P_2 in* (1.5).

(*v*) *What does it intuitively mean that the diagonal of both matrices have only zeros?*

(*vi*) *Use a computer to compute the eigenvalues of either P_1 or P_2. What do you notice?*

Exercise 1.3. *Draw a graph G for which the matrix*

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix} \tag{1.6}$$

is the transition matrix for the simple random walk on G. Use a computer to compute the eigenvalues of P. What do you notice? Compare and contrast with Exercise 1.2 (vi).

1.3 Markov chains

We are now ready to define a Markov chain. A Markov chain is, first, a type of *stochastic process*.

Definition 1.2. Given a state space Ω, a discrete-time[†] **stochastic process** on Ω is a sequence of random variables $(X_0, X_1, X_2, \ldots)$, all defined on the same probability space, such that every X_j takes values in Ω.

So a stochastic process is a random sequence $(X_0, X_1, X_2, \ldots)$ of elements of Ω. We typically think of a stochastic process having infinite length, as opposed to the random vectors of finite length that you might have encountered before. Another point of view on this definition is to think of a stochastic process as a "movie": time is unfolding in discrete intervals $t = 0, 1, 2, \ldots$, and at each new step in time we encounter a new element of Ω.

Not all stochastic processes are Markov chains. To be a Markov chain, we additionally need the "forgetfulness" property that we saw with random walks on graphs: our transition probabilities must only depend on our current state. The rigorous formulation is as follows.

Definition 1.3. Given a set Ω with n elements labeled by $\{1, 2, \ldots, n\}$ and an $n \times n$ transition matrix P, a **Markov chain** with state space Ω and

[†]It is possible to consider stochastic processes (X_t) with continuous time $t \in \mathbb{R}$, but the Markov chains we consider here always have discrete time.

transition matrix P is a stochastic process $X = (X_0, X_1, X_2, \ldots)$ on Ω such that X satisfies *the Markov property*

$$\mathbb{P}(X_{k+1} = y \,|\, X_0 = j_0, X_1 = j_1, \ldots, X_{k-1} = j_{k-1}, X_k = x\} \\ = \mathbb{P}(X_{k+1} = y \,|\, X_k = x) = P_{xy} \quad (1.7)$$

for any $k \geq 1$ and any $j_0, j_1, \ldots, j_{k-1}, x, y \in \Omega$.

The elements of Ω are called the states of the Markov chain. For a random walk on a graph, the states are the vertices of the graph. We will see later that, if we know the distribution of X_0, from the above definition we get the joint distribution of the random vector $(X_0, X_1, \ldots, X_k)$, for every $k = 1, 2, \ldots$. It follows that the probabilistic properties of a Markov chain is completely determined by three things: its state space Ω, the transition matrix P, and its *initial distribution* which is the distribution of X_0.

In Definition 1.3, the state space Ω is finite. We will describe a few examples of Markov chains for infinite but discrete Ω (say the set of natural numbers) later in the text. Markov chains with finite state space are sometimes called finite Markov chains to distinguish from the latter class. We again stress that the number of steps k need not be. In fact, later in the text we will be concerned with how X_k behaves as $k \to \infty$. Some authors describe the property (1.7) as the time homogeneous Markov property and the corresponding stochastic process as a time-homogeneous Markov chain. There are also Markov chains that are not time-homogeneous in the sense that the transition matrix can change with time. However, we will ignore this additional qualifier since all our Markov chains will be time-homogeneous.

Pay special attention to what the Markov property (1.7) is saying: the history of how we arrived at state x does not matter for the next jump - only the fact that we are now at x. This is what we saw with our random walks on graphs, and this is what makes Markov chains special among stochastic processes.

Example 1.1. Let's consider a simple example. Let $\Omega = \{1, 2\}$ and

$$P = \begin{bmatrix} p & 1-p \\ q & 1-q \end{bmatrix} \quad (1.8)$$

for some $0 < p < 1$ and $0 < q < 1$. This corresponds to a weighted random

walk on a "graph" with two vertices, labelled 1 and 2, as in Figure 1.7. Note that this "graph" has loops, or edges $\{i, i\}$ that connect the same vertex to itself, and so is not strictly a graph in our sense of the term. However, we can, of course, still draw a diagram of its behavior similar to how we would draw a graph. Our transition matrix (1.8) says

$$\begin{aligned}
\mathbb{P}(X_{j+1} = 1 \,|\, X_j = 1) &= P_{11} = p, \\
\mathbb{P}(X_{j+1} = 2 \,|\, X_j = 1) &= P_{12} = 1 - p, \\
\mathbb{P}(X_{j+1} = 1 \,|\, X_j = 2) &= P_{21} = q, \\
\mathbb{P}(X_{j+1} = 2 \,|\, X_j = 2) &= P_{22} = 1 - q.
\end{aligned}$$

Drawing a diagram, as in Figure 1.7, is often a very helpful way of keeping track of these probabilities and seeing what is going on. This is called a *graphical representation* of the Markov chain. All finite time-homogeneous Markov chains have a graphical representation as a weighted random walk on a graph with loops.

We know that if we start at vertex 1, the probability of moving to state 2 is $1 - p$. But how do we compute probabilities for more than one step? What is

$$\mathbb{P}(X_2 = 2 \,|\, X_0 = 1),$$

for instance? This is where the Markov property (1.7) comes into play. Recalling that $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B)$,

$$\begin{aligned}
\mathbb{P}(X_2 = 2|X_0 = 1) &= \mathbb{P}(X_2 = 2, X_1 = 2|X_0 = 1) + \mathbb{P}(X_2 = 2, X_1 = 1|X_0 = 1) \\
&= \mathbb{P}(X_2 = 2 \,|\, X_1 = 2, X_0 = 1)\mathbb{P}(X_1 = 2 \,|\, X_0 = 1) \\
&\quad + \mathbb{P}(X_2 = 2 \,|\, X_1 = 1, X_0 = 1)\mathbb{P}(X_1 = 1 \,|\, X_0 = 1) \\
&= \mathbb{P}(X_2 = 2 \,|\, X_1 = 2)\mathbb{P}(X_1 = 2 \,|\, X_0 = 1) \qquad (1.9)
\end{aligned}$$

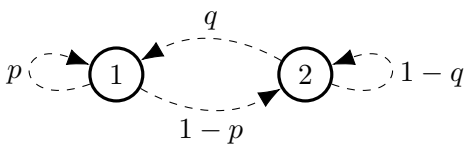

Figure 1.7: A diagram of the Markov chain in Example 1.1.

$$\begin{aligned} &+ \mathbb{P}(X_2 = 2 \,|\, X_1 = 1)\mathbb{P}(X_1 = 1 \,|\, X_0 = 1) \\ &= (1-q)(1-p) + (1-p)p, \end{aligned} \tag{1.10}$$

where we used the Markov property in (1.9) to eliminate conditioning on the first step.

Similarly, we find

$$\begin{aligned} \mathbb{P}(X_2 = 1|X_0 = 1) &= \mathbb{P}(X_2 = 1, X_1 = 2|X_0 = 1) + \mathbb{P}(X_2 = 1, X_1 = 1|X_0 = 1) \\ &= \mathbb{P}(X_2 = 1 \,|\, X_1 = 2, X_0 = 1)\mathbb{P}(X_1 = 2 \,|\, X_0 = 1) \\ &\quad + \mathbb{P}(X_2 = 1 \,|\, X_1 = 1, X_0 = 1)\mathbb{P}(X_1 = 1 \,|\, X_0 = 1) \\ &= \mathbb{P}(X_2 = 1 \,|\, X_1 = 2)\mathbb{P}(X_1 = 2 \,|\, X_0 = 1) \\ &\quad + \mathbb{P}(X_2 = 1 \,|\, X_1 = 1)\mathbb{P}(X_1 = 1 \,|\, X_0 = 1) \\ &= q(1-p) + p^2 \end{aligned} \tag{1.11}$$

□

Exercise 1.4. *Compute* $\mathbb{P}(X_2 = 2 \,|\, X_0 = 2)$. *Suppose* X_0 *has the following initial distribution* $\mathbb{P}(X_0 = 1) = \mathbb{P}(X_0 = 2) = 1/2$. *Find the joint probability* $\mathbb{P}(X_0 = 2,\ X_2 = 2)$.

1.4 A first look at asymptotic behavior

Obviously the approach outlined above will get tedious very quickly. What if we wish to find $\mathbb{P}(X_3 = 2 \,|\, X_0 = 2)$ or $\mathbb{P}(X_5 = 2 \,|\, X_0 = 2)$ or

$$\mathbb{P}(X_{1000} = 2 \,|\, X_0 = 2)? \tag{1.12}$$

This sort of question is what we mean by the **asymptotic behavior** of the Markov chain. That is, if we run the chain for a *very large* number of steps, what is the probability it is at any given state? What is its probability distribution among all the vertices? Are some vertices more likely than others? Does a limiting distribution exist? If we change the value of X_0 do the above questions give a different answer? Answering these questions is one of our key motivating problems.

Thinking concretely about (1.12), it is clear that working out all the possible chains starting and ending at 2 through 1000 steps is not feasible.

Fortunately, matrix algebra comes in to save the day in a remarkable way. Notice that the square of the matrix P in (1.8) is

$$P^2 = \begin{bmatrix} p^2 + q(1-p) & p(1-p) + (1-p)(1-q) \\ pq + q(1-q) & (1-p)q + (1-q)^2 \end{bmatrix}.$$

We observe that P^2_{12} is exactly our answer (1.10) for $\mathbb{P}(X_2 = 2 \,|\, X_0 = 1)$, while P^2_{11} is what we found in (1.11) for $\mathbb{P}(X_2 = 1 \,|\, X_0 = 1)$. This is not a coincidence: the (j,k) entry P^n_{jk} of the nth power of P is exactly what you get for summing the probabilities of all possible paths from j to k in n steps, using the Markov property. We formalize this happy fact in the following theorem.

Theorem 1.1. *For any $k \in \mathbb{N} = \{1, 2, 3, \ldots\}$ and $i, j \in \Omega$,*

$$\mathbb{P}(X_k = j \,|\, X_0 = i) = P^k_{ij}, \tag{1.13}$$

the (i,j)-entry of the kth power of the transition matrix P.

We will use the notations P^k_{ij} and $P^k(i,j)$ interchangeably. This theorem tells us that $P^k_{ij} = P^k(i,j)$ is the probability of moving from i to j in exactly k steps.

Proof. We use induction on k. By the definition of P, (1.13) holds for $k = 1$. Now suppose (1.13) holds for some $k = m \geq 1$. We need to show it holds for $k = m + 1$. We condition on the mth step and observe

$$\begin{aligned}
&\mathbb{P}(X_{m+1} = j \,|\, X_0 = i) \\
&\qquad = \sum_{l=1}^{n} \mathbb{P}(X_{m+1} = j, X_m = l \,|\, X_0 = i) \\
&\qquad = \sum_{l=1}^{n} \mathbb{P}(X_{m+1} = j \,|\, X_m = l, X_0 = i) \mathbb{P}(X_m = l \,|\, X_0 = i) \\
&\qquad = \sum_{l=1}^{n} \mathbb{P}(X_{m+1} = j \,|\, X_m = l) \mathbb{P}(X_m = l \,|\, X_0 = i) \qquad (1.14) \\
&\qquad = \sum_{l=1}^{n} P_{lj} P^m_{il} = \sum_{l=1}^{n} P^m_{il} P_{lj} \qquad (1.15)
\end{aligned}$$

$$= (P^m \cdot P)_{ij} = P_{ij}^{m+1}, \tag{1.16}$$

where we used the Markov property in (1.14) and the inductive hypothesis in the first equality of (1.15), and (1.16) is just the definition of matrix multiplication. This proves our desired formula for the case $m+1$, completing the induction argument. □

Note that (1.13) is stated in terms of the number of steps from the initial position. But what about probabilities like

$$\mathbb{P}(X_{100} = y \,|\, X_{90} = x)?$$

By the Markov property, it is as if the entire chain starts afresh at step 90. So the only thing that matters is the number of steps between the two times, and so $\mathbb{P}(X_{100} = y \,|\, X_{90} = x) = \mathbb{P}(X_{10} = y \,|\, X_0 = x) = P_{xy}^{10}$.

Example 1.2 (The 6-cycle, revisited)**.** Let's go back to our simple random walk on the 6-cycle $\Omega = \{1, 2, 3, 4, 5, 6\}$. We found its transition matrix P_2 above in (1.5). Suppose we wanted to find $\mathbb{P}(X_k = 5 \,|\, X_0 = 1)$ for large k. That is, we start our walk at vertex 1, and want to know the likelihood it is at vertex 5 after a large number of steps. Using a computer, we have

$$P_2^{50} \approx \begin{bmatrix} 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \end{bmatrix}, \tag{1.17}$$

while

$$P_2^{51} \approx \begin{bmatrix} 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \end{bmatrix}. \tag{1.18}$$

So, by Theorem 1.1, we have $\mathbb{P}(X_{50} = 5 \,|\, X_0 = 1) \approx 1/3$, while, interestingly,

$\mathbb{P}(X_{51} = 5 \,|\, X_0 = 1) \approx 0$.

What is happening? Note that if we start at vertex 1, then after an *even* number of steps, we must be at an odd-numbered vertex. Similarly, after an *odd* number of steps, we must be at an even-numbered vertex. Our P_2^{50} matrix (1.17) suggests that, after a large even number of steps, our chain is equally likely to be at any of the odd vertices, while (1.18) suggests that, after a large odd number of steps, we are equally-spaced out among the even vertices. In particular, it appears that

$$\lim_{k\to\infty} \mathbb{P}(X_k = 5 \,|\, X_0 = 1)$$

does not exist, since the value alternates between 1/3 and 0 for large values of k (assuming the pattern we see in (1.17) and (1.18) continues, which we would need to verify).

Exercise 1.5. *Find the transition matrix P for the simple random walk on the 5-cycle $\{1, 2, 3, 4, 5\}$. Use a computer to compute P^{50} and P^{51}. Compare and contrast to the case of the 6-cycle above.*

Exercise 1.6. *Recall the transition matrix P_1 in* (1.4) *for the graph in Figure 1.1. Use a computer to compute P_1^{50} and P_1^{51}. What do you find? How does this compare with the 5-cycle example in Exercise 1.5?*

While a computer can typically compute the 50th power of a transition matrix, it can be computationally expensive. And sometimes it is just not possible, as with the matrix (1.4) in Example 1.1 for general p and q (the expressions for each entry would be prohibitively nasty; try, if you dare, in Mathematica or Maple). So while we now know that

$$\mathbb{P}(X_{1000} = 2 \,|\, X_0 = 2) = P_{22}^{1000},$$

we still do not necessarily *comprehend* this probability.

Let's step back and think more generally. Suppose we have a Markov chain on $\Omega = \{1, 2, \ldots, n\}$ with $n \times n$ transition matrix P. Let's start our Markov chain at $X_0 = 1$, which we could describe as beginning with the probability distribution

$$\mu_0 = \underbrace{(1, 0, 0, \ldots, 0)}_{n}$$

on the state space Ω. For $k = 1, 2, 3 \ldots$, let

$$\mu_k = (\mu_{k1}, \mu_{k2}, \ldots, \mu_{kn})$$

be the row vector which is the probability distribution of where our chain is at after k steps. In other words,

$$(\mu_k)_j = \mu_{kj} = \mathbb{P}(X_k = j \,|\, X_0 = 1).$$

Do we have a limit $\lim_{k\to\infty} \mu_{kj}$ for each entry, or a limiting distribution $\lim_{k\to\infty} \mu_k$ as a whole?

Observe that

$$\begin{aligned}\mu_1 &= \big(\mathbb{P}(X_1 = 1 \,|\, X_0 = 1), \mathbb{P}(X_1 = 2 \,|\, X_0 = 1), \ldots, \mathbb{P}(X_1 = n \,|\, X_0 = 1)\big) \\ &= (P_{11}, P_{12}, \ldots, P_{1n}) = \mu_0 P,\end{aligned}$$

where $\mu_0 P$ is the row vector μ_0 times the matrix P (which, in this case, just picks out the top row of P). Similarly,

$$\begin{aligned}\mu_2 &= \big(\mathbb{P}(X_2 = 1 \,|\, X_0 = 1), \mathbb{P}(X_2 = 2 \,|\, X_0 = 1), \ldots, \mathbb{P}(X_2 = n \,|\, X_0 = 1)\big) \\ &= \big(P_{11}^2, P_{12}^2, \ldots, P_{1n}^2\big) \\ &= \mu_0 P^2 = (\mu_0 P) P = \mu_1 P,\end{aligned}$$

since we found $\mu_0 P = \mu_1$. Continuing this pattern, we see

$$\mu_{k+1} = \mu_k P \tag{1.19}$$

for all k. Now *suppose* a limiting vector $\pi = \lim_{k\to\infty} \mu_k$ exists. Then taking limits on both sides of (1.19) gives

$$\pi = \lim_{k\to\infty} \mu_{k+1} = \lim_{k\to\infty} (\mu_k P) = (\lim_{k\to\infty} \mu_k) P = \pi P,$$

and so such a limit would satisfy

$$\boxed{\pi = \pi P} \tag{1.20}$$

That is, π would be a row vector that is a *left eigenvector* of P with eigenvalue 1. This equation (1.20) will prove very significant in our investigation

of Markov chains and their limiting behavior. We are immediately led to a number of questions:

1. When does such a solution π of (1.20) exist? Can π be taken to a probability vector? That is, is there a solution to (1.20) that has nonnegative coordinates that add up to one? Finally, if such a solution exists, is it unique?

2. Does

$$\lim_{k\to\infty} \mu_{kj} = \lim_{k\to\infty} \mathbb{P}(X_k = j \,|\, X_0 = 1) = \pi_j? \tag{1.21}$$

   Or, more generally, does the initial value matter? Is

$$\lim_{k\to\infty} \mathbb{P}(X_k = j \,|\, X_0 = i) = \pi_j \tag{1.22}$$

   for any starting point $i \in \{1, 2, \ldots, n\}$?

3. If (1.21) or (1.22) holds, what is the rate of convergence? In other words, how many steps k does it take until the probability we are at vertex j is within a given magnitude of error of π_j?

Exercise 1.7. *Show that* $\pi = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ *is a left eigenvector for the 6-cycle transition matrix* (1.5).

Exercise 1.7 shows that answering question 1 affirmatively above is not the same as that for question 2. Indeed, we have such a π for the simple random walk on the 6-cycle, but our computations in Example 1.2 suggest that a limiting probability distribution doesn't exist, since the walk alternates between even and odd vertices.

In order to give a positive answer to both questions 1 and 2, we will see that our Markov chain needs the properties of *irreducibility* and *aperiodicity*, which we introduce in the next section. In Chapter 3, we will see that irreducible and aperiodic Markov chains converge exponentially fast to their limiting distributions π (see Theorem 3.4), answering question 3.

1.5 Irreducibility and aperiodicity

Definition 1.4. A Markov chain X on Ω with transition matrix P is **irreducible** if for any two states $x, y \in \Omega$, there exists $k \in \mathbb{N}$ such that

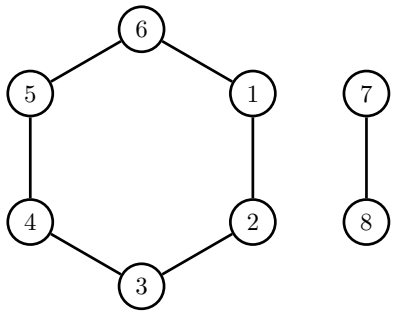

Figure 1.8: Disconnected graphs are not irreducible.

$(P^k)_{x,y} > 0$. In other words, *for any pair of states* x and y, there is a positive probability that the chain can move from x to y in some finite number of steps.

Example 1.3. The random walk on the six-cycle is irreducible. If $x = 1$ and $y = 4$, for instance, we have a path $1 \to 2 \to 3 \to 4$ which has probability $(1/2)^3 > 0$. Similarly, any two vertices have a path of positive probability connecting them.

However, if we add a disconnected line segment to the graph, as in Figure 1.8, then the random walk is no longer irreducible: there is no positive-probability path from vertex 1 to 7, for example.

The random walk on the graph in Figure 1.1 is also irreducible, as is the weighted walk on the graph in Figure 1.7, so long as $0 < p < 1$ and $0 < q < 1$. However, if either $p \in \{0, 1\}$ or $q \in \{0, 1\}$, the graph becomes irreducible. For instance, if $p = 1$, then we cannot get from $x = 1$ to $y = 2$. □

From the point of view of the graph $G = (V, E)$, irreducibility of the simple symmetric random walk on G is equivalent to the connectivity of the graph. That is, there must be a path on the graph from any one vertex to any other. This also works for weighted random walks if all edges have positive probability either way. The graph must be in "one piece".

Exercise 1.8. *Is the Markov chain in Problem 1.4 irreducible? Why or why not?*

Aperiodicity is a slightly more subtle property.

Definition 1.5. Let X be an irreducible Markov chain. For any state $x \in \Omega$, let

$$T(x) := \{\, k \ : \ P^k(x,x) > 0 \,\}.$$

That is, $T(x)$ is the collection of all times k for which it is possible to return to x in k steps. The **period of** X is

$$\gcd(T(x)),$$

the greatest common divisor of all the integers in $T(x)$. The Markov chain is **aperiodic** if the period is 1.

The first question to settle is whether or not this definition is legitimate: can two different states $x, y \in \Omega$ yield $\gcd(T(x)) \neq \gcd(T(y))$? If so, then our definition of the period of X would vary from vertex to vertex. The next theorem, however, says that the period is the same for all vertices.

Theorem 1.2. *If X is an irreducible Markov chain, then*

$$\gcd(T(x)) = \gcd(T(y))$$

for any two states $x, y \in \Omega$.

Although we omit the proof, the idea is to use the irreducibility of the chain: since X is irreducible/connected, any information about one vertex x can be "communicated" to any other vertex y by using a path from x to y. See [2, Lemma 1.6] for proof details.

The definition of aperiodicity is somewhat opaque at first sight. The following exercise should help.

Exercise 1.9. *Suppose there is a state $x \in \Omega$ such that $P(x,x) > 0$. Argue that the Markov chain must be aperiodic.*

The previous exercises can be generalized with the help of a fact from number theory to give a helpful interpretation of the period. If the period of a chain is d, then for each state x, there exists an N (that may depend on x) such that

$$P^{nd}(x,x) > 0$$

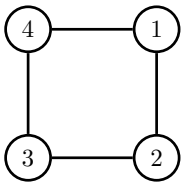

Figure 1.9: Is the walk on the 4-cycle aperiodic?

for all $n \geq N$. In other words, it is possible to return to the same state x in all sufficiently large multiples of the period. And if the chain is aperiodic, i.e. $d = 1$, this says that $P^n(x, x) > 0$ for all sufficiently large n: it is possible to return to x in *all* sufficiently large numbers of steps.

Example 1.4. The simple random walk on the 4-cycle $\{1, 2, 3, 4\}$, as in Figure 1.9, has transition matrix

$$P = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}. \tag{1.23}$$

Since the 4-cycle is a connected graph, the chain is irreducible. Is it aperiodic? Let's consider $T(1) = \{\, k \ : \ P^k(1, 1) > 0 \,\}$, the set of possible return times to 1. Some possible paths from 1 to 1 include

$$\begin{aligned} &1 \to 2 \to 1 && k = 2 \\ &1 \to 2 \to 3 \to 2 \to 1 && k = 4 \\ &1 \to 2 \to 3 \to 4 \to 1 && k = 4 \\ &1 \to 2 \to 3 \to 2 \to 3 \to 4 \to 1 && k = 6 \end{aligned}$$

We claim that *all* possible paths from 1 to 1 require an even number of steps. Indeed, given a path p, write the length k_p as

$$k_p = a_p + b_p,$$

where a_p is the number of clockwise steps in the path, and b_p the number of counter-clockwise. Numbering the paths above as 1 to 4, respectively, we

have

$$\begin{aligned} a_1 &= b_1 = 1 \\ a_2 &= b_2 = 2 \\ a_3 &= 4, b_3 = 0 \\ a_4 &= 5, b_4 = 1. \end{aligned}$$

In general, we claim that

$$a_p = b_p + 4n \tag{1.24}$$

for some $n \in \mathbb{Z}$, as we can begin to see from our examples. Why (1.24)? As we are travelling from 1 to 1, if we do a "sub-cycle" like $3 \to 2 \to 3$ in the fourth path, we add the same number of clockwise and counter-clockwise steps. The remaining difference $4n$ accounts for how many times we loop around the square.

Thus, the total number of steps is

$$\begin{aligned} k_p &= a_p + b_p \\ &= 2b_p + 4n, \end{aligned}$$

which is always even. Since a return of length 2 is possible, $\gcd(T(1)) = 2$. So we see that the walk on the 4-cycle is not aperiodic, but is rather periodic with period 2.

Exercise 1.10. *For the random walk on the 4-cycle, compute*

$$\mathbb{P}(X_{1000} = 3 \,|\, X_0 = 1) \quad \textit{and} \quad \mathbb{P}(X_{1000} = 2 \,|\, X_0 = 1).$$

Example 1.5. What about a random walk X on the 5-cycle $\{1, 2, 3, 4, 5\}$? This is again clearly irreducible, but it is also periodic? Since paths from 1 to 1 include

$$\begin{aligned} &1 \to 2 \to 1 && k = 2 \\ &1 \to 2 \to 3 \to 4 \to 5 \to 1 && k = 5, \end{aligned}$$

we see $T(1)$ contains co-prime integers 2 and 5, yielding $\gcd(T(x)) = 1$.

So while the 4-cycle walk is periodic, the 5-cycle walk is aperiodic. These arguments easily generalize to walks on n-cycles.

Theorem 1.3. *Let X be a random walk on an n-cycle. If n is even, the period of X is 2. If n is odd, X is aperiodic.*

Proof. Repeat the logic of examples 1.4 and 1.5 for the even and odd cases, respectively. □

Exercise 1.11. *Suppose that every non-diagonal element of the transition probability matrix P is strictly positive. That is, $P(x, y) > 0$, for all $x \neq y \in \Omega$. We don't assume anaything about the diagonal elements (in fact, every $P(x, x)$ could be zero). Argue that the Markov chain is aperiodic.*

Exercise 1.12. *Let G be a connected graph. Suppose there is a path of odd-length $2k + 1$, for some k, from one given vertex x to itself. Show that the random walk on G is aperiodic.*

1.6 Stationary distributions π

So far our calculations in terms of probabilities for Markov chains have all been conditional. We know how to compute things like

$$\mathbb{P}(X_1 = y \,|\, X_0 = x) \text{ or } \mathbb{P}(X_{14} = y \,|\, X_0 = x) \text{ or } \mathbb{P}(X_{350} = y \,|\, X_{125} = x)$$

using the transition matrix P and its powers. But what if we want the *unconditional* probability $\mathbb{P}(X_1 = y)$?

Suppose the **initial distribution** of X_0 is given by the vector

$$\mu_0 = (q_1, q_2, \ldots, q_n) := (\mathbb{P}(X_0 = 1), \mathbb{P}(X_0 = 2), \ldots, \mathbb{P}(X_0 = n)).$$

By the law of total probability,

$$\mathbb{P}(X_1 = y) = \sum_{j=1}^{n} \mathbb{P}(X_1 = y \,|\, X_0 = j)\mathbb{P}(X_0 = j) = \sum_{j=1}^{n} P_{jy} q_j. \tag{1.25}$$

So, if we have an explicit initial distribution $\mu_0 = (q_1, q_2, \ldots, q_n)$, we can compute unconditional probabilities. Note that we can view (1.25) as

matrix multiplication

$$\sum_{j=1}^{n} P_{jy} q_j = (\mu_0 P)_y,$$

the yth entry of row-matrix product $\mu_0 P$. Thus, denoting the *unconditional* probability row vector for the first step as μ_1,

$$\mu_1 := (\mathbb{P}(X_1 = 1), \mathbb{P}(X_1 = 2), \ldots, \mathbb{P}(X_1 = n)),$$

we have

$$\begin{aligned} \mu_1 =& ((\mu_0 P)_1, (\mu_0 P)_2, \ldots, (\mu_0 P)_n) \\ =& \mu_0 P. \end{aligned} \tag{1.26}$$

Similarly, by the law of total probability and (1.26),

$$\begin{aligned} \mu_2 :=& (\mathbb{P}(X_2 = 1), \mathbb{P}(X_2 = 2), \ldots, \mathbb{P}(X_2 = n)) \\ =& \mu_1 P = (\mu_0 P) P = \mu_0 P^2, \end{aligned}$$

and, continuing this logic, we have

$$\mu_k := (\mathbb{P}(X_k = 1), \mathbb{P}(X_k = 2), \ldots, \mathbb{P}(X_k = n)) = \mu_0 P^k. \tag{1.27}$$

So, matrix powers also give us *unconditional* probabilities after multiple steps, so long as we *left-multiply* by the initial distribution.

We may have a starting distribution μ_0 that is *deterministic*, i.e. that always begins the chain at some fixed vertex j, and so μ_0 is all zeros except for a single 1 at entry j. This is indeed what we did in the derivation leading up to (1.20). But it doesn't have to be the case. We can also randomly choose our initial position.

Example 1.6. Consider the random walk on the 6-cycle (hexagon) with $X_0 \sim \text{Unif}(1, 2, \ldots, 6)$. That is,

$$\mu_0 = \left(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6} \right). \tag{1.28}$$

What is $\mathbb{P}(X_1 = 2)$? If $X_1 = 2$, clearly X_0 could only be 1 or 3. Repeating

our calculation with the law of total probability, we find

$$\begin{aligned}\mathbb{P}(X_1 = 2) &= \sum_{j=1}^{6} \mathbb{P}(X_1 = 2 \,|\, X_0 = j)\mathbb{P}(X_0 = j) \\ &= \mathbb{P}(X_1 = 2 \,|\, X_0 = 1)\mathbb{P}(X_0 = 1) + \mathbb{P}(X_1 = 2 \,|\, X_0 = 3)\mathbb{P}(X_0 = 3) \\ &= \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{6}.\end{aligned}$$

So, note that we interestingly have

$$\mathbb{P}(X_1 = 2) = \mathbb{P}(X_0 = 2) = \frac{1}{6}$$

under the initial distribution (1.28). But a moments thought shows this computation would equally apply to *any* vertex in the 6-cycle, and so

$$\mathbb{P}(X_1 = j) = \mathbb{P}(X_0 = j) = \frac{1}{6}, \tag{1.29}$$

for *every* j. Thus if we start the chain with the distribution μ_0, (1.29) says that our distribution after one step is still μ_0. But then we can repeat this process: taking another step is just like starting over by the Markov property, and so the distribution of X_k after *any* number of steps is still μ_0!

This is rather remarkable. Even though our state X_n varies over time, from a probabilistic point of view our chain is stationary: the probability that we are at any given vertex j after k steps is always $q_j = (\mu_0)_j$. Such starting distributions, when they exist, will prove to be significant, and we give them the obvious name.

Definition 1.6. For a Markov chain X on $\Omega = \{1, 2, \ldots, n\}$ with transition matrix P, a **stationary distribution**

$$\pi = (\pi_1, \pi_2, \ldots, \pi_n)$$

is a probability distribution on Ω (arranged in a row vector) such that

$$\pi P = \pi. \tag{1.30}$$

In other words, π is a left eigenvector for P with eigenvalue $\lambda = 1$.

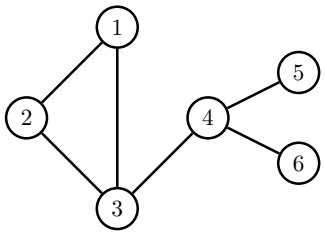

Figure 1.10: What is π for the simple random walk on this graph?

The observation we made above bears repeating. If we have π as in (1.30) and we start the chain at a random vertex distributed according to π, what is the distribution of X_k? From (1.27) and the eigenvector property (1.30), we have

$$\begin{aligned}\mu_k :=&(\mathbb{P}(X_k = 1), \mathbb{P}(X_k = 2), \ldots, \mathbb{P}(X_k = n)) \\ =&\pi P^k = (\pi P)P^{k-1} = \pi P^{k-1} = (\pi P)P^{k-2} = \pi P^{k-2} = \cdots = \pi P = \pi.\end{aligned}$$

So, if we start the chain with the stationary distribution, the distribution of the chain is stationary: X_k is *always* distributed according to π; X_k's pdf is always π.

The following exercise is a helpful observation that we will repeatedly use, and will be a good tool for you to keep in mind.

Exercise 1.13. (*a*) *Suppose a Markov chain on $\Omega = \{1, 2, \ldots, n\}$ has a transition matrix P such that every column of P sums to 1. Show that the uniform distribution $\mu = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ on Ω is a stationary distribution for P.*

(*b*) *Build a Markov chain on the state space $\Omega = \{1, 2\}$ for which the uniform distribution is not stationary.*

When does such a distribution π exist? If it exists, is it unique? These are natural questions that we seek to answer in subsequent sections. There is one class of Markov chains for which the answer is simple.

1.6.1 π for simple random walks on graphs

Theorem 1.4. *A simple random walk on a connected graph $G = (V, E)$ has a unique stationary distribution $\pi = (\pi_v)_{v \in V}$, given by the formula*

$$\pi_v = \frac{\deg(v)}{2|E|} \tag{1.31}$$

for each $v \in V$. Here $|E|$ is the number of edges of G.

Note that the assumption that G is connected is equivalent to saying that the random walk is irreducible - it is possible to travel from any one vertex to any other. Note also there is no assumption on aperiodicity.

Before we prove this, let's work through some exercises and examples.

Exercise 1.14. *Theorem 1.4 says that π is unique when G is connected. Show by a simple example that π may not be unique if G is not connected.*

Exercise 1.15. *It is a fact that the denominator in* (1.31) *satisfies the formula*

$$2|E| = \sum_{v \in V} \deg(v).$$

Give an argument for why this formula holds. Futhermore, why must this be the denominator in (1.31)*?*

Example 1.7. Consider the walk on the 6-cycle again. Here, $\deg(v) = 2$ for each vertex, and $|E| = 6$. Hence

$$\pi_v = \frac{2}{12} = \frac{1}{6}$$

for each vertex v. So, as we saw above, π is the uniform distribution on V.

Example 1.8. The graph in Figure 1.1 has six edges, and thus we have

$$\pi = \Big(\frac{1}{12}, \frac{3}{12}, \frac{3}{12}, \frac{3}{12}, \frac{2}{12}\Big) = \Big(\frac{1}{12}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}\Big).$$

Exercise 1.16. *Give an intuitive explanation for why vertices of higher degree must have more probability in π.*

Exercise 1.17. *Find the stationary distribution of the graph in Figure 1.10.*

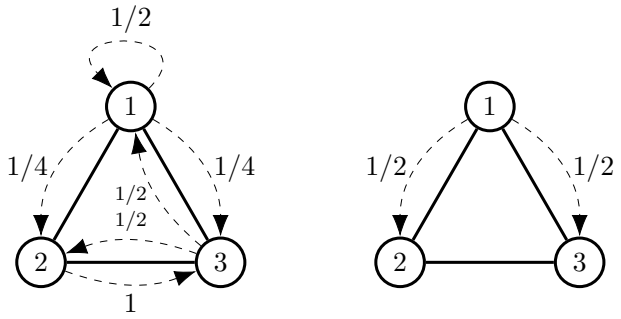

Figure 1.11: A non-simple walk and a simple random walk on a 3-cycle. Their differing transition probabilities yield different stationary distributions.

Note that the hypothesis that the Markov chain on the graph G in Theorem 1.4 is *simple* is also essential. Non-simple random walks on graphs as in §1.1.3 are completely legitimate, but Theorem 1.4 doesn't apply to them. Moreover, such a simple formula for the stationary distribution does not exist for arbitrary Markov chains. Simple random walks are special!

Example 1.9. For instance, we could consider the two Markov chains on the 3-cycle, as described in Figure 1.11. The transition matrices are

$$P_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \quad \text{and} \quad P_2 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

The latter matrix P_2 defines the simple random walk, and each entry of its stationary distribution π_2 has the same formula $\pi_2(v) = \frac{2}{2 \cdot 3} = \frac{1}{3}$. However, we find that $\pi_1 = (\frac{4}{11}, \frac{3}{11}, \frac{4}{11})$ satisfies

$$\pi_1 = \pi_1 P_1,$$

and so π_1 is a stationary distribution for the non-simple walk. In particular, the formula (1.31) does not apply to P_1.

Proof of Theorem 1.4. Recall that the entries for the transition matrix P are $P_{ij} = \frac{1}{\deg(i)}$ if $i \sim j$ and 0 otherwise. We need to show $\pi P = \pi$, and we can do this component-wise by showing

$$(\pi P)_j = \pi_j \tag{1.32}$$

for each $j \in \{1, 2, \ldots, n\}$, where $n = |V|$. Indeed, for a fixed j, $(\pi P)_j$ is the result of multiplying π with the jth column of P, and hence

$$\begin{aligned}(\pi P)_j &= \sum_{i=1}^{n} \pi_i P_{ij} \\ &= \sum_{i \sim j} \frac{\deg(i)}{2|E|} \cdot \frac{1}{\deg(i)} \\ &= \sum_{i \sim j} \frac{1}{2|E|} = \frac{\deg(j)}{2|E|} = \pi_j,\end{aligned}$$

giving (1.32). (The second-to-last equality holds because the number of vertices i adjacent to j is, by definition, $\deg(j)$.) □

Example 1.10. As a last example, we look at another *non*-simple random walk. Consider the n-cycle where we move clockwise with probability p and counter-clockwise with probability $1 - p$, $0 < p < 1$. What is the stationary distribution π_p? Does it depend on p? Surprisingly it does not, and we obtain the uniform distribution for all p,

$$\pi_p \sim \text{Unif}\{1, 2, \ldots, n\}.$$

One way to see this is to try a specific case and then extrapolate. If we do the biased walk on the 5-cycle, for instance, we have the transition matrix

$$P = \begin{bmatrix} 0 & p & 0 & 0 & 1-p \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ p & 0 & 0 & 1-p & 0 \end{bmatrix},$$

and we can simply compute that

$$\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right) P = \left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right).$$

The underlying reason is that the columns still add to 1 (why?), which we saw in 1.13 (a) yields a uniform π.

1.7 Hitting times and return times

So long as we can find a stationary distribution π, we now know how to run a Markov chain in such a way that we know the exact distribution of X_n for all n: we simply start with $X_0 \sim \pi$. There are other important questions about what will happen in the future, though. For instance, how long will it take for a chain to reach a certain state? Or how many steps does it take to return to a given state, on average? These turn out to be important questions in furthering our analysis, and these are the questions we take up in this section. We now have some groundwork of vocabulary and fundamentals and can go a bit deeper in our study of Markov chains.

Definition 1.7. The **hitting time** τ_x of a state $x \in \Omega$ is the first nonnegative time that the Markov chain is in state x. That is,

$$\tau_x := \min\{\, k = 0, 1, 2, \dots \,:\, X_k = x \,\}. \tag{1.33}$$

So note that τ_x is a *random variable*, a random amount of time. Different runs of the chain will result in different hitting times τ_x, and we can ask questions like, "What is the average $\mathbb{E}(\tau_x)$ of τ_x?" Or, "What is τ_x's distribution?" Note also that if we start the chain at $X_0 = x$, then (1.33) says $\tau_x = 0$; a hitting time of zero *is* allowed.

On the other hand, sometimes we wish to know when we first *return* to a given state. Then we exclude the possibility $k = 0$ in (1.33).

Definition 1.8. The (first) **return time** τ_x^+ of x is the first time after the chain has started that we assume the state x. That is,

$$\tau_x^+ := \min\{\, k = 1, 2, \dots, \,:\, X_k = x \,\}. \tag{1.34}$$

This is, of course, another random amount of time, just like τ_x. While $\tau_x \geq 0$, however, we have $\tau_x^+ \geq 1$. So if we start the chain at x, then $\tau_x = 0$ but we do not automatically know anything about τ_x^+, other than $\tau_x^+ \geq 1$.

We can extend these definitions in a natural way to collections of states $A \subset \Omega$. We define the **hitting and return times to a set** A as, respectively,

$$\tau_A := \min\{\, k = 0, 1, 2, \dots, \,:\, X_k \in A \,\},$$

$$\tau_A^+ := \min\{\, k = 1, 2, \ldots, \,:\, X_k \in A \,\}.$$

These are the first time, and the first time after starting, that the chain enters the set A, respectively.

Example 1.11. Consider the walk on $\Omega = \{0, 1\}$ defined by

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

See Figure 1.12. Suppose $X_0 = 0$ (equivalently, $\mu_0 = (1, 0)$). Then the hitting time τ_1 of state 1 is a random variable. What is its distribution? We compute

$$\begin{aligned}
&\mathbb{P}(\tau_1 = 0) = 0 \\
&\mathbb{P}(\tau_1 = 1) = p \\
&\mathbb{P}(\tau_1 = 2) = \mathbb{P}(0 \to 0 \to 1) = (1-p)p \\
&\mathbb{P}(\tau_1 = 3) = \mathbb{P}(0 \to 0 \to 0 \to 1) = (1-p)^2 p \\
&\qquad\qquad \vdots \\
&\mathbb{P}(\tau_1 = k) = (1-p)^{k-1} p,
\end{aligned}$$

since to hit 1 for the first time in k steps, we must stay at vertex 0 for $k-1$ steps, and then move to vertex 1. This is the familiar pmf of a $\mathrm{Geo}(p)$ random variable, and so $\tau_1 \sim \mathrm{Geo}(p)$.

What about τ_1^+? Well, since we are starting with $X_0 = 0 \neq 1$, the minima in (1.33) and (1.34) give exactly the same numbers; $k = 0$ is not a possibility. So $\tau_1^+ \sim \mathrm{Geo}(p)$ too.

Similarly, if $X_0 = 1$, then $\tau_0 = \tau_0^+ \sim \mathrm{Geo}(q)$.

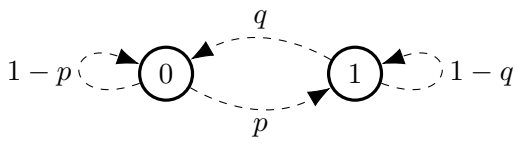

Figure 1.12: What are the hitting times for this Markov chain?

How about the distribution of τ_0^+, when $X_0 = 0$? We find

$$\begin{aligned}
\mathbb{P}(\tau_0^+ = 1) &= 1 - p \\
\mathbb{P}(\tau_0^+ = 2) &= \mathbb{P}(0 \to 1 \to 0) = pq \\
\mathbb{P}(\tau_0^+ = 3) &= \mathbb{P}(0 \to 1 \to 1 \to 0) = p(1-q)q \\
\mathbb{P}(\tau_0^+ = 4) &= \mathbb{P}(0 \to 1 \to 1 \to 1 \to 0) = p(1-q)^2 q \\
\mathbb{P}(\tau_0^+ = 5) &= \mathbb{P}(0 \to 1 \to 1 \to 1 \to 1 \to 0) = p(1-q)^3 q \\
&\vdots \\
\mathbb{P}(\tau_0^+ = k) &= p(1-q)^{k-2} q,
\end{aligned}$$

for $k \geq 2$. In particular, note that τ_0^+ is *not* a geometric random variable (even though it is similar to one); it is not one of our familiar named distributions.

1.7.1 Expected hitting times

One of simplest questions to ask about a random variable is for its average. Can we say anything about the expectations of our new random variables τ_x, τ_x^+, τ_A and τ_A^+? On average, how long does it take to hit or return to x or A?

Example 1.12. For example, suppose the vertices of our graph were all airports in the world, with $x \sim y$ if there is a flight between airport x and y. Now, imagine a virus arises in some city x. How much time would other countries, on average, have to prepare before the virus reaches them via flights? Using a Markov chain model, this simply becomes $\mathbb{E}(\tau_A)$ for the given country A.

We will be able to explicitly compute the averages using linear algebra. Let's first establish some notation. As usual, our state space is $\Omega = \{1, 2, \ldots, n\}$, and without loss of generality we may assume that the set A in question consists of the last k elements of Ω, $A = \{n-k+1, n-k+2, \ldots, n\}$ for some $1 \leq k < n$ (if A didn't consist of just the last k vertices, we can relabel the states so this is so). Now, partition the transition matrix P as

$$P = \left[\begin{array}{c|c} P_{(n-k)\times(n-k)} & P_{(n-k)\times k} \\ \hline P_{k\times(n-k)} & P_{k\times k} \end{array}\right].$$

Here $P_{(n-k)\times(n-k)}$ is the top left $(n-k)\times(n-k)$ submatrix, giving all probabilities of moving from a state in $\{1,2,\ldots,n-k\}$ to a state in $\{1,2,\ldots,n-k\}$. $P_{(n-k)\times k}$ is the top right submatrix, corresponding to the probabilities of moving from states $\{1,2,\ldots,n-k\}$ to states $\{n-k+1,k+2,\ldots,n\}$. And so on for the bottom two blocks.

The submatrix that is most important to us for τ_A is the top left block $P_{(n-k)\times(n-k)} =: Q$. This is because Q gives the probabilities of starting and remaining outside of A. Define the $(n-k)\times(n-k)$ matrix M as

$$M := (I-Q)^{-1},$$

where I is the $(n-k)\times(n-k)$ identity matrix. The next theorem says that, surprisingly, our expected hitting time comes from summing the row of M corresponding to our starting state.

Theorem 1.5. *Let X be an irreducible Markov chain on Ω with transition matrix P. With the notation as above, let*

$$i \in \{1,2,\ldots,n-k\} = \Omega\backslash A$$

be any state outside of A. Then

$$\mathbb{E}(\tau_A \,|\, X_0 = i) = \sum_{j=1}^{n-k} M_{ij}. \tag{1.35}$$

So to find $\mathbb{E}(\tau_A \,|\, X_0 = i)$, we sum the elements of the ith row of the matrix M; while the notation is somewhat complex, and it is not immediately clear why this should be true, that much is simple. In matrix Q, the ith row gives the probabilities of staying outside of the set A, starting from state $i \notin A$. We will see that the ith row of M gives the expected number of *visits* to each of the states outside of A before hitting A, starting at i.

Example 1.13. Let's consider the simple random walk on the graph in Figure 1.13, which has transition matrix

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

Suppose we want to find the expected hitting time to $A = \{3, 4\}$ starting from either vertex 1 or 2. We partition the transition matrix P as

$$P = \left[\begin{array}{cc|cc} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \hline \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right],$$

isolating the vertices corresponding to A. Our matrix Q is then the top-left corner

$$Q = \begin{bmatrix} 0 & \frac{1}{3} \\ \frac{1}{2} & 0 \end{bmatrix},$$

the probabilities that we start and remain outside of A (if our vertices were labelled differently, Q would not be in the top-left corner; we then simply select the submatrix with these probabilities). Thus

$$I - Q = \begin{bmatrix} 1 & -\frac{1}{3} \\ -\frac{1}{2} & 1 \end{bmatrix} \quad \text{and} \quad (I-Q)^{-1} = \begin{bmatrix} \frac{6}{5} & \frac{2}{5} \\ \frac{3}{5} & \frac{6}{5} \end{bmatrix},$$

and (1.35) says

$$\begin{aligned} \mathbb{E}(\tau_A \,|\, X_0 = 1) &= \frac{6}{5} + \frac{2}{5} = \frac{8}{5}, \\ \mathbb{E}(\tau_A \,|\, X_0 = 2) &= \frac{3}{5} + \frac{6}{5} = \frac{9}{5}. \end{aligned}$$

So, if we begin at vertex 1, on average it takes $8/5$ steps to enter $\{3, 4\}$. If we start at vertex 2, it takes $9/5$ steps, on average. (We are not surprised that $\mathbb{E}(\tau_A \,|\, X_0 = 1) < \mathbb{E}(\tau_A \,|\, X_0 = 2)$, as there are more routes to A from vertex 1 than from vertex 2.) □

Proof of Theorem 1.5. We first claim that, for any $j \notin A$,

$$M_{ij} = \mathbb{E}(\,\#\{\text{visits to } j \text{ before hitting } A\} \mid X_0 = i\,) \tag{1.36}$$

Before we prove this, let's see that this would give our result (1.35). Indeed,

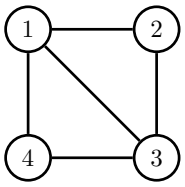

Figure 1.13: Computing $\mathbb{E}(\tau_A \,|\, X_0 = i)$ for $A = \{3, 4\}$, $i \in \{1, 2\}$.

we can decompose the hitting time τ_A into the sum

$$\tau_A = \sum_{j \notin A} \#\{\text{visits to } j \text{ before hitting } A\}$$

and thus taking expectations yields

$$\begin{aligned}
\mathbb{E}(\tau_A \,|\, X_0 = i) &= \mathbb{E}\Big(\sum_{j \notin A} \#\{\text{visits to } j \text{ before hitting } A\} \,\Big|\, X_0 = i \Big) \\
&= \sum_{j \notin A} \mathbb{E}\big(\#\{\text{visits to } j \text{ before hitting } A\} \,|\, X_0 = i \big) \\
&= \sum_{j \notin A} M_{ij}, \qquad (1.37)
\end{aligned}$$

assuming (1.36) holds. But (1.37) is precisely the claimed result (1.35). So we see that it is enough to prove (1.36).

Recall $Q = P_{(n-k)\times(n-k)}$ gives the probabilities of moving around in states outside of A, and thus

$$(Q^\ell)_{ij} = \mathbb{P}(\, X_\ell = j, \tau_A > \ell \,|\, X_0 = i \,).$$

That is, Q^ℓ gives the probabilities of moving around Ω *outside* of the set A for ℓ steps (just as P^ℓ gives the probabilities for moving anywhere in Ω in ℓ steps). Recalling the expectation formula for indicator random variables

$$\mathbb{E}(\mathbb{1}_B) = 1 \cdot \mathbb{P}(B) + 0 \cdot \mathbb{P}(B^c) = \mathbb{P}(B),$$

we see

$$(Q^\ell)_{ij} = \mathbb{P}(\, X_\ell = j, \tau_A > \ell \,|\, X_0 = i \,) = \mathbb{E}\big(\mathbb{1}_{\{X_\ell = j, \tau_A > \ell\}} \,|\, X_0 = i \,\big),$$

and so

$$\mathbb{E}(\,\#\{\text{visits to } j \text{ before hitting } A\} \mid X_0 = i\,) = \mathbb{E}\Big(\sum_{\ell=0}^{\infty} \mathbb{1}_{\{X_\ell = j, \tau_A > \ell\}} \;\Big|\; X_0 = i\Big)$$

$$= \sum_{\ell=0}^{\infty} \mathbb{E}\big(\,\mathbb{1}_{\{X_\ell = j, \tau_A > \ell\}} \,|\, X_0 = i\,\big) = \sum_{\ell=0}^{\infty} (Q^\ell)_{ij}. \tag{1.38}$$

Now, the matrix Q is *sub-stochastic*, which is to say each row of Q sums to ≤ 1 (instead of exactly 1 for the entire *stochastic* matrix P). Furthermore, since our chain is irreducible, one of the rows of Q must sum to < 1, since if every row summed to 1, it would be impossible to reach a state in A from a state in $\Omega \backslash A$, and the chain would be reducible. For such sub-stochastic matrices, we have the identity

$$\sum_{\ell=0}^{\infty} Q^\ell = (I - Q)^{-1} = M. \tag{1.39}$$

(See [1]. In essence, this is saying that the familiar geometric series formula

$$\sum_{\ell=0}^{\infty} x^\ell = \frac{1}{1-x} = (1-x)^{-1},$$

which holds for *numbers* $|x| < 1$, also holds for sub-stochastic *matrices* Q which satisfy "$|Q| < 1$," which is a beautiful fact.) Given (1.39), we have

$$\begin{aligned} M_{ij} &= \Big(\sum_{\ell=0}^{\infty} Q^\ell\Big)_{ij} \\ &= \sum_{\ell=0}^{\infty} (Q^\ell)_{ij} \\ &= \mathbb{E}(\,\#\{\text{visits to } j \text{ before hitting } A\} \end{aligned}$$

by (1.38). This is exactly (1.36), what we needed to prove. □

Exercise 1.18. *For the simple random walk on the graph in Figure 1.13, compute* $\mathbb{E}(\tau_A \mid X_0 = 4)$*, where* $A = \{2, 3\}$*.*

1.7.2 Expected return times

Theorem 1.5 tells us how to find the average hitting time for A when we start outside of A. What about for the expected *return* time $\mathbb{E}(\tau_A^+ \mid X_0 = i)$ to A? We can start with an obvious observation: if $i \notin A$, then "returning" is the same as hitting, and so $\tau_A^+ = \tau_A$ and we can use Theorem 1.5. However, this is still rather cumbersome if we wish to focus on an individual state $A = \{x\}$, since we would need to invert an $(n-1) \times (n-1)$ matrix $I - Q$, and then sum up the row corresponding to our starting vertex.

Is there a simpler way? It turns out the answer is yes, if we wish to find the return time to our starting vertex. Working out the elegant formula (1.40) is the point of this section.

Theorem 1.6. *Let π be a stationary distribution of an irreducible Markov chain X on $\Omega = \{1, 2, \ldots, n\}$. Then, for any $x \in \Omega$,*

$$\mathbb{E}(\tau_x^+ \mid X_0 = x) = \frac{1}{\pi(x)}. \tag{1.40}$$

In particular, irreducible Markov chains always have a unique stationary distribution π, given by

$$\pi(x) = \frac{1}{\mathbb{E}(\tau_x^+ \mid X_0 = x)} \tag{1.41}$$

for each $x \in \Omega$.

It is clear that if *any* stationary distribution satisfies (1.40), then irreducible chains have a unique stationary distribution, since they all must be defined by the formula (1.41).

To completely prove Theorem 1.6, we need some tools from linear algebra that we currently lack. We can immediately, however, prove the following weaker version as a "stepping stone" to Theorem 1.6. We will use Theorem 1.7 and the *Perron-Frobenius theorem* to prove Theorem 1.6 later in §3.2.3.[‡]

Theorem 1.7. *Any irreducible Markov chain has a stationary distribution π. More explicitly, let X be an irreducible Markov chain on $\Omega = \{1, 2, \ldots, n\}$*

[‡]Even there, however, we will use a part of the Perron-Frobenius theorem that we state without proof, which leaves something to be desired. We will completely tie this down by using *harmonic functions* in §5.3 to complete the proof of Theorem 1.6 only using Theorem 1.7. See the end of §5.3.

with transition matrix P. Then the row vector $\pi \in \mathbb{R}^n$ defined by (1.41) *is a probability distribution on Ω and satisfies*

$$\pi P = \pi. \tag{1.42}$$

Before diving into the proof, let's simplify our notation by defining

$$\mathbb{P}_x(X_j = k) := \mathbb{P}(X_j = k \,|\, X_0 = x), \tag{1.43}$$

$$\mathbb{E}_x(\tau_A) := \mathbb{E}(\tau_A \,|\, X_0 = x). \tag{1.44}$$

As we're usually considering conditional probabilities, this will make our expressions more succinct.

Proof of Theorem 1.7. Consider

$$\tilde{\pi}(y) := \mathbb{E}_x\big(\#\{\text{visits of } y \text{ before } \tau_x^+\}\big),$$

the average number of times we visit state y before returning to x. Note that

$$\tilde{\pi}(x) = 1, \tag{1.45}$$

since we start at x, and hence visit it once before returning. Summing these yields

$$\begin{aligned}\sum_{j=1}^n \tilde{\pi}(j) &= \sum_{j=1}^n \mathbb{E}_x\big(\#\{\text{visits of } j \text{ before } \tau_x^+\}\big) \\ &= \mathbb{E}_x\Big(\sum_{j=1}^n \#\{\text{visits of } j \text{ before } \tau_x^+\}\Big) \\ &= \mathbb{E}_x(\tau_x^+), \end{aligned} \tag{1.46}$$

since the total number of vertices visited before returning to x, if we count the starting vertex (which we do), is the time it takes to return to x. Defining the row vector $\tilde{\pi} := (\tilde{\pi}(1), \tilde{\pi}(2), \ldots, \tilde{\pi}(n))$, our goal is to show

$$\tilde{\pi} P = \tilde{\pi}, \tag{1.47}$$

and hence that

$$\frac{\tilde{\pi}}{\mathbb{E}_x(\tau_x^+)}P = \frac{\tilde{\pi}}{\mathbb{E}_x(\tau_x^+)}. \tag{1.48}$$

If we then define $\pi := \tilde{\pi}/\mathbb{E}_x(\tau_x^+)$, we have

$$\sum_{j=1}^{n} \pi(j) = \frac{1}{\mathbb{E}_x(\tau_x^+)} \sum_{j=1}^{n} \tilde{\pi}_j = 1$$

by (1.46), and so by (1.48), π will be a stationary distribution for P. So to complete the proof that a stationary distribution exists, we need to show (1.47).

We can prove this vector identity by showing the individual components of either side are identical. For instance, for the first component of the identity (1.47) says

$$\sum_{j=1}^{n} \tilde{\pi}(j)P_{j1} = \tilde{\pi}(1). \tag{1.49}$$

To prove this, define $M_j := \#\{\text{visits of } j \text{ before } \tau_x^+\}$. So $\mathbb{E}_x(M_j) = \tilde{\pi}(j)$. Then if U_{j1} is the proportion of times you jump from state j to 1 on any run of the Markov chain, we have the identity

$$M_1 = \sum_{j=1}^{n} M_j U_{j1}. \tag{1.50}$$

Note that each $M_j U_{j1}$ is a random variable: we have M_j trials where we are trying to reach vertex 1, each with probability of success P_{j1}. Hence $M_j U_{j1} | M_j = m_j \sim \text{Bin}(m_j, P_{j1})$, which has mean $m_j P_{j1}$. Averaging over all trials of the chain, and using the tower property of conditional expectation yields

$$\mathbb{E}_x(M_j U_{j1}) = \mathbb{E}_x(\mathbb{E}_x(M_j U_{j1}|M_j)) = \mathbb{E}_x(M_j P_{j1}) = \mathbb{E}_x(M_j)P_{j1} = \tilde{\pi}(j)P_{j1}$$

for each j. Taking expectations in (1.50) then yields

$$\begin{aligned}\tilde{\pi}(1) = \mathbb{E}_x(M_1) &= \sum_{j=1}^{n} \mathbb{E}_x(M_j U_{j1}) \\ &= \sum_{j=1}^{n} \tilde{\pi}(j) P_{j1},\end{aligned}$$

which is exactly (1.49). The argument for the other entries $2 \leq j \leq n$ of (1.47) is the same, and so we conclude that (1.47) holds.

As discussed above, this proves that $\pi := \tilde{\pi}/\mathbb{E}_x(\tau_x^+)$ is a stationary distribution for P. Furthermore, by (1.45),

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathbb{E}_x(\tau_x^+)} = \frac{\mathbb{E}_x\big(\#\{\text{visits of } x \text{ before } \tau_x^+\}\big)}{\mathbb{E}_x(\tau_x^+)} = \frac{1}{\mathbb{E}_x(\tau_x^+)},$$

which gives our desired formula (1.41). □

Theorem 1.7 says that the formula (1.41) always works for us as a stationary distribution. Notice, however, that we also have another useful equation: in the proof we defined the stationary distribution via $\pi = \tilde{\pi}/\mathbb{E}_x(\tau_x^+)$, where x was any fixed state. Hence for any other state $y \in \Omega$,

$$\pi(y) = \frac{\tilde{\pi}(y)}{\mathbb{E}_x(\tau_x^+)} = \frac{\mathbb{E}_x\big(\#\{\text{visits of } y \text{ before } \tau_x^+\}\big)}{\mathbb{E}_x(\tau_x^+)}. \tag{1.51}$$

Both formulas (1.41) and (1.51) will prove useful in different contexts.

Example 1.14. To see (1.41) in action, consider the "star graph" of Figure 1.14. Here there is one central vertex, labelled 0, connected to n "leaf" vertices $1, 2, \ldots, n$. If we are at 0, we go to any of the leaves with equal probability. If we are at a leaf, we move to the center with probability p, and stay at the same leaf with probability $1 - p$, for some fixed $0 < p < 1$. What is the stationary distribution for this chain?

Do find π, let's compute the return times and use (1.41). Note that, if we start with $X_0 = 0$, $\tau_0^+ \sim 1 + Y$, where $Y \sim \text{Geo}(p)$, since we immediately step away to a leaf, and then we toss independent coins with probability of

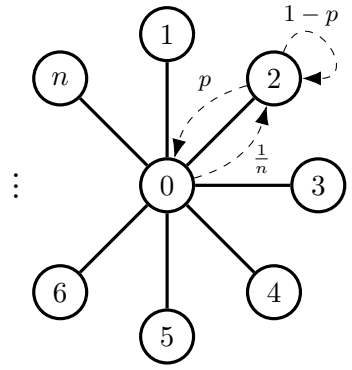

Figure 1.14: What is π for this chain?

success p to see if we return. Hence

$$\mathbb{E}_0(\tau_0^+) = 1 + \frac{1}{p} = \frac{1+p}{p},$$

and so $\pi(0) = \frac{p}{1+p}$. By symmetry, the $\mathbb{E}_j(\tau_j^+)$ are all identical for $j = 1, 2, \ldots, n$, and so $\pi(1) = \pi(2) = \cdots = \pi(n)$. We therefore have

$$1 = \sum_{j=1}^{n} \pi(j) = \pi(0) + n\pi(1) = \frac{p}{p+1} + n\pi(1),$$

and so $\pi(1) = \frac{1}{n(1+p)}$. We conclude

$$\pi = \Big(\frac{p}{p+1}, \frac{1}{n(1+p)}, \frac{1}{n(1+p)}, \ldots, \frac{1}{n(1+p)}\Big).$$

So, for instance, if there are $n = 10$ branches and we move to the center with probability $p = 1/5$. Then our stationary distribution is

$$\pi = \Big(\frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \ldots, \frac{1}{12}\Big). \tag{1.52}$$

Note two things this says:

- If we start the Markov chain X_n with our initial position X_0 randomly selected according to (1.52), then at *every* subsequent step k, the distribution of X_k on the vertices is still exactly π. For example,

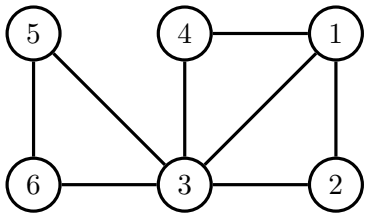

Figure 1.15

$\mathbb{P}(X_{1,000,000} = 5 \,|\, X_0 \sim \pi) = 1/12.$

- Secondly, Theorem 1.6 tells us that if we start at vertex 0, on average it will take us 6 steps to return back to 0. Similarly, if we start at any vertex $1, 2, \ldots, 10$, it will take us 12 steps, on average, to return.

Exercise 1.19. *Find* $\mathbb{E}_1(\tau_1^+)$ *and* $\mathbb{E}_5(\tau_5^+)$ *for the graph in Figure 1.15.*

1.8 Time-reversibility

Consider a biased and an un-biased walk on the 5-cycle, as illustrated in Figure 1.16. Note that both have stationary distribution

$$\pi = \Big(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\Big),$$

as we discussed in Example 1.10. Suppose we had movies of our Markov chains unfolding on each of the pentagons. What would happen if we played the movies in reverse?

For the unbiased walk, we wouldn't see any difference: when reversed, the walk would still jump CW or CCW with equal probability, just as in the forward direction. The biased walk would look different, though. In the forward direction it prefers jumping CW, so when reversed it has a preference for jumping CCW. In this section we study what conditions on a Markov chain will tell when the "movie" looks the same in both directions.

First, some notation: Recall that we defined $\mathbb{P}_x(\cdot)$ and $\mathbb{E}_x(\cdot)$ above in (1.43) and (1.44) as the conditional probability and expectation, given $X_0 = x$ (the dot just means that we can plug in any valid event or function,

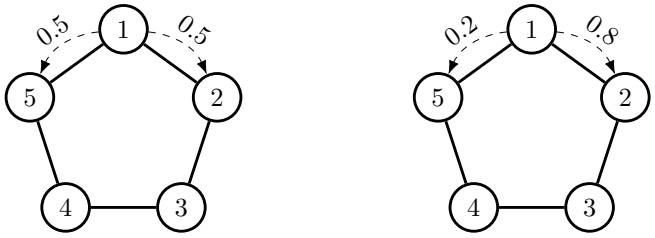

Figure 1.16: An unbiased walk and a biased walk on the 5-cycle.

respectively). We similarly define

$$\mathbb{P}_\mu(\cdot) := \mathbb{P}(\,\cdot \,|\, X_0 \sim \mu), \tag{1.53}$$

where $X_0 \sim \mu$ means that X_0 is randomly distributed on Ω, with pdf given by μ. Similarly,

$$\begin{aligned} \mathbb{E}_\mu(\cdot) :=& \mathbb{E}(\,\cdot \,|\, X_0 \sim \mu) \\ =& \sum_{x \in \Omega} \mathbb{E}_x(\cdot)\mu(x). \end{aligned} \tag{1.54}$$

1.8.1 Bayes' rule and time-reversal

Suppose our Markov chain has stationary distribution π and transition matrix P. If we start with $X_0 \sim \pi$, then we know that $X_k \sim \pi$ for each $k = 1, 2, \ldots$, and

$$\mathbb{P}_\pi(X_{k+1} = y \,|\, X_k = x) = \mathbb{P}(X_{k+1} = y \,|\, X_k = x) = P_{xy} \tag{1.55}$$

for any $k \geq 1$. To reverse time (or reverse "the movie"), we want to instead compute

$$\hat{P}_{yx} := \mathbb{P}_\pi(X_k = x \,|\, X_{k+1} = y), \tag{1.56}$$

the probability of going from y to x under time reversal. Note that this is backwards from what we have normally been computing: now we are conditioning on our position in a *later* time and asking about the probability of our position one step *earlier*. This is exactly what it means to "go forward" in reversed time.

Are (1.56) and (1.55) the same? Applying the definition of conditional probability twice, we find

$$\begin{aligned}\hat{P}_{yx} &= \frac{\mathbb{P}_\pi(X_k = x,\, X_{k+1} = y)}{\mathbb{P}_\pi(X_{k+1} = y)} \\ &= \frac{\mathbb{P}_\pi(X_{k+1} = y \,|\, X_k = x)\mathbb{P}_\pi(X_k = x)}{\mathbb{P}_\pi(X_{k+1} = y)} \\ &= P_{xy} \cdot \frac{\pi(x)}{\pi(y)},\end{aligned} \tag{1.57}$$

where we have actually re-derived Bayes' rule. Based on this computation, we make the following definition.

Definition 1.9. The **time-reversal** of an irreducible Markov chain on Ω with transition matrix P and stationary distribution π is the Markov chain on Ω whose transition matrix $\hat{P}$ has entries

$$\hat{P}_{yx} = P_{xy}\frac{\pi(x)}{\pi(y)} \tag{1.58}$$

for each $x, y \in \Omega$.

We can make two immediate observations:

- First, note that (1.58) *does* define a transition matrix. Clearly each $\hat{P}_{xy} \geq 0$, and for a fixed $y \in \Omega$, the corresponding row sum is

  $$\sum_{x\in\Omega} \hat{P}_{yx} = \frac{1}{\pi(y)} \sum_{x\in\Omega} P_{xy}\pi(x) = \frac{1}{\pi(y)}\pi(y) = 1,$$

  where the second equality follows from the fact that $\pi P = \pi$. So the transition probabilities given by (1.57) define a legitimate transition matrix, and we have a (potentially new) Markov chain, the time reversal of our original.

- Secondly, note that since the original chain is irreducible, so is the reversed chain. Indeed, for any two states x, y, there is a forward path $y \to x$ of positive probability. Reversing this gives a positive-probability path $x \to y$ for the backwards chain.

Example 1.15. Let's revisit our walk on the 5-cycle. Since the stationary distribution is always uniform (whether or not the walk is biased), (1.58) becomes

$$\hat{P}_{yx} = P_{xy}. \tag{1.59}$$

We can now make our introductory remarks regarding Figure 1.16 more rigorous. For the unbiased walk, (1.59) says

$$\hat{P}(j, j+1) = P(j+1, j) = \frac{1}{2} = P(j, j+1), \tag{1.60}$$

so we cannot tell the difference between the original and reversed movies. For the biased walk with $P(j, j+1) = p = 1 - P(j, j-1)$, however, we find

$$\hat{P}(j, j+1) = P(j+1, j) = 1 - p = 1 - P(j, j+1), \tag{1.61}$$

and so the movies are different if $p \neq 1/2$. In particular, for the example of $p = 0.8$, (1.61) says the probability that our time-reversed Markov chain $\hat{X}$ moves CW is 0.2, while the probability of moving CCW is similarly 0.8, exactly the reverse of the original walk. See Figure 1.17. So, as we expected, because the starting walk preferred CW steps, the time-reversed walk prefers to take CCW steps.

Another helpful case for the intuition is $p = 1$. In this totally-biased walk we are guaranteed to move CW each step. Equation (1.58) for the time-reversal then says

$$\hat{P}(j, j-1) = P(j-1, j) = 1,$$

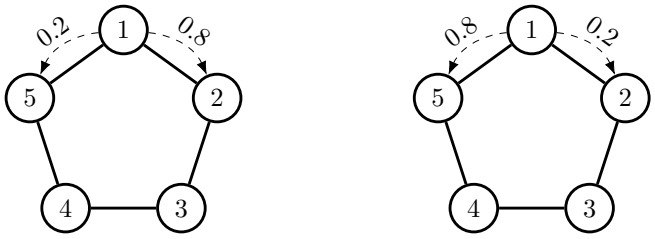

Figure 1.17: A biased walk on the 5-cycle and its time-reversal.

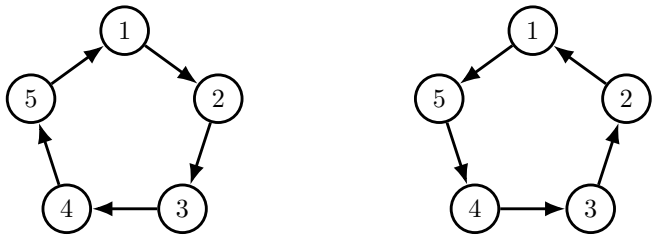

Figure 1.18: Reversing the totally-biased walk ($p = 1$). In the original walk on left we only move CW, and in the time reversal we only move CCW.

and so we move CCW with probability 1. Again, the "movie" is reversed in the obvious way. See Figure 1.18

1.8.2 Stationary distributions and reversibility; the detail balance equations

If our original chain has stationary distribution π, what can we say about the stationary distribution for the reversed chain? Does reversing the movie change π? The next theorem answers in the negative.

Theorem 1.8. *If (X_n) is an irreducible Markov chain with stationary distribution π, then π is also stationary for the reversed chain $(\hat{X}_n)$. Furthermore, writing $\hat{\mathbb{P}}$ for probabilities under the transition matrix $\hat{P}$,*

$$\hat{\mathbb{P}}_\pi(\hat{X}_0 = x_0, \hat{X}_1 = x_1, \ldots, \hat{X}_k = x_k) \\ = \mathbb{P}_\pi(X_0 = x_k, X_1 = x_{k-1}, \ldots, X_k = x_0). \quad (1.62)$$

Note that (1.62) is just the mathematical formulation that we are reversing time: the path $x_0 \to x_1 \to \cdots \to x_k$ in the reversed chain arises from "reversing the movie" of the path $x_k \to x_{k-1} \to \cdots \to x_0$ in the original chain. Also, recall the fact that π is stationary for both chains matches what happened with our biased walk on the 5-cycle: regardless of p, the uniform distribution is stationary, and so it is stationary for both the biased walk and its reversal (which can be viewed as replacing p with $1 - p$).

Proof. Let's first verify that π is still stationary for $\hat{P}$. Of course, this means

$$\pi\hat{P} = \pi, \quad (1.63)$$

which we can show component-wise, as usual. Indeed, fixing an entry k, we need

$$\sum_{j=1}^{n} \pi(j)\hat{P}_{jk} = \pi(k). \tag{1.64}$$

Recalling the definition (1.58), we see

$$\begin{aligned}\sum_{j=1}^{n} \pi(j)\hat{P}_{jk} &= \sum_{j=1}^{n} \pi(j)P_{kj}\frac{\pi(k)}{\pi(j)} \\ &= \pi(k)\sum_{j=1}^{n} P_{kj} = \pi(k)\cdot 1,\end{aligned}$$

since we are summing the entire kth row of P. Hence (1.64) holds as does (1.63); π is indeed stationary for $\hat{P}$.

The second part (1.62) then follows. We will only show

$$\hat{\mathbb{P}}_\pi(\hat{X}_0 = x_0, \hat{X}_1 = x_1) = \mathbb{P}_\pi(X_0 = x_1, X_1 = x_0),$$

as the general argument is similar and follows from induction. Indeed, we have

$$\begin{aligned}\hat{\mathbb{P}}_\pi(\hat{X}_0 = x_0, \hat{X}_1 = x_1) &= \hat{\mathbb{P}}_\pi(\hat{X}_1 = x_1 \,|\, \hat{X}_0 = x_0)\hat{\mathbb{P}}_\pi(\hat{X}_0 = x_0) \\ &= \hat{P}_{x_0,x_1}\pi(x_0) \\ &= P_{x_1,x_0}\frac{\pi(x_1)}{\pi(x_0)}\pi(x_0) \\ &= P_{x_1,x_0}\pi(x_1) \\ &= \mathbb{P}_\pi(X_1 = x_0 \,|\, X_0 = x_1)\mathbb{P}_\pi(X_0 = x_1) \\ &= \mathbb{P}_\pi(X_0 = x_1, X_1 = x_0). \qquad \square\end{aligned}$$

So we know how to time-reverse a Markov chain: we simply write down the reversed transition matrix $\hat{P}$ from the original P via (1.58). When is it that $\hat{P} = P$? That would mean that we cannot distinguish which way the movie is being played.

Definition 1.10. A Markov chain is **time-reversible** (or simply **reversible**)

with respect to its stationary distribution π if

$$\hat{P} = P. \tag{1.65}$$

Note that π is implicit in (1.65), since whenever we write $\hat{P}$ we are using (1.58) and hence the stationary distribution π. So it's important to keep in mind that reversibility is *always with respect to a given stationary distribution* π. This also should not be surprising, however: to compute the time-reversed probabilities (1.56), we had to know the *unconditional* probability of being at a vertex in (1.57). We only know that for all times k, generally speaking, if we have the stationary distribution π.

Example 1.16. For reversible chains, reversibility will *not hold* in general if we do not start with $X_0 \sim \pi$. For instance, we saw above in Example 1.15 that the simple walk on the 5-cycle is reversible with respect to its stationary distribution (which is uniform). See equation (1.60). What if we instead start with $X_0 = 1$, i.e. with the distribution $\mu = (1, 0, 0, 0, 0)$? Then, for instance, we find

$$\begin{aligned}\hat{P}(2,1) &= \mathbb{P}_\mu(X_0 = 1 \,|\, X_1 = 2) \\ &= \mathbb{P}_\mu(X_1 = 2 \,|\, X_0 = 1)\frac{\mathbb{P}_\mu(X_0 = 1)}{\mathbb{P}_\mu(X_1 = 2)} = \frac{1}{2} \cdot \frac{1}{\frac{1}{2}} = 1.\end{aligned}$$

However, $P(2,1) = \frac{1}{2}$, and so $\hat{P} \neq P$.

So how can we tell if the chain is reversible? Suppose that we do have $P = \hat{P}$, and let's consider an individual entry $P(x, y)$. From our assumption and (1.58),

$$P(x,y) = \hat{P}(x,y) = P(y,x)\frac{\pi(y)}{\pi(x)},$$

and so, multiplying the $\pi(x)$ over, we see $P = \hat{P}$ iff

$$\boxed{\pi(x)P(x,y) = \pi(y)P(y,x)} \tag{1.66}$$

for every $x, y \in \Omega$. We will call (1.66) the **detail balance equations** or **DBE's** for short, and this will prove to be a very important equation. So important, in fact, that you should memorize (1.66) to have it at your

fingertips.

We have proven the first part of the following theorem.

Theorem 1.9. *Suppose P is a transition matrix for Ω and π is a probability distribution on Ω such that the detail balance equations* (1.66) *hold for all $x, y \in \Omega$. Then*

(*i*) *P is reversible with respect to π, $P = \hat{P}$, and*

(*ii*) *π is a stationary distribution for P.*

Proof. As we mentioned, the discussion preceding the theorem statement proves part (*i*). So we only need to show (*ii*). Indeed, sum (1.66) over all $x \in \Omega$ to obtain

$$\sum_{x \in \Omega} \pi(x) P(x, y) = \sum_{x \in \Omega} \pi(y) P(y, x). \tag{1.67}$$

Note that the left-hand side is the y-component of the vector-matrix product πP. The right-hand side is

$$\pi(y) \sum_{x \in \Omega} P(y, x) = \pi(y)$$

since we are summing over a fixed row in P, and so (1.67) says

$$(\pi P)_y = \pi(y)$$

for any y. Hence $\pi P = \pi$, as claimed. □

So, note the power of the DBE's: if we verify that (1.66) holds, not only do we know the chain is reversible with respect to π, but we also automatically get that π is the stationary distribution.

So far the only reversible chains we've seen are the simple walks on n-cycles. But this is part of a much larger class of examples.

Theorem 1.10. *Simple random walks on graphs are reversible with respect to their stationary distribution* (1.31).

Proof. We simply need to verify the DBE's (1.66). Indeed, pick any vertices x and y. If $y \not\sim x$, then both $P(x, y)$ and $P(y, x)$ are zero, and the DBE

holds. If $y \sim x$, then

$$\pi(x)P(x,y) = \frac{\deg(x)}{2|E|} \cdot \frac{1}{\deg(x)} = \frac{1}{2|E|},$$

while

$$\pi(y)P(y,x) = \frac{\deg(y)}{2|E|} \cdot \frac{1}{\deg(y)} = \frac{1}{2|E|}$$

as well. □

Another class of reversible Markov chains are those with *symmetric* transition matrices.

Definition 1.11. A transition matrix is **symmetric** if $P(x,y) = P(y,x)$ for all $x, y \in \Omega$. In terms of linear algebra, this says that $P = P^T$, the transpose of the matrix P.

Theorem 1.11. *If a Markov chain on $\Omega = \{1, 2, \ldots, n\}$ has a symmetric transition matrix P, then the uniform distribution $\pi = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ is stationary and the chain is reversible with respect to π.*

Exercise 1.20. *Prove Theorem 1.11 in two different ways: verify $\pi P = \pi$, and also show that the DBE's* (1.66) *hold.*

1.8.3 Two in-depth examples

Time reversing a Markov chain can be conceptually intimidating. In this section, we work through two in-depth examples to help bolster our understanding of how reversal works in practice.

Example 1: A "geometric" walk on $\mathbb{N}$

Suppose we have the infinite state space $\Omega = \mathbb{N} = \{1, 2, 3, \ldots\}$, and $0 < p < 1$ is fixed. Consider the Markov chain with transition probabilities

$$P(j, j+1) = 1 - p \qquad \text{and} \qquad P(j, 1) = p \tag{1.68}$$

for all $j \in \Omega$. See Figure 1.19. At each step, this Markov chain moves one step right with probability $1-p$, and jumps back to 1 with probability p. So

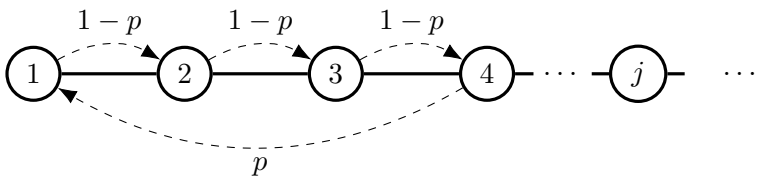

Figure 1.19: Is this Markov chain reversible?

we can think of it as working in cycles where each cycle consists of walking up the integers via repeated "failures," until it "succeeds" and resets at value 1. Starting from 1, each of these cycles has random length L, where $L \sim$Geo(p). Hence this is a "geometric-type" Markov chain on $\mathbb{N}$.

Before we explicitly compute the time reversal, it is a good idea to first try to "reverse the movie" in our minds. This will give some intuition to guide the subsequent computations.

Exercise 1.21. *Without performing any calculations, informally describe the behavior of the time-reversal of this Markov chain.*

We would like to determine the time-reversal of this chain, but of course the time-reversal is with respect to the stationary distribution π. So we first need to find π, which we will do via computing the expected return times $\mathbb{E}_x(\tau_x^+)$. We note that

$$\mathbb{P}_1(\tau_1^+ = k) = \mathbb{P}(1 \to 2 \to \cdots \to k \to 1) = (1-p)^{k-1}p,$$

and so $\tau_1^+ | X_0 = 1$ is a Geo(p) random variable. Hence $\mathbb{E}_1(\tau_1^+) = \frac{1}{p}$, and so $\pi(1) = p$ by (1.41). To compute $\pi(j)$ for $j \neq 1$, we use our other formula (1.51) with $x = 1$,

$$\pi(j) = \frac{\mathbb{E}_1\big(\#\{\text{visits of } j \text{ before } \tau_1^+\}\big)}{\mathbb{E}_1(\tau_1^+)} = p\mathbb{E}_1\big(\#\{\text{visits of } j \text{ before } \tau_1^+\}\big).$$

Noting that the chain will visit j at most once before returning to 1, we find

$$\begin{aligned}\mathbb{E}_1\big(\#\{\text{visits of } j \text{ before } \tau_1^+\}\big) &= 0 \cdot \mathbb{P}_1(\tau_1^+ < j) + 1 \cdot \mathbb{P}_1(\tau_1^+ \geq j) \\ &= \mathbb{P}_1(\tau_1^+ \geq j) \\ &= (1-p)^{j-1},\end{aligned}$$

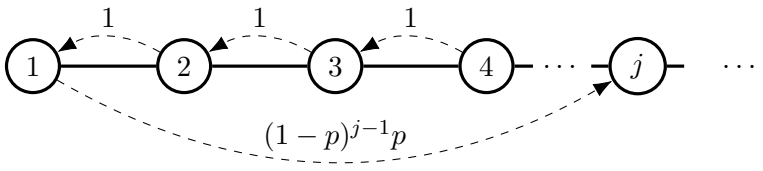

Figure 1.20: The time reversal of the geometric walk.

and thus $\pi(j) = p(1-p)^{j-1}$. As this formula also holds for $j = 1$, we see π has the pdf of a Geo(p) random variable.

We can then compute the time-reversal. Recalling that

$$\hat{P}(i,j) = P(j,i)\frac{\pi(j)}{\pi(i)},$$

we have from (1.68) that $\hat{P}(i,j) \neq 0$ iff $i = j+1$ or $i = 1$. In the former case,

$$\begin{aligned}\hat{P}(j+1,j) &= P(j,j+1)\frac{\pi(j)}{\pi(j+1)} \\ &= (1-p)\frac{(1-p)^{j-1}p}{(1-p)^j p} = 1. \end{aligned} \tag{1.69}$$

In the latter case,

$$\begin{aligned}\hat{P}(1,j) &= P(j,1)\frac{\pi(j)}{\pi(1)} \\ &= p\frac{(1-p)^{j-1}p}{p} = (1-p)^{j-1}p. \end{aligned} \tag{1.70}$$

Hence we obtain the following picture for the reversed chain: if we are at a vertex $j > 1$, (1.69) says that we decrease by 1 each step until we reach state 1, at which point (1.70) says that we jump to a new integer j, where $j \sim$ Geo(p). See Figure 1.20. (Is this what you had for Exercise 1.21?) The reversed "movie" is thus quite different than the original, and this Markov chain is *not* reversible. Don't forget from Theorem 1.8, though, that the stationary distribution for the reversed chain is still the same π as for the forward chain, $\pi \sim$ Geo(p).

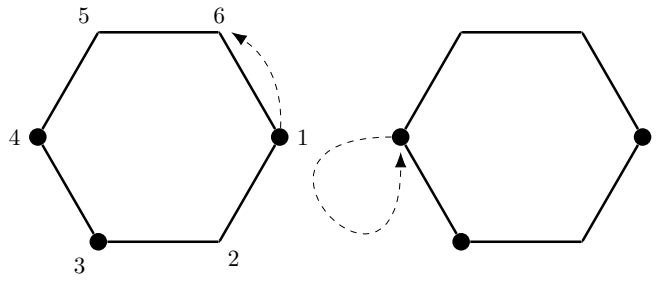

Figure 1.21: An asymmetric exclusion MC. A randomly selected particle tries to take one anti-clockwise step. If the vertex is unoccupied, it moves there (left), while if occupied, it remains in the same position (right).

Example 2: An asymmetric exclusion on an n-cycle

Consider $1 \leq k < n$ particles placed on the vertices of an n-cycle, where no two particles occupy the same vertex. Order the vertices and assign each a 0 or 1 based on whether it is empty or has a particle, respectively. The sample space Ω is then the collection of all elements of $\{0,1\}^n$ that have exactly k ones, and an element $\omega \in \Omega$ is a vector of zeroes and ones. For example, the starting position of two 6-cycles in Figure 1.21 is $\omega = (1,0,1,1,0,0)$. Since we are choosing k entries of an n-tuple to have ones, $\#\Omega = \binom{n}{k}$.

The chain proceeds by choosing a particle uniformly at random at each time and trying to move it one step CCW. It can only move if that position is unoccupied, and otherwise the move is suppressed and the particles remain in the same state. The 6-cycle on the left in Figure 1.21 shows a move from $(1,0,1,1,0,0)$ to $(0,0,1,1,0,1)$, while the chain on the right stays at $(1,0,1,1,0,0)$.

In terms of the vector ω, each step of the chain chooses one of the 1's uniformly at random. If the entry immediately to the left of this 1 is a 0, the entries switch. If it is another 1, nothing happens.

This Markov chain is an *asymmetric exclusion process* on the n-cycle.

Exercise 1.22. *Without performing any calculations, informally describe the behavior of the time-reversal of this Markov chain.*

To formally determine the time-reversal, we first need to find the sta-

tionary distribution π. We claim that this is uniform,

$$\pi(\omega) = \frac{1}{\binom{n}{k}} \tag{1.71}$$

for each $\omega \in \Omega$. Indeed, by Exercise 1.13 (a), it suffices to show that the column sums satisfy

$$\sum_{\omega' \sim \omega} P(\omega', \omega) = 1 \tag{1.72}$$

for any fixed $\omega \in \Omega$. So we need to know what states ω' lead to ω, and what each transition probability $P(\omega', \omega)$ is. For example, in Figure 1.21, the states that could lead to the given initial configuration $(1, 0, 1, 1, 0, 0)$ are

$$(0, 1, 1, 1, 0, 0), \ (1, 0, 1, 0, 1, 0) \ \text{ and } \ (1, 0, 1, 1, 0, 0),$$

corresponding to a particle moving from 2 to 1, a particle moving from 5 to 4, and particle 4 being chosen and not moving, as on in the right in the figure. Note that this is one ω' for each of the three particles. Starting from any one of these ω', we have $P(\omega', \omega) = 1/3$, since we have to pick exactly one of the particles in each case, and they are all equally likely. Hence (1.72) holds for the column corresponding to $\omega = (1, 0, 1, 1, 0, 0)$.

This logic readily extends. For a general ω, each of the k ones corresponds to exactly one ω' such that $\omega' \sim \omega$: if the entry immediately to the right of the 1 is a 0, then swapping the 1 and 0 gives ω' (the shifted particle can move one step to yield ω). If the entry is another 1, then $\omega' = \omega$ (the adjacent particle is chosen and cannot move). In either case, exactly one of the k particles must be chosen, and since each has probability $1/k$, we again have (1.72). We conclude the stationary distribution is uniform (1.71), as claimed.

For our reversal under π, we thus have

$$\hat{P}(\omega, \omega') = P(\omega', \omega)\frac{\pi(\omega')}{\pi(\omega)} = P(\omega', \omega). \tag{1.73}$$

What does this mean? Consider two cases. If $\omega' = \omega$, (1.73) says that the probability of staying at state ω in the reversed chain is the same as in

the forward chain. If $\omega' \neq \omega$, then $P(\omega', \omega) \neq 0$ iff one particle in the ω' configuration can move one step CCW to obtain ω. When we switch the order in $\hat{P}(\omega, \omega')$, the particle is then moving clockwise. Thus $\hat{P}(\omega, \omega') \neq 0$ iff one particle of ω can move one step *clockwise* to obtain the state ω'.

So time reversal in our exclusion process just switches the orientation of particle movement: we still uniformly choose a particle at each step, but now try to move clockwise instead of anti-clockwise, staying put if there is no open vertex. This is obviously different than the original chain, and so we see the asymmetric exclusion is *not* time-reversible.

Exercise 1.23. *Use the DBE's* (1.66) *to give an alternate proof that this chain is not time-reversible.*

Problems for chapter 1

Problem 1.1. Consider the transition matrix

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

(*a*) Draw a graphical representation for how this chain moves (see Figures 1.7 and 1.11 for inspiration if needed).

(*b*) What do the rows of P sum to? What do the columns of P sum to?

(*c*) Find the characteristic polynomial of P.

(*d*) Find the eigenvalues of P and a basis for each eigenspace. What do you notice about the eigenvalues of P?

(*e*) Is P diagonalizable? If so, diagonalize it as QDQ^{-1} (that is, find all three of these matrices and verify that their product QDQ^{-1} gives P). If not, explain why not.

Problem 1.2. Consider the transition matrix

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

(*a*) Draw a graphical representation for how this chain moves (see Figures 1.7 and 1.11 for inspiration if needed).

(*b*) What do the rows of P sum to? What do the columns of P sum to?

(*c*) Find all the (right) eigenvalues for P. (Even though we care primarily about left-eigenvalues and vectors, here find the usual *right* ones from linear algebra.) What do you observe about the eigenvalues?

(*d*) Show that the right-eigenvalues and left-eigenvalues for any square matrix are always the same. (*Hint*: think properties of the determinant.)

(*e*) Show that for our P, the right eigenspace for $\lambda = 1$ and the left eigenspace for $\lambda = 1$ are *not* the same.

(f) Find a stationary distribution π for P.

(g) Compute P^{50}. What do you observe?

Problem 1.3. Consider the following simple model for a Seattle weather forecast. The weather can be either 'rain' or 'no rain'. Suppose that the chance of rain tomorrow only depends on whether it rains today or not, and not on past weather conditions. If it rains today, the chance it will be rain tomorrow again is $\alpha \in (0,1)$; and if it does not rain today, the probability it will rain tomorrow is $\beta \in (0,1)$. Starting from some day zero, let $X_0, X_1, X_2, \ldots$ be a sequence where

$$X_k = \begin{cases} 1, & \text{if it rains on day } k \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the state space and the transition probability matrix of the Markov chain $(X_k,\ k = 0, 1, 2, \ldots)$.

(b) If it rains on day zero, find the probability it will rain on day 3.

(c) If there is only 50% chance that it rains on day zero, find the probability it will rain on day 3.

(d) If it rains on day zero, what is the probability that it is going to rain nonstop from days 1 through 7?

Problem 1.4. Consider the so-called random walk with *absorbing* barriers on $\Omega = \{1, 2, 3, 4, 5\}$. This is also a Markov chain on $\Omega = \{1, 2, 3, 4, 5\}$ with the following transition probability matrix

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

(a) Draw a graphical diagram that shows how this chain moves.

(b) Is this chain irreducible? Why or why not?

(c) Find $\mathbb{E}(X_5 \mid X_3 = 2, X_0 = 3)$.

(*d*) Find a row vector π such that π has every coordinate nonnegative and the sum of coordinates is one and $\pi P = \pi$. Is π unique?

(*e*) Use a computer to compute P^{50}, and round to the nearest hundredth. What do you observe? Explain what this intuitively means.

Problem 1.5. Consider the so-called random walk with *reflecting boundary* on $\Omega = \{1, 2, 3, 4, 5\}$. This is a Markov chain with the following transition probability matrix

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

(*a*) Draw a graphical diagram that shows how this chain moves.

(*b*) Is this chain irreducible? Why or why not?

(*c*) If this chain is irreducible, what is its period?

(*d*) Find $\mathbb{E}(X_2 \mid X_0 = 1)$.

(*e*) Find a stationary distribution for this Markov chain.

(*f*) Estimate $\mathbb{P}(X_{2200} = 2 \,|\, X_{1200} = 4)$ and $\mathbb{P}(X_{2201} = 2 \,|\, X_{1200} = 4)$.

Problem 1.6. Consider the random walk on the graph in Figure 1.22.

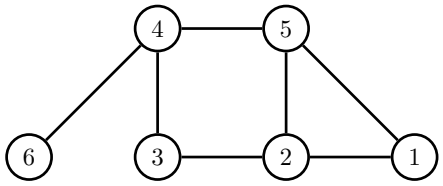

Figure 1.22

(*a*) Write down the transition matrix P of this Markov chain.

(*b*) Is the random walk irreducible? Why or why not?

(*c*) What is the period of this chain? Is it aperiodic?

(*d*) Suppose $X_0 = 6$. What is

$$\mathbb{P}(X_4 = 1, X_3 \neq 1, X_2 \neq 1, X_1 \neq 1 \mid X_0 = 6)?$$

That is, what is the probability that, starting from 6, you reach 1 for the first time in 4 time steps?

(*e*) Use a computer to compute P^{50}. What do you observe?

Problem 1.7. Consider the 3×3 matrix

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

(*a*) Draw a diagram for the Markov chain whose transition probabilities are given by A.

(*b*) Find the characteristic equation for this matrix.

(*c*) Find all the eigenvalues of this matrix. What do you notice about their values? (Even though we care primarily about left-eigenvalues and vectors, here find the *right*-eigenvalues, as usual in linear algebra. Can you show that the right- and left-eigenvalues are always the same?)

(*d*) Find a solution to the linear system of equations $\pi A = \pi$ where $\pi = (\pi_1, \pi_2, \pi_3)$ is a vector in $\mathbb{R}^3$ satisfying the constraint $\pi_1 + \pi_2 + \pi_3 = 1$. This is the stationary distribution for the Markov chain with transition matrix A.

(*e*) Use a computer to compute A^{50}. What do you observe?

Problem 1.8. Modify the random walk from Problem 1.5 by considering the so-called *lazy* random walk with reflecting barriers on $\Omega = \{1, 2, 3, 4, 5\}$. This is a Markov chain with the same state space and a new transition matrix

$$\widetilde{P} = \frac{1}{2}(I + P),$$

where I represents the 5×5 identity matrix. Answer parts (*a*) through (*f*) of Problem 1.5 for this new Markov chain.

Problem 1.9. Consider a Markov chain with state space $\Omega = \{1, 2, 3, 4, 5\}$ and the following transition probability matrix

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

That is, at every state $\{1, 2, 3, 4\}$, the chain either goes up by one with probability $2/3$ or returns to 1 with probability $1/3$. Once it reaches 5, it jumps to 1, and continues again.

(a) Is this chain irreducible? Why or why not?

(b) If this chain is irreducible, what is its period?

(c) Find a stationary distribution π for this chain.

(d) If you start from state 1, find the exact distribution of the first return to 1.

(e) Compute $\mathbb{E}(\tau_1^+ \,|\, X_0 = 1)$.

(f) Use a computer to compute P^{50}. What do you observe? Explain what this intuitively means.

Problem 1.10. Consider a Markov chain with state space $\Omega = \{0, 1, 2, 3, 4\}$ and the following transition probability matrix

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

(a) Is this chain irreducible? Why or why not?

(b) If this chain is irreducible, what is its period?

(c) Find a stationary distribution π for this chain.

(*d*) If you start from state 0, find the expected hitting time to the states $\{1, 2, 4\}$.

(*e*) If you start from state 0, find the expected number of times you will visit 1 before you hit the states $\{2, 4\}$. Leave the answer as an indicated coordinate of a matrix inverse. That is, say something like "the answer is the $(1, 2)$ coordinate of M^{-1}" where M is written explicitly. You don't need to invert M.

(*f*) Find the expected return time to zero $\mathbb{E}_0\left(\tau_0^+\right)$.

(*g*) Use a computer to compute P^{50}. What do you observe? Explain what this intuitively means.

Problem 1.11. A community of N individuals is in the middle of a zombie outbreak. Fortunately, the community has managed to discover a serum that can cure zombies and prevents them from further infection. Suppose at the beginning of each day, every individual is in one of three possible conditions: non-infected, zombie, and cured. If during day t, a non-infected person becomes infected, he or she will become a zombie the next day $(t+1)$, but will get cured from the following day $(t + 2)$ onwards. Let X_t and Y_t denote the number of zombies and the number of non-infected persons on day t, respectively. During each day, the probability that a given non-infected person comes in contact with a given zombie is $0 < p < 1$, independently for every zombie and every other non-infected person.

(*a*) If $X_t = i$, what is the probability that a given non-infected person will come in contact with a zombie during day t?

(*b*) Is the pair (X_t, Y_t), $t = 0, 1, 2, \ldots$, a Markov chain? If so, give an expression for its state space and transition probabilities.

(*c*) Suppose $X_0 = 1, Y_0 = N - 1$, find the distribution of X_2. Leave your answer in summation notation.

(*Remark:* This is the *Reed-Frost model*, a simple epidemiological model for the spread of disease. It is part of a larger class of *SIR* models, which track the number of *susceptible*, *infected*, and *recovered* individuals. For an interactive app showing disease spread in this model, made by Jamie Forschmiedt, visit `https://forscj.shinyapps.io/ReedFrostModel/`.)

Problem 1.12. A math professor possesses r umbrellas that he uses in going between his home and his office. If he is at his home at the beginning of the day and it is raining, then he will take an umbrella with him to his office, provided there is one at home to be taken. On his way back from his office, he will bring back an umbrella if it is raining and there is (at least) one umbrella at his office. If it is not raining, the professor does not use an umbrella. Assume that it rains at the beginning (or at the end) of each day with probability 1/2, independently of the past. Let X_n be the number of umbrellas at home at the beginning of the day $n = 1, 2, \ldots$.

(*a*) Is X_n a Markov chain? If so, find its state space and transition probabilities.

(*b*) Is this chain irreducible? Aperiodic ?

(*c*) Find a stationary distribution for this Markov chain for $r = 3$.

(*d*) Suppose $r = 3$. If the professor finds one day that there are no umbrellas left at home, what is the expected number of days after which he will find himself in a similar situation?

Problem 1.13. Consider the biased random walk X on the 6-cycle where at each step the random walker turns one step CCW with probability $\frac{1}{4}$ and turns onfe step CW with probability $\frac{3}{4}$.

(*a*) Find the stationary distribution π for this Markov chain.

(*b*) Describe the transition probabilities of the time-reversal of X under π.

(*c*) Suppose $X_0 \sim \pi$. What is the probability $\mathbb{P}_\pi(X_{n-2} = 1 \mid X_n = 1)$?

Problem 1.14. Consider the complete bipartite graph. This is a graph with two kinds of vertices n red vertices, $R_1, \ldots, R_n$ and m blue vertices $B_1, \ldots, B_m$. There is an edge for every pair of red and blue vertices $\{R_i, B_j\}$, $1 \le i \le n$ and $1 \le j \le m$, and so a total of mn edges. Consider the simple random walk on this graph.

(*a*) What is the stationary distribution of this chain?

(*b*) Let τ_1 denote the hitting time of vertex R_1. Starting from vertex R_2, what is the probability mass function and expectation of τ_1?

(*c*) Let τ_1^+ denote the return time to vertex R_1. Starting from vertex R_1, what is the probability mass function and expectation of τ_1^+?

Problem 1.15. Let K_n denote the complete graph on n vertices. That is, the vertex set is $V = \{1, 2, \ldots, n\}$ and any pair of possible edges $\{i, j\}$ for $i \neq j$ is present (in particular, note there are no loops). Consider the random walk on this graph.

(*a*) What is the stationary distribution of this Markov chain?

(*b*) Let τ_1 denote the hitting time of state 1. What is the the distribution of τ_1 if we start the chain at $k \neq 1$? What is $\mathbb{E}_k(\tau_1)$?

(*c*) Let τ denote the hitting time of the set $\{1, 2\}$. Find $P_k(X_\tau = 1)$ and $\mathbb{E}_k(\tau)$ for $k = 3, 4, \ldots, n$.

(*d*) The **cover time** of a Markov chain is the time it takes the chain to visit every state. Find the expected cover time for this random walk, starting from any state.

Problem 1.16. Consider the following Markov chain on all permutations of the numbers $\{1, 2, 3, 4\}$. There are $4! = 24$ such permutations. For any given permutation, randomly choose two positions and switch the two numbers appearing in those positions. As an example, suppose the current state of the Markov chain is 1243. Choose randomly a pair, say 1 and 3. After switching those two we obtain 3241. Say, now we randomly pick 2 and 1. After switching, we obtain 3142. And so on.

(a) Assume that this Markov chain is irreducible. Show that it is reversible with respect to the uniform distribution on the 24 permutations.

(b) Call rank of 4 to be the position from the left where 4 appears in the permutation. That is, if the permutation is 4321, the rank of 4 is 1. If the permutation is 1432, the rank of 4 is 2, and so on. Show that the rank of 4 in the Markov chain of permutations in part (a) is itself a Markov chain with state space $\{1, 2, 3, 4\}$. Find its transition probabilities and stationary distribution.

(c) Suppose we start from the permutation 1243. Let τ be the stopping time when 4 is either the leftmost number or the rightmost number in

the permutation (say 4321 or 4123 or 1234 etc.). That is, τ is the first time the rank of 4 is either 1 or 4. What is $\mathbb{E}_{1243}(\tau)$? Hint: use linear algebra.

(d) Starting at 1243 again, what is the probability that at τ, 4 is in the rightmost position (i.e., the rank of 4 is 4)? Hint: use recursion.

Problem 1.17. Consider the random walk on the 5-cycle. That is $\Omega = \{1, 2, 3, 4, 5\}$ and the transition probability matrix is

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

(*a*) Let τ_1 denote the hitting time of the state 1. Let $f_k = \mathbb{E}_k(\tau_1)$. Note that $f_1 = 0$. Write down a recursive equation for $f_k, k = 1, 2, \ldots, 5$.

(*b*) Solve the above recursion to get a formula for f_k.

(*c*) Now let τ denote the hitting time of $\{3, 4\}$. Find $\mathbb{P}_1(X_\tau = 3)$ and $\mathbb{E}_1(\tau)$ by using known formulas about simple symmetric random walks on the integers.

Problem 1.18. Consider a Markov chain with state space $\Omega = \{1, 2, 3, 4\}$ and the following transition matrix

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

(*a*) Is this chain irreducible? Why or why not?

(*b*) Is this chain aperiodic? Why or why not?

(*c*) If the chain is irreducible find a stationary distribution for it.

(*d*) Find $\mathbb{E}_4(\tau_1)$, where τ_1 is the first hitting time of 1.

(*e*) Find the expected number of times the Markov chain will visit 3 before τ_{12}, starting from 4. That is, compute

$$\mathbb{E}_4\left[\sum_{i=0}^{\tau_{12}-1} 1\{X_i = 3\}\right].$$

(*f*) Find the time reversal transition matrix $\hat{P}$ for this Markov chain.

(*g*) For the time-reversed Markov chain, find $\hat{\mathbb{E}}_1(\tau_4)$, where τ_4 is the first hitting time of 4.

Problem 1.19. Consider Ω to be the set of all sequences of $+1$ or -1 of length $2N$ such that the sum of the coordinates is exactly zero. In other words, there are exactly N many $+1$s and N many -1s in each such sequence. If $N = 2$, one such sequence is $u = (+1, +1, -1, -1)$, while another one is $v = (+1, -1, +1, -1)$.

(*a*) How many elements does Ω have?

(*b*) Consider a Markov chain $\{S_0, S_1, S_2, \ldots\}$ with state space Ω that proceeds as follows. At each step, pick a $+1$ at random, and independently pick a -1 at random and swap their positions. For example, u can change to v by switching the second and third coordinates. Is this Markov chain irreducible?

(*c*) What is the stationary distribution of this Markov chain?

(*d*) Let X_n denote the number of $+1$s in the first N coordinates of S_n. For example, this number for u is 2 and for v is 1. Then $\{X_0, X_1, \ldots\}$ is also a Markov chain. Find its state space and transition probabilities.

(*e*) Find the stationary distribution of this chain.

(*f*) Are both X and S time reversible? Why or why not?

Problem 1.20. (*Random walk bridge*) Consider an urn with N-many $+1$'s and N-many -1's. Randomly sample, without replacement, each of these $2N$ numbers one by one. This gives us a random sequence $w = (w_1, w_2, \ldots, w_{2N})$ of exactly N-many $+1$'s and N-many -1's.

Consider a stochastic process by declaring $S_0 = 0$ and then successively adding coordinates of w, i.e.,

$$S_k = w_1 + w_2 + \ldots + w_k, \quad k = 1, 2, \cdots, 2N.$$

This process is called the **random walk bridge** because of the property in the first part below.

(a) Show that $S_{2N} = 0$.

(b) Find the probability mass function and expectation of each S_k.

(c) Show that the bridge has the following time-reversal symmetry:

$$(S_{2N}, S_{2N-1}, \ldots, S_1, S_0)$$

is again distributed as a random walk bridge. *Hint:* Time-reverse the random sequence w.

Problem 1.21. Consider the so-called *top to random* shuffling. Suppose you have a pack of 10 cards labeled $\{0, 1, 2, \ldots, 9\}$ arranged in a permutation from left to right. There are 10! such arrangements. For example, arrangement 5038761249 means the leftmost card is 5, next comes card 0, and so on, the rightmost two cards are 4 and 9. Consider the following Markov chain where at each step you pick the leftmost card, choose a number I uniformly between 1 and 10, and insert the card at position I from the left. For the example arrangement, you'd pick 5 and pick a uniformly distributed number between 1 and 10, say 3. Then insert 5 as the third card from the left and get the new arrangement 0358761249. The rank of card numbered 0 is the position from the left where it appears. In the example shown, the rank of 0 is initially 2 (it appears as the second card from the left), and then becomes 1 after card 5 jumps.

(a) Show that the rank of card 0 on $\Omega = \{1, 2, \ldots, 10\}$ is a Markov chain by describing its transition probabilities.

(b) Find the stationary distribution for the Markov chain in (a).

(c) Describe the time-reversal of this Markov chain in (a) by writing down the explicit transition probabilities.

Problem 1.22. Consider the following Markov chain on $\Omega = \{1, 2, \ldots, n\}$. If you are currently at i, for $1 \leq i \leq n-1$, then you sample, uniformly at random, one of the numbers from $\{i+1, \ldots, n\}$ and jump there. If you are at n, you move down to 1 with probability one.

(*a*) Find $\mathbb{E}_1(\tau_n)$, where τ_n is the hitting time of state n.

(*b*) Find the stationary distribution of this Markov chain.

(*c*) Starting from $X_0 = 1$, what is the probability that you will visit i before hitting n for the first time? That is, what is the probability that the number of visits of i before τ_n is positive? Express as a formula in $2 \leq i \leq n-1$.

Problem 1.23. Suppose the weather on any day depends on the weather conditions of the previous two days. More precisely, suppose the weather can only be sunny (S) or cloudy (C). If it was sunny today and yesterday, it will be sunny tomorrow with probability 0.8. If it is sunny today and cloudy yesterday, it will be sunny tomorrow with probability 0.6. If it is cloudy today but sunny yesterday, it will be sunny tomorrow with probability 0.4, and if it cloudy for the last two days, it will be sunny tomorrow with probability 0.1.

Such a model can be turned to a Markov chain whose current state is the weather both today and yesterday. Consider $\Omega = \{(S,C), (S,S), (C,S), (C,C)\}$, where (S,C) means sunny yesterday but cloudy today.

(*a*) Find the transition probability matrix of this Markov chain?

(*b*) What is the stationary distribution of this Markov chain?

Problem 1.24. Let A and B be two walkers independently doing the simple random walk on the 5-cycle $\Omega = \{1, 2, 3, 4, 5\}$. Consider the graph distance between A and B, i.e., the length of the shortest path joining the two. For example, if A is at 1 and B is at 5, the graph distance is 1 and not 4. In particular, the graph distance can only take values in $\{0, 1, 2\}$.

(*a*) Let X_t denote the graph distance between the two walkers at time t. Show that X_t is a Markov chain by describing its transition probabilities.

(*b*) Suppose walker A starts at 2 while walker B starts at 5. How many steps will it take for them to be at the same vertex, on average?

(*c*) If A and B both start at 2, how many times, on average, will they be one unit apart before they reach two units apart?

Problem 1.25. A sociologist is studying how socio-economic class changes from generation to generation in a certain country, and assumes that the class of the children depends only upon the class of the parents. The sociologist uses a simple model of three classes {lower, middle, higher} and comes up with the transition matrix

$$P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}$$

For example, entry $P_{13} = 0.1$ means that the probability the children from a lower-class family enter the upper class is 0.1 (for simplicity we assume all the children are in the same class). Round all numerical results to nearest thousandth, and feel free to use a computer.

(*a*) Is this chain irreducible? Aperiodic?

(*b*) Find a stationary distribution π.

(*c*) Use a computer to compute P^{50}. What do you observe?

(*d*) For a couple in the lower class, determine the expected number of generations for their descendants to enter the highest class.

(*e*) For middle-class parents, what is the expected number of generations before their descendants are in the middle class again?

(*f*) Is this chain reversible? Explain what your conclusion intuitively means about the classes and the economy in this country.

(*g*) A couple comes to the sociologist and asks for the probability that their distant descendants, hundreds of years in the future, will be in the lower class. What should the sociologist say?

Problem 1.26. (*Random walk with geometric waiting times*) Start with a simple random walk $(X_n)_{n\geq 0}$ on the vertices of a graph $G = (V, E)$, and let $p : V \to [0, 1]$ be a collection of probabilities. That is, there is one probability p_x assigned to each vertex x. Now modify the chain as follows: if we arrive at a given vertex x at step k, we sample an independent $\text{Geo}(p_x)$ random variable $T \in \{1, 2, \ldots\}$, and we wait at vertex x for an additional $T - 1$ units of time before jumping to the next vertex. That is, $X_k = X_{k+1} = \cdots = X_{k+T-1} = x$, and X_{k+T} is one of the adjacent vertices $y \sim x$. For example, if we arrive at x at step k and $T = 3$, then

$$X_k = x, \quad X_{k+1} = x, \quad X_{k+2} = x, \quad X_{k+3} = y,$$

where $y \sim x$ is a randomly-chosen neighboring vertex. Is this new process $Y = (Y_k)_{k\geq 0}$ a Markov chain? Explain.

Chapter 2

Classical models

2.1 Random walks on $\mathbb{Z}$ and Gambler's ruin

An important class of Markov chains are the random walks on the integers $\mathbb{Z}$ or subsets $\{a, a+1, a+2, \ldots, a+n\}$ of the integers. In this section we look at several such chains and establish some famous formulas describing their behavior. Section 2.1.1 looks at the probabilities for hitting either end of a subset $\{a, a+1, a+2, \ldots, a+n\} \subset \mathbb{Z}$, while Section 2.1.2 looks at the average hitting times for the endpoints. We conclude in Section 2.1.3 by looking at walks on all of $\mathbb{Z}$ and their large-scale behavior, where we observe the emergence of *Brownian motion*, the most fundamental stochastic process in modern probability theory.

2.1.1 Boundary hitting probabilities

Consider the simple walk on $\Omega = \{0, 1, 2, \ldots, n\}$ with **absorbing boundaries**, by which we mean that once the chain reaches either 0 or n, it is permanently stuck there (recall you've already seen this walk in Problem 1.4). We call $\{0, n\}$ the **boundary** $\partial\Omega$ of Ω and write

$$\partial\Omega = \{0, n\}. \tag{2.1}$$

Formally, the transition probabilities for the absorbing walk are

$$P(j, j-1) = P(j, j+1) = \frac{1}{2}, \qquad 1 \leq j \leq n-1,$$

$$P(0,0) = P(n,n) = 1. \tag{2.2}$$

Now, suppose we start with $1 \leq X_0 = k \leq n-1$. Let $\tau = \tau_{\partial\Omega}$ be the hitting time to the boundary of Ω. Then either $X_\tau = 0$ or $X_\tau = n$. What is $\mathbb{P}_k(X_\tau = n)$? It is clear that $\mathbb{P}_0(X_\tau = n) = 0$, since we cannot move away from zero if we begin there. It is likewise clear that $\mathbb{P}_n(X_\tau = n) = 1$. It also seems that we should have

$$0 < \mathbb{P}_1(X_\tau = n) < \mathbb{P}_2(X_\tau = n) < \cdots < \mathbb{P}_{n-1}(X_\tau = n) < 1,$$

since the closer we begin to n, the more likely it should be that we hit $\partial\Omega$ at n rather than 0.

Can we say anything more? How much do the probabilities increase each step? Is there an explicit formula? The famous *Gambler's ruin* theorem answers these questions and says that the probabilities are, in fact, a *linear* function in the starting position k.

Theorem 2.1 (Gambler's ruin)**.** *For the Markov chain on $\Omega = \{0, 1, 2, \ldots, n\}$ described above,*

$$\mathbb{P}_k(X_\tau = n) = \frac{k}{n} \tag{2.3}$$

$$\mathbb{P}_k(X_\tau = 0) = 1 - \frac{k}{n} \tag{2.4}$$

for any $k \in \{0, 1, 2, \ldots, n\}$.

One way to think about (2.3) is that the probability of terminating at the right endpoint is the ratio of the distance to the *left* endpoint to the entire interval length, $\mathbb{P}_k(X_\tau = n) = \frac{k-0}{n-0}$. That is, we take the ratio of how far away the "bad" endpoint is to the length of the entire playing field. This, of course, makes a lot of sense: the further away the bad point is, the more likely we are to end at the "good" boundary point. The surprising thing about Theorem 2.1 is that it says the probability is exactly this ratio, and not some complicated function of it.

The proof exploits a very deep relationship between Markov chains and differential equations.

Proof. Set $p(k) := \mathbb{P}_k(X_\tau = n)$. We have that $p(0) = 0$ and $p(n) = 1$, and we need to show that p is linear with slope $\frac{1}{n}$. We first claim that p satisfies

the *mean-value property*

$$p(k) = \frac{1}{2}p(k-1) + \frac{1}{2}p(k+1) \tag{2.5}$$

for each $1 \le k \le n-1$. That is, we claim that each value of p (outside of $\partial\Omega$) is just the average of its two neighbors. We see this from conditioning on our first step (this simple idea, called **the method of recursion**, is often very useful):

$$\begin{aligned}
p(k) &= \mathbb{P}(X_\tau = n \,|\, X_0 = k) \\
&= \mathbb{P}(X_\tau = n, X_1 = k-1 \,|\, X_0 = k) + \mathbb{P}(X_\tau = n, X_1 = k+1 \,|\, X_0 = k) \\
&= \mathbb{P}(X_\tau = n \,|\, X_1 = k-1, X_0 = k)\mathbb{P}(X_1 = k-1 \,|\, X_0 = k) \\
&\qquad + \mathbb{P}(X_\tau = n \,|\, X_1 = k+1, X_0 = k)\mathbb{P}(X_1 = k+1 \,|\, X_0 = k) \\
&= \mathbb{P}(X_\tau = n \,|\, X_1 = k-1)\frac{1}{2} + \mathbb{P}(X_\tau = n \,|\, X_1 = k+1)\frac{1}{2} \qquad (2.6) \\
&= \mathbb{P}(X_\tau = n \,|\, X_0 = k-1)\frac{1}{2} + \mathbb{P}(X_\tau = n \,|\, X_0 = k+1)\frac{1}{2} \qquad (2.7) \\
&= \frac{1}{2}p(k-1) + \frac{1}{2}p(k+1),
\end{aligned}$$

as claimed. Here we have used the Markov property twice, first in (2.6) to ignore the earlier information on X_0, and then again in (2.7) to say that the chain "starts over" after the first step.

Next, define the increment $\Delta_k := p(k) - p(k-1)$, $1 \le k \le n$. If p really is linear, as claimed, then the increment should be constant, independent of k. We note subtracting $\frac{1}{2}p(k)$ and $\frac{1}{2}p(k-1)$ from both sides of (2.5) yields

$$\frac{1}{2}(p(k) - p(k-1)) = \frac{1}{2}(p(k+1) - p(k)),$$

or $\Delta_k = \Delta_{k+1}$, showing that all adjacent Δ's are equal. Hence all the Δ's are equal to Δ_1, say. Note that we thus have the telescopic sum

$$\begin{aligned}
n\Delta_1 &= \Delta_1 + \Delta_2 + \Delta_3 + \cdots + \Delta_n \\
&= p(1) - p(0) + p(2) - p(1) + p(3) - p(2) + \cdots + p(n) - p(n-1) \\
&= p(n) - p(0) = 1 - 0,
\end{aligned}$$

and so $\Delta_1 = \frac{1}{n} = \Delta_k$ for each k. Therefore, p is linear, as claimed, with

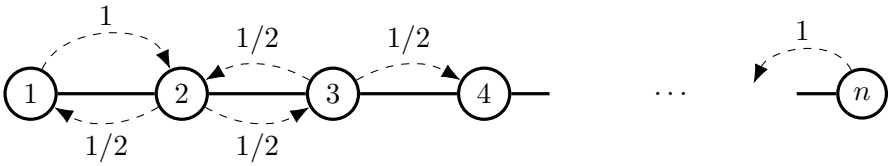

Figure 2.1: The simple walk on $\{0, 1, \ldots, n\}$ with reflecting boundary.

slope $\frac{1}{n}$, and

$$p(k) = p(0) + \Delta_1 + \Delta_2 + \cdots + \Delta_k = \frac{k}{n},$$

yielding (2.3). The other formula (2.4) immediately follows, as

$$\mathbb{P}_k(X_\tau = 0) = 1 - \mathbb{P}_k(X_\tau = n). \qquad \square$$

Sometimes we want to consider different behaviors once we reach $\partial\Omega$, and it is clear that Theorem 2.1 will still hold, since the walk does not visit the boundary until the hitting time. In particular, (2.3) and (2.4) hold for the simple random walk on $\Omega = \{0, 1, \ldots, n\}$ with **reflecting boundary**, which has transition probabilities

$$P(0, 1) = 1, \quad P(n, n-1) = 1,$$

as in Figure 2.1. The difference, of course, is that the walk with the reflected boundaries continues going after hitting one of the endpoints.

We can also eliminate the boundary altogether, and consider the simple random walk (X_n) on the infinite state space $\mathbb{Z}$, we still have

$$\mathbb{P}_k(X_\tau = n) = \frac{k}{n}$$

for $\tau = \tau_{\{0,n\}}$, and $0 \le k \le n$.

Thinking geometrically, it is clear that the principle behind Theorem 2.1 should continue to hold even if we choose a different portion of the integers $\{a, a+1, \ldots, b\}$ to do our walk on, and hence have boundary of $\{a, b\}$ instead of $\{0, 1\}$. We ask the reader to use Theorem 2.1 to prove the following general version of the hitting probability formulas.

Theorem 2.2 (Gambler's ruin for a general interval)**.** *Let (X_n) be a simple random walk on $\mathbb{Z}$, and let $\tau = \tau_{\{a,b\}}$ be the hitting time to integers $a < b$. Then for any integer $a \le k \le b$,*

$$\mathbb{P}_k(X_\tau = b) = \frac{k-a}{b-a} \qquad \text{and} \qquad \mathbb{P}_k(X_\tau = a) = \frac{b-k}{b-a} \tag{2.8}$$

Exercise 2.1. *Prove* (2.8) *by using Theorem 2.1 and a change of variables.*

We will see another elegant way to show (2.8) in Chapter 5 once we have the tools of *martingales* and the *Optional Sampling Theorem*. See Problem 5.8.

2.1.2 Expected hitting times

We completely understand the boundary hitting *probabilities* via Theorem 2.2: the probability we hit the right boundary is a linear function of the distance we are from the left boundary, with boundary values of 0 and 1. What about the *time* it takes, on average, to hit the boundary? How many steps will the random walk take? We again begin with the simple setting $\Omega = \{0, 1, \ldots, n\}$. Let $\tau = \tau_{\{0,n\}} = \tau_{\partial\Omega}$. We can use the same ideas as in the proof of Theorem 2.1 to obtain an explicit formula.

Exercise 2.2. (*i*) *What are $\mathbb{E}_0(\tau)$ and $\mathbb{E}_n(\tau)$?*

(*ii*) *Before you look at the next theorem statement, sketch your guess of the shape of the graph of the function $k \mapsto \mathbb{E}_k(\tau)$ for $0 \le k \le n$.*

Theorem 2.3. *For the simple random walk on $\Omega = \{0, 1, \ldots, n\}$, the hitting time $\tau := \tau_{\partial\Omega}$ satisfies*

$$\mathbb{E}_k(\tau) = k(n-k), \qquad 0 \le k \le n. \tag{2.9}$$

Note that the function $k \mapsto k(n-k)$ in (2.9) is maximized at $k = n/2$ and doesn't change when k is replaced with $n - k$. To give a concrete example, if we are doing the walk on $\{0, 1, \ldots, 11\}$, we see that the average hitting time would be largest if we could start at $k = 5.5$. Of course that doesn't make sense - here the largest value will occur at both $k = 5$ and $k = 6$, the integers surrounding the midpoint. Also, starting at 2 will have the same

average hitting time as starting at 9, as will starting at 3 and 8. We see that the symmetries in the hitting times reflect the symmetry of the interval.

Proof. Write $f_k := \mathbb{E}_k(\tau)$. We first claim that we have the averaging property

$$f_k = 1 + \frac{1}{2}f_{k-1} + \frac{1}{2}f_{k+1} \tag{2.10}$$

for each $1 \leq k \leq n-1$. Note that (2.10) is intuitively clear: we take one step (+1) and with equal probability arrive at either $k-1$ or $k+1$. The chain starts over, and we need the average number of steps from that new vertex to hit the boundary.

To make this rigorous, we can use the tower property of conditional expectation and the method of recursion again:

$$\begin{aligned}
f_k &= \mathbb{E}_k(\tau) \\
&= \mathbb{E}_k(\mathbb{E}_k(\tau \,|\, X_1)) \\
&= \mathbb{E}_k(\tau \,|\, X_1 = k-1)\mathbb{P}_k(X_1 = k-1) + \mathbb{E}_k(\tau \,|\, X_1 = k+1)\mathbb{P}_k(X_1 = k+1) \\
&= \mathbb{E}_k(1 + \tau - 1 \,|\, X_1 = k-1)\frac{1}{2} + \mathbb{E}_k(1 + \tau - 1 \,|\, X_1 = k+1)\frac{1}{2} \\
&= \big(1 + \mathbb{E}_k(\tau - 1 \,|\, X_1 = k-1)\big)\frac{1}{2} + \big(1 + \mathbb{E}_k(\tau - 1 \,|\, X_1 = k+1)\big)\frac{1}{2} \\
&= 1 + \mathbb{E}_k(\tau - 1 \,|\, X_1 = k-1)\frac{1}{2} + \mathbb{E}_k(\tau - 1 \,|\, X_1 = k+1)\frac{1}{2} \\
&= 1 + \mathbb{E}_{k-1}(\tau)\frac{1}{2} + \mathbb{E}_{k+1}(\tau)\frac{1}{2} \qquad (2.11) \\
&= 1 + \frac{1}{2}f_{k-1} + \frac{1}{2}f_{k+1},
\end{aligned}$$

where in (2.11) we have used the fact that the number of remaining steps until X_1 hits the boundary is $\tau - 1$ (we have taken one step), and hence

$$\begin{aligned}
&\mathbb{E}_k(\tau - 1 \,|\, X_1 = k-1) = \mathbb{E}(\tau \,|\, X_0 = k-1) = \mathbb{E}_{k-1}(\tau) \quad \text{and} \\
&\mathbb{E}_k(\tau - 1 \,|\, X_1 = k+1) = \mathbb{E}(\tau \,|\, X_0 = k+1) = \mathbb{E}_{k+1}(\tau).
\end{aligned}$$

Now that we have (2.10) we are basically done. We first claim that this

implies

$$f_k = k + \frac{k}{k+1} f_{k+1}, \qquad 0 \le k \le n-1, \tag{2.12}$$

which we prove with induction. The base case $k = 0$ is clear since $f_0 = 0$. Now suppose (2.12) holds for some $0 \le m \le n-2$. We need to show the $m+1$ case holds too, which is to say,

$$f_{m+1} = m + 1 + \frac{m+1}{m+2} f_{m+2}. \tag{2.13}$$

By (2.10) and our inductive assumption,

$$\begin{aligned} f_{m+1} &= 1 + \frac{1}{2} f_m + \frac{1}{2} f_{m+2} \\ &= 1 + \frac{1}{2}\left(m + \frac{m}{m+1} f_{m+1}\right) + \frac{1}{2} f_{m+2} \qquad \Leftrightarrow \\ \frac{m+2}{2(m+1)} f_{m+1} &= 1 + \frac{1}{2} m + \frac{1}{2} f_{m+2} \qquad \Leftrightarrow \\ f_{m+1} &= m + 1 + \frac{m+1}{m+2} f_{m+2}. \end{aligned}$$

Hence we have (2.13), as claimed, which completes the induction proof for (2.12). And now we can use our knowledge of the other boundary value $f_n = 0$ to work backwards to get (2.9). This would be another induction proof, but we just work out the first several cases and let the reader fill in the details. Identity (2.12) with $k = n-1$ gives

$$f_{n-1} = n - 1 + \frac{n-1}{n} f_n = n - 1.$$

Feed this back into (2.12) to see

$$f_{n-2} = n - 2 + \frac{n-2}{n-1} f_{n-1} = 2(n-2),$$

which similarly implies

$$f_{n-3} = n - 3 + \frac{n-3}{n-2} f_{n-2} = 3(n-3).$$

In general, we find

$$f_{n-k} = (n-k)k, \qquad 0 \le k \le n, \tag{2.14}$$

which is equivalent to (2.9) (replace k with $n-k$). □

Exercise 2.3. *Give a formal induction proof for* (2.14)*. Start with verifying the base case $k = 0$. Then make the inductive assumption that* (2.14) *holds for some $0 \le m \le n-1$, and show this implies it holds for $k = m+1$.*

Just as we extended our hitting probabilities to general intervals of integers in Theorem 2.2, so we can do so with Theorem 2.3.

Theorem 2.4 (Gambler's ruin hitting times for a general interval)**.** *Let (X_n) be a simple random walk on $\mathbb{Z}$, and let $\tau = \tau_{\{a,b\}}$ be the hitting time to integers $a < b$. Then for any integer $a \le k \le b$,*

$$\mathbb{E}_k(\tau) = (k-a)(b-k) \tag{2.15}$$

Exercise 2.4. *Prove* (2.15) *by using Theorem 2.3 and a change of variables.*

As with the hitting probability, you will revisit the average hitting time in Problem 5.9 once we have *martingales* and the *Optional Sampling Theorem* to work with.

2.1.3 The simple random walk on $\mathbb{Z}$ and Brownian motion

It is also important to consider the case where there are no boundary conditions, and our walk is on all of $\mathbb{Z}$. Let X_n be such a simple walk on $\mathbb{Z}$ starting at $X_0 = 0$. We can think of this as flipping a fair coin at each step and moving up one unit if we get heads, and down one if we get tails. Plotting X_n as a function of the step n and linearly interpolating between the points (that is, drawing a straight line between successive points) gives a random piece-wise linear function. Thirty steps of one sampling are shown in Figure 2.2. We call this random function X_t, where now t no longer needs to be an integer.

An interesting question is what happens with this graph as we "zoom out" and display more and more steps. What if we plot up to $n = 300$ or $n = 3000$, instead of $n = 30$? Since we have a random graph, some care

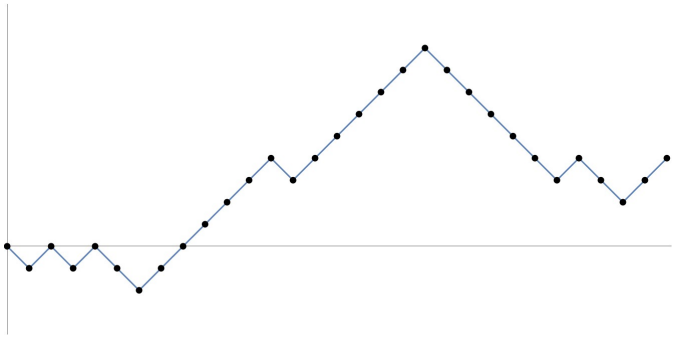

Figure 2.2: Thirty steps of a simple random walk (X_n) on $\mathbb{Z}$, with linear interpolation between the values.

is needed to know how to "zoom out" in a reasonable way. One way is to consider X_{nt} as $n \to \infty$, which compresses more and more steps of X_n into the same amount of time. At $t = 1$, for instance, $X_{n\cdot 1} = X_n$, and in one time unit we have already gone n steps in our random walk.

This pre-composition scales time, but we also need to rescale space. We can see this by computing the average and variance of the random variable X_{nt}. As $n \to \infty$, $\mathbb{E}(X_{nt})$ and $\mathrm{Var}(X_{nt})$ should be well behaved for fixed t if we are going to have a nice limit. Let's consider $t = 1$, for example. Writing $X_n = \sum_{j=1}^n Y_j$, where the Y_j are our iid coin flips, $\mathbb{P}(Y_j = 1) = \mathbb{P}(Y_j = -1) = 1/2$, we have

$$\mathbb{E}(X_n) = \sum_{j=1}^{n} \mathbb{E}(Y_j) = 0 \qquad \text{and} \tag{2.16}$$

$$\mathrm{Var}(X_n) = \sum_{j=1}^{n} \mathrm{Var}(Y_j) = n\,\mathrm{Var}(Y_1) = n\mathbb{E}(Y_1^2), \tag{2.17}$$

as $\mathrm{Var}(Y_1) = \mathbb{E}(Y_1^2) - \mathbb{E}(Y_1)^2 = \mathbb{E}(Y_1^2)$. Since $\mathbb{E}(Y_1^2) = 1$, we see $\mathrm{Var}(X_n) = n$, which says that as n increases, X_{nt} will have wilder and wilder oscillations at $t = 1$, even though its average is always zero, (2.16). It would make sense to normalize to control for this, and if we consider $\frac{1}{\sqrt{n}} X_{nt}$ instead, re-doing (2.17) shows that the variance at $t = 1$ is then exactly 1 for all n. In fact,

the Central Limit Theorem says

$$\frac{1}{\sqrt{n}}X_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} Y_n \Rightarrow Z \sim N(0,1). \tag{2.18}$$

This is the correct scaling of space, and under it, $\frac{1}{\sqrt{n}}X_{nt}$ at $t = 1$ is asymptotically a standard normal random variable. The two graphs in Figure 2.3 show $\frac{1}{\sqrt{n}}X_{nt}$ with $n = 10$ and $n = 100$ for the same sampling of the random walk as in Figure 2.2. The time axis goes up to $t = 30$, thus displaying 300 and 3000 steps of the original walk, respectively.

Exercise 2.5. *What happens if we consider other scalings* $\frac{1}{n^{1/2+\beta}}X_{nt}$ *instead of* $\frac{1}{\sqrt{n}}X_{nt}$*? Show that:*

(a) *If* $\beta > 0$*, then the variance at* $t = 1$ *tends to zero as* $n \to \infty$*. Thus the limiting graph ceases to be random and is flattened to be identically zero.*

(b) *If* $\beta < 0$*, then the variance at* $t = 1$ *tends to* ∞ *as* $n \to \infty$*. Thus "zooming out" gives more and more extreme oscillations, and there is no limiting distribution at* $t = 1$*.*

Together, parts (a) *and* (b) *show that using* $\beta = 0$*, corresponding to our* $1/\sqrt{n}$ *scaling, is necessary to obtain a non-trivial distributional limit.*

The jagged, fractal-like random function we get in the distributional limit is of fundamental importance in probability theory, economics and physics, and is called *Brownian motion* B_t. Our work above in (2.18) shows $B_1 \sim N(0,1)$, and for any time $t \geq 0$, one can similarly show $B_t \sim N(0,t)$. Brownian motion is, in fact, a *continuous-time* Markov chain. Further study of its properties is beyond the scope of our text, but we want you to be aware of its existence, as well as whet your appetite for deeper study.

One technique for analyzing Brownian motion B_t is to start with the piece-wise linear functions $\frac{1}{\sqrt{n}}X_{nt}$. One can study their properties and then attempt to take a distributional limit as $n \to \infty$ to obtain information about B_t. This boils down to understanding the simple random walk (X_n), our original Markov chain. A natural starting place for understanding X_n is to determine its distribution for all n.

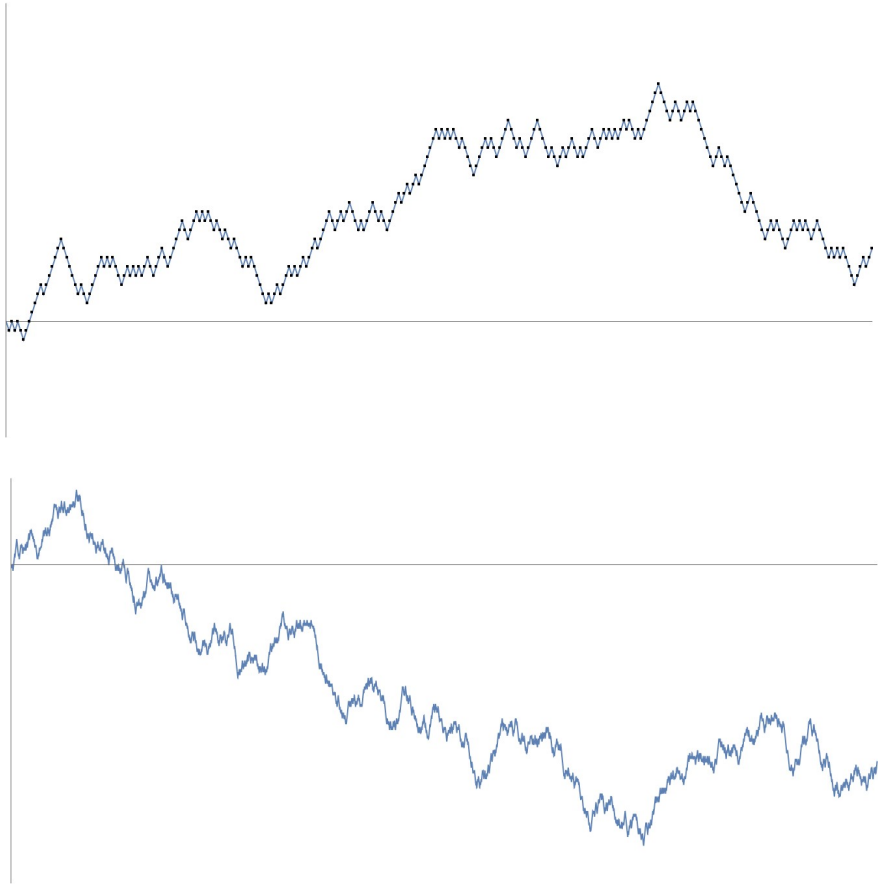

Figure 2.3: The same sampling of the random walk in Figure 2.2, zoomed out to show 300 and 3000 steps, respectively.

Lemma 2.5. *For $n \in \mathbb{N}$, the pdf of X_n is*

$$\mathbb{P}(X_n = k) = \binom{n}{\frac{n+k}{2}} \frac{1}{2^n} \quad \textit{for} \quad k = -n, -n+2, \ldots, n-2, n, \tag{2.19}$$

while $\mathbb{P}(X_n = k) = 0$ for all other k.

Let's parse what this is saying. The condition on k says that the parity of our position X_n must match the parity of the step number n. For instance, if $n = 5$, then X_n is odd and $X_n \in \{-5, -3, -1, 1, 3, 5\}$. Also, the formula is actually very intuitive: there are n steps, and hence 2^n total ways to choose $+1$'s and -1's for each step. If we take $\frac{n+k}{2}$ steps of $+1$, then the number of -1 steps is

$$n - \frac{n+k}{2} = \frac{n-k}{2},$$

and our ending position is $1 \cdot \frac{n+k}{2} - 1 \cdot \frac{n-k}{2} = k$. Hence the binomial coefficient counts the number of ways to pick out the correct number of $+1$ steps from the n total so that we reach position k. This sketches one proof of (2.19); we offer another slightly more formal argument below.

Proof. We may write $X_n = \sum_{j=1}^n Y_j$ for iid coin flips $Y_1, \ldots, Y_n$, $\mathbb{P}(Y_j = 1) = \mathbb{P}(Y_j = -1) = 1/2$. Let

$$H_n = \sum_{j=1}^n Y_j \mathbb{1}_{\{Y_j = 1\}} \quad \text{and} \quad T_n = \sum_{j=1}^n -Y_j \mathbb{1}_{\{Y_j = -1\}}$$

be the number of heads and tails through n steps, respectively. Since $T_n = n - H_n$, we have

$$X_n = H_n - T_n = 2H_n - n.$$

Noting that $H_n \sim \text{Bin}(n, 1/2)$, we thus see

$$\begin{aligned} \mathbb{P}(X_n = k) &= \mathbb{P}(2H_n - n = k) \\ &= \mathbb{P}\big(H_n = \frac{n+k}{2}\big) = \binom{n}{\frac{n+k}{2}} \frac{1}{2^n} \end{aligned}$$

when $\frac{n+k}{2} \in \{0, 1, \ldots, n\}$, which is equivalent to $k \in \{-n, -n+2, -n+$

$4, \ldots, n\}$. □

2.1.4 Birth and death chains

The simple random walk on $\mathbb{Z}$ is fundamentally important, but other variations occur in theory and practice. One class of such models are the *birth and death chains*, which model population dynamics. Consider a population between 0 and n that increases or decreases by at most one during each unit of time. We can view this as a Markov chain on the state space $\Omega = \{0, 1, 2, \ldots, n\}$ of size $n + 1$ with parameters

$$p_k := p(k, k+1), \quad q_k := p(k, k-1), \quad r_k := p(k, k), \tag{2.20}$$

where $p_k + q_k + r_k = 1$, $q_0 = 0$ and $p_n = 0$. Here the p_k, q_k and r_k can all vary with k. We call such a Markov chain a **birth and death chain**. Note that the chain is irreducible iff

(*i*) $p_k, q_k > 0$ for all $1 \leq k \leq n-1$, and

(*ii*) $p_0 > 0$ and $q_n > 0$,

since it is only under these conditions that we can move between any two states.

What is the stationary distribution of these chains? Are they reversible? The next theorem answers in the affirmative for both questions.

Theorem 2.6. *Every irreducible birth and death chain is time reversible under an explicit stationary distribution.*

Proof. Note that if we can build a stationary distribution π which satisfies the detail balance equations, then the MC is automatically time reversible under π. So, it suffices to find π_k, $k = 0, 1, \ldots, n$ with $\pi_k > 0$ and $\sum_{k=1}^{n} \pi_k = 1$ satisfying the detail balance equations

$$\pi_k p(k, j) = \pi_j p(j, k)$$

for all j, k. Since $p(k, j) = 0$ except when $j = k-1, k, k+1$, there are only three cases to check:

(*i*) $\pi_k p(k, k+1) = \pi_{k+1} p(k+1, k)$,

(ii) $\pi_k p(k,k-1) = \pi_{k-1} p(k-1,k)$, and

(iii) $\pi_k p(k,k) = \pi_k p(k,k)$.

Case (iii) is trivially true, and (ii) is encompassed by (i) through replacing k with $k+1$. So, recalling our notation (2.20), we just need build π for which

$$\pi_k p_k = \pi_{k+1} q_{k+1}, \qquad k = 0, 1, \ldots, n-1. \tag{2.21}$$

To that end, define "weights" w_k, $k = 0, 1, \ldots, n$, via $w_0 := 1$ and, for $1 \le k \le n$,

$$\begin{aligned} w_k :&= \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} \\ &= \frac{\mathbb{P}(X_1 = 1, X_2 = 2, \ldots, X_k = k | X_0 = 0)}{\mathbb{P}(X_1 = k-1, X_2 = k-2, \ldots, X_k = 0 | X_0 = k)} \end{aligned} \tag{2.22}$$

and set

$$\pi_k = \frac{w_k}{\sum_{j=0}^n w_j}, \quad k = 0, 1, \ldots, n. \tag{2.23}$$

Exercise 2.6. *Verify that π is a probability distribution on Ω and that π satisfies* (2.21)*.*

We conclude from the exercise that $(\pi_k)_{k=0}^n$ is the unique reversible stationary distribution for the birth and death chain. □

Remark 2.7. Note that the expression (2.22) can serve as a helpful mnemonic for the w_k formula: the numerator is the probability of starting at 0 and continually increasing until you reach k, and the denominator is the probability of the reversed journey.

Example 2.1. The simplest case is when $p_k = p$ for all $0 \le k \le n-1$ and $q_k = q$ for $1 \le k \le n$, where $p + q = 1$ (and so $r_k = 0$ for $k = 1, 2, \ldots n-1$, but $r_0 = 1 - p$ and $r_n = 1 - q$). That is, when all the p's and q's are the same. What is our stationary distribution (2.23) here? We find

$$w_0 = 1 = \left(\frac{p}{q}\right)^0, \quad w_k = \frac{p_0 p_1 \cdots p_{k-1}}{q_1 q_2 \cdots q_k} = \left(\frac{p}{q}\right)^k,$$

and thus

$$\pi_k = \frac{(p/q)^k}{\sum_{j=0}^{n}(p/q)^j}, \qquad k = 0, 1, \ldots, n.$$

Exercise 2.7. *Make this example even simpler by setting $p = q$. What kind of random walk is this? What named distribution do we get for π?*

2.2 The Ehrenfest urn

The Ehrenfest urn is a Markov chain modelling the diffusion of particles through a porous membrane. It was developed by the physicist Paul Ehrenfest (1880-1933) and his wife Tatyana (1876-1964) as a way of explaining the second law of thermodynamics. It is an intrinsically interesting model, but also important for us because it is a Markov chain which is not a simple random walk on a graph, but rather a *projection* of such a walk. We explain precisely what this means below.

The model starts with two urns having a total of N identical balls. Among all the balls (in both urns), pick one uniformly at random and switch its urn. If we track the number of balls X_n in the first urn after n steps, then we have a Markov chain on $\Omega = \{0, 1, \ldots, N\}$, the **Ehrenfest urn model**.

Exercise 2.8. *Show that the transition probabilities for X_k are are*

$$p(k, k+1) = \mathbb{P}_k(\textit{adding a ball from urn 2 to urn 1}) = 1 - \frac{k}{N},$$
$$p(k, k-1) = \mathbb{P}_k(\textit{adding a ball from urn 1 to urn 2}) = \frac{k}{N}.$$

Informally speaking, what state does the chain have a "preference" for?

The chain is clearly irreducible. What is its stationary distribution π? The question is simplified by broadening our viewpoint to see this chain as a projection of a Markov chain on a hypercube.

2.2.1 The Ehrenfest urn as a projection from $\mathbb{R}^N$

Let

$$\Omega' = \{0, 1\}^N = \big\{\, \{s_1, s_2, \ldots, s_N\} \mid s_j \in \{0, 1\} \text{ for all } j \,\big\}$$

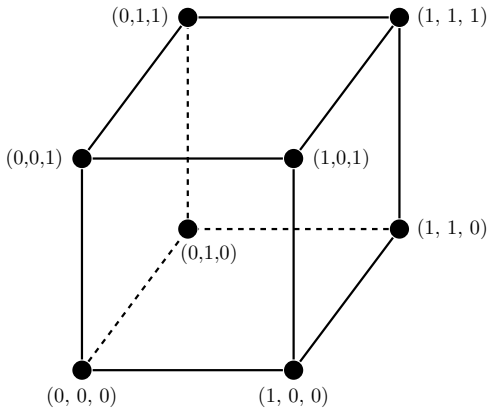

Figure 2.4: The hypercube $\{0,1\}^3$ in $\mathbb{R}^3$

be the set of all sequences of 0's and 1's of length N. This is simply the collection of vertices of the (hyper)cube in $\mathbb{R}^N$, and as such, it has a natural graph structure: say that two vertices u and v are adjacent,

$$u = (u_1, u_2, \ldots, u_N) \sim (v_1, v_2, \ldots, v_N) = v,$$

if you can go from u to v by switching exactly one coordinate from 0 to 1 or from 1 to 0. See Figure 2.4 for the case $N = 3$. More formally, $u \sim v$ if

$$\|u - v\|_1 := \sum_{j=1}^{N} |u_j - v_j| = 1.$$

Thus, each step of the random walk on this graph randomly switches exactly one of the coordinates.

Exercise 2.9. *What are the degrees of the vertices of the hypercube in $\mathbb{R}^3$? How many edges are there? Generalize these two questions to the hypercube in $\mathbb{R}^N$.*

The hypercube is clearly connected, and so its random walk is irreducible. Note that it is also *regular*, which is to say every vertex u has the same degree N. By our formula for random walks on graphs, we see that its stationary

distribution is

$$\pi'(u) = \frac{\deg(u)}{2|E|} = \frac{N}{2^N N} = \frac{1}{2^N},$$

which is simply the uniform distribution on Ω'.

How does this all relate to the Ehrenfest urn? We claim that we can "project" the random walk on the hypercube onto the urn chain. If so, we can simply project the stationary distribution too, and we'll have the desired stationary distribution π for the urn model.

Intuitively, the first urn will hold those coordinates which are 1, and the other urn will hold those coordinates which are zero. So, the count of balls in the first urn is just the sum of the coordinates of a vertex u.

To formalize this, consider the function $F : \Omega' \to \{0, 1, \ldots, N\}$ given by

$$F(u) = \#\{\text{ coords of } u \text{ which are } 1\} = \sum_{j=1}^{N} u_j.$$

Now, if $(X'_1, X'_2, \ldots)$ is the simple random walk on the hypercube, we claim that $(F(X'_1), F(X'_2), \ldots)$ is the Ehrenfest Markov chain. We can see this by looking at the transition probabilities of $F(X'_j)$. If $F(X'_j) = k$, then our current coordinate has k ones and $N - k$ zeros. Hence

$$\mathbb{P}(F(X'_{j+1}) = k + 1 | F(X'_j) = k) = \frac{N - k}{N},$$

the probability that we switch one of the zeros. Similarly,

$$\mathbb{P}(F(X'_{j+1}) = k - 1 \,|\, F(X'_j) = k) = \frac{k}{N},$$

the probability that we switch one of the ones. These are exactly the answers we saw in Exercise 2.8, showing $(F(X'_1), F(X'_2), \ldots)$ is exactly the Ehrenfest urn, as claimed.

2.2.2 Projecting to get the stationary distribution

Therefore, if $X' \sim \pi' = \text{Unif}(\Omega')$, then $X = F(X') \sim \pi$, the stationary distribution on the urn. The uniform distribution on Ω' just assigns a 1 or 0 to each coordinate with equal probability, and then X totals the 1's.

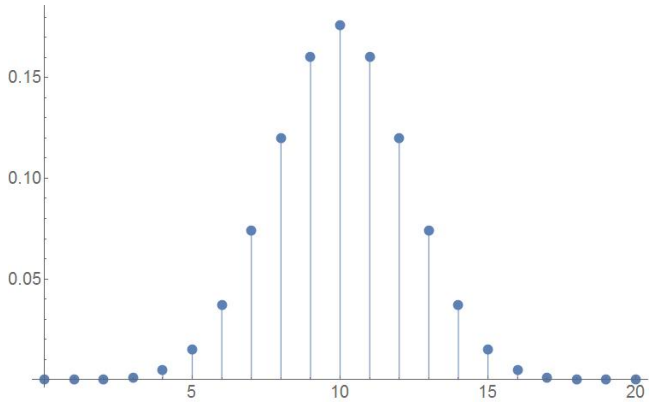

Figure 2.5: The stationary distribution π for the Ehrenfest urn with $N = 20$ particles. We see π is symmetric and that there is a strong preference for a balance of particles between the two containers. (Is that what you guessed in Exercise 2.8?)

Thinking of this as the number of "successes" in N trials, we see that $X \sim \text{Bin}(N, 1/2)$. Thus our stationary distribution π is

$$\pi(k) = \binom{N}{k} \frac{1}{2^N}, \quad k = 0, 1, \ldots, N.$$

We can verify stationarity by checking the DBE's (1.66). Indeed, on the one hand,

$$\pi(k)p(k, k+1) = \binom{N}{k} \frac{1}{2^N} \left(1 - \frac{k}{N}\right) = \frac{(N-1)!}{k!(N-k-1)!2^N}, \tag{2.24}$$

while on the other hand,

$$\pi(k+1)p(k+1, k) = \binom{N}{k+1} \frac{1}{2^N} \frac{k+1}{N} = \frac{(N-1)!}{k!(N-k-1)!2^N} \tag{2.25}$$

as well. Thus $\pi \sim \text{Bin}(N, 1/2)$ is stationary, and the Ehrenfest urn is time-reversible under π. See a plot of π for $N = 20$ in Figure 2.5.

As an aside, note again the value of the DBE's: by simply verifying the equality of (2.24) and (2.25), not only do we verify that proposed π is the stationary distribution, we also show that the urn is reversible under π.

2.3 Bernoulli-Laplace diffusion

A related urn model, *Bernoulli-Laplace diffusion*, starts with two urns again, but now with N red and N blue balls. The evolution is identical to the Ehrenfest model: take a ball uniformly and switch its urn. But now keep track of the number X_n of *red* balls in the first urn after n steps, instead of the total number of balls. Then $(X_n, n = 0, 1, \ldots)$ is a Markov chain on $\{0, 1, \ldots, N\}$.

Exercise 2.10. (*a*) *Compute the transition probabilites* $p(k, k-1), p(k, k)$ *and* $p(k, k+1)$ *for Bernoulli-Laplace diffusion chain.*

(*b*) *Show that the hypergeometric distribution*

$$\pi(k) = \frac{\binom{N}{k}^2}{\binom{2N}{N}}, \quad k = 0, 1, \ldots, N$$

is stationary for the chain by verifying the DBE.

2.4 The Pólya urn

2.4.1 Negative feedback vs positive feedback

In the Ehrenfest urn model, the preference is to return the system to the state of having the same number of balls in both urns, as is evident from the stationary distribution in Figure 2.5. If one urn has more balls than the other, it is more likely that that urn loses a ball rather than gains one in the next turn. This is a type of *negative feedback*, where the system pushes against unbalances. *Positive feedback* models have the opposite behavior: growth in one direction encourages further growth in the same way. These are models where "the rich get richer."

2.4.2 The Pólya urn

The **Pólya urn model** is a positive feedback system. Here's how it works: start with one black ball and one red ball in an urn. At each stage, pull out a ball uniformly at random, and then return it to the urn with another ball of the same color. Let X_n be the number of black balls after n turns. Then $(X_n, n = 0, 1, \ldots)$ is a Markov chain on $\Omega = \mathbb{N}$.

The transition probabilities are easy enough: X_n can either stay constant or increase by one at each stage. After n steps there are a total of $n+2$ balls, and we observe

$$\begin{aligned}\mathbb{P}(X_{n+1} = k+1 \,|\, X_n = k) \\ &= \mathbb{P}(\text{choose black in step } n+1 \,|\, X_n = k) = \frac{k}{n+2},\end{aligned}$$

and hence $\mathbb{P}(X_{n+1} = k \,|\, X_n = k) = 1 - \frac{k}{n+2}$. Note that these probabilities are **time-inhomogeneous**, that is, they change with each step of time. In other words, for each distinct n, we need a different (infinite) matrix $P = P(n)$ to describe the transition probabilities.

2.4.3 Long-term behavior of the Pólya urn

How a particular sampling of this Markov chain plays out heavily depends on the first few steps. Figure 2.6 shows several samples of the urn for 1000 steps, plotting the proportion of black balls. We see that, perhaps surprisingly, this proportion appears to tend to a constant. What is the distribution of this constant? In Theorem 2.8 and 2.9 we will rigorously show the (perhaps surprising) answer that the limiting ratio is uniformly distributed in $[0,1]$.

We begin to see why this might be the case in the following theorem.

Theorem 2.8.

$$\mathbb{P}(X_n = k) = \frac{1}{n+1}, \qquad k = 1, 2, \ldots, n+1. \tag{2.26}$$

That is, $X_n \sim \text{Unif}(\{1, 2, \ldots, n+1\})$.

Proof. Let $B_n = X_n - 1$ be the number of black balls added to the urn after n steps. We wish to calculate $P(B_n = k)$, and observe that there are many ways to add k black balls in n steps, such as

$$\underbrace{B, B, \ldots, B}_{k}, \underbrace{R, R, \ldots, R}_{n-k} \quad \text{or} \quad B, R, \underbrace{B, B, \ldots, B}_{k-1}, \underbrace{R, R, \ldots, R}_{n-k-1}$$

$$\text{or} \quad \underbrace{R, R, \ldots, R}_{n-k}, \underbrace{B, B, \ldots, B}_{k}.$$

However, the surprising thing is that each of these $\binom{n}{k}$ ways all have the same probability. For example, for immediately adding k black balls, we have

$$\begin{aligned}\mathbb{P}(B, B, \ldots, B, R, R, \ldots, R) &= \underbrace{\frac{1}{2} \cdot \frac{2}{3} \cdots \frac{k}{k+1}}_{k \text{ black}} \cdot \underbrace{\frac{1}{k+2} \cdot \frac{2}{k+3} \cdots \frac{n-k}{n+1}}_{n-k \text{ red}} \\ &= \frac{k!(n-k)!}{(n+1)!}\end{aligned}$$

Exercise 2.11. *Show that we also have*

$$\mathbb{P}(B, R, \underbrace{B, B \ldots, B}_{k-1}, \underbrace{R, R, \ldots, R}_{n-k-1}) = \frac{k!(n-k)!}{(n+1)!} \quad \textit{and}$$
$$\mathbb{P}(R, R, \ldots, R, B, B, \ldots, B) = \frac{k!(n-k)!}{(n+1)!}.$$

Extrapolating from this pattern, we thus find, for $k = 0, 1, \ldots, n$,

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

which is independent of k. Thus, for $k = 1, 2, \ldots, n+1$,

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

as claimed in (2.26). □

We can now prove that the limiting proportion of black balls is uniform in $[0, 1]$.

Theorem 2.9. *The proportion $X_n/(n+2)$ of black balls converges in distribution to a* Unif$(0, 1)$ *random variable. That is, for every $x \in [0, 1]$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{X_n}{n+2} \leq x\right) = x. \tag{2.27}$$

Note that we can view this as a type of law of large numbers for the Pólya urn.

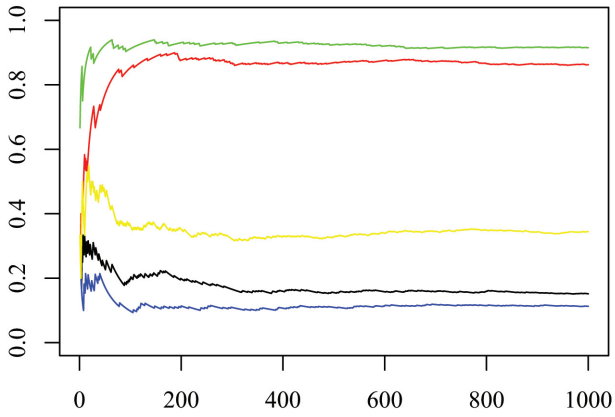

Figure 2.6: Five samples of 1000 steps of the Pólya urn, beginning with one black and one red ball. The y-axis is the proportion of black balls $\frac{X_n}{n+2}$. We numerically see that this proportion eventually stabilizes, limiting in a random proportion $P \in (0,1)$. Theorem 2.9 gives the distribution of P. Simulations by Celeste Zeng. For more simulations, the R code behind them and further analysis, visit her github page at `https://celeste-zeng.github.io/Polya_Urn/`.

Proof. We have by (2.26) that

$$\mathbb{P}\left(\frac{X_n}{n+2} \leq x\right) = \mathbb{P}(X_n \leq (n+2)x) = \frac{\lfloor (n+2)x \rfloor}{n+1}, \tag{2.28}$$

where $\lfloor (n+2)x \rfloor$ is the "floor" of $(n+2)x$, the greatest integer less than or equal to $(n+2)x$. For large values of n, we would like to replace the floor with $(n+2)x$ and say that it makes very little difference. Indeed, observe that

$$\begin{aligned}\frac{\lfloor (n+2)x \rfloor}{(n+2)x} &= \frac{(n+2)x - \big((n+2)x - \lfloor (n+2)x \rfloor\big)}{(n+2)x} \\ &= 1 - \frac{(n+2)x - \lfloor (n+2)x \rfloor}{(n+2)x} \to 1\end{aligned}$$

as $n \to \infty$ since $0 \leq |(n+2)x - \lfloor (n+2)x \rfloor| \leq 1$. Hence sending $n \to \infty$ in (2.28) is the same as

$$\lim_{n\to\infty} \frac{(n+2)x}{n+1} = x. \qquad \square$$

You will work with a more general Pólya urn scheme, starting with a black balls and b red balls, in Problem 2.6.

2.4.4 The Pólya urn and Bayesian statistics

Pólya urns arise naturally in Bayesian statistics. Consider the following simple statistical problem: we have a coin with unknown probability p of being heads, and we start flipping it. How do we estimate the unknown parameter p? If Y_n is the count of heads after n tosses, then a good estimate of p is $\widehat{p} := Y_n/n$, the proportion of heads.*

But what if p itself is random? In Bayesian statistics one assumes that p has its own specific distribution, the *prior belief*. If we assume that $p \sim \text{Unif}(0,1)$, then the sequence of the number of heads $(Y_1, Y_2, \ldots)$ turns out to be the same stochastic process as the count of added black balls $(B_1, B_2, \ldots)$ for the Pólya urn, another beautiful result.

Theorem 2.10. *Sample $P \sim \text{Unif}(0,1)$. Given that $P = p$, let I_n, $n = 1, 2, \ldots$, be iid Bernoulli random variables with probability of success p. Let $Y_n := \sum_{k=1}^{n} I_k$ be the number of successes after n trials. Then the process $(1 + Y_n, n = 1, 2, \ldots)$ is a Pólya urn starting with one black and one red ball.*

This result is a consequence of De Finetti's Theorem, which is sometimes called the "fundamental theorem of Bayesian statistics." We will tackle the proof from a more elementary angle, though, and will not invoke the De Finetti machinery.

Proof. Let $(X_n, n = 1, 2, \ldots)$ be the Pólya urn. Note that

$$\mathbb{P}(X_{n+1} = k+2 \,|\, X_n = k+1) = \frac{k+1}{n+2},$$

and thus we want to show

$$\begin{aligned} \mathbb{P}(Y_{n+1} = k+1 \,|\, Y_n = k) &= \mathbb{P}(1 + Y_{n+1} = k+2 \,|\, 1 + Y_n = k+1) \\ &= \frac{k+1}{n+2}. \end{aligned} \tag{2.29}$$

*Indeed, in the statistics terminology this is the *maximum likelihood estimator*, or the value of p which has highest probability of being correct, given Y_n is binomial.

In fact, this is all we need to show: since $Y_{n+1} - Y_n \in \{0, 1\}$, this would immediately imply

$$\mathbb{P}(Y_{n+1} = k \,|\, Y_n = k) = 1 - \frac{k+1}{n+2} = \mathbb{P}(X_{n+1} = k+1 \,|\, X_n = k+1),$$

completing the proof.

How do we obtain (2.29)? Start with the definition

$$\mathbb{P}(Y_{n+1} = k+1 \,|\, Y_n = k) = \frac{\mathbb{P}(Y_{n+1} = k+1, Y_n = k)}{\mathbb{P}(Y_n = k)}, \tag{2.30}$$

and note that $\mathbb{P}(Y_n = k)$ is obvious if we are given P:

$$\mathbb{P}(Y_n = k \,|\, P = p) = \binom{n}{k} p^k (1-p)^{n-k},$$

as $Y_n | P = p \sim \text{Bin}(n, p)$. The tower property of conditional expectations thus gives

$$\mathbb{P}(Y_n = k) = \mathbb{E}(\mathbb{P}(Y_n = k \,|\, P)) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp,$$

where the expectation is taken with respect to P, and we are integrating against the density $f_P(p) \equiv 1$ of P. We similarly have

$$\mathbb{P}(Y_{n+1} = k+1, Y_n = k) = \int_0^1 \binom{n}{k} p^{k+1} (1-p)^{n-k} dp.$$

Combining (2.29) and (2.30), we thus need to show

$$\frac{\int_0^1 \binom{n}{k} p^{k+1} (1-p)^{n-k} dp}{\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp} = \frac{k+1}{n+2}. \tag{2.31}$$

The pressing question, then, is how (on earth) to evaluate integrals like $\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp$. Two clever ideas come to the rescue: first, introduce a parameter $\lambda > 0$, and then note that

$$\sum_{k=0}^{n} \binom{n}{k} \lambda^k p^k (1-p)^{n-k} = (\lambda p + 1 - p)^n$$

by the binomial theorem. Now we have something that we can easily integrate on the right-hand side:

$$\sum_{k=0}^{n} \lambda^k \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \int_0^1 (\lambda p + 1 - p)^n dp$$
$$= \frac{1}{n+1} \frac{\lambda^{n+1} - 1}{\lambda - 1} = \frac{1}{n+1} \sum_{k=0}^{n} \lambda^k.$$

Since λ is arbitrary, the coefficients of λ^k on both sides must agree:

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \frac{1}{n+1} \tag{2.32}$$

for all $k = 0, 1, \ldots, n$. Hence we have the denominator of (2.31). But also the numerator, if we note

$$\int_0^1 \binom{n}{k} p^{k+1} (1-p)^{n-k} dp = \frac{\binom{n}{k}}{\binom{n+1}{k+1}} \int_0^1 \binom{n+1}{k+1} p^{k+1} (1-p)^{n+1-(k+1)} dp$$
$$= \frac{\binom{n}{k}}{\binom{n+1}{k+1}} \cdot \frac{1}{n+2}$$

by (2.32) with n replaced by $n+1$ and k by $k+1$. Therefore,

$$\frac{\int_0^1 \binom{n}{k} p^{k+1} (1-p)^{n-k} dp}{\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp} = \frac{\frac{\binom{n}{k}}{\binom{n+1}{k+1}} \cdot \frac{1}{n+2}}{\frac{1}{n+1}}$$
$$= \frac{n!}{k!(n-k)!} \cdot \frac{(k+1)!(n-k)!}{(n+1)!} \cdot \frac{n+1}{n+2}$$
$$= \frac{k+1}{n+2},$$

which was exactly our goal. □

Problems for chapter 2

Problem 2.1. Consider the random walk on $\{1,2,3,4,5\}$ with the following transition matrix:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

(*a*) Let $f_k = \mathbb{E}_k(\tau_4)$ be the expected hitting time of 4 starting at k. Find f_k for $k = 1,2,3,4,5$.

(*b*) Let τ be the first hitting time of $\{1,5\}$. Compute $\mathbb{E}_3(\tau)$.

Problem 2.2. Consider an urn with balls of three colors: red, blue, and green. The total number of balls is N. Run the following Markov chain on the state space

$$\Omega = \{(i,j,k) \in \mathbb{Z}^3 : \ i \geq 0, j \geq 0, k \geq 0, \ i+j+k = N\}.$$

Pick a ball at random. If it is red, replace it with a blue or a green ball with equal probability. Similarly, if it is blue, replace it with a red or a green with equal probability. Finally, if it is green, replace it with a red or a blue with equal probability. Find a reversible stationary distribution for this chain. (*Hint:* Think of a standard named distribution that is a good guess for π. Then verify the DBE to check for reversibility under π.)

Problem 2.3. Consider the Gambler's ruin problem for the biased random walk on the line. That is, consider the chain $(X_0, X_1, \ldots)$ on the integers $\mathbb{Z}$ where the transition probabilities are $P(j, j+1) = p$ and $P(j, j-1) = 1-p$, for some $0 < p < 1$, $p \neq 1/2$. Fix integers $0 < n$ and let $\tau = \tau_{\{0,n\}}$ be the hitting time of $\{0,n\}$.

(*a*) Show that

$$\mathbb{P}_k(X_\tau = n) = \frac{(q/p)^k - 1}{(q/p)^n - 1}, \quad \text{where } q = 1-p.$$

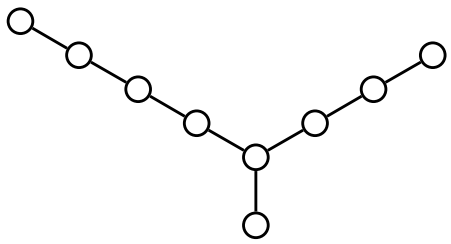

Figure 2.7: A star-like graph considered in problem 2.4.

(*b*) Show that

$$\lim_{p\to 1/2} \frac{(q/p)^k - 1}{(q/p)^n - 1} = \frac{k}{n},$$

and explain the significance of this limit in this context.

Problem 2.4. Consider the star graph that has one center vertex 0 that is connected to N arms of lengths $l_1, l_2, \ldots, l_N$, for positive integers l_i's. See Figure 2.7 for an example with $N = 3$, $l_1 = 3, l_2 = 4$ and $l_3 = 1$. Suppose you start a random walk on this graph starting at the center 0. What is the probability that it will reach the end of the first arm before it reaches the end of any other arm? Write it as a formula in $l_1, \ldots, l_N$. Your formula should give $\frac{4}{19}$ for the graph in Figure 2.7.

Problem 2.5. Consider the random walk on the n-cycle. Suppose the random walk starts at 1. We are interested in the expectation of the cover time τ_n which is the first time when the walk has visited every vertex of the cycle.

(*a*) Let τ_{n-1} be the first time when all but exactly one of the vertices of the n-cycle has been visited. That is, $X_{\tau_{n-1}}$ visits the last but one new vertex to be visited. Use Gambler's Ruin to show that

$$\mathbb{E}\left(\tau_n - \tau_{n-1}\right) = n - 1.$$

Hint: Look at the set of all already visited vertices at τ_{n-1}. Suppose $X_{\tau_{n-1}} = 5$ and vertex 6 is the last vertex to be yet visited. Then the random walk can either go to 6 from 5, or go all the way in the reverse direction and get to 6 from 7.

(*b*) Repeat the argument above to show that if τ_k is the stopping time which is the first time k vertices in the cycle have been visited, then

$$\mathbb{E}(\tau_k - \tau_{k-1}) = k - 1.$$

(*c*) Show that the expected cover time is given by the formula

$$\mathbb{E}(\tau_n) = \binom{n}{2}.$$

Problem 2.6. Fix $a, b \in \mathbb{N}$ and consider a Pólya urn starting with a black balls and b red balls. As usual, for each turn you pick a ball at random and then return it to the urn along with a ball with the same color. Let X_n denote the number of black balls in the urn after the nth turn is completed.

(*a*) Show that, for all $n = 1, 2, 3, \ldots$,

$$\mathbb{E}\left(\frac{X_n}{n+a+b} \mid X_{n-1}\right) = \frac{X_{n-1}}{n+a+b-1}.$$

(*b*) Use the above (or otherwise) to compute $\mathbb{E}(X_n)$ for every n. Your answer should only depend on a, b and n.

(*c*) Use the above (or otherwise) to show the amazing formula

$$\mathbb{P}(n\text{th pick is black}) = \frac{a}{a+b}, \quad \text{for all } n \geq 1.$$

(*d*) Show that, similarly,

$$\mathbb{P}(n\text{th pick is black} \mid \text{first two picks are black}) = \frac{a+2}{a+b+2}, \quad \text{for all } n \geq 3.$$

Problem 2.7. Imagine the queue in front of an ice cream shop. Suppose there are currently k people standing in the queue. In the next time period, the first person in the queue gets his/her ice cream and leaves with probability 1/2. Independently, another person joins the queue with probability 1/2. If the queue is currently empty, there is 1/2 probability that someone will join the queue in the next time period. Suppose the shop can only ac-

commodate at most a queue of N people after which no more new customers are allowed to join the queue until someone leaves.

(a) Let X_n be the number of people standing in the queue at time n. This is a Birth-and-Death chain. Find its state space and transition probabilities. Is it irreducible? Aperiodic?

(b) What is the probability that starting with a k person queue, the shop will become full before it completely empties?

(c) Find the stationary distribution for this chain.

(d) Suppose $N = \infty$, i.e., there is no upper bound on the length of the queue. Consider a lazy random walk $\{S_k,\ k = 0, 1, 2, 3, \ldots\}$ on the integers with transition probabilities

$$p(j, j+1) = \frac{1}{4},\ p(j, j-1) = \frac{1}{4},\ p(j, j) = \frac{1}{2}.$$

Show that the absolute value process $\{|S_k|,\ k = 0, 1, 2, 3, \ldots\}$ is the same Birth-and-Death chain as X by find its transition probabilities.

(e) Find the expected time it will take from an empty queue (0 people) to become full (N people) for the first time.

Problem 2.8. Consider a sequence of coin flips with $\mathbb{P}(H) = p$, for some $0 < p < 1$. Define a Markov chain with state space $\{TT, TH, HT, HH\}$ which shows the outcome of the previous toss and the current one. For example, $X_n = TH$ means that the $(n-1)$th toss was T and the nth toss was H.

(a) Start flipping the coin. How many tosses, on average, does it take to get two heads in a row? Use Markov chain methods.

(b) As we flip, how many times do we expect to see HT before getting two heads in a row?

Chapter 3

Asymptotic behavior of Markov chains

3.1 Asymptotics of Markov chains

In many of the problems in Chapter 1 you were asked to compute high powers of a transition matrix. The point was to get some intuition for the long-term behavior of the chain. You probably noticed that often, although not always, these became close to the matrix where every row was the stationary distribution π.

In this chapter we make this observation rigorous, and give conditions on the chain for when we can ensure that each row converges to π. The conditions are simple enough, and perhaps you have already guessed them: the chain must be irreducible and aperiodic. Our first main theorem in this chapter, Theorem 3.4, says that such chains converge to their stationary distribution. More than that, it says that the convergence occurs *exponentially fast*. That is why we can often see the behavior of the limit as $n \to \infty$ after only 50 steps or so, if not sooner.

What does it mean, however, for a Markov chain to *converge* to a given distribution? We need some way of describing how far two probability distributions are from each other to talk about convergence. We develop such a measurement, the *total variance distance*, in §3.1.2.

And what if our chain is *not* aperiodic? Recall the example of the 6-cycle, which has period 2. We saw in Example 1.2 that the 50th and 51st

powers of the transition matrix are

$$P^{50} \approx \begin{bmatrix} 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \end{bmatrix}$$

and

$$P^{51} \approx \begin{bmatrix} 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \\ 0 & 0.333 & 0 & 0.333 & 0 & 0.333 \\ 0.333 & 0 & 0.333 & 0 & 0.333 & 0 \end{bmatrix}.$$

There is no hope for a limiting distribution here over *all* steps, because on each even step the chain can only be at half of the vertices, and on each odd step it can only be on the other half. So there is no limit. (This is analogous to the fact that the sequence $\{(-1)^n\}_{n=1}^{\infty}$ has no limit as $n \to \infty$.) We have a simple work-around to deal with this, though, which you have also already encountered in the problems for chapter 1 (Problem 1.8). We begin in §3.1.1 by showing the process of "lazification" turns a periodic chain into an aperiodic one, without altering the stationary distribution. Hence by our upcoming convergence theorem, the lazy version of the walk will still converge to stationarity.

3.1.1 Lazy random walks

Definition 3.1. For a Markov chain X with $n \times n$ transition matrix P, the **lazy version of** X is the Markov chain $\tilde{X}$ with transition matrix

$$\tilde{P} := \frac{1}{2}(I_n + P), \tag{3.1}$$

where I_n is the $n \times n$ identity matrix.

Example 3.1. Let's think about the 6-cycle again, whose transition matrix

is

$$P = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 \end{bmatrix},$$

and hence whose lazy version has transition matrix

$$\tilde{P} = \begin{bmatrix} 1/2 & 1/4 & 0 & 0 & 0 & 1/4 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/2 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 & 1/2 & 1/4 \\ 1/4 & 0 & 0 & 0 & 1/4 & 1/2 \end{bmatrix}.$$

What has happened? You can see the effect of adding the identity matrix is to make each diagonal entry $\tilde{P}_{jj} > 0$. That is, now we have the possibility of remaining at the same state for a unit of time; loops are added to the graph. We could say that we now have to flip a coin at the start of each step. Heads, we stay at the same vertex. Tails, we move according to the original transition probabilities.

The lazy walk has the following properties. The key properties are (*iii*) and (*iv*): our lazy walk is *always* aperiodic, which is obvious from the loops, and further we have not disrupted the stationary distribution.

Lemma 3.1. *The lazy random walk defined by* (3.1) *satisfies the following.*

(*i*) $\tilde{P}$ *is a transition matrix.*

(*ii*) *If the original chain* X *is irreducible, so is the lazy chain* $\tilde{X}$*.*

(*iii*) $\tilde{X}$ *is aperiodic.*

(*iv*) *If* π *is stationary for* P*,* π *is also stationary for* $\tilde{P}$*.*

Proof. For (*i*), all the entries of $\tilde{P}$ are obviously non-negative. So we just need the rows to sum to one. Setting $\mathbf{1} := (1, 1, \ldots, 1)$, the row vector of all

ones, we have

$$\tilde{P}\mathbf{1}^T = (\frac{1}{2}I + \frac{1}{2}P)\mathbf{1}^T = \frac{1}{2}\mathbf{1}^T + \frac{1}{2}P\mathbf{1}^T = \frac{1}{2}\mathbf{1}^T + \frac{1}{2}\mathbf{1}^T = \mathbf{1}^T,$$

as needed.

For (ii), we have added loops to the chain, which does not disconnect it.

To see (iii), observe that we can always return to a given state in one step. □

It is a good (and simple) exercise for you to prove part (iv).

Exercise 3.1. *Suppose P has stationary distribution π. Show that π is also stationary for $\tilde{P}$.*

How do hitting times change under $\tilde{P}$? If you have to flip a coin and get heads to be allowed to move each step, how would that affect the number of steps you need to reach $A \subset \Omega$? The following theorem answers this, but we encourage you to guess for yourself before reading it. What do you intuitively think?

Theorem 3.2. *Let $A \subset \Omega$, and let τ_A be the hitting time of A. Let $f_k := \mathbb{E}_k(\tau_A)$ be the expected hitting time of A under P starting from k, and $\tilde{f}_k := \tilde{\mathbb{E}}_k(\tau_A)$ the corresponding expectation under $\tilde{P}$. Then*

$$\tilde{f}_k = 2f_k \quad \text{for all } k.$$

Proof. Clearly $f_k = 0 = 2\tilde{f}_k$ for $k \in A$. For $k \notin A$, we examine the recursions satisfied by f_k and $\tilde{f}_k$. Similar to what we saw for Gambler's ruin in §2.1.2,

$$f_k = 1 + \sum_{j=1}^{n} p_{kj} f_j$$

for each k. Subtract the $\sum_{j=1}^{n} p_{kj} f_j$ term over to the left, and note that we can collect all the equations together in matrix form as

$$(I - P_{\Omega\setminus A})\mathbf{f}^T = \mathbf{1}^T. \tag{3.2}$$

Here the $\Omega\setminus A$ subscript means that we only include the rows and columns not corresponding to elements of A. Similarly, $\mathbf{f}$ is the row vector whose

components are the f_k, but where we leave out entries k corresponding to A. Similarly, $\tilde{f}_k = 1 + \frac{1}{2}\tilde{f}_k + \sum_{j=1}^{n} \frac{1}{2} p_{kj} \tilde{f}_j$, or equivalently

$$(I - P_{\Omega \setminus A}) \frac{1}{2} \tilde{\mathbf{f}}^T = \mathbf{1}^T \tag{3.3}$$

Comparing (3.2) and (3.3), we see that we have two solutions to the same linear system. One can argue from linear algebra that since P is a stochastic matrix (all its rows sum to one), $I - P_{\Omega \setminus A}$ is invertible. Therefore, there is a unique solution to this linear system, forcing $\mathbf{f} = \frac{1}{2}\tilde{\mathbf{f}}$. This says $f_k = \frac{1}{2}\tilde{f}_k$ for each $k \notin A$, as we needed to show. □

3.1.2 Total variation distance of probability distributions

We wish to say that the probability distribution

$$P_{xk}^n := \mathbb{P}(X_n = k \mid X_0 = x), \qquad k \in \{1, 2, \ldots, N\}$$

of the location of our chain after n steps converges to some fixed probability distribution as $n \to \infty$. However, we currently lack a precise meaning for the *convergence* of a probability distribution. We address this by introducing the *total variation norm*, which gives us a way to measure the "distance" between two probability distributions. Once we have this, we can say that a sequence of probability distributions $\{p_n\}$ converges to a distribution q if the total variation between the p_n and q goes to zero.

Definition 3.2. Let $\Omega = \{1, 2, \ldots, n\}$ be a sample space with probability mass functions p and q on Ω. That is, p and q are functions from Ω to $[0, 1]$ with $1 = \sum_{k=1}^{n} p_k = \sum_{k=1}^{n} q_k$. The **total variation (TV) distance** between p and q is

$$\|p - q\|_{TV} := \frac{1}{2} \sum_{k=1}^{n} |p_k - q_k|. \tag{3.4}$$

Here $p_k = p(k)$ and $q_k = q(k)$ are, of course, the values p and q give to vertex $k \in \Omega$. We note that we will interchangeably use "TV-distance" and "TV-norm" for $\|p - q\|_{TV}$, although "distance" is the more precise term.

The total variance distance satisfies some nice properties. In particular, part (iv) of the following lemma gives us some nice intuition: the TV-

distance is the biggest discrepancy we can see when comparing the probability of the same event under both distributions.

Lemma 3.3. *Let p, q and r be probability distributions on Ω. The total variation norm has the following properties:*

(*i*) *It is a* metric. *That is, $\|p-q\|_{TV} \geq 0$, $\|p-q\|_{TV} = \|q-p\|_{TV}$, and*

$$\|p-q\|_{TV} \leq \|p-q\|_{TV} + \|q-r\|_{TV}\,.$$

(*ii*) $\|p-q\|_{TV} \leq 1$.

(*iii*) $|p_k - q_k| \leq 2\,\|p-q\|_{TV}$ *for all* k.

(*iv*)

$$\|p-q\|_{TV} = \max_{A\in 2^\Omega} |\mathbb{P}_p(A) - \mathbb{P}_q(A)|, \tag{3.5}$$

where $\mathbb{P}_p(A)$ is the probability of the set A under the distribution p, $\mathbb{P}_q(A)$ its probability under q, and 2^Ω the powerset of Ω.

Proof. For (*i*), the first two properties are clear from the definition (3.4). The third property follows from the fact that $|p_k - r_k| \leq |p_k - q_k| + |q_k + r_k|$, the familiar *triangle inequality* for absolute values.

For (*ii*), set

$$B := \{k \in \Omega \,:\, p_k > q_k\}. \tag{3.6}$$

Then $B^c = \{k \in \Omega \,:\, p_k \leq q_k\}$, and we observe

$$\begin{aligned}
\|p-q\|_{TV} &= \frac{1}{2}\sum_{k\in B} |p_k - q_k| + \frac{1}{2}\sum_{k\in B^c} |p_k - q_k| \\
&= \frac{1}{2}\sum_{k\in B} (p_k - q_k) - \frac{1}{2}\sum_{k\in B^c} (p_k - q_k) \qquad (3.7)\\
&= \frac{1}{2}(\mathbb{P}_p(B) - \mathbb{P}_q(B)) - \frac{1}{2}(\mathbb{P}_p(B^c) - \mathbb{P}_q(B^c)) \\
&= \frac{1}{2}(\mathbb{P}_p(B) - \mathbb{P}_q(B)) - \frac{1}{2}(1 - \mathbb{P}_p(B) - 1 + \mathbb{P}_q(B)) \\
&= \mathbb{P}_p(B) - \mathbb{P}_q(B) \qquad (3.8)
\end{aligned}$$

$$\leq \mathbb{P}_p(B) \leq 1.$$

We (iii), we note

$$|p_k - q_k| \leq \sum_{j=1}^{n} |p_j - q_j| \leq 2 \|p - q\|_{TV}.$$

Lastly, for (iv), consider the set B in (3.6) and note

$$\mathbb{P}_p(B) - \mathbb{P}_q(B) = \frac{1}{2} \sum_{k \in B} (p_k - q_k) \geq 0$$

by definition of B, and so $\mathbb{P}_p(B) - \mathbb{P}_q(B) = |\mathbb{P}_p(B) - \mathbb{P}_q(B)|$. Hence by (3.8),

$$\|p - q\|_{TV} = |\mathbb{P}_p(B) - \mathbb{P}_q(B)| \leq \max_{A \subset 2^{\Omega}} |\mathbb{P}_p(A) - \mathbb{P}_q(A)|. \tag{3.9}$$

If we can also show the reverse inequality

$$\|p - q\|_{TV} \geq \max_{A \subset 2^{\Omega}} |\mathbb{P}_p(A) - \mathbb{P}_q(A)|,$$

our desired equality (3.5) follows. To that end, fix any set $D \subset 2^{\Omega}$. We first show that

$$|\mathbb{P}_p(D) - \mathbb{P}_q(D)| \leq \|p - q\|_{TV}. \tag{3.10}$$

Using $\mathbb{P}(D) = \mathbb{P}(D \cap B) + \mathbb{P}(D \cap B^c)$, we see

$$\begin{aligned}
\mathbb{P}_p(D) - \mathbb{P}_q(D) &= \mathbb{P}_p(D \cap B) - \mathbb{P}_q(D \cap B) + \mathbb{P}_p(D \cap B^c) - \mathbb{P}_q(D \cap B^c) \\
&\leq \mathbb{P}_p(B) - \mathbb{P}_q(B) + \mathbb{P}_p(D \cap B^c) - \mathbb{P}_q(D \cap B^c) && (3.11) \\
&\leq \mathbb{P}_p(B) - \mathbb{P}_q(B) && (3.12) \\
&= \|p - q\|_{TV}, && (3.13)
\end{aligned}$$

where (3.11) holds since

$$\mathbb{P}_p(D \cap B) - \mathbb{P}_q(D \cap B) = \sum_{k \in D \cap B} (p_k - q_k) \leq \sum_{k \in B} (p_k - q_k) = \mathbb{P}_p(B) - \mathbb{P}_q(B)$$

since $p_k - q_k \geq 0$ for $k \in B$. Similarly, by definition of B^c, $\mathbb{P}_p(D \cap B^c) -$

$\mathbb{P}_q(D \cap B^c) \leq 0$, and so dropping it from (3.11) only increases the value, yielding (3.12).

We would like this inequality with an absolute value, though, as in (3.10). So we also need to show that $-(\mathbb{P}_p(D) - \mathbb{P}_q(D)) \leq \|p - q\|_{TV}$. To do this, we reverse the roles of p and q and define the set C via

$$C := \{k \in \Omega \ : \ p_k < q_k\}, \qquad C^c = \{k \in \Omega \ : \ p_k \geq q_k\}.$$

Compare (3.6). Since we have $\|p - q\|_{TV} = \|q - p\|_{TV}$, we can simply repeat the equations beginning at (3.7) with all the p's and q's reversed to conclude that

$$\|q - p\|_{TV} = \mathbb{P}_q(C) - \mathbb{P}_p(C).$$

(Work out the details if you need to convince yourself.) In particular,

$$\mathbb{P}_q(C) - \mathbb{P}_p(C) = \|p - q\|_{TV}\,.$$

Now, just repeat the argument above to obtain

$$\begin{aligned} \mathbb{P}_q(D) - \mathbb{P}_p(D) &= \mathbb{P}_q(D \cap C) - \mathbb{P}_p(D \cap C) + \mathbb{P}_q(D \cap C^c) - \mathbb{P}_p(D \cap C^c) \\ &\leq \mathbb{P}_q(C) - \mathbb{P}_p(C) \\ &= \|p - q\|_{TV}\,. \end{aligned}$$

Combined with (3.13), we thus have (3.10). Since D was an arbitrary set, the inequality persists in the max,

$$\max_{A \in 2^\Omega} |\mathbb{P}_p(A) - \mathbb{P}_q(A)| \leq \|p - q\|_{TV}\,.$$

Combined with (3.9), this gives (3.5), as claimed. □

While part (*iv*) of the lemma gives some nice intuition, the TV-norm can still seem abstract. Let's do an explicit computation to see how things can work out in practice.

Example 3.2. How close to being uniformly distributed is a walk on the 5-cycle after three steps? That is, suppose we start a simple random walk on the 5-cycle at vertex 1, say. What is the TV-norm between the resulting

distribution p after three steps from the uniform distribution

$$q := \Big(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\Big)?$$

To answer this, we recall the transition matrix P for the 5-cycle is

$$P = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix},$$

and so the three-step transition matrix is

$$P^3 = \begin{bmatrix} 0 & \frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} \\ \frac{3}{8} & 0 & \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{3}{8} & 0 & \frac{3}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{3}{8} & 0 & \frac{3}{8} \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} & 0 \end{bmatrix},$$

which shows that the distribution p_3 of X_3, given that $X_0 = 1$, is the first row of P^3,

$$p_3 = \Big(0, \frac{3}{8}, \frac{1}{8}, \frac{1}{8}, \frac{3}{8}\Big).$$

Using the definition (3.4), we have that the TV-norm to the uniform distribution is thus

$$\|p_3 - q\|_{TV} = \frac{1}{2}\Big(\Big|0 - \frac{1}{5}\Big| + \Big|\frac{3}{8} - \frac{1}{5}\Big| + \Big|\frac{1}{8} - \frac{1}{5}\Big| + \Big|\frac{1}{8} - \frac{1}{5}\Big| + \Big|\frac{3}{8} - \frac{1}{5}\Big|\Big) = \frac{7}{20}.$$

Note that we have $\|p_3 - q\|_{TV} \leq 1$, as must be the case from Lemma 3.3(ii). By part (iv) of the lemma, we also know that the difference between the probability that X_3 is in any collection B of vertices and the probability that a uniformly-chosen vertex Y is in the same collection B is at most $7/20 = 0.35$. So, for instance, for the set of vertices $B = \{1, 3, 5\}$, we see

$$|\mathbb{P}(X_3 \in B \,|\, X_0 = 1) - \mathbb{P}(Y \in B)| \leq \max_{A \subset V} |\mathbb{P}(X_3 \in A \,|\, X_0 = 1) - \mathbb{P}(Y \in A)|$$

$$= \|p_3 - q\|_{TV} = 0.35,$$

where the first equality is by (3.5). Equivalently,

$$\mathbb{P}(Y \in B) - 0.35 \leq \mathbb{P}(X_3 \in B \,|\, X_0 = 1) \leq \mathbb{P}(Y \in B) + 0.35.$$

So beyond giving us a way to measure distance between two probability distributions, we also see part of the power of the TV-norm is that it gives an upper bound on the difference in probability of the same event under the two distributions.

We will see below in Theorem 3.4 that as $n \to \infty$, the distribution p_n of X_n after n steps converges to the uniform distribution q on the vertices.

Excursus: other norms on finite probability distributions

As an aside, we note two other details that you should be aware of to be a well-educated human but that we will not be using. First, there are other ways that we could assign a distance between p and q. If we think of $p = (p_1, \ldots, p_N)$ and $q = (q_1, \ldots, q_N)$ as vectors in $\mathbb{R}^N$, we could take the Euclidean distance

$$\|p - q\|_2 = \Big(\sum_{k=1}^{N} (p_k - q_k)^2 \Big)^{1/2}, \tag{3.14}$$

which is how far these vectors are from each other in N-dimensional space. The TV-distance (3.4) is not identical and will not give the same number, in general. While the Euclidean norm is more geometrically natural, the TV-norm is probabilistically more suitable, as we get properties like Lemma 3.3 (iv), which do not hold for the Euclidean norm.

We could also replace the "2" in (3.14) with any power $r \geq 1$ to obtain $\|p - q\|_r$, with $r = 1$ corresponding to the TV-distance up to the factor of $1/2$ in front. All of these give distinct measurements of distance between p and q, but there is a sense in which they are all "comparable" to the TV-distance. Precisely, for any $r \geq 1$ there exists a constant $C = C(r)$ such that

$$\frac{1}{C} \|p - q\|_r \leq \|p - q\|_{TV} \leq C \|p - q\|_r$$

for any $p, q \in \mathbb{R}^N$. In particular, this inequality shows that if $\|p_n - q\|_{TV} \to 0$ as $n \to \infty$ if and only if $\|p_n - q\|_r \to 0$ for any $r \geq 1$. So while we choose the TV-distance for its nice properties, convergence in TV-distance is equivalent to convergence in any other natural sense of "distance" between finite probability distributions.

3.1.3 The first key convergence theorem: exponential convergence to π

For a chain with transition matrix P, consider

$$P^n_{xk} = \mathbb{P}(X_n = k \,|\, X_0 = x),$$

the probability distribution on Ω generated by the position of the Markov chain after n steps (note that this is also the (x,k)-entry of the matrix P^n). For a fixed n and a fixed x, this is a pdf p on Ω, which we will write as

$$P^n(x, \cdot).$$

The "dot" notation means we can plug in any element of Ω and get the probability we are there after n steps, starting from x. Note that if $\Omega = \{1, 2, \ldots, N\}$, then collecting these probabilities give the vector

$$P^n(x, \cdot) = (P^n(x,1), P^n(x,2), \ldots, P^n(x,N)) \in \mathbb{R}^N.$$

For instance, we saw in Example 3.2 that for the simple walk on a 5-cycle started from the first vertex,

$$P^3(x, \cdot) = p = \Big(0, \frac{3}{8}, \frac{1}{8}, \frac{1}{8}, \frac{3}{8}\Big).$$

Now that we have the total variation norm we can talk about the limit $P^n(x, \cdot)$ as $n \to \infty$, because we have a way to talk about how far probability distributions are from each other. The following theorem says that the limiting distribution is the stationary distribution π. Recall from Theorem 1.6 that irreducible chains have a unique stationary distribution.

Theorem 3.4. *Let X be an irreducible and aperiodic Markov chain with stationary distribution π. Then there exists constants $C > 0$ and $0 < \alpha < 1$*

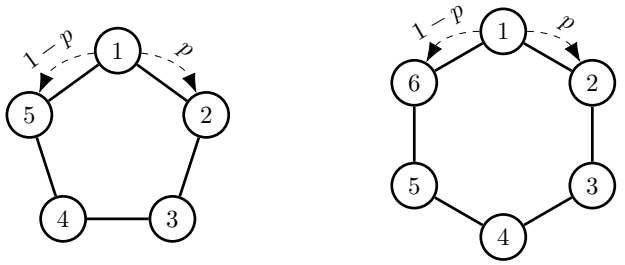

Figure 3.1: A random walk on a 5-cycle and a 6-cycle. Even though the transition probabilities at any vertex are the same, only one has a distribution which converges to stationarity.

such that, for any initial state $x \in \Omega$,

$$\|P^n(x,\cdot) - \pi\|_{TV} \leq C\alpha^n. \tag{3.15}$$

Furthermore, as the right-hand side of (3.15) *is independent of* x*, we have*

$$\max_{x\in\Omega} \|P^n(x,\cdot) - \pi\|_{TV} \leq C\alpha^n. \tag{3.16}$$

For instance, if $\alpha = 1/2$ and $x = 2$ then this says

$$0 \leq \|P^n(2,\cdot) - \pi\|_{TV} \leq \frac{C}{2^n} \to 0,$$

which is exponentially fast. So, Theorem 3.4 is saying that *irreducible and aperiodic Markov chains converge exponentially fast to their stationary distributions, regardless of where they are started.* While the proof of this theorem is beyond the scope of the text, we will spend a while trying to digest and understand it.

A first comment is that we should not be surprised about the hypotheses. Stationary distributions only make sense for irreducible chains, and so that assumption is natural. Let's again consider why aperiodicity is also vital.

Example 3.3. Consider a biased walk X_n on the vertices of a 5-cycle $\Omega_1 = \{1, 2, \ldots, 5\}$ and another biased walk Y_n on the vertices of a 6-cycle $\Omega_2 = \{1, 2, \ldots, 6\}$, as in Figure 3.1. The stationary distributions for these walks

are both uniform,

$$\pi_1 = (1/5, 1/5, \ldots, 1/5) \quad \text{and} \quad \pi_2 = (1/6, 1/6, \ldots, 1/6).$$

Since Y_n has period 2, for a given n we always have that $Y_n \in \{1, 3, 5\}$ or $Y_n \in \{2, 4, 6\}$, and so there is no possibility that $\|P_Y^n(x, \cdot) - \pi_2\|_{TV} \to 0$. We thus see that the aperiodicity assumption is necessary.

On the other hand, the X_n walk fulfills the assumptions of Theorem 3.4, and the theorem says that its distribution after n steps converges exponentially fast to π_1. For instance,

$$\|P_X^n(5, \cdot) - \pi_1\|_{TV} \leq C\alpha^n. \tag{3.17}$$

We considered the unbiased example $p = 1/2$ in Example 3.2, and saw that $\|P_X^3(1, \cdot) - \pi_1\|_{TV} = 0.35$ (recall the biased and unbiased walks have the same stationary distribution, the uniform distribution). The inequality (3.17) says that this quickly gets *much* smaller as n increases. For example, using a computer we find

$$P^{30}(1, \cdot) \approx (0.2007, 0.1994, 0.2002, 0.2002, 0.1994),$$

and thus using (3.4) we compute

$$\|P_X^{30}(1, \cdot) - \pi_1\|_{TV} \approx 0.0011.$$

That is, we are extremely close to uniform after just 30 steps.

Even though the periodic 6-cycle example shows Theorem 3.4 cannot possibly hold for chains with period ≥ 2, we recall from Section 3.1.1 that this is not insurmountable. We may simply form the corresponding lazy chain $\tilde{P} = \frac{1}{2}I + \frac{1}{2}P$ which is aperiodic and has the same stationary distribution π as P (recall Exercise 3.1). Theorem 3.4 then applies to $\tilde{P}$ and says

$$\max_{x \in \Omega} \|\tilde{P}_X^n(x, \cdot) - \pi\|_{TV} \leq C\alpha^n.$$

Exercise 3.2. (*a*) *Form the lazy walk from the walk on the 6-cycle in Figure 3.1. Update the diagram with the new transition probabilities.*

(*b*) *For $p = 0.3$, write the corresponding transition matrix $\tilde{P}$ and use a computer to find $\tilde{P}^{25}, \tilde{P}^{40}$ and $\tilde{P}^{55}$. What do you observe? How does this contrast with what would be happening for P^{25}, P^{40} and P^{55}?*

While Theorem 3.4 gives us exponential converge to π, it does not say anything about the precise rate, or what α is in (3.15). We will take up that question when we discuss *mixing times* in Section 3.2 below; see Theorem 3.8.

Two perspectives on exponential convergence

To enrich our understanding and intuition for Theorem 3.4, let's consider it from both a linear algebra and a probabilistic point of view, thinking in terms of *functions* f on our state space $\Omega = \{1, 2, \ldots, N\}$. This will also provide a natural bridge to our next convergence theorem.

Up to this point we have not thought very much about functions f on Ω, but note that there is an exact correspondence between such functions and vectors $\mathbf{f} \in \mathbb{R}^N$. Indeed, a function $f : \Omega \to \mathbb{R}$ just assigns a number $f(i)$ to each vertex $i \in \Omega$, and so we can list these as a vector $\mathbf{f} = (f(1), f(2), \ldots, f(N)) \in \mathbb{R}^N$. Conversely, a row vector $\mathbf{f}$ in $\mathbb{R}^N$ defines a function on Ω by considering the value at state $i \in \Omega$ to be the ith component $\mathbf{f}_i$. So functions f on Ω and vectors $\mathbf{f} \in \mathbb{R}^N$ are equivalent.

Now let's consider Theorem 3.4 in terms of linear algebra and the product $P^n\mathbf{f}^T$, which is the $N \times N$ matrix P^n with the column vector $\mathbf{f}^T$ (recall that by default our vectors are row vectors, like π). Theorem 3.4 says that, in particular, *each row* of P^n converges to π,

$$P^n = \begin{bmatrix} P^n(1,\cdot) \\ P^n(2,\cdot) \\ \vdots \\ P^n(n,\cdot) \end{bmatrix} \underset{n\to\infty}{\longrightarrow} \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix} = \mathbf{1}^T\pi. \tag{3.18}$$

Returning to the 5-cycle, for instance, we see that

$$P^{30} = \begin{bmatrix} 0.2007 & 0.1994 & 0.2002 & 0.2002 & 0.1994 \\ 0.1994 & 0.2007 & 0.1994 & 0.2002 & 0.2002 \\ 0.2002 & 0.1994 & 0.2007 & 0.1994 & 0.2002 \\ 0.2002 & 0.2002 & 0.1994 & 0.2007 & 0.1994 \\ 0.1994 & 0.2002 & 0.2002 & 0.1994 & 0.2007 \end{bmatrix},$$

and each row is manifestly close to $\pi = (0.2, 0.2, 0.2, 0.2, 0.2)$. Given (3.18), we have the convergence

$$P^n \mathbf{f}^T \to \mathbf{1}^T \pi \mathbf{f}^T.$$

So, after many steps, multiplying P^n by $\mathbf{f}^T$ is very close to multiplying $\mathbf{f}^T$ by the matrix where every row is π. Looking at this row by row, we have

$$\sum_{j=1}^{N} P^n_{xj} f_j \to \sum_{j=1}^{N} \pi_j f_j. \tag{3.19}$$

for each $x \in \Omega$.

Secondly, let's re-cast this in terms of probability. Given a function $f : \Omega \to \mathbb{R}$, suppose we wish to find $\mathbb{E}_x(f(X_n))$. That is, we start from x and run the chain for a large number n of steps, and then evaluate $f(X_n)$. What is this value, on average? By definition,

$$\mathbb{E}_x(f(X_n)) = \sum_{j=1}^{N} f(j)\, \mathbb{P}_x(X_n = j) = \sum_{j=1}^{N} P^n_{xj} f(j) \underset{n\to\infty}{\longrightarrow} \sum_{j=1}^{N} \pi_j f(j)$$

as we noted above in (3.19). However, we also have that

$$\sum_{j=1}^{N} \pi_j f(j) = \mathbb{E}_\pi(f(X_0)), \tag{3.20}$$

the expectation of $f(X_0)$ when X_0 is distributed on Ω according to π. We could also simply write this as $\mathbb{E}_\pi(f)$. Hence we see Theorem 3.4 implies

$$\lim_{n\to\infty} \mathbb{E}_x(f(X_n)) = \mathbb{E}_\pi(f). \tag{3.21}$$

In words, the average of a function evaluated after many steps is basically its average against the stationary distribution. This should be intuitively obvious, because the theorem says the distribution of X_n is very close to π after many steps.

Seeing things in this light leads us to our second main convergence theorem. What if, instead of $\mathbb{E}_x(f(X_n))$, we averaged over all the values $f(X_1), f(X_2), \ldots, f(X_n)$ up to the current state? In other words, can we say anything about

$$\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)? \tag{3.22}$$

Exercise 3.3. *Explain in words how the average $\mathbb{E}_x(f(X_n))$ in the left-hand side of* (3.21) *differs from the average in* (3.22)*.*

3.1.4 The second key convergence theorem: ergodicity

Our next main theorem says that we can compute (3.22), asymptotically, and that we get the same answer as in (3.21).

Theorem 3.5 (Ergodic Theorem)**.** *Let X be an irreducible Markov chain on Ω and $f : \Omega \to \mathbb{R}$ a function. Then, with probability 1,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f(X_j) = \mathbb{E}_\pi(f), \tag{3.23}$$

where $\mathbb{E}_\pi(f) = \mathbb{E}_\pi(f(X_0))$ is as in (3.20)*.*

Let's start by making some remarks and parsing what this is saying.

- First, note that the average on the left is *temporal*, the average value the function takes over all the states visited through n steps. The average on the right is *spatial* - the average of the function over all the states in Ω. So the theorem says that the temporal averages converge to the π-spatial average,

$$\lim_{n\to\infty} \underbrace{\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)}_{\textit{temporal average}} = \underbrace{\sum_{j=1}^{N} \pi_j f(j)}_{\textit{spatial average}}$$

- Secondly, observe the connection with the Law of Large Numbers, which states that an iid sequence $Y_0, Y_1, \ldots$ with $\mathbb{E}(Y_j) = \mu$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} Y_j = \mu \tag{3.24}$$

  with probability 1. The similarity with (3.23) is striking; on both left-hand sides we have a temporal average, and on both right-hand sides we have a fixed mean. We can therefore regard the Ergodic Theorem as a generalized "law of large numbers" for irreducible Markov chains.

- Lastly, it is worth pointing out a contrast with Theorem 3.4. While that theorem required the chain to be both irreducible and aperiodic, Theorem 3.5 only requires irreducibility. So the walk on the 6-cycle, for instance, satisfies this new averaging limit (3.23), even though it does not satisfy the exponential convergence of (3.15).

Example 3.4. One of the things we can do with the ergodic theorem is compute the long-term proportion of time that we spend on a given vertex or collection of vertices. For instance, for the simple walk X_j on the 6-cycle $\Omega = \{1, 2, \ldots, 6\}$, what is the long-term proportion of time we spend on vertex 2?

Exercise 3.4. *Intuitively reason out what the answer should be before reading further.*

Let $f(x) = \mathbb{1}_2(x)$, the function that gives a zero if we don't feed it the 2nd vertex, but gives 1 when we do. This is the **indicator function** for state 2. In vector form $\mathbf{f} = (f_j)_{j=1}^6 = (0, 1, 0, 0, 0, 0)$. Again, X_n is not aperiodic, but since it is irreducible Theorem 3.5 says

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(X_j) &= \lim_{n\to\infty} \frac{\#(\text{visits to vertex 2 in steps } 1, 2, \ldots, n)}{n} \\
&= \lim_{n\to\infty} (\text{Proportion of time } X_j \text{ spends on vertex 2}) \\
&= \mathbb{E}_\pi f = \sum_{k=1}^{6} f_k \pi_k = \pi_2 = \frac{1}{6}.
\end{aligned}$$

This is, of course, exactly what we would expect: in the long run, the simple walk should spend an equal proportion of time at any vertex. The Ergodic Theorem gives us a way to rigorously show this.

Exercise 3.5. *Suppose instead we have a biased walk X_j on the 6-cycle, where the probabilities of CW and CCW steps are 0.9 and 0.1, respectively. What is the long-term proportion of time that X_j spends on vertex 2?*

Exercise 3.6. *Consider a simple random walk Y_j on the vertices of the 13-cycle $\Omega := \{1, 2, \ldots, 13\}$. What is the long-term proportion of time that X_j spends on vertices 5, 6 and 13?*

3.1.5 Proof of the Ergodic Theorem

We need to show (3.23) holds for *all* functions $f : \Omega \to \mathbb{R}$, and we claim that it suffices to show it for the indicator functions

$$\mathbb{1}_i(x) = \begin{cases} 1 & x = i, \\ 0 & x \neq i, \end{cases} \qquad i = 1, 2, \ldots, N$$

Indeed, note that we can write any function $f : \Omega \to \mathbb{R}$ as

$$f(x) = \sum_{i=1}^{N} f(i)\mathbb{1}_i(x), \tag{3.25}$$

as is easy to see by evaluating both sides at any $j \in \Omega$. So, suppose for the moment that we have proved

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}_i(X_j) = \mathbb{E}_\pi(\mathbb{1}_i) = \pi_i \tag{3.26}$$

for each $i = 1, 2, \ldots, N$. Then, using the expression (3.25), we see

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(X_j) &= \lim_{N\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \sum_{i=1}^{N} f(i)\mathbb{1}_i(X_j) \\ &= \sum_{i=1}^{N} f(i) \Big(\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}_i(X_j) \Big) \end{aligned}$$

$$= \sum_{i=1}^{N} f(i)\pi_i = \mathbb{E}_\pi(f),$$

where the second-to-last equality is our assumption (3.26). So, all we need to prove is (3.26).

As an aside, note that we can re-cast everything we are doing in terms of linear algebra: the indicators $\mathbb{1}_i$ correspond to the standard basis vectors

$$\mathbf{e}_1 = (1, 0, 0, \ldots, 0),\ \mathbf{e}_2 = (0, 1, 0, \ldots, 0),\ \ldots,\ \mathbf{e}_N = (0, 0, \ldots, 0, 1),$$

and (3.25) writes the vector $\mathbf{f}$ as a linear combination of the $\mathbf{e}_j$'s. What we just argued is that it suffices to prove the theorem for the basis vectors.

We proceed to show (3.26) holds for any fixed $i = 1, 2, \ldots, N$. Our main tool will be the Law of Large Numbers (3.24), using a sequence of *return times* to a given vertex. Indeed, given $X_0 = x$, set $\tau_0^+ := 0$ and

$$\begin{aligned}
\tau_1^+ &:= \text{the first return time to } x \\
&= \min\{\, t > 0 \ :\ X_t = x \,\}, \\
\tau_2^+ &:= \text{the second return time to } x \\
&= \min\{\, t > \tau_1^+ \ :\ X_t = x \,\}, \\
&\vdots \\
\tau_k^+ &:= \text{the } k\text{th return time to } x \\
&= \min\{\, t > \tau_{k-1}^+ \ :\ X_t = x \,\}.
\end{aligned}$$

By the Markov property, the chain restarts each time we return to x, and so in particular,

$$\tau_k^+ - \tau_{k-1}^+, \qquad k = 1, 2, \ldots$$

is an iid sequence of random variables. The Law of Large Numbers (3.24) therefore tells us

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (\tau_k^+ - \tau_{k-1}^+) = \mathbb{E}_x(\tau_1^+ - \tau_0^+) = \mathbb{E}_x(\tau_1^+). \tag{3.27}$$

The sum above, however, "telescopes" according to

$$\sum_{k=1}^{n}(\tau_k^+ - \tau_{k-1}^+) = \tau_n^+ - \tau_0^+ = \tau_n^+,$$

and so (3.27) also says

$$\mathbb{E}_x(\tau_1^+) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(\tau_k^+ - \tau_{k-1}^+) = \lim_{n\to\infty} \frac{\tau_n^+}{n}. \tag{3.28}$$

Now back to our indicator $\mathbb{1}_i$. Let S_k be the number of visits to i between the $(k-1)$th and kth return of the chain to x,

$$S_k := \sum_{j=\tau_{k-1}^+}^{\tau_k^+ - 1} \mathbb{1}_i(X_j). \tag{3.29}$$

As with the hitting times, the Markov property implies that $S_1, S_2, \ldots$ is an i.i.d. sequence, and so invoking the Law of Large Numbers again yields

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} S_k = \mathbb{E}(S_1) = \mathbb{E}_x(\#\text{ visits to } i \text{ before returning to } x). \tag{3.30}$$

But recalling our formula (1.51) from the proof of Theorem 1.7,

$$\pi_i = \frac{\mathbb{E}_x(\#\text{ visits to } i \text{ before returning to } x)}{\mathbb{E}_x(\tau_1^+)},$$

(3.30) says

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} S_j = \pi_i \mathbb{E}_x(\tau_1^+). \tag{3.31}$$

The clever idea is to combine this all as follows. We have by (3.29) that

$$\sum_{j=0}^{t_n^+ - 1} \mathbb{1}_i(X_j) = \sum_{j=1}^{n} S_j,$$

since we are counting all the visits to state i up to the nth return to x.

Hence

$$\frac{1}{\tau_n^+}\sum_{j=0}^{\tau_n^+-1}\mathbb{1}_i(X_j)=\frac{1}{\tau_n^+}\sum_{j=1}^{n}S_j=\frac{\frac{1}{n}\sum_{j=1}^{n}S_j}{\frac{\tau_n^+}{n}}\underset{n\to\infty}{\longrightarrow}\frac{\pi_i\mathbb{E}_x(\tau_1^+)}{\mathbb{E}_x(\tau_1^+)}=\pi_i$$

by (3.28) and (3.31). Hence we have

$$\lim_{n\to\infty}\frac{1}{\tau_n^+}\sum_{j=0}^{\tau_n^+-1}\mathbb{1}_i(X_j)=\pi_i, \tag{3.32}$$

which is almost our desired limit (3.26). To finish the argument, we simply note that for any $\tau_n^+\le m\le\tau_{n+1}^+$,

$$\sum_{j=0}^{\tau_n^+-1}\mathbb{1}_i(X_j)\le\sum_{j=0}^{m-1}\mathbb{1}_i(X_j)\le\sum_{j=0}^{\tau_{n+1}^+-1}\mathbb{1}_i(X_j)$$

since the number of visits is non-decreasing, and therefore

$$\frac{1}{\tau_{n+1}^+}\sum_{j=0}^{\tau_n^+-1}\mathbb{1}_i(X_j)\le\frac{1}{m}\sum_{j=0}^{m-1}\mathbb{1}_i(X_j)\le\frac{1}{\tau_n^+}\sum_{j=0}^{\tau_{n+1}^+-1}\mathbb{1}_i(X_j). \tag{3.33}$$

Note carefully the indices. We would like to replace $1/\tau_{n+1}^+$ on the left with $1/\tau_n^+$, and similarly replace $1/\tau_n^+$ on the right with $1/\tau_{n+1}^+$, and then invoke (3.32). For the left, we observe

$$\frac{1}{\tau_{n+1}^+}=\frac{\tau_n^+}{\tau_{n+1}^+}\cdot\frac{1}{\tau_n^+}\overset{d}{=}\frac{\tau_n^+}{\tau_n^++\tau_1^+}\cdot\frac{1}{\tau_n^+}=\frac{1}{1+\frac{\tau_1^+}{\tau_n^+}}\cdot\frac{1}{\tau_n^+}\underset{n\to\infty}{\longrightarrow}1\cdot\frac{1}{\tau_n^+}$$

since $\tau_n^+\to\infty$ and τ_1^+ is fixed, and where "$\overset{d}{=}$" means equal in distribution. Thus, with probability 1, $1/\tau_{n+1}^+$ is very close to $1/\tau_n^+$ for large n, and similarly for $1/\tau_n^+$ and $1/\tau_{n+1}^+$ for the right-hand side. Our inequality (3.33) therefore sandwiches $\frac{1}{m}\sum_{j=0}^{m-1}\mathbb{1}_i(X_j)$ between quantities that are very close to π_i by (3.32), which yields the desired limit (3.26). □

3.2 Mixing times

One of the primary uses of Markov chains is to simulate a random sample from a given distribution π, as we will explore further in the next chapter. We can now begin to see how this can work, though, via Theorem 3.4: if we run an irreducible and aperiodic chain for a large number of steps n, (3.15) says that

$$\frac{1}{2}\sum_{j=1}^{N} |\mathbb{P}(X_n = j) - \pi_j)| < C\alpha^n. \tag{3.34}$$

In other words, X_n is very close to π in distribution, and so X_n is itself an approximate sample of π. However, can we quantify how close to π we are? Suppose we want, for instance,

$$\sum_{j=1}^{n} |\mathbb{P}(X_N = j) - \pi_j)| < \frac{1}{100}. \tag{3.35}$$

How many steps n must we take? In this section we will see that for reversible chains, we can answer this in terms of the eigenvalues and eigenvectors of the transition matrix P.

We will prove in the following sections that the exponential rate of convergence α in (3.34) is controlled by the second-largest eigenvalue of P (see Theorem 3.8 below), and we will give an upper bound for how many steps we need to reach a given threshold of error, as in (3.35).

Exercise 3.7. *Note that in* (3.34) *and* (3.35)*, we wrote* $\mathbb{P}(X_n = j)$ *instead of specifying a starting point* $\mathbb{P}_x(X_n = j)$*. Why that imprecision?*

3.2.1 Definition of mixing time

We begin by giving a name to first time when we achieve an error threshold as in (3.35). Let X_n, $n = 0, 1, \ldots$, be an irreducible and aperiodic chain with transition matrix P, and taking values in $\Omega = \{1, 2, \ldots, N\}$, as usual. Theorem 3.4 tells us that

$$\max_{x\in\Omega} \|P^n(x, \cdot) - \pi\|_{TV} \underset{n\to\infty}{\longrightarrow} 0.$$

Definition 3.3 (Mixing time)**.** For $0 < \epsilon < 1$, the **ϵ-mixing time** $t_{\text{mix}}(\epsilon)$ is the minimal time $n \in \mathbb{N}$ such that

$$\max_{x \in \Omega} \|P^n(x, \cdot) - \pi\|_{TV} \leq \epsilon.$$

So the mixing time $t_{\text{mix}}(\epsilon)$ tells us the earliest time n when we are guaranteed to be at most ϵ-far away from the stationary distribution in total variation, no matter what state x we begin at. Usually we will not be able to exactly compute $t_{\text{mix}}(\epsilon)$, but upper bounds will suffice in practice. We will need linear algebra techniques to obtain upper bounds, and so we first turn our attention to developing (or reviewing, depending on your background) some tools for dealing with powers of matrices.

3.2.2 The spectral decomposition

Recall that a square matrix $M \in \mathbb{R}^{N \times N}$ is orthogonally diagonalizable iff it is symmetric, $M^T = M$. So, starting with a symmetric M, we have the **spectral decomposition**

$$M = U^T \Lambda U \tag{3.36}$$

for some matrix U^T with orthonormal columns and a diagonal matrix

$$\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}.$$

The λ_j are the eigenvalues of M, and the columns u_j^T of U^T are the corresponding eigenvectors,

$$M u_j^T = \lambda_j u_j^T \tag{3.37}$$

for all j.* Note that this is exactly what you get if you multiply both sides of (3.36) by U^T and compare the columns. Let's comment on a possible

*Recall our convention is that vectors are row vectors, and hence a column vector is written via the transpose.

point of confusion: note from (3.37) that the spectral decomposition gives us the *right*-eigenvalues and *right*-eigenvectors of M, while up to this point we have primarily been interested in solving $\pi P = \pi$ for the *left*-eigenvector π of P with *left*-eigenvalue 1. This distinction does not matter for the eigenvalues: the characteristic polynomial of a matrix M and its transpose M^T are the same, which implies the left- and right-eigenvalues of M are identical. However, the left- and right-*eigenvectors* for a fixed eigenvalue λ are generally not identical for non-symmetric matrices. See Exercise 3.8 below.

If we multiply out the right-hand side of (3.36), we find

$$M = \sum_{j=1}^{N} \lambda_j u_j^T u_j, \tag{3.38}$$

which implies

$$M^k = \sum_{j=1}^{N} \lambda_j^k u_j^T u_j \tag{3.39}$$

by orthogonality and the fact that the u_j are unit vectors. It is a good idea to convince yourself of these last two assertions.

Exercise 3.8. *Show that* (3.38) *follows from* (3.36)*, and derive* (3.39) *from* (3.38) *using induction on* k*.*

Exercise 3.9. (*a*) *Show that the left- and right- eigenvalues of a matrix are identical. That is, for any given* $\lambda \in \mathbb{R}$*, there is some non-zero row vector* v *such that* $vP = \lambda v$ *if and only if there is a non-zero column vector* w^T *such that* $Pw^T = \lambda w^T$*. (*Hint*: use the characteristic polynomial.)*

(*b*) *The stationary distribution* π *for the simple walk on the graph in Figure 1.1 is given in Example 1.8. We know* π *is a left-eigenvector for* P *with* $\lambda = 1$*. Show that* π^T *is not a right-eigenvector for* P *for any* λ*.*

3.2.3 The Perron-Frobenius Theorem and convergence rate for symmetric P

If we want to understand the high powers of a symmetric matrix P, it is clear from (3.39) that we need to understand the eigenvalues λ of P. We

already know that 1 is an eigenvalue. This is clear from $\pi P = \pi$, when P has a stationary distribution π, and also from the equation $P\mathbf{1}^T = \mathbf{1}^T$, where $\mathbf{1} = (1, 1, \ldots, 1)$ is the vectors of one's in $\mathbb{R}^N$.

Let's think further about a specific example. If P is symmetric, then the columns of P sum to 1 and $\pi = (\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}) = \frac{1}{N}\mathbf{1}$ by Exercise 1.13(a). Hence $\pi^T = \frac{1}{N}\mathbf{1}^T$ is also a right-eigenvector of P. Noting that $\|\pi\|_2 = \frac{1}{\sqrt{N}}$, we see the unit-vector version of π is $u_1 = \sqrt{N}\pi$, and (3.39) then says

$$P^n = N\pi^T\pi + \sum_{j=2}^{N} \lambda_j^n u_j^T u_j. \tag{3.40}$$

If all the remaining eigenvalues are smaller than 1,

$$|\lambda_j| < 1 \quad \text{for } j = 2, 3, \ldots, N, \tag{3.41}$$

then $\lambda_j^n \to 0$, and taking the limit in (3.40) yields

$$\begin{aligned} \lim_{n\to\infty} P^n &= N\pi^T\pi + 0 \\ &= \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & & \ddots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix}. \end{aligned} \tag{3.42}$$

This is thus the simplest situation: a one-dimensional eigenspace for $\lambda = 1$, and all other eigenvalues satisfying (3.41). The next theorem tells us this is always the case when our chain is irreducible and aperiodic.

Theorem 3.6 (Perron-Frobenius)**.** *Let P be a transition matrix.*

(i) *If λ is an eigenvalue of P, then $|\lambda| \le 1$.*

(ii) *If P is irreducible, then the eigenspace corresponding to $\lambda_1 = 1$ has dimension 1.*

(iii) *If P is irreducible and aperiodic, then -1 is not an eigenvalue of P.*

Proof. Parts (i) and (ii) are a version of the *Perron-Frobenius Theorem*, a result in linear algebra that we will not prove here. We will argue for (iii),

however. Suppose P is irreducible and aperiodic, and $Pv^T = -v^T$ for some column vector v^T. We need to show that $v = \mathbf{0} = (0, 0, \ldots, 0)$. Note that $Pv^T = -v^T$ implies $P^n v^T = (-1)^n v^T$. However, by Theorem 3.4, we know that *each row* of P^n converges to π (recall the discussion around (3.18)), and so

$$P^n v^T \underset{n\to\infty}{\longrightarrow} \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix} v^T = \begin{bmatrix} \mathbb{E}_\pi(v) \\ \mathbb{E}_\pi(v) \\ \vdots \\ \mathbb{E}_\pi(v) \end{bmatrix}. \tag{3.43}$$

So, if we take the limit only using even n, we have

$$\begin{bmatrix} \mathbb{E}_\pi(v) \\ \mathbb{E}_\pi(v) \\ \vdots \\ \mathbb{E}_\pi(v) \end{bmatrix} = \lim_{\substack{n\to\infty \\ n \text{ even}}} P^n v^T = \lim_{\substack{n\to\infty \\ n \text{ even}}} v^T = v^T,$$

while if we go along odd n, we find

$$\begin{bmatrix} \mathbb{E}_\pi(v) \\ \mathbb{E}_\pi(v) \\ \vdots \\ \mathbb{E}_\pi(v) \end{bmatrix} = \lim_{\substack{n\to\infty \\ n \text{ odd}}} P^n v^T = \lim_{\substack{n\to\infty \\ n \text{ odd}}} (-1)^n v^T = -v^T.$$

However, both of these limits are the same by (3.43), and so $v = -v$ implying $v = \mathbf{0}$, as claimed. We conclude that P does not have an eigenvalue of -1. □

Order the eigenvalues of P in decreasing size $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and repeat them according to multiplicity (so if the eigenspace of λ_1 has dimension 2, then we would have $\lambda_1 = \lambda_2 > \lambda_3$). The Perron-Frobenius theorem then says

$$\lambda_1 = \underbrace{1 > \lambda_2}_{\text{if irreducible}} \geq \lambda_3 \geq \cdots \geq \underbrace{\lambda_n > -1}_{\text{if irreducible \& aperiodic}}$$

In particular, powers P^n of the matrix satisfy the limit (3.42) when P is

symmetric, irreducible and aperiodic, an observation we collect into the following theorem.

Theorem 3.7. *Let P be a symmetric transition matrix for an irreducible and aperiodic chain, with eigenvalues $\lambda_1 = 1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N > -1$. Then the exponential rate of convergence α in Theorem 3.4 is given by*

$$\max_{j \geq 2} |\lambda_j|. \tag{3.44}$$

Even though $\lambda_2 \geq \lambda_j$ for $j \geq 2$, note that $\max_{j \geq 2} |\lambda_j|$ is not necessarily $|\lambda_2|$ since the λ_j can be negative.

Proof. We have seen that the spectral theorem gives

$$P^n = N\pi^T\pi + \sum_{j=2}^{N} \lambda_j^n u_j^T u_j.$$

The first matrix $N\pi^T\pi$ is the constant $N \times N$ matrix with entries $1/N$. By Theorem 3.6, $|\lambda_j| < 1$ for $j = 2, 3, \ldots, N$, and hence

$$\sum_{j=2}^{N} \lambda_j^n u_j^T u_j \underset{n \to \infty}{\longrightarrow} 0$$

exponentially fast as $n \to \infty$, with rate controlled by (3.44). Each row of what is left, the matrix $N\pi^T\pi$, is the stationary distribution. □

The Perron-Frobenius theorem also allows us to prove uniqueness of the stationary distribution π for irreducible Markov chains. Recall that in Theorem 1.7 we proved the *existence* of π for any irreducible chain, along with the explicit formula

$$\pi(x) = \frac{1}{\mathbb{E}(\tau_x^+ \mid X_0 = x)}. \tag{3.45}$$

What remains to finish the proof of Theorem 1.6 is to show that *any* stationary distribution satisfies the formula (3.45).

Proof of Theorem 1.6. Let ρ be any stationary distribution. Then $\rho P = \rho$ and so ρ is a left-eigenvector of P corresponding to $\lambda_1 = 1$. Note that the

left- and right-eigenspaces for λ_1 have identical dimension; this follows from the argument in Exercise 3.9(a), as the characteristic polynomials for P and P^T are identical. Hence Theorem 3.6 (ii) says that the left-eigenspace for λ_1 is spanned by the π given by (3.45). In particular, there exists $c \in \mathbb{R}$ such that $\rho = c\pi$, and so summing the entries ρ_j yields

$$\sum_{j=1}^{N} \rho_j = c \sum_{j=1}^{N} \pi_j = c.$$

Since ρ is a probability distribution, $c = 1$ and so $\rho = \pi$. □

Exercise 3.10. *Consider the simple random walk (X_n) on the 7-cycle $\Omega = \{1, 2, 3, 4, 5, 6, 7\}$. Write out the transition probability matrix P, and use a computer to take large powers of P. Numerically verify that*

$$\lim_{n\to\infty} P^n = \begin{bmatrix} 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{bmatrix}.$$

What is the exponential rate of convergence?

3.2.4 Convergence rate for reversible P

Having *symmetric* transition probabilities P in Theorem 3.7 is a very restrictive condition. Can we anything about our convergence rate for aperiodic, irreducible and reversible P that are not symmetric? The answer is yes, via a heavy dose of linear algebra (as we are dealing with high powers of the transition matrix P, there is no other route). We will find that the rate of convergence in Theorem 3.7 also holds in this case.

Suppose P is reversible with respect to π,

$$\pi_x P_{xy} = \pi_y P_{yx}, \qquad x, y \in \Omega. \tag{3.46}$$

If π is not uniform, then P is not symmetric, and we can't immediately

apply the spectral decomposition as above. However, if we build the right symmetric matrix out of P, using reversibility, we can still gain useful information.

To that end, we define diagonal matrices

$$D_\pi := \mathrm{diag}(\pi_1, \pi_2, \ldots, \pi_n) = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n \end{bmatrix}$$

and $D_\pi^p := \mathrm{diag}(\pi_1^p, \pi_2^p, \ldots, \pi_n^p)$ for some power p, and consider

$$A := D_\pi^{1/2} P D_\pi^{-1/2}.$$

Noting that the $(x, y)-$component of A is

$$A_{xy} = \frac{\sqrt{\pi_x}}{\sqrt{\pi_y}} P_{xy}, \tag{3.47}$$

we claim that A is symmetric, $A_{xy} = A_{yx}$.

Exercise 3.11. *Convince yourself of* (3.47)*, and use* (3.46) *and* (3.47) *to show that* $A_{xy} = A_{yx}$*.*

Symmetric matrices have the spectral decomposition (3.38), and so

$$A = D_\pi^{1/2} P D_\pi^{-1/2} = \sum_{j=1}^{N} \lambda_j u_j^T u_j \tag{3.48}$$

for some eigenvalues λ_j and corresponding orthonormal eigenvectors u_j^T. Note that when you multiply out $D_\pi^{1/2} P D_\pi^{-1/2}$ with itself k times, the inner $D_\pi^{\pm 1/2}$ matrices cancel, and hence

$$A^k = D_\pi^{1/2} P^k D_\pi^{-1/2},$$

and so we see by (3.39) that

$$A^k = D_\pi^{1/2} P^k D_\pi^{-1/2} = \sum_{j=1}^{N} \lambda_j^k u_j^T u_j.$$

We interested in $P^k D_\pi^{-1}$, and the above shows

$$\begin{aligned} P^k D_\pi^{-1} = D_\pi^{-1/2} A^k D_\pi^{-1/2} &= \sum_{j=1}^N \lambda_j^k D_\pi^{-1/2} u_j^T u_j D_\pi^{-1/2} \\ &= \sum_{j=1}^N \lambda_j^k (D_\pi^{-1/2})^T u_j^T u_j D_\pi^{-1/2} \\ &= \sum_{j=1}^N \lambda_j^k (u_j D_\pi^{-1/2})^T u_j D_\pi^{-1/2} \\ &= \sum_{j=1}^N \lambda_j^k v_j^T v_j \end{aligned} \tag{3.49}$$

for $v_j := u_j D_\pi^{-1/2}$. Note that the v_j^T are eigenvalues of P with eigenvalue λ_j, since

$$\begin{aligned} P v_j^T = P D_\pi^{-1/2} u_j^T &= D_\pi^{-1/2} A D_\pi^{1/2} D_\pi^{-1/2} u_j^T \\ &= D_\pi^{-1/2} A u_j^T = \lambda_j D_\pi^{-1/2} u_j^T = \lambda_j v_j^T. \end{aligned}$$

In particular, by Theorem 3.6 (ii), the eigenvector v_1^T for $\lambda_1 = 1$ must be a multiple of the constant vector $\mathbf{1}$ (P is a transition matrix, so $\mathbf{1}^T$ is an eigenvector, and by Theorem 3.6, the eigenspace is one-dimensional). So $v = (c, c, \ldots, c)$ for some c.

We claim, however, that $c = \pm 1$, i.e. $v = \pm \mathbf{1}$. Indeed, we have

$$v_1 = (c, c, \ldots, c) = u_1 D_\pi^{-1/2} = \left(\frac{u_{11}}{\sqrt{\pi_1}}, \frac{u_{12}}{\sqrt{\pi_2}}, \ldots, \frac{u_{1N}}{\sqrt{\pi_N}} \right),$$

and so $u_{1j} = c\sqrt{\pi_j}$ for each j. However, u_1 is a unit vector, as that's what the spectral decomposition gave us for A in (3.48). Therefore

$$1 = \|u_1\|^2 = \sum_{j=1}^N u_{1j}^2 = c^2 \sum_{j=1}^N \pi_j = c^2 \cdot 1,$$

and so $c = \pm 1$, as claimed.

Now suppose that P is irreducible and aperiodic. Picking out the $(x, y)-$entry

of (3.49), we have

$$(P^k D_\pi^{-1})_{xy} = \frac{P^k_{xy}}{\pi_y} = \sum_{j=1}^{N} \lambda_j^k v_{j,x} v_{j,y} = \lambda_1 v_{1,x} v_{1,y} + \sum_{j=2}^{N} \lambda_j^k v_{j,x} v_{j,y}$$
$$= 1 \cdot 1 + \sum_{j=2}^{N} \lambda_j^k v_{j,x} v_{j,y}. \qquad (3.50)$$

By Theorem 3.6 (i) and (iii), we know $|\lambda_j| < 1$ for $j = 2, 3, \ldots, N$, and thus taking limits in the above yields

$$\lim_{k\to\infty} \frac{P^k_{xy}}{\pi_y} = 1. \qquad (3.51)$$

We have recovered our convergence result, Theorem 3.4. Note that (3.51) is independent of x, as we would expect from (3.16). Furthermore, we see from (3.50) that the exponential rate of convergence is again controlled by the largest of $\lambda_2, \lambda_3, \ldots, \lambda_N$. In other words, we have proved the following extension of Theorem 3.7 to more general P.

Theorem 3.8. *Let P be the transition matrix of an irreducible, aperiodic and reversible chain with eigenvalues $1, \lambda_2, \lambda_3, \ldots, \lambda_N$. Then the exponential rate of convergence α in Theorem 3.4 is given by*

$$\max_{j\geq 2} |\lambda_j|.$$

3.2.5 The relaxation time

Motivated by Theorems 3.8 (and its more restrictive counterpart 3.7), we set

$$\lambda^* := \max_{j\geq 2} |\lambda_j|,$$

the largest of the non-identity eigenvalues of P. We define the *absolute spectral gap* as $\gamma^* := 1 - \lambda^*$.

Definition 3.4. The **relaxation time** t_{rel} is the reciprocal of the absolute

spectral gap,

$$t_{\text{rel}} = \frac{1}{\gamma^*} = \frac{1}{1-\lambda^*} = \frac{1}{1-\max_{j\geq 2}|\lambda_j|}$$

The name comes from the following theorem, which should seem plausible in light of the exponential decay we've seen in (3.50). (We won't give a proof, however.)

Recall that the *mixing time* $t_{\text{mix}}(\epsilon)$ is the minimal time t we need to guarantee that our chain is ϵ-close to stationarity, regardless of where we start. In other words, it's the minimal t such that

$$\max_{x\in\Omega} \left\| P^t(x,\cdot) - \pi \right\|_{TV} < \epsilon.$$

Theorem 3.9. *Let P be irreducible, aperiodic and reversible with respect to π. Let π_m be the smallest of the entries of π,*

$$\pi_m := \min_{j=1,2,\ldots,N} \pi_j.$$

Then, for any $0 < \epsilon < 1$,

$$(t_{rel} - 1)\log\left(\frac{1}{2\epsilon}\right) \leq t_{mix}(\epsilon) \leq \log\left(\frac{1}{\epsilon\pi_m}\right) t_{rel}. \tag{3.52}$$

Note what this theorem says practically: if we know the eigenvalues of P (and we can figure these out on a computer), and hence the relaxation time t_{rel}, we have bounds (3.52) on the number of steps we need to run the chain to "mix" ϵ-close to stationarity. In other words, we have control on how many steps we have to take in order to gain any level of precision in sampling our stationary distribution. Note that the logarithm in (3.52) is base e, or the natural logarithm.

The proof, unfortunately, is beyond the scope of our text; see [2, Ch.4].

3.3 Two examples of mixing times

Let's work out the bounds in Theorem 3.9 in a couple of concrete examples.

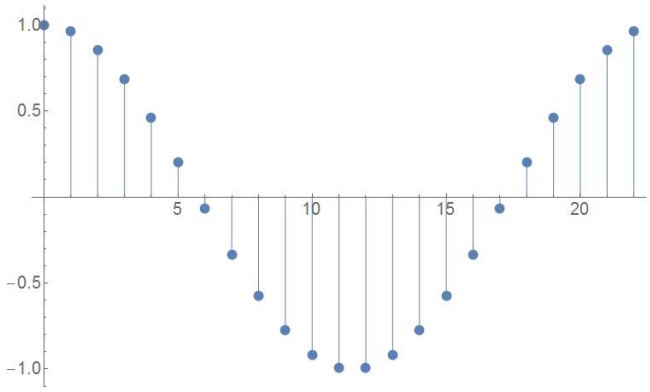

Figure 3.2: The eigenvalues λ_j for the transition matrix P of the simple walk on the 23-cycle.

3.3.1 Random walk on an n-cycle

Consider a random walk on the n-cycle $\Omega = \{0, 1, \ldots, n-1\}$. The transition matrix has eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$, where we claim

$$\lambda_j = \cos\Big(\frac{2\pi j}{n}\Big), \qquad j = 0, 1, \ldots, n-1. \tag{3.53}$$

See Figure 3.2 the case of the $n = 23$. One way we can prove that these are the eigenvalues is to explicitly give the eigenvectors. We claim that each λ_j has an eigenvector $f_j^T \in \mathbb{R}^n$ that satisfies the formula

$$f_j(k) = \cos\Big(\frac{2\pi j}{n}k\Big), \qquad k = 0, 1, \ldots, n-1. \tag{3.54}$$

In particular, when $j = 0$ we recover the eigenvalue $\lambda_0 = 1$ and the corresponding (right) eigenvector is $f_0^T = \mathbf{1}^T$.

Exercise 3.12. *Prove* (3.54) *by showing that, for each j, the kth entry of Pf_j^T is the same as the kth entry of $\lambda_j f_j^T$, $k = 0, 1, \ldots, n-1$. Hint: use* (3.53)*, the kth row of P (which only has two non-zero entries) and the trig formula*

$$\cos(a)\cos(b) = \frac{1}{2}(cos(a+b) + \cos(a-b)).$$

Does the n-cycle have an eigenvalue of -1? From (3.53), we would need

$$\frac{2\pi j}{n} = \pi \quad \Leftrightarrow \quad j = \frac{n}{2},$$

which is possible for integer j if and only if n is even. Note that this consistent with the Perron-Frobenius Theorem, Theorem 3.6(*iii*): aperiodic chains do not have -1 as an eigenvalue, and for the n-cycle aperiodicity is equivalent to n being odd. Since our mixing time bounds (3.52) are for aperiodic chains, we restrict our attention to odd n.

We can read off the second-largest eigenvalue from our formula (3.53). Indeed, note that the argument of cosine is between 0 and 2π, and so the second-largest value is when we are the second-closest to 0 or 2π, yielding

$$\lambda^* = \max_{j \neq 0} |\lambda_j| = \lambda_1 = \cos\Big(\frac{2\pi}{n}\Big) = \lambda_{n-1} = \cos\Big(\frac{2\pi(n-1)}{n}\Big),$$

and so the spectral gap is $\gamma^* = 1 - \cos(2\pi/n)$ and the relaxation time is

$$t_{\text{rel}} = \frac{1}{1 - \cos(2\pi/n)}. \tag{3.55}$$

Since $\cos(x) = 1 - \frac{1}{2}x^2 + O(x^4)$ for x close to 0, we find

$$\gamma^* \approx 1 - \left(1 - \frac{1}{2} \cdot \frac{4\pi^2}{n^2}\right) = \frac{2\pi^2}{n^2}$$

for large n, and thus $t_{\text{rel}} = \frac{1}{\gamma^*} \approx \frac{n^2}{2\pi^2}$. The stationary distribution is uniform, $\pi = (1/n, 1/n, \ldots, 1/n)$, and so we have the upper bound

$$t_{\text{mix}}(\epsilon) \leq \log\left(\frac{1}{\epsilon \pi_m}\right) t_{\text{rel}} \approx \log\Big(\frac{n}{\epsilon}\Big) \frac{n^2}{2\pi^2}. \tag{3.56}$$

If we run the chain for this many steps, we are guaranteed to be past the ϵ-mixing point to stationarity. The lower bound is similarly

$$(t_{\text{rel}} - 1) \log\Big(\frac{1}{2\epsilon}\Big) \approx \left(\frac{n^2}{2\pi^2} - 1\right) \log\Big(\frac{1}{2\epsilon}\Big)$$

If we only run for fewer than this many steps, we know that we have not reached ϵ-mixing.

For both the upper and lower bounds, we do not have to use the Taylor series expansion for cosine; we can simply use the exact formula (3.55) for the relaxation time instead. The point of using the Taylor series is to see how the bounds grow as n increases, which becomes explicit in (3.56) but is harder to immediately see if we use (3.55).

Exercise 3.13. *You are running a simple random walk (X_n) on the vertices of a 23-gon. Start the chain from any given vertex. Approximately how many steps N should you run to ensure that the distribution of X_N is within $1/1000$ of the uniform distribution on the vertices of the 23-gon, in total variation norm? What is the minimum possible number of steps we could use?*

Exercise 3.14. *Answer the same questions for the 5-cycle.*

3.3.2 Random walk on the hypercube

Consider the random walk on the vertices of the hypercube $\{-1,1\}^N$ in $\mathbb{R}^N$. If $N = 1$, this is a walk on the two vertices of a line segment, if $N = 2$, on the four vertices of a square, if $N = 3$, on the eight vertices of the cube. See Figure 2.4 for the case of $N = 3$ for the hypercube $\{0,1\}^3$. (We will use $\{-1,1\}^N$ instead because our formulas work out particularly nicely.)

Are these walks aperiodic? You were asked to compute the number of edges for the hypercube in $\mathbb{R}^N$ in Exercise 2.9. This is always an even number, and it is not too hard to see that the walk always has period 2, just like the walk on the n-cycle when n is even. For example, when $N = 2$ we have the 4-cycle. We therefore consider the lazy chain with transition matrix $\tilde{P} = \frac{1}{2}(I + P)$ to get an aperiodic walk.

What happens to our eigenvalues? Note that if λ is an eigenvalue for P with eigenvector v, then

$$\tilde{\lambda} = \frac{1}{2}(1 + \lambda) \tag{3.57}$$

is an eigenvalue for $\tilde{P}$ with eigenvector v. Indeed,

$$\tilde{P}v = \frac{1}{2}\Big(v + Pv\Big) = \frac{1}{2}\Big(v + \lambda v\Big) = \Big(\frac{1}{2} + \frac{1}{2}\lambda\Big)v,$$

as claimed. But if $-1 \le \lambda \le 1$, as is the case for transition matrices, then $0 \le \frac{1}{2}(1+\lambda) \le 1$, and so all the eigenvalues of $\tilde{P}$ are positive. In particular,

there is no -1 eigenvalue, which is again as we expect from Theorem 3.6 (*iii*) once we consider the lazy walk.

We start the eigenvalues and eigenvectors for P, before switching to the lazy walk via (3.57). The vertices of our hypercube are all vectors $\omega = (\omega_1, \omega_2, \ldots, \omega_N)$, where each $\omega_j = \pm 1$. Each vertex ω has a corresponding row and column on P, and we can describe the eigenvectors v in terms of their components $v(\omega)$ for each of the 2^N vertices ω.

We can parametrize the eigenvalues the eigenvectors in terms of the 2^N subsets J of $\{1, 2, \ldots, N\}$. For each $J \subset \{1, 2, \ldots, N\}$, there is an eigenvalue

$$\lambda_J = 1 - \frac{2 \cdot \#(J)}{N},$$

with corresponding eigenvector v_J, which has the 2^N components

$$v_J(\omega) = \prod_{j \in J} \omega_j \tag{3.58}$$

for each vertex $\omega \in \{-1, 1\}^N$.

An example will clarify, and so let's return to $N = 3$. Here the vertices are

$$\begin{aligned} \omega_1 &= (1, 1, 1), \\ \omega_2 &= (-1, 1, 1), \\ \omega_3 &= (1, -1, 1), \\ &\vdots \\ \omega_8 &= (-1, -1, -1). \end{aligned}$$

Our statement about the eigenvalues and eigenvectors says that if we pick $J = \{1, 3\} \subset \{1, 2, 3\}$, for instance, we get a corresponding eigenvalue

$$\lambda_{\{1,3\}} = 1 - \frac{2 \cdot 2}{3} = -\frac{1}{3},$$

with a formula (3.58) for the eight entries of $v_{\{1,3\}}$,

$$\begin{aligned} v_{\{1,3\}}(\omega_1) &= \omega_{11} \cdot \omega_{13} = 1 \cdot 1 = 1, \\ v_{\{1,3\}}(\omega_2) &= \omega_{21} \cdot \omega_{23} = -1 \cdot 1 = -1, \end{aligned}$$

$$\vdots$$

$$v_{\{1,3\}}(\omega_8) = \omega_{81} \cdot \omega_{83} = (-1)(-1) = 1.$$

Moving to the lazy walk $\tilde{P}$ for aperiodicity, we have eigenvalues

$$\tilde{\lambda}_J = \frac{1}{2}(1 + \lambda_J) = 1 - \frac{\#(J)}{N} \in [0, 1]$$

for each subset $J \subset \{1, 2, \ldots, N\}$. The largest is therefore $\lambda = 1$ for $J = \emptyset$, and the second largest is $\lambda^* = 1 - \frac{1}{N}$ for any singleton set J. Thus the spectral gap is $\gamma^* = 1 - (1 - \frac{1}{N}) = \frac{1}{N}$, and our relaxation time is $t_{\text{rel}} = \frac{1}{\gamma^*} = N$. It is not too hard to see that the stationary distribution is again uniform, $\pi = (\frac{1}{2^N}, \ldots, \frac{1}{2^N})$, and so our estimate (3.52) says an upper bound for mixing is

$$t_{\text{mix}}(\epsilon) \leq \log\left(\frac{1}{\epsilon \pi_m}\right) t_{\text{rel}} = \log\left(\frac{2^N}{\epsilon}\right) N = N^2 \log\left(\frac{2}{\sqrt[N]{\epsilon}}\right). \tag{3.59}$$

For instance, if we are in $\mathbb{R}^3$ and want to guarantee that our walk is within a millionth of the stationary distribution $(1/8, 1/8, \ldots, 1/8)$ in TV-norm, we need to run for at most

$$3^2 \log(2 \cdot \sqrt[3]{10^6}) < 48$$

steps. Pretty fast, right? That's an example of the exponential convergence (3.15) to stationarity.

When $\epsilon \leq 1$ (as we want for small deviations from stationarity),

$$t_{\text{mix}}(\epsilon) \leq N^2 \log\left(\frac{2}{\sqrt[N]{\epsilon}}\right) \leq N^2 \log\left(\frac{2}{\epsilon}\right),$$

and so we see for fixed ϵ that our upper bound grows like N^2. This is slower than for the N-cycle, where our upper bound (3.56) grows like $\log(N)N^2$. In other words, we need fewer steps to mix in a high-dimensional hypercube than in a large N-cycle.

Exercise 3.15. *Give an intuitive explanation for why the random walk on the hypercube $\{-1, 1\}^N$ mixes faster than the the random walk on the N-cycle.*

Problems for chapter 3

Problem 3.1. Toss a fair coin repeatedly. Let S_n be the number of heads after n tosses. Show that there is a limiting value for the proportion of times that S_n is divisible by 7, and compute the value for this limit. (*Hint*: Find an appropriate Markov chain that models this phenomenon.)

Problem 3.2. Consider the complete graph K_n on n vertices with loops. That is, $\Omega = \{1, 2, \ldots, n\}$ and every edge $\{i, j\}$ is present, including $\{i, i\}$. A random walker starts from a vertex and randomly jumps to any other vertex with equal probability. The transition matrix of the chain is

$$P = \frac{1}{n}\mathbf{1}^T\mathbf{1},$$

where $\mathbf{1}$ is the row vector of all ones.

(*a*) Show that P has exactly one eigenvalue that is one, and all other eigenvalues are zeroes.

(*b*) Find the absolute spectral gap and the relaxation time for this chain. Estimate an upper bound on the mixing time $t_{\mathbf{mix}}(\epsilon)$ using the relaxation time.

(*c*) Show that $t_{\mathbf{mix}}(\epsilon) = 1$ for any $\epsilon > 0$.

Problem 3.3. (Random walks on edges) Consider a connected graph $G = (V, E)$. Instead of considering simple symmetric RW on vertices of this graph, consider a new random walk on the edges. That is, consider an edge $e = \{x, y\}$ between two vertices x and y. Think of two *directed edges*, $x \to y$ and $y \to x$, depending on whether you go from x to y or the other direction. Let Ω be the set of all directed edges $x \to y$ and $y \to x$, where $\{x, y\}$ is an edge in the graph. Run a Markov chain with the following rule: if you are currently at an edge $x \to y$, then pick uniformly at random one of the neighbors of y (including x). Call this neighbor z and jump to the edge $y \to z$.

(*a*) Is this new RW irreducible?

(*b*) In the long run, what fraction of time do you see the new random walk visiting a directed edge $x \to y$? Use ergodicity of the simple symmetric RW on the vertices of the graph to answer this question.

(*c*) Use part (*b*) to identify the stationary distribution of the new RW on the directed edges.

(*d*) Is the new RW on the directed edges reversible with respect to the stationary distribution? Why or why not?

(*e*) Compare and contrast your results with the simple walk on the vertices of the graph.

Problem 3.4. Consider $\Omega = \{0,1\}^n$, the hypercube in dimension n. Consider the following Markov chain $\{X_0, X_1, X_2, \ldots\}$ on Ω. Suppose the current state is $w = (w_1, w_2, \ldots, w_n)$. In the next step turn w to the new vector $(w_0, w_1, w_2, \ldots, w_{n-1})$, where w_0 is uniformly chosen to be 0 or 1. That is, drop w_n, shift all coordinates by one step to the right and add a random 0 or 1 at the beginning.

(*a*) Is this chain irreducible? Why or why not?

(*b*) Find its stationary distribution π.

(*c*) Suppose $X_0 = \mathbf{0} = (0, 0, \ldots, 0)$, the vector of all zeros. Find the *exact* total variation distance between $P^k(\mathbf{0}, \cdot)$ and π as a function of k.

(*d*) Find the smallest k when $\max_x \left\|P^k(x,\cdot) - \pi\right\|_{\text{TV}} \leq 1/2$.

(*e*) Find the smallest k when $\max_x \left\|P^k(x,\cdot) - \pi\right\|_{\text{TV}} = 0$.

(*f*) Describe its time-reversed chain by computing its transition probabilities.

Problem 3.5. For parameters $0 < p, q < 1$ consider the following stochastic matrix

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

Consider a Markov chain with state space $\Omega = \{1, 2\}$ and transition matrix P. The stationary distribution for this chain is $\pi = (q/(p+q), p/(p+q))$.

(*a*) Show that this chain is reversible.

(*b*) Find the eigenvalues of P.

(c) Find the relaxation time t_{rel} of this chain and an upper bound on the mixing time $t_{\text{mix}}(\epsilon)$ for $\epsilon = \frac{1}{10}$ using the relaxation time.

Problem 3.6. Consider again the Markov chain from Problem 3.5. Let μ_t denote $\mathbb{P}_1(X_n = 1)$ and define

$$\Delta_n = \mu_n - \frac{q}{p+q}.$$

(a) Show that the following recursion holds:

$$\Delta_{n+1} = (1 - p - q)\Delta_n.$$

(b) Compute the total variation distance

$$\|P^n(1, \cdot) - \pi\|_{\text{TV}}.$$

(c) Compute the mixing time $t_{\text{mix}}(\epsilon)$ for $\epsilon = \frac{1}{10}$ and compare with the upper bound obtained in Problem 3.5.

Problem 3.7. Consider a random walk X_t, $t = 0, 1, 2, 3, \ldots$, on the hypercube $\Omega = \{0, 1\}^N$. For w in the hypercube let $|w| := \sum_{i=1}^N w_i$ denote the number of ones in w.

(a) In the long run, what fraction of time does the random walker have exactly k many ones, for $k = 0, 1, 2, \ldots, N$?

(b) Let $s_t = |X_t|$, and let $\bar{s}_t$ be the empirical average $t^{-1} \sum_{i=0}^{t-1} s_i$. Compute

$$\lim_{t \to \infty} \bar{s}_t.$$

(c) Evaluate the limit of the empirical variance,

$$\lim_{t \to \infty} \frac{1}{t} \sum_{i=0}^{t-1} (s_i - \bar{s}_t)^2.$$

Problem 3.8. Consider the Ehrenfest urn model with $N \geq 2$ balls in two urns A and B. Start with a balls in urn A. At each step, pick a ball at random and switch its urn. Let X_n be the number of balls in urn A after n

steps, and let $Y_n = N - 2X_n$ be the difference of the number of balls in urn B and urn A.

(a) Show that

$$\mathbb{E}(Y_{n+1} \mid Y_n) = Y_n \left(1 - \frac{2}{N}\right).$$

(b) Compute $\mathbb{E}(Y_n)$ for each n as a function of $X_0 = a$ and show that $\lim_{n\to\infty} \mathbb{E}(Y_n) = 0$.

(c) Let τ_N be the first time that all N balls are in urn A (i.e., first hitting time of N for X_n). Find $\mathbb{E}_{N-1}(\tau_N)$. (Hint: There is a reason we are not asking for $\mathbb{E}_k(\tau_N)$ for all k.)

(d) Start with $X_0 = N - 2$. Before τ_N, there will be a *last time* that urn A will have $(N - 2)$ balls. That is, consider the (random) time

$$L_{N-2} = \max\{0 \le t \le \tau_N : X_t = N - 2\}.$$

Find $\mathbb{E}_{N-2}(L_{N-2})$. (Hint: compute $\mathbb{E}_{N-2}(\tau_N)$).

(e) The stationary distribution of the Ehrenfest model is $\mathrm{Bin}(N, 1/2)$. Given a number $0 < p < 1$, $p \neq 1/2$, modify the Ehrenfest model suitably so that the new Markov chain has a stationary distribution $\mathrm{Bin}(N, p)$.

Chapter 4

Monte Carlo Methods

4.1 An introduction to sampling algorithms

Very often working statisticians will need to *simulate* random variables on their computers. Our computers can only generate a random number uniformly distributed in the unit interval $(0, 1)$. In reality, computers do slightly worse since they can only generate pseudo-random numbers that will pass statistical tests for randomness but are actually deterministic. It is a fascinating philosophical discussion to distinguish between true randomness in nature, such as those arising from quantum mechanics, and those generated by pseudo-random generating algorithms, such as the linear congruential generator (look it up!), but in practice this difference is mostly inconsequential (as they say, "If you can't tell, does it matter?"). But one thing is certain, that humans left to their own devices are terrible at trying to simulate a random sequence of zeroes and ones in their mind.

Let's assume that you have a *random number generator*, i.e. a computer that generates an i.i.d. sequence of perfect Uni$(0, 1)$ random variables. The problem of simulations is to use this sequence to create other random variables and stochastic processes that are required in your model. The objective is to see the output your model produces when the simulated random variables are fed to it. You can then, say, compare the results with real data to verify the efficacy of your model, or perhaps, predict a future outcome.

Thus, simulation is big business and a number of standard algorithms are frequently used. The simplest example is to simulate a coin toss with a probability of H given by $0 < p < 1$. Here is a solution. Generate a

Uni$(0,1)$ distributed random number U using your generator. If $U < p$, declare $X = 1$ (or $X = H$), otherwise, if $U \geq p$, declare $X = 0$ (or $X = T$). It is easy to see that X is a Bernoulli(p) random variable. In other words, I have simulated the outcome of a coin toss with probability p of Heads.

More generally, the easiest case is the simulation of one-dimensional random variables with explicitly invertible cumulative distribution functions.

Let's start with a basic observation. Suppose X is a random variable with a cumulative distribution function (cdf) F. Recall that the cdf is a function from $\mathbb{R}$ to $[0,1]$ given by

$$F(t) = P\left(X \leq t\right).$$

Assume that F is continuous, i.e., X is a continuous random variable. Since X is a real-valued, and F acts on real numbers, one gets a new transformed random variable $U = F(X)$.

I want to warn you that the following statement is nonsense: "Since, $F(t) = P\left(X \leq t\right)$, then $F(X) = P(X \leq X) = 1$". Instead, let's see what we mean by an example. Consider X to be an Exp(1) random variable. Then

$$F(t) = \begin{cases} 0, \text{ if } t \leq 0, \\ 1 - e^{-t}, \text{ if } t > 0. \end{cases}$$

In this case $U = 1 - e^{-X}$, since X is always positive. Note that F is strictly increasing on $(0, \infty)$, where it can be inverted. In fact, verify that, for $u > 0$, $F^{-1}(u) = -\log(1-u)$.

Theorem 4.1. *Assume that F is continuous and strictly increasing at all t such that $0 < F(t) < 1$. Hence the inverse function F^{-1} is well-defined. Then, the random variable $U = F(X)$ is distributed as Uni$(0,1)$. Conversely, if U is a Uni$(0,1)$ random variable, then $X = F^{-1}(U)$ has cdf F.*

Proof. Let us compute the cdf of $U = F(X)$. Since F has a range between zero and one, clearly $0 \leq U \leq 1$. Pick an $0 < a < 1$. Since F is invertible with an inverse F^{-1},

$$P(U \leq a) = P\left(F(X) \leq a\right) = P\left(X \leq F^{-1}(a)\right) = F\left(F^{-1}(a)\right) = a.$$

Thus, the cdf of U coincides with that of Uni$(0,1)$, and, hence, by the uniqueness of cdf, U must be uniformly distributed over $(0,1)$.

Conversely, let U be a Uni$(0,1)$ random variable. Define $X = F^{-1}(U)$. Then

$$P(X \le t) = P\left(F^{-1}(U) \le t\right) = P(U \le F(t)) = F(t).$$

The final equality is due to the cdf of U and since $0 \le F(t) \le 1$. □

The assumption in the previous theorem can be relaxed a bit by simply defining an inverse when F is continuous but perhaps not strictly increasing. We will not prove the following statement (but you are welcome to try). Theorem 4.1 continues to hold for all continuous F if we define $F^{-1}(t) = \inf\{x : F(x) \ge t\}$, the left-continuous generalized inverse of F.

Theorem 4.1 gives us our first algorithm to simulate. Suppose we wish to simulate a random variable X with a cdf F whose inverse F^{-1} is explicitly computable. Then, to get a sample of X, simply ask your computer to output a Uni$(0,1)$ random number U, and declare $X = F^{-1}(U)$. This is called the inverse cdf method of sampling.

Exercise 4.1. *Describe how you can use the inverse cdf method to sample any Exp(λ) random variable for $\lambda > 0$. Write a program to execute it.*

Unfortunately, the inverse cdf method does not work for some standard distributions. One prominent example is the standard normal whose cdf (and also the inverse cdf) is not explicit. Before we move on to more generally applicable methods of sampling, let us generalize the inverse cdf method to sample discrete random variables.

Suppose X is a discrete random variable taking values in $\mathbb{N}$ with a probability mass function π. That is, $P(X = i) = \pi_i$, $i = 1, 2, \ldots$, with $\pi_i \ge 0$ and $\sum_{i=1}^{\infty} \pi_i = 1$. Here's a method to simulate a sample of X: divide up the closed unit interval $[0,1]$ in successive pieces of sub-intervals of length π_i, as in Figure 4.1. Generate a uniformly distributed random number U in $(0,1)$ and note in which sub-interval U falls. Call X to be the number of sub-intervals to the left of U, including the one in which it falls. That is, if $U \le \pi_1$, declare $X = 1$. If $\pi_1 < U \le \pi_1 + \pi_2$, declare $X = 2$. Generally, if

$$\sum_{i=1}^{k-1} \pi_i < U \le \sum_{i=1}^{k} \pi_i, \tag{4.1}$$

declare $X = k$. Clearly X takes values in natural numbers.

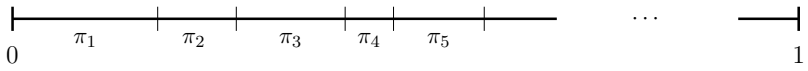

Figure 4.1: Sampling from a discrete distribution π.

We claim that X has pmf π. To see this, note from (4.1), for any $k = 1, 2, \ldots,$

$$P(X = k) = P\left(\sum_{i=1}^{k-1} \pi_i < U \leq \sum_{i=1}^{k} \pi_i\right) = \sum_{i=1}^{k} \pi_i - \sum_{i=1}^{k-1} \pi_i = \pi_k.$$

This completes the proof.

4.1.1 Rejection sampling

Rejection sampling (sometimes called acceptance-rejection sampling) lets you sample from a probability distribution π, assuming that you can generate a sample from another probability distribution p. Both p and π could be either probability mass functions or probability density functions (pdf), as long as the following assumption holds.

Assumption 1. $\pi(x) = 0$ if $p(x) = 0$ and for some $M > 0$ the ratio $\pi(x)/p(x) \leq M$, if $p(x) > 0$.

Note that, one can take $M = \max_{x:p(x)>0} \frac{\pi(x)}{p(x)}$ if this quantity is finite. But you don't have to. Any number larger than this maximum will also do as M. The point is to find an M such that the ratio $0 < \pi(x)/Mp(x) < 1$, whenever the ratio is well-defined. There is also an implicit assumption here: for any x one can compute the quantity $\pi(x)/p(x)$ explicitly.

The algorithm for rejection sampling to get a sample from π under Assumption 1 goes like this.

Step 1. Generate a sample Y from p.
Step 2. Generate an independent Uni$(0, 1)$ random variable U.
Step 3. If $U \leq \frac{\pi(Y)}{Mp(Y)}$, declare $X = Y$. Stop and return the output. Otherwise, discard the sample Y. Go back to Step 1. Continue until you stop, i.e., an output X is generated. X is clearly a random variable.

Theorem 4.2. *The distribution of the final output X is π.*

Proof. Suppose Y is distributed according to p. For simplicity, we are going to assume that p (and π) is a pmf. Let I be the indicator random variable $I = 1\{U \leq \pi(Y)/Mp(Y)\}$. Then, note that X is only outputted when $I = 1$. Thus

$$P(X = x) = P\left(Y = x \mid I = 1\right).$$

(By the way, ask yourself, why is $P(X = x) \neq P\left(Y = x, I = 1\right)$).

Thus

$$\begin{aligned} P(X = x) &= \frac{P(Y = x, I = 1)}{P(I = 1)} = \frac{P(Y = x)P(I = 1 \mid Y = x)}{P(I = 1)} \\ &= \frac{p(x)P\left(U \leq \pi(x)/Mp(x)\right)}{P(I = 1)} \\ &= \frac{p(x)\pi(x)/Mp(x)}{P(I = 1)} = \frac{\pi(x)}{MP(I = 1)}. \end{aligned}$$

At this point, we will be done if we show that $MP(I = 1) = 1$. But this has to be so. Just add over x on both sides to get

$$1 = \sum_x P(X = x) = \sum_x \frac{\pi(x)}{MP(I = 1)} = \frac{\sum_x \pi(x)}{MP(I = 1)} = \frac{1}{MP(I = 1)}.$$

Thus $MP(I = 1) = 1$, and we are done. □

Note that the probability of *acceptance*, i.e., $P(I = 1)$, is exactly $\frac{1}{M}$. Since this quantity must be at most one, $M \geq 1$ always. $M = 1$ if and only if $p = \pi$, when the algorithm is not needed. So $M > 1$ in all nontrivial cases. The larger the value of M is, the more frequently the algorithm will reject. That is, the run time of the algorithm grows linearly with M. In fact, convince yourself that the number of times the algorithm runs through Steps 1–3 before finally accepting a sample is Geo$(1/M)$.

Example 4.1. Let's see an example. Let π be the density

$$\pi(x) = 30x^2(1 - x)^2, \quad 0 < x < 1.$$

This is the beta$(3, 3)$ probability density. Take p to be the Uni$(0, 1)$ density. Then, clearly $p(x) = 0$ if x is outside $(0, 1)$, and, for such an x, $\pi(x) = 0$ as

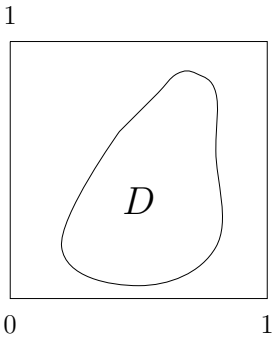

Figure 4.2: How can we sample a point uniformly from the region D?

well. When $x \in (0,1)$,

$$\frac{\pi(x)}{p(x)} = \frac{30x^2(1-x)^2}{1} = 30(x(1-x))^2 \leq 30\left(\frac{1}{4}\right)^2 = \frac{30}{16} = \frac{15}{8}.$$

The only inequality above follows from the bound $x(1-x) \leq \frac{1}{4}$, for all $0 \leq x \leq 1$ (why?). Thus, we can take $M = 15/8$.

Thus the rejection sampling algorithm to generate a sample from beta$(3,3)$ follows these steps.

Step 1. Generate a Uni$(0,1)$ random variable Y.

Step 2. Generate an independent Uni$(0,1)$ random variable U.

Step 3. If

$$U \leq \frac{30Y^2(1-Y)^2}{15/8} = 16Y^2(1-Y)^2,$$

output $X = Y$ as your desired sample and stop. Otherwise, discard Y and U, and go back to Step 1.

When the algorithm stops, your output X is a sample from beta$(3,3)$.

Example 4.2. For a second example, let's consider a geometric probability problem. Suppose I wish to sample uniformly from the region $D \subseteq [0,1]^2$. See Figure 4.2.

So, π is the uniform density over D, i.e., $\pi(x) = 1/\text{Area}(D)$, for $x \in D$, and zero outside. What is an easy density to sample from? A natural choice is Uni$\big([0,1]^2\big)$, since this is just sampling a pair of independent Uni$(0,1)$ random variables Y_1, Y_2 and representing them as a vector $Y = (Y_1, Y_2)$.

Thus $p(y) = 1$, for $y = (y_1, y_2) \in [0,1]^2$. Hence,

$$\frac{\pi(y)}{p(y)} \leq \frac{1}{\text{Area}(D)}.$$

Hence, we can take $M = 1/\text{Area}(D)$, assuming that we can compute it. Thus

$$\frac{\pi(y)}{Mp(y)} = \begin{cases} \frac{1/\text{Area}(D)}{1/\text{Area}(D)} = 1, & \text{if } y \in D, \\ \frac{0}{1/\text{Area}(D)} = 0, & \text{otherwise.} \end{cases}$$

Note that this ratio is either zero or one. So, if Y is sampled from p and U is independently sampled from $\text{Uni}(0,1)$, then $U \leq \pi(Y)/Mp(Y)$ if and only if $Y \in D$.

Hence, the rejection sampling algorithm is quite simple here. Generate a vector $Y = (Y_1, Y_2)$, where Y_1, Y_2 are independent $\text{Uni}(0,1)$ random variables. If Y is in D, accept the sample and declare $X = Y$. Otherwise generate a fresh sample of Y. Repeat until you get a sample that lies in D.

As you can guess from here, the smaller $\text{Area}(D)$ is, the longer it takes to generate a sample from π. This becomes a serious problem in high-dimensions. In the next section we will see how Markov chains combined with the rejection algorithm can lead to much more efficient sampling algorithms that work even in high-dimensions.

Exercise 4.2. *Describe how you will sample from the following density using rejection sampling*

$$\pi(x) = \frac{e^{-x}}{1 - e^{-1}}, \quad 0 \leq x \leq 1,$$

and zero otherwise. Give at least two choices for the density p that you can use to simulate π.

Exercise 4.3. *Given a random number generator, describe any method to sample $Geo(p)$ distribution for any $0 < p < 1$. Explain why the rejection sampling method would fail if we try to sample $\pi = Poi(1)$, using $p = Geo(1/2)$.*

4.2 Markov Chain Monte Carlo

Consider the set Ω of all permutations of the set $\{1, 2, \ldots, n\}$. How many permutations are there? $n!$, a number so large that it grows faster than all exponentials. There is a classic estimate, called Stirling's approximation, that shows $n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$. Thus, Ω might be finite but it is gigantic even for moderate values of n, say $n = 15$. The problem is to simulate a uniformly distributed random permutation from this set. I will describe two solutions to this problem, one not using Markov chains, and another using a Markov chain.

Solution 1. Generate n i.i.d. Uni$(0, 1)$ random variables $U_1, U_2, \ldots, U_n$. Sort them from the smallest to the largest values. That is, let $U_{(1)} = \min_{1 \le i \le n} U_i$, denote the smallest of the values, $U_{(2)}$ be the second smallest, and so on, until $U_{(n)} = \max_{1 \le i \le n} U_i$ is the largest value. For example, if our U_i's are, to three decimal places,

$$U_1 = 0.520, \quad U_2 = 0.512, \quad U_3 = 0.765, \quad U_4 = 0.428,$$

then

$$U_{(1)} = 0.428, \quad U_{(2)} = 0.512, \quad U_{(3)} = 0.520, \quad U_{(4)} = 0.765.$$

This sorting generates a permutation σ: the value of σ_1 is the *rank* of U_1, that is, the index i such that $U_{(i)} = U_1$ (why is i unique?). Similarly, σ_2 is the rank of U_2, the index i such that $U_{(i)} = U_2$, and so on. In our example, our permutation $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ is

$$\sigma = (3, 2, 4, 1).$$

In general, the vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ is a random element of Ω, and it follows from symmetry that the distribution of σ must be uniform over all permutations.

The complexity of this algorithm is therefore determined by the complexity of the sorting algorithm used. Common sorting algorithms typically have a time complexity of the order of $n \log n$. This is, of course, a much much smaller number than the size of Ω itself, which is $n!$. Our next algorithm is a Markov chain based algorithm which also achieves similar time-complexity.

Solution 2. The goal is to create a reversible Markov chain on Ω that moves by "small steps". There are many such candidate Markov chains which are often visualized best by card shuffling. We will talk about a particularly nice one called the *random transposition* chain.

Imagine a deck of cards marked 1 through n on top of each other. The goal is to shuffle these cards. Although many of you have shuffled cards in real life, in mathematics this has a precise meaning. The goal of a card shuffling technique is to arrange the cards randomly such that every arrangement of the n cards is equally likely to appear from top to bottom. Clearly every arrangement from top to bottom of the n cards corresponds to a permutation and vice versa.

Now, given a deck of cards, here is one method of shuffling. Pick two random numbers I and J independently from $\{1, 2, \ldots, n\}$. Suppose $I = i, J = j$. Pick the ith card from the top and swap its position with the jth card from the top. There is a trivial case, if $I = J$, when the deck does not change at all. Otherwise the deck changes by a "small step" due to the swapping of the two cards. Repeat this procedure, by sampling a fresh pair of I, J independently.

It is clear that this creates a Markov chain on Ω, given by the arrangement of the cards. In fact, if $I = i, J = j$, you hop from one permutation to another by multiplying with the transposition (i, j). There is a result in combinatorics which says that every permutation can be obtained as a product of transpositions. It follows that this Markov chain is irreducible. Since $P(I = J) = 1/n > 0$, the chain is aperiodic as well.

Notice that, if you go from one arrangement σ to another arrangement $\sigma' \neq \sigma$ by swapping two cards, you can also go from σ' to σ by swapping the same pair of cards again (with the same probability $1/n^2$). Thus, the Markov chain is reversible and its unique stationary distribution is the uniform distribution over Ω. *Et volia!*

Note that each step of the Markov chain simply requires generating two (discrete) uniform random variables. Convince yourself that this can be done easily by a random number generator. Thus, equipped with a random number generator, one can run this Markov chain for as many steps as one wishes. How many steps? The following is a deep result that is proved by using an area of mathematics called *representation theory*.

Theorem 4.3. *For $c > 0$, let $k = \frac{1}{2}n \log n + cn$. Then,*

$$\max_{\sigma \in \Omega} \left\| P^k(\sigma, \cdot) - \text{Uni} \right\|_{TV} \leq ae^{-2c},$$

for some universal constant a that does not depend on n or c.

In other words, $t_{\text{mix}}(\epsilon)$ is about $\frac{1}{2}n \log n$ (plus a multiple of n depending on ϵ, which is of a smaller order). Thus, this random procedure will give you an approximately uniformly distributed permutation at about the same order of efficiency as the sorting algorithm.

This is our first example of Markov Chain Monte Carlo or MCMC for short, a workhorse of modern data science. MCMC is a procedure where one uses a Markov chain to generate a sample from a specified distribution.

Let's start with a situation very similar to the above algorithm. Suppose Ω is a large finite set, the operative word being large, and we wish to sample an element of Ω uniformly at random. Create a connected graph with vertex set Ω and a relatively small number of edges per vertex. If possible, make the graph regular. Then the lazy random walk on this graph is reversible with respect to the uniform distribution. Run this random walk until it mixes, and the terminal value will be an approximate sample from the uniform distribution over Ω. If you want multiple samples, run the entire chain multiple times, independently.

In the example of random transpositions, Ω is the set of permutations. The graph is the following. Two permutations σ and σ' are neighbors, if one can be obtained by multiplying the other by a transposition (i, j). This is a regular graph where every vertex has degree exactly $n(n-1)/2$. You can verify that the random transposition chain is similar to a lazy random walk on this graph.

How do I create a Markov chain to simulate from a distribution that is not uniform? There are as many choices as your creativity. But there are some choices that are standard and popular for various reasons. The next section describes the grand-daddy of such Markov chains that runs by combining a random walk with rejection sampling at each step. This is the famous Metropolis algorithm.

4.3 The Metropolis-Hastings algorithm

Suppose Ω is a finite state space. Assume that we can run an irreducible Markov chain on Ω with a symmetric transition matrix $q(x, y) = q(y, x)$. The stationary distribution is uniform over Ω, and hence, after running this chain for a large number of steps, the terminal value is approximately a sample from the uniform distribution. But what if we are interested in sampling from another probability distribution π on Ω that is not uniform? This question led to one of the top five impactful algorithms of the 20th century, the Metropolis-Hastings algorithm. For brevity, we will refer to this algorithm as simply Metropolis.

The Metropolis algorithm creates a reversible Markov chain X on Ω with distribution π. Suppose currently $X_t = x$. Let us describe the probability $P(X_{t+1} = y \mid X_t = x)$.

Step 1. Generate a step from the symmetric transition matrix q (which is called the *base chain*). That is, create an auxiliary random variable Y_{t+1} such that

$$P\left(Y_{t+1} = y \mid X_t = x\right) = q(x, y).$$

Step 2. Suppose $Y_{t+1} = y$. If $\pi(y) \geq \pi(x)$, *accept* this step and declare $X_{t+1} = y$.

Otherwise, $\pi(y) < \pi(x)$, in which case you accept this step with probability $\pi(y)/\pi(x)$. That is to say, generate a coin toss with probability of H given by $\pi(y)/\pi(x)$. If the coin lands Heads, declare $X_{t+1} = y$, otherwise, if the coin lands Tails, *reject* the step by keeping $X_{t+1} = X_t = x$.

Theorem 4.4. *X_t, $t = 0, 1, 2, \ldots$ is a reversible Markov chain with stationary distribution π.*

Proof. Assume, for simplicity, that $q(x, x) = 0$ for all x. Suppose for every distinct pair $(x, y) \in \Omega \times \Omega$, I have a probability $a(x, y)$ (between zero and one) to accept the move $Y_{t+1} = y$ from $X_t = x$. That is,

$$P\left(Y_{t+1} = y \mid X_t = x\right) = q(x, y),\ P\left(X_{t+1} = y \mid X_t = x, Y_{t+1} = y\right) = a(x, y),$$

while

$$P\left(X_{t+1} = x \mid X_t = x, Y_{t+1} = y\right) = 1 - a(x, y).$$

Thus the transition probabilities for the chain X are

$$p(x,y) = q(x,y)a(x,y),\ y \neq x, \quad p(x,x) = 1 - \sum_y q(x,y)a(x,y).$$

What kind of choices for the function $a(\cdot,\cdot)$ will lead to p satisfying DBE for the distribution π,

$$\pi(x)p(x,y) = \pi(y)p(y,x), \tag{4.2}$$

for all $x \neq y$? Suppose (4.2) holds, then, for all $x \neq y$, $\pi(x)q(x,y)a(x,y) = \pi(y)q(y,x)a(y,x)$. Thus, necessarily

$$\pi(x)a(x,y) = \pi(y)a(y,x).$$

Since $0 \leq a(x,y) \leq 1$, we must have $\pi(x)a(x,y) \leq \pi(x)$. On the other hand, from the last display, $\pi(x)a(x,y) = \pi(y)a(y,x) \leq \pi(y)$, since $0 \leq a(y,x) \leq 1$. Therefore we conclude that

$$\begin{aligned} \pi(x)a(x,y) &\leq \min\left(\pi(x), \pi(y)\right), \text{ or, equivalently,} \\ a(x,y) &\leq \min\left(1, \frac{\pi(y)}{\pi(x)}\right). \end{aligned}$$

It is in our interest to maximize $a(x,y)$ while maintaining (4.2) in order to minimize run time. Thus, the largest possible value of $a(x,y)$ we can take is $a(x,y) = \min\left(1, \pi(y)/\pi(x)\right)$. In other words,

$$a(x,y) = \begin{cases} 1, & \text{if } \pi(y) \geq \pi(x), \\ \pi(y)/\pi(x), & \text{if } \pi(y) < \pi(x). \end{cases} \tag{4.3}$$

Verify directly that this formula for $a(x,y)$ satisfies (4.2):

$$\begin{aligned} \pi(x)a(x,y) &= \pi(x)\min\left(1, \pi(y)/\pi(x)\right) = \min(\pi(x), \pi(y)) \\ &= \pi(y)\min\left(\pi(x)/\pi(y), 1\right) = \pi(y)a(y,x). \end{aligned}$$

This completes the proof. □

In Theorem 4.4, the base chain is symmetric and we wish to simulate from a non-uniform distribution π. But, what about the case when it is easy

to simulate from a non-uniform distribution, but we wish to simulate from π equal to the uniform distribution on Ω. Here is a such an example: suppose we are given a graph that is not regular, i.e., vertices do not have a constant degree. It is easy to run a random walk on this graph, but the stationary distribution will not be uniform. Recall that the stationary distribution at a vertex v is in fact $\deg(v)/2\,|E|$, where $\deg(v)$ is the degree of the vertex v and $|E|$ is the total number of edges in the graph. How can I modify this random walk to obtain a uniform pick from the vertex set? The generalized version of the Metropolis algorithm allows you to do just that.

Metropolis algorithm for a general base chain. Suppose $q(x, y)$ is the transition probability from x to y for an arbitrary irreducible Markov chain on the finite state space Ω. Let π be any probability distribution on Ω. The Metropolis chain is given by the following steps. Suppose, currently, $X_t = x$.

Step 1. Generate a step from q, i.e., let Y_{t+1} be given by

$$P\left(Y_{t+1} = y \mid X_t = x\right) = q(x, y).$$

Step 2. Suppose $Y_{t+1} = y$. Accept $X_{t+1} = y$ with probability

$$a(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right). \tag{4.4}$$

Otherwise, reject the sample and declare $X_{t+1} = X_t = x$.

Theorem 4.5. *$(X_t,\ t = 0, 1, 2, \ldots)$ is a Markov chain that is reversible with stationary distribution π.*

Since the proof is exactly the same as before and we omit it (fill in the details, based on the previous proof, if you need to convince yourself). Let us instead go back to our problem of sampling uniformly from a non-regular graph. There, q is the transition probability of the simple random walk. Thus

$$q(x, y) = \frac{1}{\deg(x)} 1\{y \sim x\}.$$

Take π to be the uniform distribution. Thus $\pi(x) = \frac{1}{|\Omega|}$, where $|\Omega|$ is the

size of Ω, the vertex set. Thus,

$$a(x,y) = \min\left(1, \frac{1/|\Omega|}{1/|\Omega|}\frac{1/\deg(y)}{1/\deg(x)}\right) 1\{y \sim x\} = \min\left(1, \frac{\deg(x)}{\deg(y)}\right) 1\{y \sim x\}.$$

Thus, the Metropolis algorithm will run in the following manner. Suppose the chain is currently at vertex x. Choose one of the neighbors of x uniformly at random. Say, the chosen neighbor is y. If $\deg(y) \leq \deg(x)$, the chain jumps to y. Otherwise, if $\deg(y) > \deg(x)$, the chain jumps to y with probability $\deg(x)/\deg(y)$ and stays at x with probability $1-\deg(x)/\deg(y)$.

Thus the chain tends to move towards lower degree vertices to counterbalance the tendency of the simple random walk to move towards higher degree vertices. This changes the stationary distribution of the simple random walk, which gives more mass to vertices of higher degree, to the uniform distribution over all vertices. That's it! That's the secret!

4.4 Sampling from the Gibbs distribution

The acceptance probabilities appearing in (4.3) and (4.4) only depend on the ratio $\pi(y)/\pi(x)$. This simple observation has an enormous advantage in practice. Let us demonstrate via an example.

Suppose we want to generate a random graph (a social network, say) between N individuals. Here the vertices are given by the set $V = \{1, 2, \ldots, N\}$ and there is an edge $\{i, j\}$ if individuals i and j are friends of each other. How many possible such graphs are there? There are $N(N-1)/2$ many possible pairs of friends. Each such pair could be friends or could not be friends. This tells us there are $2^{N(N-1)/2}$ many possible graphs with vertices V. Let Ω denote the set of all such possible graphs. Thus $|\Omega| = 2^{N(N-1)/2}$, growing super-exponentially with N. Any respectable social network should have at least thousands of users which gives you an idea of how large Ω can be.

A natural probability distribution on Ω is the uniform distribution. There is a nice way to generate a sample from this uniform distribution, called the Erdős-Rényi random graph model. For each possible pair $\{i, j\}$, toss a fair coin. If the coin turn H, draw the edge, and if the coin turns T, don't draw the edge. I will leave it to you to convince yourself that the resulting random graph is distributed uniformly on Ω. The uniform distri-

bution is also the stationary distribution of the Markov chain on Ω where at each step you choose a random pair $\{i, j\}$. If there is an edge $\{i, j\}$, remove that edge, while if there is no such edge, draw that edge. As you can see this creates a symmetric transition matrix and the resulting Markov chain has a uniform stationary distribution. In fact, the Markov chain is nothing but the random walk on the hypercube in dimension $2^{N(N-1)/2}$.

Real social networks are not uniformly distributed. If A is a friend of B and B is a friend of C, then it is quite likely that A is a friend of C. This dependence is not captured in the Erdős-Rényi model where the three edges $\{A, B\}, \{B, C\}$ and $\{C, A\}$ are independent. How would one go about modeling such random graphs?

One commonly proposed solution is the following. Given a graph $G = (V, E)$, count the number of triangles in the graph. That is, let

$$\Delta(G) = \#\left\{\{i, j, k\} : \ i \sim j, j \sim k, k \sim i\right\}.$$

For a constant β, create a probability distribution on Ω given by the formula

$$\pi_\beta(G) = \frac{e^{\beta\Delta(G)}}{\sum_{G' \in \Omega} e^{\beta G'}} = \frac{e^{\beta\Delta(G)}}{Z_\beta},$$

where $Z_\beta = \sum_{G' \in \Omega} e^{\beta G'}$, where the sum is taken over all possible graphs in Ω. Z_β is called the normalizing constant and is not explicitly computable.

Clearly π_β is a probability mass function on Ω. How does it behave? If $\beta > 0$, the larger $\Delta(G)$ is, the bigger is the weight $e^{\beta\Delta(G)}$. Thus, for $\beta > 0$, π_β favors graphs with more triangles. In fact, the larger $\beta > 0$ is, graphs with more triangles become more likely under π_β.

On the other hand, if $\beta < 0$, the weight $e^{\beta\Delta(G)}$ gets smaller as $\Delta(G)$ gets larger. Thus, π_β favors graphs with fewer triangles. When $\beta = 0$, π_β is simply the uniform distribution on Ω.

Models such as π_β are called Gibbs distributions and are commonly used to incorporate dependency structures in data. Originally, such models come from a field called statistical physics where $1/\beta$ is called the temperature. The primary difficulty in trying to sample from π_β is the non-computability of the normalizing constant Z_β.

Fortunately, Metropolis does not require Z_β. Take the base chain to be the symmetric chain on Ω described above. You can turn this chain aperiodic

by choosing I and J i.i.d. uniformly at random from $\{1, 2, \ldots, N\}$. If $I = J$, stay where you are. Otherwise if $I = i \neq J = j$, then erase the edge $\{i, j\}$ if it exists, or add it if it does not. Thus G can jump to G' if and only if these two graphs differ by at most one edge, and $q(G, G') = q(G', G) = 2/N^2$.

In order to modify this chain according to Metropolis, we compute $a(G, G')$ from (4.3), Notice

$$\frac{\pi(G')}{\pi(G)} = \frac{e^{\beta\Delta(G')}/Z_\beta}{e^{\beta\Delta(G)}/Z_\beta} = e^{\beta(\Delta(G')-\Delta(G))}.$$

This ratio does not depend on the normalizing constant at all!

Recall we are interested in the case of $\beta > 0$ (favors more triangles). Thus, $\pi(G') \geq \pi(G)$ if and only if $\Delta(G') \geq \Delta(G)$. Thus, if you add an edge (thereby increasing the number of triangles), jump from G to G'. If you remove an edge (thus possibly decreasing the number of triangles), jump from G to G' with probability $e^{\beta(\Delta(G')-\Delta(G))}$. This Markov chain run for a large number of steps will be approximately distributed according to π_β.

4.5 Gibbs sampling.

So far from Section 4.1 we learned that simulating one-dimensional probability distributions are often easy. What if we can simulate a multidimensional probability distribution $\pi(x_1, \ldots, x_n)$ by reducing the problem to one dimension? This is the idea behind Gibbs sampling which is also called Glauber dynamics in certain contexts. I want to warn about the nomenclature: Gibbs sampling is a Markov chain Monte Carlo algorithm while the Gibbs distribution introduced in the last section is a probability distribution. They are both named after the American physicist Josiah Willard Gibbs (1839–1903), one of the founders of the field called *statistical mechanics*.

We describe the Gibbs sampling algorithm below for a joint pmf, but a similar algorithm works for a joint pdf as well. So, let π be a pmf in $\mathbb{R}^n$. Given a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, let

$$x^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), \quad i = 1, 2, \ldots, n.$$

Thus $x^i \in \mathbb{R}^{n-1}$ is simply the vector x whose ith coordinate is dropped. Consider now the conditional pmf, under π, of X_i at x_i, given the rest of

the values $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n$,

$$\pi^i_{x^i}(y) := P\left(X_i = y \mid X^i = x^i\right), \quad i = 1, 2, \ldots, n.$$

These are n one-dimensional conditional distributions. Assume that it is possible to sample from each of these, no matter the value of the conditioned variables. We will combine these distributions to run an MCMC algorithm.

Gibbs sampling algorithm. Start from an arbitrary vector $X(0)$. Suppose $X(t) = x = (x_1, \ldots, x_n)$. Pick an independent random variable I uniformly distributed among $\{1, 2, \ldots, n\}$. Suppose $I = i$. *Update* the ith coordinate by drawing a sample from the conditional distribution $\pi^i_{x^i}$. Let the value of this fresh sample be y. Then, define $X(t+1)$ by

$$X_i(t+1) = y, \text{ and } X_j(t+1) = X_j(t), \text{ for all } j \neq i.$$

Repeat.

Theorem 4.6. *The Markov chain X is reversible with respect to the stationary distribution π.*

Proof. Let's check Detail Balance Equations (DBE). We only consider one case, $I = 1$. The rest are similar. Suppose x and x' are two vectors that differ only in the first coordinate. That is $x = (z, x_2, \ldots, x_n)$ and $x' = (y, x_2, \ldots, x_n)$. Then $X(t) = x, X(t+1) = x'$ is only possible if $I = 1$. Thus

$$\pi(x)p(x, x') = \pi(x)P(I = 1)\pi^1_{x^1}(y) = \pi(x)\frac{1}{n}\pi^1_{x^1}(y).$$

But, by definition of conditional pmf, $\pi(x) = P(X^1 = x^1)\pi^1_{x^1}(z)$. Thus

$$\pi(x)p(x, x') = \frac{1}{n}P(X^1 = x^1)\pi^1_{x^1}(z)\pi^1_{x^1}(y) = \pi(x')p(x', x),$$

where the last equality follows by symmetry. This verifies DBE and we are done! Convince yourself that this chain is irreducible and aperiodic. □

Let us do an example.

Example 4.3. Fix a natural number N. Let S_N denote the set of natural

numbers (m, k) such that $m + k \leq N$. That is

$$S_N = \{(m, k) : \ m \geq 1, \ k \geq 1, \ m + k \leq N\}.$$

Let us use Gibbs sampling to pick a uniformly distributed sample (X, Y) from S_N.

In order to do this we have to identify the two conditional distributions. Clearly, if $X = m$, then Y can be any of $\{1, 2 \ldots, N - m\}$ with equal probability. Thus the conditional distribution of Y, given $X = m$, is Uni$\{1, \ldots, N - m\}$, i.e.,

$$P(Y = k \mid X = m) = \frac{1}{N - m}, \quad k = 1, 2, \ldots, N - m.$$

Exactly in the same way, the conditional distribution of X, given $Y = k$, is Uni$\{1, \ldots, N - k\}$. Hence, Gibbs sampling proceeds by picking a random integer I uniformly from the pair $\{0, 1\}$. If $I = 0$, update the current value of X by sampling a uniform natural number between 1 and the $N - k$, where k is the current value of Y. If $I = 1$,update the current value of Y by sampling a uniform natural number between 1 and the $N - m$, where m is the current value of X. Repeat several time to get an appprximate sample from the uniform distribution over S_N.

Notice here that the algorithm can be made slightly more efficient. If $I = 0$ twice in a row, we are updating the value of X twice successively while keeping the current value of Y. This means that the first update is thrown away and the only the second update is retained. This is a waste of computational resources. Thus, it is more efficient to not choose I at random at all but alternative between 0 and 1. That is $I = 0$ in the first step, followed by $I = 1$, followed by $I = 0$, followed by $I = 1$, and so on. This is a common variant of the Gibbs sampling algorithm.

Example 4.4. For our second example let us assume that Gibbs sampling works for continuous densities (it does). Sample from the joint density

$$f(p, q, r) = \frac{1}{W} pr, \quad 0 < p < q < 1, \quad 0 < r < 1,$$

where W is the normalizing constant. W is unknown and I urge you to keep it that way and not compute it. It is unimportant for Gibbs sampling.

Let (X, Y, Z) be random variables with joint density f. We need to compute the three conditional densities of X, given $Y = q, Z = r$, Y, given $X = p, Z = r$, and Z, given $X = p, Y = q$.

Let us start from the conditional density of Y at q given the other two. Note that the joint density f does not involve q explicitly in the formula. Thus

$$f_Y(q \mid X = p, Z = r) = c,\ p < q < 1,$$

where c is a constant that does not depend on q. This is only possible if this conditional density is uniform over the interval $[p, 1]$. Since uniform densities are explicit, c must be $1/(1 - p)$. Thus

$$f_Y(q \mid X = p, Z = r) = \frac{1}{1-p},\ p < q < 1,$$

Similarly, consider $f_Z(r \mid X = p, Y = q)$. The joint density f as a function of r is simply a constant times r. Thus

$$f_Z(r \mid X = p, Y = q) = cr,\ 0 < r < 1,$$

for some normalizing constant c. It is easy to see by integrating the above that $c = 2$. Thus

$$f_Z(r \mid X = p, Y = q) = 2r,\ 0 < r < 1,$$

which is a beta$(2, 1)$ density and does not depend on p and q. Thus Z is independent of (X, Y).

Finally

$$f_X(p \mid Y = q, Z = r) = cp, \quad 0 < p < q.$$

By integrating both sides we must get one. Thus

$$c \int_0^q p dp = 1,$$

or that $c = 2/q^2$. Note that, although both $f_X(\cdot \mid Y = q, Z = r)$ and $f_Z(\cdot \mid X = p, Y = q)$ are linear functions of the argument, they have very different supports where they are positives. This is why Z is independent from X, Y while X is dependent on Y. Nevertheless, with these three explicit

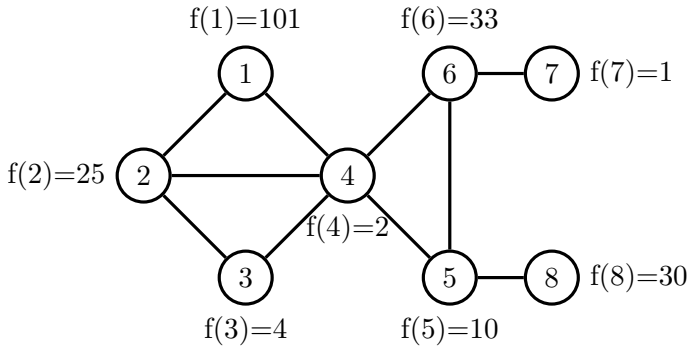

Figure 4.3: Optimization of a function over a graph.

densities at hand one can run Gibbs sampling as usual. A little bit of efficiency can be gained by separately generating Z, which is independent, while simulating only (X, Y) via Gibbs sampling.

4.6 Stochastic optimization

Suppose $G = (V, E)$ is a finite graph where, without loss of generality, we take the set of vertices $V = [n]$. Suppose there is a function $f : V \to \mathbb{R}$. We wish to find the minimum of the function $\min_{v \in V} f(v)$ and the minimizing vertex v^* such that $f(v^*) = \min_{v \in V} f(v)$. You might think what is the big problem here. After all I have a bunch of numbers and I can simply go through them one by one until I find the minimum. The problem is that n could be a very large number, let's say in the order of millions. Evaluating the function at every vertex and sorting them is computationally expensive. One might also consider a greedy "gradient descent" algorithm. Start from any vertex v_0 and evaluate $f(v_0)$. Now look at the neighbors of v_0. Among those neighbors find the vertex v_1 that minimizes f.

This doesn't quite work due to local minimum. Take the example of the graph shown in Figure 4.3. Suppose v_0 is vertex 8. It has only one neighbor 5 with a lower f-value, $f(5) = 10 < f(8) = 30$. So you jump to $v_1 = 5$. v_1 has two neighbors 4 and 6. The minimum value of f among them is achieved at 4 and $f(4) = 2 < f(5) = 10$. Thus $v_2 = 2$. Now we have nowhere to go since all the neighbors of v_2 have a higher value of f making

v_2 a local minimum of the function f. However, the absolute minimum of f is at vertex 7 which our gradient descent algorithm cannot reach. To avoid getting trapped in local minimums we will add randomness by creating a probability distribution on the vertices.

For a parameter $\lambda \geq 0$, consider the so-called *Gibbs probability distribution at temperature* $T = 1/\lambda$:

$$\pi_\lambda(i) = \frac{1}{Z_\lambda} e^{-\lambda f(i)}, \quad i \in V = [n].$$

Here Z_λ is called the normalizing constant given by

$$Z_\lambda = \sum_{j=1}^{n} e^{-\lambda f(j)}.$$

Note that this is the unique choice of Z_λ that turns π_λ to be a probability mass function.

The idea comes from statistical physics (hence the concept of temperature). When $\lambda = 0$, i.e., the temperature $T = \infty$ (very hot),

$$Z_0 = \sum_{j=1}^{n} e^{-0 \cdot f(i)} = n, \quad \pi_0(i) = \frac{1}{n},$$

making π_0 the uniform distribution on $[n]$. On the other hand, when the temperature gets very cold (i.e., $\lambda \to \infty$ or $T \to 0$), π_λ puts almost its entire mass on the minimizer of f. To see what I mean, assume that f has a unique minimizer, say vertex 1. That is $f(1) < f(v)$ for every vertex $v \neq 1$. Then, rewrite π_λ as follows

$$\pi_\lambda(i) = \frac{1}{Z_\lambda} e^{-\lambda f(i)} = \frac{e^{-\lambda f(i)}}{\sum_{j=1}^{n} e^{-\lambda f(j)}} = \frac{e^{-\lambda (f(i) - f(1))}}{\sum_{j=1}^{n} e^{-\lambda (f(j) - f(1))}}.$$

Consider the denominator $\sum_{j=1}^{n} e^{-\lambda (f(j) - f(1))}$. Either $j = 1$, in which case the corresponding term $e^{-\lambda (f(1) - f(1))} = e^0 = 1$ or $j > 1$, in which case $f(j) - f(1) > 0$, and hence $\lim_{\lambda \to \infty} e^{-\lambda (f(j) - f(1))} = 0$. Thus, summing up over all j, we get

$$\lim_{\lambda \to \infty} \sum_{j=1}^{n} e^{-\lambda (f(j) - f(1))} = 1.$$

Applying the same logic to the numerator of π_λ, we get

$$\lim_{\lambda\to\infty} \pi_\lambda(i) = \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{if } i > 1. \end{cases}$$

In other words the probability distribution π_λ converges to the trivial distribution that puts mass one at the minimizer of f. Notice that the same proof shows that if f has more than one minimizer, then π_λ, as $\lambda \to \infty$, converges to the uniform distribution on the set of minimizers of f.

The idea of stochastic optimization of f is that instead of greedily minimizing f, one looks to sample from the probability distribution π_λ for a large value of λ. Will such a sample always give us the minimizer? No, not unless $\lambda = \infty$. But with high probability it will sample a vertex whose f values is close to the minimum. So the next question is how do we sample from π_λ and one possible answer is via the Metropolis algorithm.

Hill Climb algorithm for stochastic optimization. Fix a base chain, say simple random walk on the graph. Thus the transition probability for the base chain is given by

$$q(x, y) = \frac{1}{\deg(x)}, \text{ if } y \text{ is a neighbor of } x, \text{ and zero otherwise.}$$

Recall that this is a reversible Markov chain but not symmetric unless the graph is regular.

Now run Metropolis with this base chain. Suppose, currently $X_t = x$, pick Y_{t+1} uniformly from one of the neighbors of x. Then accept this choice as X_{t+1} with probability

$$\begin{aligned} \min\left(\frac{\pi_\lambda(y)q(y,x)}{\pi_\lambda(x)q(x,y)}, 1\right) &= \min\left(\frac{e^{-\lambda f(y)}/\deg(y)}{e^{-\lambda f(x)}/\deg(x)}, 1\right) \\ &= \min\left(\frac{\deg(x)}{\deg(y)} e^{-\lambda(f(y)-f(x))}, 1\right). \end{aligned}$$

Otherwise, remain where you are by declaring $X_{t+1} = x$.

In the special case of regular graphs, $\deg(x) = \deg(y)$ and the algorithm becomes simpler. Accept y uniformly chosen among neighbors of x with probability $\min\left(e^{-\lambda(f(y)-f(x))}, 1\right)$. Thus, if $f(y) < f(x)$, you'd always accept

y and hence decrease f by doing so. But if $f(y) > f(x)$, you may still accept x with probability $e^{-\lambda(f(y)-f(x))}$. This step lets you escape the dreaded local minima which cannot be avoided in a gradient descent.

Simulated Annealing. But what if I insist on getting the actual minimum instead of a sample from a near-minimum with high probability? A variation of the hill climb algorithm called simulated annealing lets you do that. The idea is that you grow $\lambda = \lambda(t)$ *slowly* with time t, typically of the order of $\log(t)$. There is a theorem that, under suitable assumptions, the Markov chain converges to the minimizer of f but there are plenty of other ad hoc choices that appear to work in practice. See the paper by Dimitris Bertsimas and John Tsitsiklis titled Simulated Annealing that appeared in the journal *Statistical Science*, volume 8, number 1, pages 10-15, 1993.

Problems for chapter 4

Problem 4.1. Suppose you are trying to sample from some distribution π on a sample space Ω via the Metropolis-Hastings algorithm.

(*a*) Define your base chain to have transition probabilities $q(x, y) = \pi(y)$. What are your acceptance probabilities and the transition probabilities for the metropolis chain with this base chain?

(*b*) What is the mixing time for metropolis chain with this choice of q?

(*c*) Explain why this is not the approach we take in §4.3, and why this is not used in practice.

Problem 4.2. We wish to find a reversible Markov chain on the hypercube $\Omega = \{0, 1\}^N$ with a stationary distribution that favors vertices with more ones than zeroes in the following sense. For w in Ω, let $|w| = \sum_{i=1}^{N} w_i$ denote the number of ones in $w = (w_1, \ldots, w_N)$. Define the probability distribution

$$\pi(w) = \frac{|w|^2}{Z}, \quad \text{where} \quad Z = \sum_{v \in \Omega} |v|^2$$

is the normalizing constant.

(*a*) Describe a reversible Markov chain with stationary distribution π by writing its transition probabilities. Simplify your expressions for the transition probability $p(w_1, w_2)$ so that they are only in terms of $|w_1|$.

(*b*) Explicitly verify that your new chain satisfies the DBE's for π.

Problem 4.3. For each problem clearly describe the conditional distribution of each coordinate given the others. Then describe the procedure for running Gibbs sampling to sample from the joint distribution. Assume that Gibbs sampling works for continuous densities as well as discrete distributions. *Guess* the conditional from the structure of the joint distribution. *Avoid* doing integration as much as possible. Use your knowledge of the all the named one dimensional distributions/ densities.

(*a*) Sample from the joint density

$$f(x, y, z) = \frac{1}{W}, \quad 0 < x < y < z < 1,$$

and zero elsewhere. Here W is the normalizing constant.

(b) Sample from the mixed joint pmf/pdf:

$$p(n,t) = \frac{1}{Z}(1-p)^{n-1} n e^{-nt}, \quad t > 0,\ n = 1, 2, \ldots,$$

where Z is the normalizing constant.

Problem 4.4. For a number $s > 1$, consider the Riemann zeta function $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$. This number is finite and therefore results in a probability distribution on integers

$$\pi(n) = \frac{1}{n^s \zeta(s)}, \quad n \geq 1.$$

Develop a MCMC scheme based on Metropolis algorithm to sample a random integer according to the following probability distributions.

(a) The conditional distribution of $\pi(\cdot \mid X \leq N)$. That is,

$$P(X = n) = \frac{\pi(n)}{\sum_{i=1}^{N} \pi(i)}, \quad n = 1, 2, \ldots, N.$$

Explicitly give the transition probabilities $p(j,k)$ for your new chain.

(b) The entire distribution π on $\mathbb{N}$. Explicitly give the transition probabilities $p(j,k)$ for your new chain and verify that the chain satisfies the DBE's for π.

Problem 4.5. Consider the joint density

$$f(x,y,z) = \frac{1}{W} e^{-xyz - x - 2y - 3z}, \quad x > 0,\ y > 0,\ z > 0,$$

where W is the normalizing constant.

(a) Clearly identify each of the three conditional densities (i) X, given Y, Z; (ii) Y, given X, Z; and (iii) Z, given X, Y, in terms of standard named distribution (such as normal, exponential, geometric, Poisson etc.). (*Hint*: Do not try to integrate.)

(b) Describe how you will use Gibbs sampling to generate a sample from the joint density

Problem 4.6. For each problem clearly describe the conditional distribution of each coordinate given the others. Then describe the procedure for running Gibbs sampling to sample from the joint distribution. Assume that Gibbs sampling works for continuous densities as well as discrete distributions. **Guess** the conditional from the structure of the joint distribution. **Avoid** doing integration as much as possible. Use your knowledge of the all the named one dimensional distributions/ densities.

(*a*) Sample from the mixed joint pmf/pdf:

$$f(p,n) = p(1-p)^{n-1}, \quad 0 < p < 1,\ n = 1, 2, \ldots.$$

You will need one integration to describe the density of p given $N = n$.

(*b*) Sample from the joint density:

$$f(p,q,r) = \frac{1}{Z} pqr, \quad 0 < p, q, r < 1.$$

You will need one easy integration. What is special about this joint density? Why do you actually not need Gibb's algorithm to sample from this?

Problem 4.7. Consider the following joint distributions of random variables. Describe how you will sample from them using *any method* of sampling.

(*a*) $(X_1, X_2, \ldots, X_n)$ is Mult$(N, 1/n, \ldots, 1/n)$. Explain why using Gibb's sampling with only updating coordinate at a time will not work.

(*b*) (X, Y) is a standard bivariate normal with means 0, variances 1, and correlation ρ.

Problem 4.8. Consider a connected graph $G = (V, E)$ with $V = \{1, 2, \ldots, n\}$. For $-\infty < \alpha < \infty$, consider a probability distribution

$$\pi^\alpha(x) = \frac{1}{Z_\alpha} \left(\deg(x)\right)^\alpha,$$

where Z_α is the normalizing constant. Describe a reversible Markov chain with stationary distribution π^α. Describe in words how π^α is qualitatively different for different values of α.

Problem 4.9. Let $\Omega = \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ be the $n \times n$ grid. For $\beta \geq 0$, consider the probability distribution

$$\pi^\beta(i, j) = \frac{1}{Z} \exp\left(-|i - j|^\beta\right), \quad \text{for } (i, j) \in \Omega.$$

Describe a reversible Markov chain with stationary distribution π^β. Describe in words how π^β is different from the uniform distribution on the grid.

Problem 4.10. Let $\Omega = \{1, 2, \ldots, n\}$. Define a probability distribution

$$\pi(i) = \frac{1}{Z} i(n + 1 - i), \quad 1 \leq i \leq n,$$

where Z is the normalizing constant. Describe any method to get a sample from this probability distribution.

Chapter 5

Martingales and harmonic functions

5.1 Martingales: intuition, definition and first examples

Sometimes the most effective way to analyze a process is to consider *functions* of the process instead of merely the process itself. This point of view is common in areas of mathematics such as modern algebra, for instance, where functions ("morphisms") of algebraic objects often shed significant light on the structure of the original objects themselves. The "First Isomorphism Theorem" from your introductory course in abstract algebra, for example, is a nice instance of this principle.

In this chapter we take up this approach in analyzing Markov chains $(X_n)_{n\geq 0}$. We define new functions $(Y_n)_{n\geq 0}$ from $(X_n)_{n\geq 0}$, and we will see that when (Y_n) has a nice averaging property it yields rich information about the original chain itself. We will call these new processes *martingales*. While martingales can seem abstract, the underlying intuition is simple and we will be sure to inundate our exposition with a large amount of concrete examples. Moreover, martingales prove to be an important stepping stone to more advanced topics in probability, and they are utilized in a number of deep proofs in modern theory. While much of this will fall beyond the scope of our text, we hope you will nevertheless begin to appreciate the value of martingales in this chapter, as well as maintain the awareness that the

intellectual effort you may need to expend to internalize these new concepts will be well worth the effort.

5.1.1 Adapted processes and martingales

Let $(X_n)_{n\geq 0}$ be a Markov chain. The functions $(Y_n)_{n\geq 0}$ that we will build from (X_n) will be such that, for each n, Y_n is a function of the chain history $X_0, X_1, \ldots, X_n$ through the first n steps as well as the step number n,

$$Y_n = f(n, X_0, X_1, \ldots, X_n) = f_n(X_0, X_1, \ldots, X_n). \tag{5.1}$$

Since $X_1, \ldots, X_n$ are random, Y_n is a new random variable, and $(Y_n)_{n\geq 0}$ is thus a new stochastic process built from our original (X_n). Note that the function f_n defining Y_n can change with each n. We call a stochastic process $(Y_n)_{n\geq 0}$ satisfying (5.1) for all $n \geq 0$ an **adapted process** for (X_n). This means that if we know the values of $X_0, X_1, \ldots, X_n$, we know the value of Y_n, as is clear from the formula (5.1) (the functions f_n are given).

Example 5.1. Suppose $(X_n)_{n\geq 0}$ is the simple random walk on $\mathbb{Z}$. Examples of adapted processes on (X_n) include $Y_n = X_n$ (which is rather boring) and, say, $Y_n = X_n - 10X_{n-2} + \cos(\pi n^2)$ for $n \geq 2$.

While all the functions (Y_n) of chains (X_n) we will consider are adapted processes, we are particularly interested in a subclass of adapted processes which possess a special averaging property.

Definition 5.1. A process $Y = (Y_n)_{n\geq 0}$ adapted to the Markov chain $(X_n)_{n\geq 0}$ is a **martingale** if

$$\mathbb{E}(Y_{n+1} \,|\, X_0, X_1, \ldots, X_n) = Y_n \tag{5.2}$$

for each $n \geq 1$.

So, given all the information up to step n, we expect the next value Y_{n+1} of our martingale to be identical to the current value Y_n. The key intuition here is that martingales are thus "fair games." If (Y_n) were gambling winnings, for instance, this says that our expected net change in each stage is zero; on average we neither win nor lose anything on each turn and the game is exactly fair.

5.1.2 Examples of martingales

Let's carefully walk through four examples to get a better sense of what Definition 5.1 is all about.

Example 5.2. Consider a simple random walk $(X_n)_{n\geq 0}$ on $\mathbb{Z}$ starting from $X_0 = 0$. We claim

$$Y_n = X_n \tag{5.3}$$

is itself a martingale. Here our functions f_n in (5.1) are very simple,

$$f_n(X_0, X_1, \ldots, X_n) = f_n(X_n) = X_n,$$

which is the *projection* of the random vector $(X_0, X_1, \ldots, X_n)$ to its last coordinate. While this function is not terribly exciting, it is instructive to see that it does satisfy the desired averaging property (5.2). Indeed, we have

$$\begin{aligned}
\mathbb{E}(Y_{n+1} \,|\, X_0, X_1, \ldots, X_n) &= \mathbb{E}(X_{n+1} \,|\, X_0, X_1, \ldots, X_n) \\
&= \mathbb{E}(X_{n+1} \,|\, X_n) \\
&= \frac{1}{2}(X_n + 1) + \frac{1}{2}(X_n - 1) \\
&= X_n = Y_n,
\end{aligned}$$

where we have used the Markov property in the second line. We conclude that (X_n) itself is a martingale. This should not be surprising: we can think of (X_n) as a game where you are flipping a fair coin and win a dollar for tossing heads and lose a dollar for tails, which is fair.

Other martingales, however, can be less obvious.

Example 5.3. Let $(X_n)_{n\geq 0}$ be the simple random walk on $\mathbb{Z}$ again, and consider the process $(Y_n)_{n\geq 0}$ defined by

$$Y_n = X_n^2 - n. \tag{5.4}$$

In terms of coin flipping, in this game your total earnings after n turns is the square of the difference X_n of heads and tails flipped, less the number of flips.

Our process (Y_n) is certainly adapted: the function $f(n, X_0, \ldots, X_n)$ is given by the right-hand side of (5.4). It is not immediately obvious, however, if this is a fair game. Let's see what we can do with the expectation (5.2) in the definition of a martingale:

$$\begin{aligned}
\mathbb{E}(Y_{n+1} \,|\, X_0, X_1, \ldots, X_n) &= \mathbb{E}(X_{n+1}^2 - (n+1) \,|\, X_0, X_1, \ldots, X_n) \\
&= \mathbb{E}(X_{n+1}^2 - (n+1) \,|\, X_n) \\
&= \mathbb{E}(X_{n+1}^2 \,|\, X_n) - n - 1 \\
&= \frac{1}{2}(X_n + 1)^2 + \frac{1}{2}(X_n - 1)^2 - n - 1 \\
&= X_n^2 - n = Y_n,
\end{aligned}$$

where we have again used the Markov property in the second line. Thus we do have a fair game and (Y_n) is indeed a martingale. This stochastic process is called the *quadratic martingale* for the simple random walk.

Both the martingales in Examples 5.2 and 5.3 have nice applications: you will use them in Problems 5.8 and 5.9, respectively, to re-derive the gambler's ruin hitting probability and time formulas from §2.1.1 and §2.1.2. That is, these martingales give us a completely different tool to arrive at the same formulas. Problem 5.1 explores two other martingales for the simple random walk, the *cubic martingale* and *exponential martingale*.

Example 5.4. Let's re-visit the Pólya urn of §2.4 with a black balls and b red balls. Recall that in each step of the chain, you pick a uniformly-random ball from the urn and then return it to the urn with one additional ball of the same color. Let X_n be the number of black balls after n steps. In Problem 2.6 we saw that

$$\mathbb{E}\Big(\frac{X_{n+1}}{n+1+a+b} \,\Big|\, X_n \Big) = \frac{X_n}{n+a+b}.$$

Hence if

$$Y_n = \frac{X_n}{n+a+b}, \tag{5.5}$$

we have $\mathbb{E}(Y_{n+1}|X_n) = Y_n$, showing (Y_n) is a martingale. Note that Y_n is just the proportion of blacks balls after n steps, and thus we see that this proportion is a "fair game" of the Pólya urn.

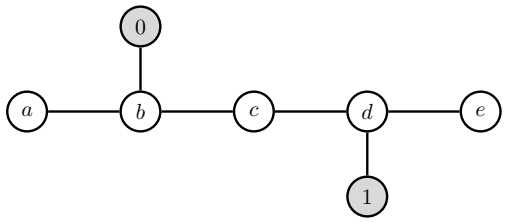

Figure 5.1: A graph G with seven vertices. Defining a martingale for the simple random walk on G can give information about the hitting probabilities of the shaded vertices.

Exercise 5.1. *Show that the proportion of red balls is also a martingale.*

Example 5.5. Consider the simple random walk $(X_n)_{n\geq 0}$ on the graph $G = (V, E)$ of Figure 5.1. Suppose we start with

$$X_0 = a \tag{5.6}$$

and we run the walk until the first time τ that we reach a gray vertex, $\tau = \min\{n \,:\, X_n \in \{0, 1\}\}$. What is $\mathbb{P}(X_\tau = 1)$? This is a "gambler's ruin" flavor of question, but now no longer for the simple walk on $\mathbb{Z}$. Martingales, it turns out, give us a slick method to answer this question.

We start by defining a function $h : V \to \mathbb{R}$ on the vertices as in Figure 5.2. That is, $h(a) = h(b) = 0.25$, $h(c) = 0.5$, $h(d) = h(e) = 0.75$, and $h(0) = 0$, $h(1) = 1$. We claim that the new process defined by the composition

$$Y_n := h(X_n) \tag{5.7}$$

is a martingale while X_n is in the white vertices $\{a, b, c, d, e\}$. From (5.7) it is clear that Y_n is adapted (once we know X_n we know the value of Y_n), and we just need to check the averaging property. Suppose, for instance, that $X_n = b$. Then

$$\begin{aligned}\mathbb{E}(Y_{n+1} \,|\, X_0, X_1, \ldots, X_n) &= \mathbb{E}(h(X_{n+1}) \,|\, X_n = b) \\ &= \frac{1}{3}(h(a) + h(0) + h(c)) \qquad (5.8) \\ &= \frac{1}{3}(0.25 + 0 + 0.5) = 0.25 = h(b) = h(X_n) = Y_n.\end{aligned}$$

Hence the averaging property holds here, and we can similarly check that it

holds at all the vertices a, b, c, d and e (as you should verify). Thus $(Y_n)_{n\geq 0}$ is a martingale so long as $n < \tau$.*

The power of martingales becomes evident when we observe the following: taking expectations in the equation $\mathbb{E}(Y_{n+1} \,|\, X_0, X_1, \ldots, X_n) = \mathbb{E}(Y_n)$ yields

$$\mathbb{E}(Y_{n+1}) = \mathbb{E}(\mathbb{E}(Y_{n+1} \,|\, X_0, X_1, \ldots, X_n)) = \mathbb{E}(\mathbb{E}(Y_n)) = \mathbb{E}(Y_n).$$

Since we similarly have $\mathbb{E}(Y_n \,|\, X_0, X_1, \ldots, X_{n-1}) = \mathbb{E}(Y_{n-1})$, we again take expectations and see $\mathbb{E}(Y_n) = \mathbb{E}(Y_{n-1})$, and hence that $\mathbb{E}(Y_{n+1}) = \mathbb{E}(Y_n) = \mathbb{E}(Y_{n-1})$. Repeating this over and over shows that the expectations are constant (on average, the value of our martingale is always the same!), and so, in particular,

$$\mathbb{E}(Y_{n+1}) = \mathbb{E}(Y_0) = h(X_0) = h(a) = 0.25 \tag{5.9}$$

by (5.6), whenever $X_n \in \{a, b, c, d, e\}$. But eventually X_{n+1} will be either vertex 0 or 1 for the first time (i.e. $\tau = n + 1$), and so we have

$$\begin{aligned}\mathbb{E}(Y_{n+1}) = \mathbb{E}(h(X_\tau)) &= h(0)\mathbb{P}(X_\tau = 0 \,|\, X_0 = a) + h(1)\mathbb{P}(X_\tau = 1 \,|\, X_0 = a) \\ &= 0 + 1 \cdot \mathbb{P}(X_\tau = 1 \,|\, X_0 = a),\end{aligned}$$

by definition of the random variable $h(X_\tau)$. However, since $\mathbb{E}(Y_{n+1}) = 0.25$ by (5.9), we have answered our question: $\mathbb{P}(X_\tau = 1 \,|\, X_0 = a) = 0.25 = h(a)$. This martingale enables us to immediately compute the hitting probability of vertex 1 before 0, starting from a; it is simply the initial value Y_0.†

Similarly, we can repeat this argument for starting at other vertices to see that $h(v) = \mathbb{P}(X_\tau = 1 \,|\, X_0 = v)$ for all $v \in \{a, b, c, d, e\}$. So a martingale solves the gambler's ruin problem for our new graph! Our new martingale $h(X_n)$ shows that our hitting probabilities are the values of h in Figure 5.2

*Note that this last restriction is necessary: the averaging property does not hold at vertices 0 and 1. For example, if $X_n = 0$, then $X_{n+1} = b$, and so

$$\mathbb{E}(Y_{n+1} | X_n = 0) = h(b) = 0.25 \neq 0 = h(0) = Y_n.$$

So we only have a martingale so long as the walk remains in the non-shaded vertices.

†While we know $\mathbb{E}(Y_n) = \mathbb{E}(Y_0)$ for any fixed, *deterministic* time n, τ is a *random* time, and so there is a slight issue in immediately concluding $\mathbb{E}(Y_0) = \mathbb{E}(Y_\tau)$. This is indeed the case, though, and we will make this argument rigorous below in §5.5.

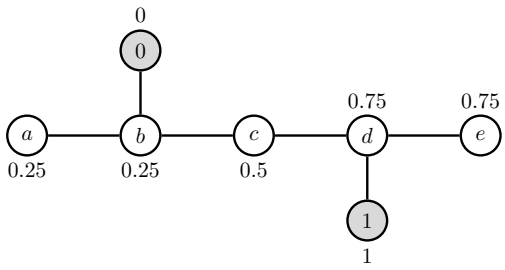

Figure 5.2: Defining a function h on the vertices that yields a martingale $Y_n = h(X_n)$. We conclude from the martingale property that the probability we hit vertex 1 before 0 starting from vertex x is precisely $h(x)$.

as our starting position X_0 varies. (Note these values pass the sanity check that the probabilities for a and b must be equal, as well as those for d and e (why?). Furthermore, by symmetry, we should have $1/2$ probability if we start at c.)

But where on earth did this function h come from? The answer is that we had to build a function that satisfies the averaging property in (5.8) at each of $\{a, b, c, d, e\}$. That leads to system of five linear equations with a unique solution, which then yields the martingale $Y_n = h(X_n)$. Functions with this averaging property are called **harmonic**, and we will study them in much greater depth in §5.3 below.

The graph in this example is relatively simple, but this martingale idea extends to *any* connected graph, and thus gives a *universal* approach for the generalized gambler's ruin hitting probability problem. We thus obtain an exciting generalization of the work we did in §2.1.1.

5.1.3 Martingale expectations

The argument in Example 5.5 which yielded $\mathbb{E}(Y_n) = \mathbb{E}(Y_0)$ for all n works for any martingale, as it only uses the averaging property (5.2). We formalize this as a useful theorem statement.

Theorem 5.1. *Martingales have constant expectation and mean-zero conditional increments. That is, if $(Y_n)_{n\geq 0}$ is a martingale for $(X_n)_{n\geq 0}$, then for all $n \geq 0$,*

$$\mathbb{E}(Y_n) = \mathbb{E}(Y_0), \tag{5.10}$$

and furthermore

$$\mathbb{E}(Y_{n+1} - Y_n \,|\, X_0, \ldots, X_n) = 0. \tag{5.11}$$

Notice that both statements of the following theorem are very consistent with our intuition of martingales as fair games.

Proof. The argument for (5.10) is as above in Example 5.5: take expectations in (5.2) and use induction. To see (5.11), observe that

$$\begin{aligned}\mathbb{E}(Y_{n+1} - Y_n \,|\, X_0, \ldots, X_n) &= \mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) - \mathbb{E}(Y_n \,|\, X_0, \ldots, X_n) \\ &= Y_n - \mathbb{E}(Y_n \,|\, X_0, \ldots, X_n)\end{aligned}$$

since (Y_n) is a martingale. In the remaining expectation, we are given $X_0, \ldots, X_n$, and since $Y_n = f_n(X_0, \ldots, X_n)$ is adapted, this means we know the value of Y_n itself. Hence $\mathbb{E}(Y_n \,|\, X_0, \ldots, X_n) = Y_n$, and the above difference is zero. □

Exercise 5.2. *Show that we also have "orthogonality of martingale increments." That is, if (Y_n) is a martingale and $n_1 < n_2 \leq n_3 < n_4$, then*

$$\mathbb{E}\big((Y_{n_4} - Y_{n_3})(Y_{n_2} - Y_{n_1})\big) = 0.$$

Let's close by seeing what Theorem 5.1 says about our example martingales from Section 5.1.2.

- In Example 5.2 we saw that the simple random walk (X_n) on $\mathbb{Z}$ starting from 0 is itself a martingale, and thus it has constant expectation

  $$\mathbb{E}(X_n) = \mathbb{E}(X_0) = 0$$

  for any n. Of course this is not the only way to see this, as we could argue from symmetry, write X_n as a sum of n i.i.d. random variables, or attempt an explicit computation using the pdf in Lemma 2.5.

- Similarly, we also see that (perhaps surprisingly) the process $Y_n = X_n^2 - n$ in Example 5.3 has constant average

  $$\mathbb{E}(Y_n) = \mathbb{E}(Y_0) = \mathbb{E}(X_0^2 - 0) = 0.$$

Hence $\mathbb{E}(X_n^2) = n$ for all n, and we thus have the nice consequence that the simple random walk after n steps has variance

$$\mathrm{Var}(X_n) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = n - 0 = n.$$

Again, this is nothing profound, as X_n is the sum of n i.i.d. random variables, each with variance 1. The point is that we get this with almost no work from the martingale.

- For the Pólya urn martingale (Y_n) of Example 5.4, we have

$$\mathbb{E}(Y_n) = \mathbb{E}(Y_0) = \frac{a}{a+b},$$

the initial proportion of black balls. Here the consequence is quite surprising: the mean proportion of black (or red) balls in the Pólya urn is always the same.

Exercise 5.3. *Explain what Theorem 5.1 says about the martingale $h(X_n)$ of Example 5.13.*

5.2 Adapted processes, martingales from eigenvalues and eigenvectors, and the Markov operator P

Now that we have an initial taste for martingales through some preliminary examples and properties, we delve further into the theory. We first re-visit adapted processes and see an important example as well as a non-example. Next, we ask if there are any reliable methods for constructing martingales. We will answer affirmatively and see, perhaps surprisingly, that we can use the eigenvalues and eigenvectors of the probability transition matrix to build them. (We will see even more robust methods later in this chapter, via harmonic functions, but this gives us a preliminary method.) Finally, we step back and define the "Markov operator," which averages a fixed function on our vertices over one step of the chain, yielding a new function. The Markov operator naturally arises when we consider the averaging property (5.2) in the definition of a martingale.

5.2.1 More on adapted processes

Martingales are a very special subclass of all adapted processes Y_n to (X_n), and one should keep in mind that not all adapted processes are martingales, and furthermore that not all processes (Y_n) are even adapted.

Example 5.6. We first consider an example of an adapted process which is not a martingale. Consider a proper subset $D \subsetneq \Omega$ and let τ be the hitting time of D^c,

$$\tau = \tau_{D^c} = \min\{n \geq 0 \,:\, X_n \in D^c\}, \tag{5.12}$$

and now define the process $(Y_n)_{n\geq 0}$ by

$$Y_n = \mathbb{1}_{\{\tau \leq n\}}$$

First, we show (Y_n) is adapted. This is intuitively clear, because the hitting time (5.12) only depends on the first n states, and so Y_n also only depends on these states; knowledge of $(X_0, X_1, \ldots, X_n)$ determines Y_n. For a formal proof, we verify the definition by constructing the functions f_n in (5.1). Indeed, writing

$$\mathbb{1}_{D^c}(x) = \begin{cases} 1 & x \in D^c, \\ 0 & x \in D, \end{cases}$$

we then have

$$\begin{aligned} Y_n &= \mathbb{1}_{\{X_0 \in D^c \text{ or } X_1 \in D^c \text{ or } \cdots \text{ or } X_n \in D^c\}} \\ &= \max\{\mathbb{1}_{D^c}(X_0), I_{D^c}(X_1), \ldots, I_{D^c}(X_n)\} \\ &= f_n(X_0, X_1, \ldots, X_n), \end{aligned}$$

proving (Y_n) is adapted.

However, (Y_n) is not a martingale. One way to see this is by writing

$$\begin{aligned} &\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) \\ &\qquad = \mathbb{E}(\mathbb{1}_{\{\tau \leq n+1\}} \,|\, X_0, \ldots, X_n) \\ &\qquad = \mathbb{P}(\tau \leq n+1 \,|\, X_0, \ldots, X_n) \\ &\qquad = \mathbb{1}_{\{X_0 \in D^c \text{ or } \cdots \text{ or } X_n \in D^c\}} + \mathbb{1}_{\{X_0 \in D, \ldots, X_n \in D\}} \mathbb{P}(X_{n+1} \in D^c \,|\, X_n). \end{aligned}$$

Since Y_n equals just the first term in this sum, we do not always have

$$\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) = Y_n,$$

and so (Y_n) is not a martingale.

Exercise 5.4. *Use Theorem 5.1 to give another argument for why the process (Y_n) in Example 5.6 is not a martingale.*

It is also helpful for the intuition to have negative examples in mind, as well.

Example 5.7. Another process we could consider is

$$Z_n := \inf\{m > n \ : \ X_m \in D\},$$

the *next* time that our walk lies within D. Is this adapted process? While (Z_n) is a function of (X_n) and n, it clearly does not only depend on the first n steps of the walk. We need to know future steps of the walk to determine Z_n. Thus $(Z_n)_{n\geq 0}$ is not adapted, and since martingales are always adapted, in particular it is not a martingale.

5.2.2 Martingales from eigenvalues and eigenvectors

Given an irreducible Markov chain on $\Omega = \{1, 2, \ldots, N\}$ with $N \times N$ transition matrix P, recall from Theorem 3.6 that the eigenspace for eigenvalue $\lambda_1 = 1$ of P is one-dimensional, and that all the other eigenvalues λ_j satisfy $|\lambda_j| < 1$. Moreover, we know a left eigenvector of λ_1 is π,

$$\pi P = \pi,$$

and a right eigenvector is the column vector $\mathbf{1}^T = (1, 1, \ldots, 1)^T$,

$$P\mathbf{1}^T = \mathbf{1}^T.$$

(Note this is an example of different left- and right-eigenspaces for the same eigenvalue when π is not uniform.) The only use we've seen for the other eigenvalues, so far, is that the spectral gap $\gamma^* = 1 - \max_{2\leq n\leq N} |\lambda_j|$ gives information about the mixing time for the chain; if the second-largest eigen-

value is close to 1 in absolute value, the chain is "close" to irreducible and mixing takes longer, as we saw in §3.2.

As we alluded to in the introduction to this section, though, another use for the other eigenvalues and eigenvectors is to construct martingales, as laid out in the following theorem.

Theorem 5.2. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with transition matrix P. If $\lambda \neq 0$ is a eigenvalue of P with right-eigenvector v^T, then the process*

$$Y_n = \lambda^{-n} v(X_n) \tag{5.13}$$

is a martingale.

So the setting is that we have a non-zero row vector $v \in \mathbb{R}^N$ which satisfies $Pv^T = \lambda v^T$. The notation $v(X_n)$ means that we choose the X_n-component of v (or of v^T - of course either way gives us the same component). The theorem says that when we consider the random component $v(X_n)$ of our eigenvector, scaled by λ^{-n}, we obtain a martingale. The scaling factor λ^{-n} is entirely deterministic, as λ is a fixed eigenvalue, and so the only random part in the definition of Y_n is which component of the eigenvector we consider.

It will be instructive to consider an example before diving into the proof.

Example 5.8. Recall the Ehrenfest urn model of §2.2: we have B identical balls split up between two urns, and at each step we pick one of the B balls uniformly at random and switch its urn. If $X_n \in \{0, 1, \ldots, B\}$ is the number of balls in the first urn, then we have the transition probabilities

$$P(k, k+1) = \frac{B-k}{B}, \qquad P(k, k-1) = \frac{k}{B}, \tag{5.14}$$

and $P(k, j) = 0$ when $j \notin \{k-1, k+1\}$. For example, if $B = 5$, our 6×6 transition matrix P is

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/5 & 0 & 4/5 & 0 & 0 & 0 \\ 0 & 2/5 & 0 & 3/5 & 0 & 0 \\ 0 & 0 & 3/5 & 0 & 2/5 & 0 \\ 0 & 0 & 0 & 4/5 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{5.15}$$

Can we find an eigenvector-eigenvalue pair for P to build a martingale? Consider the vector $v = (B, B-2, B-4, \ldots, -B)$, which entry-wise is $v(k) = B - 2k$ for $k \in \{0, 1, \ldots, B\}$. In the above case $B = 5$ above we have

$$v = (5, 3, 1, -1, -3, -5).$$

We claim that v^T is a right eigenvector for P with eigenvalue $\lambda = 1 - 2/B$.

Exercise 5.5. *Verify that this holds for the $B = 5$ case, and then verify the general case.*

We highly suggest you do this computation first on your own, but let us walk through it so you can check your work. We show the claimed identity $Pv^T = \lambda v^T$ entry-wise. Indeed, for the first row of P, we have

$$(Pv^T)(1) = 1 \cdot v^T(2) = B - 2 = \Big(1 - \frac{2}{B}\Big)B = \lambda v^T(1),$$

as needed. The computation for the last row $N = B + 1$ is similar; for a general row $k \in \{2, \ldots, B\}$, we use (5.14) and obtain

$$\begin{aligned}(Pv^T)(k) &= \frac{k}{B} \cdot v^T(k-1) + \frac{B-k}{B} \cdot v^T(k+1) \\ &= \frac{k}{B} \cdot (B - 2(k-1)) + \frac{B-k}{B} \cdot (B - 2(k+1)) \\ &= \frac{1}{B}(B^2 - 2B(k+1) + 4k) = \Big(1 - \frac{2}{B}\Big)(B - 2k) = \lambda v^T(k),\end{aligned}$$

as claimed. We conclude v^T is indeed a right eigenvector for P with eigenvalue $1 - 2/B$.

Now we can build our martingale. According to Theorem 5.2, the process

$$Z_n := \lambda^{-n} v(X_n) = \Big(1 - \frac{2}{B}\Big)^{-n}(B - 2X_n) \tag{5.16}$$

is a martingale (why we are calling it Z_n instead of Y_n will be apparent in a moment). Note that

$$v(X_n) = B - 2X_n = B - X_n - X_n \tag{5.17}$$

is the difference of the number of balls in the second urn and the first. Does that look familiar? We saw in Problem 3.8 that the process $v(X_n)$, there

called Y_n, satisfies

$$\mathbb{E}(v(X_{n+1}) \,|\, v(X_n)) = \mathbb{E}(v(X_n) \,|\, X_n) = \Big(1 - \frac{2}{B}\Big) v(X_n), \tag{5.18}$$

where the second equality holds because knowing $v(X_n)$ is equivalent to knowing X_n by (5.17). We can use this to verify that Z_n is, indeed, a martingale. We observe

$$\begin{aligned}
\mathbb{E}(Z_{n+1}|X_0, \ldots, X_n) &= \mathbb{E}(Z_{n+1} \,|\, X_n) \\
&= \Big(1 - \frac{2}{B}\Big)^{-(n+1)} \mathbb{E}(v(X_{n+1}) \,|\, X_n) \\
&= \Big(1 - \frac{2}{B}\Big)^{-(n+1)} \Big(1 - \frac{2}{B}\Big) v(X_n) \\
&= \Big(1 - \frac{2}{B}\Big)^{-n} v(X_n) = Z_n.
\end{aligned}$$

Here for the first equality we use the Markov property, for the second we use the definition of Z_n in (5.16), and for the third we use (5.18). We conclude (Z_n) has the needed averaging property and so is indeed a martingale.

Of course, once we know v^T is a right eigenvector of P with eigenvalue λ, we immediately know that the process defined by (5.16) is a martingale. The point here is that we can also verify this by hand using our earlier computations, which is satisfying and helps make everything more believable.

This example illustrates the power of Theorem 5.2. If you are "playing" the Ehrenfest urn with your friends (undoubtedly a common occurrence), how would you create a fair game? The theorem says that the difference in balls between the two urns, scaled by $(5/3)^n$ if there are 5 total balls, is a fair game. This is not super intuitive and would probably be hard to arrive at entirely on your own.

Exercise 5.6. *Explain what Theorem 5.1 says about the martingale* (5.16).

Having an example under our belts, we proceed with the proof of Theorem 5.2.

Proof. We recall our process is $Y_n = \lambda^{-n} v(X_n)$, where λ is an eigenvalue of P with corresponding right eigenvector v^T. First, note that it is clear that (Y_n) is adapted: given X_n, we know what Y_n is because we have λ and v.

To check the averaging property (5.2), we compute

$$\begin{aligned}\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) &= \mathbb{E}(\lambda^{-(n+1)} v(X_{n+1}) \,|\, X_0, \ldots, X_n) \\ &= \lambda^{-(n+1)} \mathbb{E}(v(X_{n+1}) \,|\, X_0, \ldots, X_n) \\ &= \lambda^{-(n+1)} \mathbb{E}(v(X_{n+1}) \,|\, X_n) \qquad (5.19) \\ &= \lambda^{-(n+1)} \sum_{j=1}^{N} v(j) P(X_n, j),\end{aligned}$$

where we used the Markov property in (5.19). But notice that the sum in the last line is the dot product of v^T with the row of P corresponding to X_n, which is exactly the X_n-entry of the matrix-vector product Pv^T. Since v^T is an eigenvector, we thus have

$$\begin{aligned}\lambda^{-(n+1)} \sum_{j=1}^{N} v(j) P(X_n, j) &= \lambda^{-(n+1)} (\lambda v^T)(X_n) \\ &= \lambda^{-n} v^T(X_n) = Y_n.\end{aligned}$$

We conclude (Y_n) is a martingale. □

5.2.3 The Markov operator P

Given the averaging property of martingales, it is probably not hard to believe that we will frequently encounter steps like (5.19) (if you didn't read that proof, you should do so now). The following formalism will help. Let $f : \Omega \to \mathbb{R}$ be a function on our sample space. We can simply think of f as a row vector in $\mathbb{R}^N$, $f = (f_1, f_2, \ldots, f_N)$, with f_j the value of f at state j. Given a chain (X_n) with transition matrix P, we define the **Markov operator**

$$P : \mathbb{R}^N \to \mathbb{R}^N \qquad (5.20)$$

as

$$\begin{aligned}(Pf)(x) :=& \mathbb{E}(f(X_{n+1}) \,|\, X_n = x) \qquad (5.21) \\ =& \sum_{j=1}^{N} P(x, j) f_j \qquad (5.22)\end{aligned}$$

$$=(Pf^T)(x), \tag{5.23}$$

the x-entry of the matrix-vector product Pf^T. Here (5.21) is the definition of the Markov operator, and (5.22) is the computation of the conditional expectation using the transition matrix.

So what is happening? We see the Markov operator takes a starting function f and outputs a new function Pf on our space, where the value of Pf at each state x is the P-average (5.22) of the values of f of all adjacent states in the chain. In other words, the Markov operator is averaging out our function according to P. If we start at state x and win f_j dollars each time we reach state j, $Pf(x)$ computes our expected one-step winnings.

Example 5.9. Suppose we have the function $f = (-2, 0, 3, 6, 1, 1)$ on the Ehrenfest urn with $B = 5$ balls. This means, for instance, that if there are two balls in the first urn, we "win" \$3. If there are no balls, we lose \$2. Applying the Markov operator means multiplying by the transition matrix P,

$$Pf = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/5 & 0 & 4/5 & 0 & 0 & 0 \\ 0 & 2/5 & 0 & 3/5 & 0 & 0 \\ 0 & 0 & 3/5 & 0 & 2/5 & 0 \\ 0 & 0 & 0 & 4/5 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -2 \\ 0 \\ 3 \\ 6 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 3.6 \\ 2.2 \\ 5 \\ 1 \end{bmatrix},$$

and we have averaged f according to P. So, for example, if we start with four balls in the first urn, we expect to win \$5 in one step.

Exercise 5.7. *What happens if you keep on averaging? Compute P^2f, P^3f, P^4f (use a computer if needed). Does P^nf appear to have a limit as $n \to \infty$?*

So concretely, the Markov operator is rather simple, as it is just the matrix multiplication of P with $f^T \in \mathbb{R}^N$. We can also view it abstractly, though, as a function on functions (!), since its input is a function f and its output a new function Pf. We will interchangeably switch between these perspectives, and you should keep both in mind. We observe that the concrete perspective immediately gives the following lemma.

Lemma 5.3. *The Markov operator P is linear. That is, given two functions $f, g : \Omega \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$, the function $P(\alpha f + \beta g)$ on Ω is the same as the function $\alpha(Pf) + \beta(Pg)$ on Ω.*

Proof. Since we can consider the function $P(\alpha f + \beta g)$ as the column vector $P(\alpha f^T + \beta g^T)$, the result is clear by the linearity of matrix multiplication. □

We conclude with the following remark as an aside. If P is a linear operator on *functions* via right-multiplication, what objects does it naturally act upon via left-multiplication? Recall that if $\mu = (q_1, \ldots, q_N)$ is a probability distribution on Ω, then so is μP. More concretely, if $X_0 \sim \mu$, then μP is the distribution of X_1 (review the beginning of §1.6 if you are rusty here). Thus, while P takes functions f to new functions Pf through right-multiplication, it is a linear operator on *probability distributions* on Ω through left-multiplication.

5.3 Harmonic functions

We saw in the previous section how to construct martingales using eigenvalues and eigenvectors of P. In this section we begin developing another method, using *harmonic functions*. Harmonic functions have a rich structure and we will spend the remainder of the chapter studying them and their relation to martingales.

5.3.1 Space-time harmonic and harmonic

All of our examples of martingales so far, (5.3), (5.4), (5.5), (5.7) and (5.13), have been functions of the form

$$Y_n = f_n(X_n) = f(n, X_n). \tag{5.24}$$

That is, even though the definition of a martingale permits Y_n be a function of *all* the preceeding states $X_0, X_1, \ldots, X_n$, so far our examples have only used the most-recent value X_n. Seeing that processes of this form (5.24) are evidently common, we begin by asking when they are martingales.

It is obvious that such Y_n are adapted, and so we just have to see when they satisfy the averaging property. We compute

$$\begin{aligned}\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) &= \mathbb{E}(f_{n+1}(X_{n+1}) \,|\, X_0, \ldots, X_n) \\ &= \mathbb{E}(f_{n+1}(X_{n+1}) \,|\, X_n) \\ &= (Pf_{n+1})(X_n). \end{aligned} \tag{5.25}$$

That is, the conditional expectation is the average payout by f_{n+1} after taking one step from X_n. So in order for (Y_n) to be a martingale, we need

$$\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) = (Pf_{n+1})(X_n) = Y_n = f_n(X_n)$$

for all n and any state X_n. As an identity of functions, this says $Pf_{n+1} = f_n$, and we make the following definition.

Definition 5.2. A function $f : \mathbb{Z}_{\geq 0} \times \Omega \to \mathbb{R}$ is **space-time harmonic** if, for all $n \geq 0$,

$$Pf_{n+1} = f_n, \tag{5.26}$$

where $f_n(x) := f(n, x)$.

Note that this definition is closely related to our notion of a "fair game," as component-wise it says $Pf_{n+1}(x) = f_n(x)$ for all $x \in \Omega$. That is, if our payout at stage n and state x is $f_n(x)$, then our expected payout after taking one random step in our chain is exactly $f_n(x)$, what we began with. That is, the process $Y_n = f_n(X_n)$ is a martingale. We record this for future reference as a theorem.

Theorem 5.4. *If $f : \mathbb{Z}_{\geq 0} \times \Omega \to \mathbb{R}$ is space-time harmonic, then the process (Y_n) defined by*

$$Y_n := f(n, X_n) = f_n(X_n)$$

is a martingale.

Exercise 5.8. *Prove Theorem 5.4 by showing $\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) = Y_n$ (this suffices for the proof since (Y_n) is clearly adapted).*

Example 5.10. We have already seen many example of space-time harmonic functions.

- In Example 5.3, we saw that $X_n^2 - n$ is a martingale for the simple random walk X_n on $\mathbb{Z}$. Here the underlying space-time harmonic function is $f_n(x) = x^2 - n$. The fact that $Pf_{n+1} = f_n$ is exactly the computation carried out in that example.

- For the Pólya urn, Example 5.4, we saw that the proportion $X_n/(n + a + b)$ of black balls is a martingale. Here the space-time harmonic function is $f_n(x) = x/(n + a + b)$.

- Our first martingale, Example 5.2, was based on the elementary observation that the simple walk X_n on $\mathbb{Z}$ is itself a martingale. We can also consider this as coming from a space-time harmonic function (albeit not a very interesting one), namely, $f_n(x) = x$ for all n, x. Thus there is no dependence on n and the function is always the same, the identity function $f(x) = x$.

  We can also view Example 5.5 as having an underlying space-time harmonic function, namely $f_n(x) = h(x)$, where again there is no dependency upon n.

Exercise 5.9. *What is the underlying space-time harmonic function f_n for the Ehrenfest urn martingale Z_n of Example 5.8?*

As in the last bullet point in Example 5.10, the simplest sub-class of space-time harmonic functions are those where the sequence $f_0, f_1, \ldots$ is constant in n; that is, where f_n is just a fixed function $f : \Omega \to \mathbb{R}$ for all n. Even though it is an easier case, this turns out to correspond to an extremely important class of functions, the *harmonic* functions.

Definition 5.3. A function $h : \Omega \to \mathbb{R}$ is **harmonic** on Ω (with respect to P) if $Ph = h$, which is to say,

$$(Ph)(x) = h(x) \tag{5.27}$$

for all $x \in \Omega$. We say h is **harmonic on a subset** $D \subset \Omega$ if (5.27) holds for all $x \in D$.

Note that (5.27) says

$$h(x) = \sum_{j=1}^{N} P(x,j)h(j) = (Ph)(x). \tag{5.28}$$

That is, when we start at x and average the "pay out" $h(j)$ over one random step in our chain, the result is exactly the same as our starting value $h(x)$. We call this the **mean-value property** at x, and (5.27) says that harmonic functions are precisely those functions on Ω which have the mean-value property at all $x \in \Omega$. From the abstract operator point of view, (5.27) says h is a *fixed point* of the Markov operator P; applying the Markov operator gives the same function.

Exercise 5.10. *Interpret* (5.27) *through the linear algebra lens. What does it translate to in terms of the transition matrix P and the column vector h^T?*

Just as we can build martingales through space-time harmonic functions, Theorem 5.4, so we also can through individual harmonic functions. Composing any chain (X_n) with a harmonic function h creates a "fair game" $(h(X_n))$ since the value of a harmonic function equals its average over the next step in the chain, (5.27).

Theorem 5.5. *If (X_n) is a Markov chain on Ω and $h : \Omega \to \mathbb{R}$ is harmonic, then the process (Y_n) defined by*

$$Y_n = h(X_n)$$

is a martingale.

This theorem is already proved because it is a special case of Theorem 5.4, where the space time harmonic function f_n is just given by the harmonic function h at every step n. However, because martingales of this form are especially important, we review the steps. We hope the following is very similar to what you already did for Exercise 5.8.

Proof. By definition, (Y_n) is adapted; once we know X_n we know the value of Y_n. So we check the averaging property, noting

$$\mathbb{E}(Y_{n+1} \,|\, X_0, \ldots, X_n) = \mathbb{E}(h(X_{n+1}) \,|\, X_n)$$

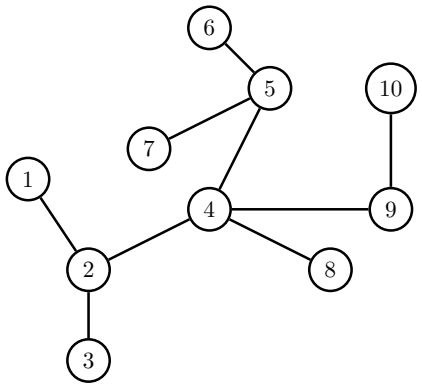

Figure 5.3: What are the harmonic functions h on this graph?

$$
\begin{aligned}
&= \sum_{j=1}^{N} P(X_n, j) h(j) \\
&= (Ph)(X_n) = h(X_n) = Y_n,
\end{aligned}
$$

where the second-to-last equality uses the harmonicity of h. We conclude that $h(X_n)$ is a martingale. □

Example 5.11. Let's concretely look at what (5.27) means for the graph in Figure 5.3, where the transition matrix P corresponds to the simple random walk on the graph. Since $P(x, j) = 1/\deg(x)$, (5.28) says that each function value $h(x)$ must equal the average of all its neighboring values. Hence,

$$
\begin{aligned}
h(1) &= h(2), \\
h(2) &= \frac{1}{3}(h(1) + h(3) + h(4)), \\
h(3) &= h(2), \\
h(4) &= \frac{1}{4}(h(2) + h(8) + h(9) + h(5)), \\
h(5) &= \frac{1}{3}(h(6) + h(7) + h(4)),
\end{aligned}
$$

and so on. One easy candidate would be a constant function, $h(x) \equiv c$ for all $x \in \Omega$ and some $c \in \mathbb{R}$. It is less obvious how to form non-constant examples, or whether such h even exist.

Exercise 5.11. *Construct a non-constant harmonic function h for the 3-cycle with matrix P given by the simple random walk, or show that no such h exists.*

What you see in Exercise 5.11 is characteristic of all harmonic functions for irreducible Markov chains.

Theorem 5.6. *If P is the transition matrix for an irreducible Markov chain on $\Omega = \{1, \ldots, N\}$ and h is harmonic with respect to P, then h is constant.*

Proof. We give the proof idea for when the chain is a simple random walk on a graph $G = (V, E)$. Suppose h is harmonic on G, and let x_0 be a vertex such that

$$h(x_0) = M := \max_{x \in V} h(x). \tag{5.29}$$

By harmonicity, we also have

$$h(x_0) = \frac{1}{\deg(x_0)} \sum_{y \sim x_0} h(y) \tag{5.30}$$

$$\leq \frac{1}{\deg(x_0)} \sum_{y \sim x_0} M = M, \tag{5.31}$$

since each $h(y) \leq M$ by (5.29). If we had some $y \sim x_0$ such that $h(y) < M$, then the inequality in (5.31) would be strict, yielding $M = h(x_0) < M$, a contradiction. Hence $h(y) \equiv M$ for all $y \sim x_0$. That is, h has the same (maximum) value on x_0 and all its neighbors.

Now we iterate on each of these adjacent vertices, taking further steps away from x_0: the same argument shows that h must also have the same value on all of *their* neighbors (see Figure 5.4). Since the Markov chain is irreducible, G is connected, and so after finitely-many iterations of this argument we see that $h(x) \equiv M$ for all $x \in V$. □

As the last line of the proof suggests, the irreducibility assumption in Theorem 5.6 is necessary.

Exercise 5.12. *Construct a non-constant harmonic function h on the disconnected graph in Figure 1.8.*

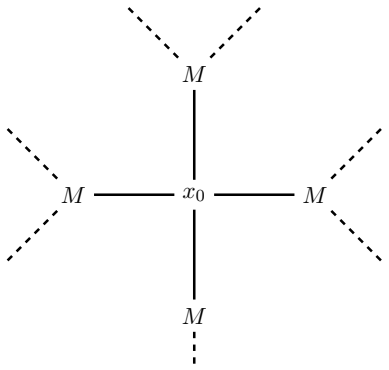

Figure 5.4: The proof idea for Theorem 5.6: since h attains its maximum M at x_0, and it is also the average of all of its neighbors, all the neighbors must likewise have value M. Then move out one step and apply this same argument to all $y \sim x_0$. Iterating shows that h is constant on Ω.

5.3.2 A bit of dirty laundry

Harmonic functions and Theorem 5.6 enable us to (finally) finish our self-contained proof that an irreducible Markov chain has a unique stationary distribution, Theorem 1.6. Recall that in Theorem 1.7 we proved that any irreducible chain has at least one stationary distribution π, given by the formula

$$\pi(x) = \frac{1}{\mathbb{E}(\tau_x^+ \,|\, X_0 = x)}. \tag{5.32}$$

We used the Perron-Frobenius theorem in §3.2.3 to prove the uniqueness of π, completing the proof of Theorem 1.6. The unsatisfying part, though, is that we did not fully prove the Perron-Frobenius theorem. Since we desire a self-contained argument, we proceed to show that any stationary distribution is, in fact, given by (5.32).

Proof of Theorem 1.6. Any stationary distribution ρ is a left-eigenvector of P for $\lambda_1 = 1$, $\rho P = \rho$. So it suffices to show that the dimension of the left-eigenspace L_1 for λ_1 is one, since then we must have

$$L_1 = \{c\pi \,|\, c \in \mathbb{R}\}$$

for π defined by (5.32). As $c\pi$ is only a probability measure when $c = 1$, we have uniqueness. Hence we show $\dim(L_1) = 1$.

It follows from the fact that any matrix is similar to its transpose that the dimension of L_1 is the same as the dimension of the *right* eigenspace R_1 for $\lambda_1 = 1$. If $v^T \in R_1$, then

$$Pv^T = v^T,$$

and v^T is thus harmonic on Ω with respect to P (compare Exercise 5.10). Hence by Theorem 5.6, v^T is a constant function/vector. In other words,

$$v^T = \begin{pmatrix} c \\ c \\ \vdots \\ c \end{pmatrix} = c \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \text{span}\{\mathbf{1}^T\},$$

and R_1 is one-dimensional. □

Note again the difference between the left- and right-eigenspaces for $\lambda = 1$ with respect to our transition matrix P. The left eigenspace is spanned by the stationary distribution π for P, whereas the right eigenspace consists of the harmonic functions on all of Ω (which are all constant functions by Theorem 5.6).

5.4 Harmonic functions on subsets $D \subsetneq \Omega$

We learned in the previous section that we can build martingales using space-time harmonic and harmonic functions. Namely, if f_n is space-time harmonic for our chain, then $Y_n := f_n(X_n)$ is a martingale, and if h is harmonic, then $Z_n := h(X_n)$ is also a martingale.

You might be wondering why even bother to mention that $h(X_n)$ is a martingale, when the only harmonic functions on an irreducible chain are constant by Theorem 5.6. The reason is that if we relax the condition that h is harmonic on *all* of Ω and instead just focus on a proper subset $D \subsetneq \Omega$, the situation dramatically changes. We study such harmonic functions in this section and the next section. We will see this class of harmonic functions is quite rich, and we will work towards classifying all such harmonic functions.

Recall that we say $h : \Omega \to \mathbb{R}$ is harmonic on $D \subsetneq \Omega$ if it has the mean-value property

$$(Ph)(x) = \sum_{j=1}^{N} P(x,j)h(j) = h(x) \tag{5.33}$$

for all $x \in D$. In particular, note that we have no requirement on the values of h at states y outside of D. Even though h is defined for $y \in D^c$, no averaging property needs to hold there.

Example 5.12. To see that this makes a difference, consider the simple random walk on the 4-cycle in Figure 5.5, and take D to be vertices 1 and 2. That is, h only must satisfy the mean-value property at vertices 1 and 2. We (arbitrarily) assign the values $h(3) = 1$ and $h(4) = 0$ for the two vertices outside of D. What values does h take on D? From (5.33), we must have

$$\begin{aligned} h(1) &= \frac{1}{2}(h(2) + h(4)) = \frac{1}{2}h(2), \\ h(2) &= \frac{1}{2}(h(1) + h(3)) = \frac{1}{2}(h(1) + 1). \end{aligned}$$

Solving this linear system yields the (unique) values $h(1) = 1/3$ and $h(2) = 2/3$. So our function h is harmonic on D but is no longer constant, even though the 4-cycle is irreducible.

It is also instructive to note that the mean value property *does not* hold at vertices $y \in \Omega \backslash D = \{3, 4\}$. Indeed, we find the average at $x = 3$ is

$$\frac{1}{2}(h(2) + h(4)) = \frac{1}{3} \neq 1 = h(3),$$

and at $x = 4$ we have

$$\frac{1}{2}(h(1) + h(3)) = \frac{2}{3} \neq 0 = h(4).$$

This is not any problem, though, as we only asked h to be harmonic on D.

This example immediately raises a number of questions. How can we think about such harmonic functions? Is there any meaning to the specific values for h in this example? Do we have a method for building other harmonic functions on D? Can we describe all such functions? We spend

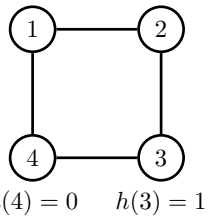

Figure 5.5: Set $D = \{1, 2\}$ and assign values 0 and 1 for h at $x = 4$ and $x = 3$, respectively. What are h's values at vertices 1 and 2, if h is harmonic there?

the remainder of this section and the next section answering these questions.

As a staring point of our analysis, write Ω as the disjoint union $\Omega = D \cup D^c$, and note that we can partition the transition matrix P in terms of vertices in D and vertices outside of D,

$$P = \begin{array}{c} \\ D \\ D^c \end{array} \begin{array}{c} \begin{array}{cc} D & D^c \end{array} \\ \begin{pmatrix} A & B \\ C & Q \end{pmatrix} \end{array}$$

If $\#\Omega = N$ and $\#D = m$, then, for instance, A is the $m \times m$ submatrix of probabilities for starting and ending in D, and B is the $m \times (N - m)$ submatrix of probabilities for starting in D and leaving D.

Now consider tweaking P to

$$\tilde{P} = \begin{array}{c} \\ D \\ D^c \end{array} \begin{array}{c} \begin{array}{cc} D & D^c \end{array} \\ \begin{pmatrix} A & B \\ \mathbf{0} & I \end{pmatrix}. \end{array} \tag{5.34}$$

Here $\mathbf{0}$ is a $(N-m) \times m$ matrix of all zeros, and I is the $(N-m) \times (N-m)$ identity matrix. It is not immediately obvious why we would want to do this, but we claim that $\tilde{P}$ is still a transition matrix. Clearly $\tilde{P}(x, y) \geq 0$ for all $x, y \in \Omega$, and so we need to verify that each row sums to one. A moment's thought shows this is obvious: if $x \in D$, then

$$\sum_{j=1}^{N} \tilde{P}(x, j) = \sum_{j=1}^{N} P(x, j) = 1,$$

while if $x \notin D$, then the only non-zero entry in row x of $\tilde{P}$ is that coming

from the identity matrix block, which is 1. That is,

$$\sum_{j=1}^{N} \tilde{P}(x,j) = \tilde{P}(x,x) = 1,$$

and so $\tilde{P}$ is indeed a transition matrix.

What is the Markov chain $\tilde{X}$ corresponding to $\tilde{P}$? Since $\tilde{P}(x,y) = P(x,y)$ for $x \in D$, $\tilde{X}$ behaves the same as X when in D. However, for $x \in D^c$,

$$\mathbb{P}(\tilde{X}_{n+1} = y \,|\, \tilde{X}_n = x) = \mathbb{1}_{\{x=y\}},$$

and so $\tilde{X}$ is permanently fixed at the first vertex it reaches outside of D. In other words, $\tilde{P}$ is the transition matrix for the chain which is *absorbed in* D^c. Another way to write this is

$$\tilde{X}_n = X_{n \wedge \tau}, \tag{5.35}$$

where $\tau = \tau_{D^c}$ is the hitting time of D^c and

$$n \wedge \tau := \min\{n, \tau\}.$$

So (5.35) says that $\tilde{X}$ is the same as X until we reach D^c, at which point $\tilde{X}$ becomes frozen in the same state. Our modified transition matrix $\tilde{P}$ thus turns each $y \in D^c$ into an absorbing boundary state for the Markov chain.

What we are really interested in, though, is harmonic functions. How do harmonic functions with respect to $\tilde{P}$ relate to harmonic functions on D for the original chain? The following lemma tells us that harmonic functions on D for P become harmonic *on all of* Ω with respect to $\tilde{P}$.

Lemma 5.7. *If h is a harmonic function on D with respect to P, then h is harmonic on all of Ω with respect to the modified transition matrix $\tilde{P}$ in* (5.34).

We encourage you to think through what is happening here through the following exercises.

Exercise 5.13. *Explain why the lemma is intuitively clear.*

Exercise 5.14. *Prove the lemma by showing* (5.33) *holds for all* $x \in \Omega$ *when* P *is replaced by* $\tilde{P}$*.*

Exercise 5.15. *We had a non-constant harmonic function* h *on* D *in Example 5.12, and Lemma 5.7 says that this* h *is harmonic on all of* Ω *with respect to* $\tilde{P}$*. Why is this not a contradiction of Theorem 5.6?*

So we now understand harmonic functions on $D \subset \Omega$ are harmonic on all of Ω for the absorbing chain $X_{n\wedge\tau}$. We still do not understand, however, how to build these harmonic functions, other than solving a linear system of equations (as in Example 5.12). Is there a more systematic approach? The following theorem says one method is to take averages of function values upon hitting D^c.

Theorem 5.8. *Let* P *be a transition matrix on* Ω *and let* $D \subsetneq \Omega$ *be a proper subset of states. Let* $\tau = \tau_{D^c}$ *be the hitting time of* D^c *and let* $f : D^c \to \mathbb{R}$ *be any function. Then the function* $h : \Omega \to \mathbb{R}$ *defined as*

$$h(x) := \mathbb{E}_x(f(X_\tau)), \qquad x \in \Omega, \tag{5.36}$$

is harmonic with respect to P *on* D *and satisfies* $h(x) = f(x)$ *for all* $x \in D^c$*.*

In this context we often say that h is harmonic on D with *boundary values* f.

The intuition here is to think of f as the "pay-out" function: we start by running the chain from some $x \in D$, and we are playing the game that we "earn" $f(y)$ dollars if we first exit D at state $y \in D^c$. Theorem 5.8 says that the function $h(x)$ which gives our average earnings from starting at x is harmonic on D.

Proof. It is clear from the definition (5.36) that $h(x) = f(x)$ for $x \in D^c$, since $\tau = 0$ for such x.

To show h is harmonic elsewhere, we need to show (5.33) holds in D. Suppose first that $f : D^c \to \mathbb{R}$ is an indicator of a single vertex $y \in D^c$,

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

Then $f(X_\tau) = \mathbb{1}_y(X_\tau)$, and to show that $\mathbb{E}_x(\mathbb{1}_y(X_\tau))$ is harmonic we condition on the first step of the walk. Indeed, for $x \in D$,

$$\begin{aligned} h(x) = \mathbb{E}_x(\mathbb{1}_y(X_\tau)) &= \mathbb{P}_x(X_\tau = y) \\ &= \sum_{j=1}^{N} \mathbb{P}_x(X_\tau = y \,|\, X_1 = j)\mathbb{P}_x(X_1 = j) \\ &= \sum_{j=1}^{N} \mathbb{P}_j(X_\tau = y)P(x,j) \\ &= \sum_{j=1}^{N} \mathbb{E}_j(\mathbb{1}_y(X_\tau))P(x,j) \\ &= \sum_{j=1}^{N} h(j)P(x,j), \end{aligned}$$

which is exactly the mean-value property, showing h is harmonic on D as claimed.

For functions $f : D^c \to \mathbb{R}$ which are not necessarily indicators, we can write f as a linear combination of indicators,

$$f(z) = \sum_{j \in D^c} f(j)\mathbb{1}_j(z).$$

Then our function h is

$$h(x) = \mathbb{E}_x(f(X_\tau)) = \sum_{j \in D^c} f(j)\mathbb{E}_x(\mathbb{1}_j(X_\tau))$$

since expectation is linear. By our above argument, each of the functions $\mathbb{E}_x(\mathbb{1}_j(X_\tau))$ in this sum is harmonic on D. Since harmonic functions are closed under linear combinations (see Problem 5.3), we conclude that h is itself harmonic. □

Exercise 5.16. *Revisit the 4-cycle in Figure 5.5 with $D = \{1, 2\}$ and the boundary values f given by $f(3) = 1$ and $f(4) = 0$. Explicitly compute $\mathbb{E}_x(f(X_\tau))$ for $x \in D$. How does this compare with the function h in Example 5.12?*

Let's see Theorem 5.8 at work in a couple of examples.

Example 5.13. Consider the simple random walk (X_n) on $\{0, 1, \ldots, N\}$ with $D = \{1, \ldots, N-1\}$ and let τ be the hitting time of $D^c = \{0, N\}$. Suppose we start the walk at some $x \in D$. What is

$$\mathbb{P}_x(X_\tau = N)?$$

We already know the answer, of course, from our considerations of gambler's ruin in §2.1.1; Theorem 2.1 says $\mathbb{P}_x(X_\tau = N) = x/N$. The new insight we gain from Theorem 5.8 is that this function

$$h(x) = \mathbb{P}_x(X_\tau = N) = \mathbb{E}_x(\mathbb{1}_N(X_\tau))$$

is actually harmonic. While the theorem guarantees this, it is also easy to explicitly verify: indeed, for $x \in D$,

$$\begin{aligned}(Ph)(x) &= \frac{1}{2}h(x-1) + \frac{1}{2}h(x+1) \\ &= \frac{x-1}{2N} + \frac{x+1}{2N} \\ &= \frac{x}{N} = h(x),\end{aligned}$$

and hence $Ph = h$. Note also that

$$h(0) = 0 = \mathbb{1}_N(0) \quad \text{and } h(N) = 1 = \mathbb{1}_N(N),$$

and so h agrees with the boundary values given by $\mathbb{1}_N$ on D^c.

Example 5.14. Just as at the end of the proof of Theorem 5.8, we can use linearity to compute our expected "earnings" $\mathbb{E}_x(f(X_\tau))$ for other "pay-out" functions f on D^c. For the same chain as in Example 5.13, consider the function $f : D^c \to \mathbb{R}$ given by

$$f(0) = -10 \quad \text{and} \quad f(N) = 40.$$

That is, we lose 10 dollars if we first exit $\{1, \ldots, N-1\}$ at 0, and win 40 dollars if we exit at N. What is our expected earnings $\tilde{h}(x) = \mathbb{E}_x(f(X_\tau))$

now? We have

$$\begin{aligned}\tilde{h}(x) &= \mathbb{E}_x(f(X_\tau)) \\ &= -10 \cdot \mathbb{P}_x(X_\tau = 0) + 30 \cdot \mathbb{P}_x(X_\tau = N) \\ &= -10 \cdot (1 - h(x)) + 30 \cdot h(x) \\ &= -10\Big(1 - \frac{x}{N}\Big) + \frac{30x}{N} = -10 + \frac{40x}{N}.\end{aligned}$$

It is easy to explicitly verify that this is a harmonic function on D which satisfies $\tilde{h}(0) = -10$ and $\tilde{h}(N) = 40$.

In Problem 5.4 you will characterize all harmonic functions on intervals of the integers for the simple random walk on $\mathbb{Z}$.

5.5 Optional sampling and harmonic functions

We are well on our way to understanding all harmonic functions on $\Omega = \{1, \dots, N\}$ with respect to transition probabilities P. Functions harmonic on all of Ω for irreducible chains are simply the constant functions, Theorem 5.6. Functions harmonic on a subset $D \subsetneq \Omega$ form a much richer class, and we know that we can build such h with arbitrary "boundary values" $f : D^c \to \mathbb{R}$ through the formula (5.36), that is,

$$h(x) = \mathbb{E}_x(f(X_\tau)) \tag{5.37}$$

for $\tau = \tau_{D^c}$. Note that we also recover the constant functions with this construction when f is constant on D^c.

But does (5.37) give *all* the harmonic functions on D? Or are there others that we cannot write as the expected "pay-out" of our Markov chain at the hitting time? In this section we bring all of our tools together to answer in the negative: there are no harmonic function on D that are not of the form (5.37) for some $f : D^c \to \mathbb{R}$. This is a beautiful and important result, and is a fitting capstone to our text.

We will need an important piece of machinery, the *optional sampling theorem*, to prove this. This will be a generalization of the averaging property for martingales to random times. So while we initially introduced harmonic functions to help us study martingales, in this last section we turn the tables

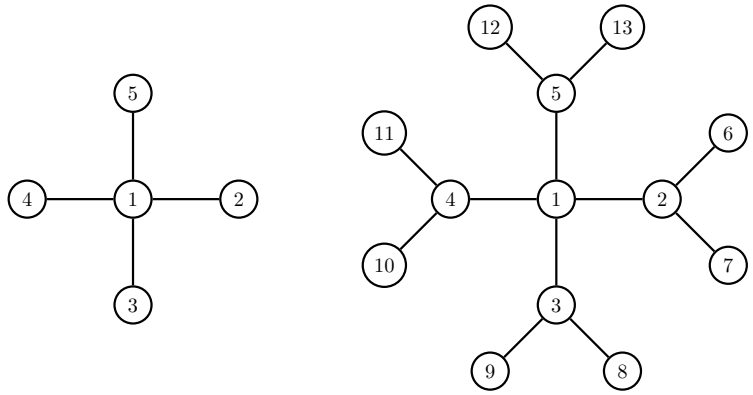

Figure 5.6: Two examples where we can explicitly see that the mean-value property implies the averaging formula $h(x) = \mathbb{E}_x(f(X_\tau))$.

and use properties of martingales to help us classify harmonic functions.

We begin by looking at how simple examples make the universality of (5.37) very plausible. The upshot will be that this is a direct consequence of the mean-value property $Ph = h$ of harmonic functions.

Example 5.15. Consider the two graphs in Figure 5.6 with transition probabilities given by the simple random walk. Start with the graph G_1 on the left, and suppose $D_1 = \{1\}$, just the center vertex. If h is harmonic on D_1, then

$$h(1) = \sum_{j=2}^{5} \frac{1}{4} h(j) = \sum_{j=2}^{5} \mathbb{P}_1(X_\tau = j) h(j) = \mathbb{E}_1(f(X_\tau))$$

if $f = h$ on $D_1^c = \{2, 3, 4\}$ is the boundary-value function. The second equality is since $\mathbb{P}_1(X_\tau = j) = 1/4$ for each $j \in D_1^c$. So we see that the mean value property immediately forces our function to be of the form (5.37).

Now expand the graph as in the right of Figure 5.6; call this G_2. Expand the set of harmonic states to include all the neighbors, $D_2 = \{1, 2, 3, 4, 5\}$, and suppose that h is any harmonic function on D_2. Then, as above,

$$h(1) = \frac{1}{4} h(2) + \frac{1}{4} h(3) + \frac{1}{4} h(4) + \frac{1}{4} h(5). \tag{5.38}$$

Since h is also harmonic at 2, we have $h(2) = \frac{1}{3}(h(1) + h(6) + h(7))$, with similar expressions at vertices 2,4 and 5. Plugging these all back into the

right-hand side of (5.38) yields

$$h(1) = \frac{4}{12}h(1) + \frac{1}{12}\sum_{j=6}^{13} h(j) \quad \Leftrightarrow \quad h(1) = \sum_{j=6}^{13} \frac{1}{8}h(j).$$

By symmetry of G_2 with reference to our starting point $x = 1$, hitting D^c at any of the vertices $6, 7, \ldots, 13$ should be equally likely. Hence the above says

$$h(1) = \sum_{j=6}^{13} \mathbb{P}_1(X_\tau = j)h(j) = \mathbb{E}_1(f(X_\tau))$$

for $f = h$ on D^c. Thus we again see that (iterations of) the mean-value property lead to (5.37).

Of course, this is no proof that all harmonic functions satisfy (5.37). And we were significantly helped by the symmetry of G_1 and G_2 in Figure 5.6 to see our expressions gave the average $\mathbb{E}_x(f(X_\tau))$. We will need a more robust approach for a general proof, provided via the *optional sampling* (or *optional stopping*) theorem.

Theorem 5.9 (Optional sampling)**.** *Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on $\Omega = \{1, \ldots, N\}$, and let $(Y_n)_{n\geq 0}$ be a martingale for (X_n). Then if τ is the hitting time of any non-empty subset $D^c \subset \Omega$,*

$$\mathbb{E}(Y_\tau) = \mathbb{E}(Y_0). \tag{5.39}$$

We saw in Theorem 5.1 that for any *fixed*, *deterministic* time n, $\mathbb{E}(Y_n) = \mathbb{E}(Y_0)$. Theorem 5.9 says something much stronger: we may replace n with a random time τ, thus obtaining the composition of random variables Y_τ, and still have the identity (5.39). To gain an appreciation for some of the subtlties involved here, we start by considering two examples.

Example 5.16. We first note that we need the assumption of irreducibility in order for the left-hand side of (5.39) to make sense. For instance, consider a simple random walk on the reducible graph of Figure 1.8 (a six-cycle plus a disconnected two-cycle). Start at a vertex in the six-cycle and let τ be the hitting time of the two-cycle. Since we will never reach the two-cycle,

$\tau = \infty$ and Y_τ is not well defined. You will show in Problem 5.10 that if (X_n) is an irreducible chain, then $\tau < \infty$ with probability one. Hence in this case it makes sense to talk about the random variable Y_τ.

Example 5.17. We secondly note that the finiteness assumption on Ω is also necessary. Indeed, consider the simple random walk (X_n) on $\mathbb{Z}$, started from $x = 0$. We saw in Example 5.2 that $Y_n = X_n$ is a martingale. Let $\tau = \tau_{\{1\}}$ be the hitting time of the vertex 1. If (5.39) held in this case, we would have

$$1 = \mathbb{E}(Y_\tau) = \mathbb{E}(Y_0) = 0,$$

which is absurd. The reason for the failure here is that the infiniteness of the state space $\mathbb{Z}$ makes τ a "nasty" stopping time.[‡] This also helps us appreciate that it is not obvious that (5.39) should hold. Unlike for deterministic times, this is not always the case, and so this is a non-trivial statement.

Proof. Our proof is based on the fact (5.11) that the conditional increments have zero mean. Note that for a deterministic time n, we can use a telescoping sum to write

$$Y_n - Y_0 = \sum_{j=1}^{n} (Y_j - Y_{j-1}) = \sum_{j=1}^{\infty} (Y_j - Y_{j-1}) \mathbb{1}_{\{j \le n\}}.$$

Similarly, for our random time τ,

$$Y_\tau - Y_0 = \sum_{j=1}^{\tau} (Y_j - Y_{j-1}) = \sum_{j=1}^{\infty} (Y_j - Y_{j-1}) \mathbb{1}_{\{j \le \tau\}}.$$

Taking expectations on both sides of the above yields

$$\mathbb{E}(Y_\tau) - \mathbb{E}(Y_0) = \sum_{j=1}^{\infty} \mathbb{E}((Y_j - Y_{j-1}) \mathbb{1}_{\{j \le \tau\}}). \tag{5.40}$$

[‡]We note that there are versions of Theorem 5.9 for infinite state spaces, but they include assumptions on the nature of τ. For instance, it suffices to have some fixed $M < \infty$ such that $\tau \le M$ with probability 1, or to have $|Y_{t \wedge \tau}| \le M$ with probability 1. That is, the hitting time is bounded or the martingale up until time τ is bounded. Both of these fail for the τ in this example: τ can be arbitrarily large, and Y_n may become very large negatively before coming back and hitting 1.

(In general, some care needs to be taken when moving an expectation through an infinite sum. Careful treatment of this is beyond the scope of our text, but the interchange is justified here.) We claim that each of the expectations in the sum is zero,

$$\mathbb{E}((Y_j - Y_{j-1})\mathbb{1}_{\{j\leq\tau\}}) = 0 \tag{5.41}$$

for all $j \geq 1$. To see this, note that

$$\mathbb{E}((Y_j - Y_{j-1})\mathbb{1}_{\{j\leq\tau\}}) = \mathbb{E}\big(\mathbb{E}((Y_j - Y_{j-1})\mathbb{1}_{\{j\leq\tau\}} \,|\, X_0, \ldots, X_{j-1})\big),$$

and so it suffices to show that

$$\mathbb{E}((Y_j - Y_{j-1})\mathbb{1}_{\{j\leq\tau\}} \,|\, X_0, \ldots, X_{j-1}) = 0$$

for each j. If we are given the values of $X_0, \ldots, X_{j-1}$, then we know the value of $\mathbb{1}_{\{j\leq\tau\}}$ (it is 1 if each $X_k \in D$, $0 \leq k \leq j-1$, and 0 otherwise) and may treat it as a constant, pulling it outside the expectation:

$$\begin{aligned}\mathbb{E}((Y_j - Y_{j-1})\mathbb{1}_{\{j\leq\tau\}} \,|\, X_0, \ldots, X_{j-1}) &= \mathbb{1}_{\{j\leq\tau\}}\mathbb{E}(Y_j - Y_{j-1} \,|\, X_0, \ldots, X_{j-1}) \\ &= \mathbb{1}_{\{j\leq\tau\}} \cdot 0 = 0\end{aligned}$$

by the conditional martingale increments, (5.11). We conclude the right-hand side of (5.40) is zero, completing the proof. □

With the optional stopping theorem in hand, we can prove that all harmonic functions are averaged "pay outs" as in (5.37).

Theorem 5.10. *If h is a function on Ω that is harmonic on $D \subsetneq \Omega$ with respect to some transition matrix P, where D^c includes states in each irreducible component of the chain, then there exists a function $f : D^c \to \mathbb{R}$ such that*

$$h(x) = \mathbb{E}_x(f(X_\tau)) \tag{5.42}$$

where $\tau = \tau_{D^c}$.

There is no mystery about what the function f is; we will see it is simply h on D^c. That is, the theorem is saying that a harmonic function on D is always equal to its expected pay-out upon leaving D!

Before the proof, a word about the technical condition on D^c: we need a state in each irreducible component because of chains like that in Figure 1.8. Indeed, suppose for that graph we start a random walk in the six-cycle and τ is the hitting time of the two-cycle. Then $\tau = \infty$ and it does not make sense to talk about X_τ, and so the right-hand side of (5.42) doesn't make sense. Our assumption on D^c in the theorem statement prevents this; D^c would have to have a vertex in both the six-cycle and the two-cycle in this case (that is, in each of the two irreducible components of the chain).

Proof. Fix an initial state $x \in \Omega$. By considering the irreducible component of the chain containing x, we see that it suffices to assume that our chain is irreducible to begin with. (For example, if x belonged to the six-cycle in Figure 1.8, we could proceed by ignoring the two-cycle, as it is impossible to enter. We would separately consider the two-cycle when we started at one of its two states.) The point is that we want to be able to apply the optional sampling theorem, and by restricting to the irreducible component containing x we may do so.

Recall from §5.4 that if h is harmonic on D for (X_n), then h is harmonic on all of Ω for the chain $(\tilde{X}_n)$ where D^c becomes an absorbing boundary, defined by

$$\tilde{X}_n = X_{\tau \wedge n}.$$

Hence $Y_n := h(\tilde{X}_n)$ is a martingale, Theorem 5.5, and we may apply the optional sampling theorem. Starting the chain (X_n) at $X_0 = x \in D$, this yields

$$\begin{aligned} \mathbb{E}(Y_\tau) &= \mathbb{E}(Y_0) \\ &= \mathbb{E}_x(h(\tilde{X}_0)) = \mathbb{E}_x(h(X_0)) = \mathbb{E}_x(h(x)) = h(x), \end{aligned}$$

as $h(x)$ is a non-random constant. On the other hand, however,

$$\mathbb{E}(Y_\tau) = \mathbb{E}_x(h(\tilde{X}_\tau)) = \mathbb{E}_x(h(X_\tau)) = \mathbb{E}_x(f(X_\tau)),$$

where $f(y) := h(y)$ for $y \in D^c$. Putting the last two equations together yields (5.42). □

We close with two applications of Theorem 5.10. The first is a new proof

of Theorem 5.6.

Alternative proof of Theorem 5.6. Let h be a function harmonic on all of Ω; we wish to show that h is constant. Fix an $x \in \Omega$ and let y be any other state; we show $h(x) = h(y)$. Indeed, we have that $Y_n := h(X_n)$ is a martingale, and for $D^c = \{y\}$, $\tau = \tau_{D^c}$, the optional sampling theorem gives

$$h(y) = \mathbb{E}_x(h(X_\tau)) = \mathbb{E}_x(Y_\tau) = \mathbb{E}_x(Y_0) = \mathbb{E}_x(h(X_0)) = h(x),$$

and hence h is constant. □

The second application is that harmonic functions with given "boundary values" f on a subset D^c are unique. That is, the boundary values of a harmonic function determine its values everywhere. Initially this seems like a very strong statement, but in light of (5.42) we see that it has to be the case. This formula says that the boundary values "communicate inwards," and thus determine all the other values of the harmonic function.

Theorem 5.11. *Let h_1 and h_2 be functions on Ω that are harmonic on $D \subsetneq \Omega$. If $h_1(y) = h_2(y)$ for all $y \in D^c$, then $h_1 = h_2$.*

Proof. We need to show $h_1(x) = h_2(x)$ for all $x \in D$. Since $X_\tau \in D^c$, (5.42) immediately gives

$$h_1(x) = \mathbb{E}_x(h_1(X_\tau)) = \mathbb{E}_x(h_2(X_\tau)) = h_2(x).$$ □

It is instructive to revisit Example 5.12 and Exercise 5.16 in light of Theorem 5.11. In the exercise, you were asked to compute $\mathbb{E}_x(f(X_\tau))$ where f has the same values on D^c as h in Example 5.12. So by Theorem 5.11, we immediately know these are the same functions. In the example we found $h(1) = 1/3$, and hence

$$\frac{1}{3} = \mathbb{E}_1(f(X_\tau)) = 0 \cdot \mathbb{P}_1(X_\tau = 4) + 1 \cdot \mathbb{P}_1(X_\tau = 3) = \mathbb{P}_1(X_\tau = 3).$$

Similarly, we had $h(2) = 2/3$, and the same logic shows that $\mathbb{P}_2(X_\tau = 3) = 2/3$. We can thus see that harmonic functions give us a powerful method to attack the gambler's ruin problem of §2.1.1 for general graphs, using boundary values of 0 and 1 and (5.42).

Problems for chapter 5

Problem 5.1. Let $(X_n)_{n\geq 0}$ be a simple random random walk on $\mathbb{Z}$ with $X_0 = 0$.

(a) Show that the process

$$Y_n := X_n^3 - 3nX_n$$

is a martingale. This is the *cubic martingale* for the simple random walk.

(b) Show that

$$Y_n := e^{\alpha X_n - \frac{\alpha^2}{2}n}$$

is a martingale, where $\alpha \in \mathbb{R}$ is a constant. This process (Y_n) is the *exponential martingale* for the simple random walk.

(c) For any $n \in \mathbb{N}$ and $t \in \mathbb{R}$, compute the moment generating function $M_{X_n}(t) = \mathbb{E}(e^{tX_n})$ for the simple walk after n steps.

Problem 5.2. Fix a proper subset $D \subsetneq \Omega$ of the state space Ω of a chain and let $\tau = \tau_D$ be the first hitting time of D. We saw in Examples 5.6 and 5.7 that the processes defined by $Y_n := \mathbb{1}_{\{\tau \leq n\}}$ and $Z_n := \inf\{m > n \ : \ X_m \in D\}$ are both not martingales. Show that we can fix this by taking our "best guess" for these random variables with the information on hand through step n. That is, show:

(i) The process $(\tilde{Y}_n)_{n\geq 0}$ defined by $\tilde{Y}_n := \mathbb{P}(\tau \leq n \mid X_0, \ldots, X_n)$ is a martingale.

(ii) The process $(\tilde{Z}_n)_{n\geq 0}$ defined by

$$\tilde{Z}_n := \mathbb{E}(\inf\{m > n \ : \ X_m \in D\} \mid X_0, \ldots, X_n)$$

is a martingale.

Problem 5.3. Suppose $h_1, \ldots, h_n$ are harmonic functions on $D \subset \Omega$ with respect to P. Show that, for any constants $c_1, \ldots, c_n$, the linear combination $\sum_{j=1}^n c_j h_j$ is likewise harmonic on D.

Problem 5.4. In this problem you will characterize all harmonic functions on an interval of integers $D = \{M, M+1, \ldots, M+N\}$ for the simple random walk on $\mathbb{Z}$.

(*a*) Show that any linear function $h(x) = mx + b$, where $m, b \in \mathbb{R}$, is harmonic on D. Explain intuitively why this is the case.

(*b*) Conversely, show that if h is harmonic on D, then there exist $m, b \in \mathbb{R}$ such that $h(x) = mx + b$. Hence all harmonic functions for the simple random walk on an interval of $\mathbb{Z}$ are linear.

Comment: Note that the arguments here do not depend on the interval $\{M, M+1, \ldots, M+N\}$ in question. That is, linear functions are harmonic on *all* of $\mathbb{Z}$. If you've been reading the text carefully, this might sound a bit suspicious: didn't we say in Theorem 5.6 that functions h that are harmonic on all of Ω are constant? The key here is that $\mathbb{Z}$ is not a *finite* state space, as we were assuming in Theorem 5.6 (and is usually the case for our Markov chains). In the proof of that theorem, it is instructive to note that we started by choosing a state x_0 where h achieves a maximum. That isn't necessarily possible if Ω is infinite, as for $\mathbb{Z}$ and (non-constant) linear functions h! So there is no contradiction between that theorem and this problem.

Problem 5.5. Consider the following Markov chain on the state space $\Omega = \{2^k,\ k = \ldots, -2, -1, -0, 1, 2, \ldots\}$. If your current position is $x = 2^k$, then the next step you jump to $2x = 2^{k+1}$ or $x/2 = 2^{k-1}$ with equal probabilities.

(*a*) Let $D = \{2^{-N}, 2^{-N+1}, \ldots, 1, 2, \ldots, 2^N\}$, for some $N \geq 1$. Suppose you absorb the chain when it hits D^c. Show that $f(x) = \log(x)$ is a harmonic function at any x in D.

(*b*) Let τ denote the hitting time of D^c. Find $\mathrm{E}_2(\tau)$ (i.e., the expected hitting time of D^c starting from 2).

Problem 5.6. Consider a connected graph $G = (V, E)$ and a function $f : V \to \mathbb{R}$. Assume that the random walk on this graph is aperiodic. Define a sequence of functions $\{f_0, f_1, \ldots\}$ on the vertices in the following way. Start with $f_0(v) = f(v)$ for all vertices v. Then, inductively,

$$f_{k+1}(v) = \frac{1}{\deg(v)} \sum_{y \sim v} f_k(y), \quad k = 0, 1, 2, \ldots.$$

That is, at each step replace the value of the function at any vertex by the average value of all its neighbors in the previous step. Find $\lim_{k\to\infty} f_k(v)$ for each v.

Problem 5.7. For the graph G in Figure 5.7, let $D = \{1, 2, 3, 4, 5\}$ and let $\tau = \tau_{D^c}$ be the hitting time of D^c for the simple random walk (X_n) on G.

(*a*) Find $\mathbb{P}_x(X_\tau = 8)$ for all vertices $x \in D$.

(*b*) Find $\mathbb{P}_x(X_\tau \in \{6, 7, 9, 10\})$ for all vertices $x \in D$.

(*c*) Find $\mathbb{P}_x(X_\tau \in \{6, 7, 8\})$ for all vertices $x \in D$.

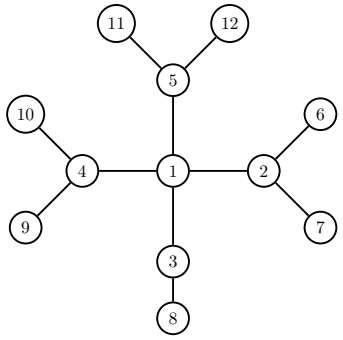

Figure 5.7: Find the hitting probabilities for certain subsets of the boundary.

Problem 5.8. Use optional sampling for the simple random walk (X_n) on integers $\{a, a+1, \ldots, b\}$ to re-derive the gambler's ruin hitting probabilities of Theorem 2.2.

Problem 5.9. Use optional sampling and the quadratic martingale $Y_n = X_n^2 - n$ for the simple random walk on integers $\{a, a+1, \ldots, b\}$ to re-derive the gambler's ruin expected hitting times of Theorem 2.4.

Problem 5.10. Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain on $\Omega = \{1, \ldots, N\}$, and let τ be the hitting time of any non-empty subset $D^c \subset \Omega$. Start the chain at some $x \in \Omega$. Show that, with probability 1, $\tau < \infty$.

Problem 5.11. A casino is designing a new game called "The Gauntlet," and, knowing your expertise in Markov chains, hires you as a consultant. The game is based on the graph in Figure 5.8 and works as follows. The

player starts at S, and in the first step immediately moves right one step. Then they roll a die. If they roll a 1,2 or 3, they move one step left. Getting a 4,5 or 6 moves them one step right. The same rule applies whenever they are at a vertex with only two neighbors. If they are at a vertex with three neighbors, then a roll of 1 or 2 moves them left, a 3 or 4 moves them vertically, and a 5 or 6 moves them right. If they ever return to S, they automatically go right one step in the next turn.

The game is free to play. It ends when the player first reaches a vertex marked L or W. If they first reach W, they win \$1,000,000. If they first reach an L vertex, they must pay the casino some amount of money x.

(a) What is the probability of winning?

(b) The casino would like to make \$1, on average, each time someone plays this game. As the consultant, what do you tell them for the value of x?

(c) The casino is worried that if x is too large then the game will be unpopular. They think it will make a big difference if they only want to earn \$0.01 on average for this game. What should the new value of x be here? Does it make a big difference?

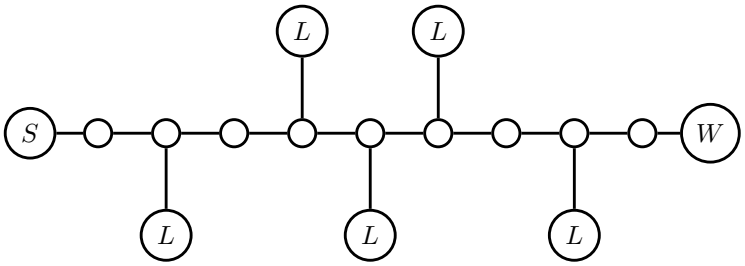

Figure 5.8: *The Gauntlet* casino game.

Appendix 1: Notation

Notation	*Description*	*First location in text*
$\mathbb{N}$	The natural numbers $\{1, 2, 3, \ldots\}$	Theorem 1.1
$\mathbb{Z}$	The integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$	Example 1.4
iid	Independent and identically distributed random variables. That is, $X_1, X_2, \ldots$ are iid if they are mutually independent and each X_j has the same distribution.	Review problem
$\text{Exp}(\lambda)$	The exponential λ distribution. $X \sim \text{Exp}(\lambda)$ if X has pdf $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.	Review problem
$\text{Geo}(p)$	A geometric random variable X with parameter p. X is the number of trials necessary to obtain the first success, where the probability of success in each trial is p. X has pmf $p_X(k) = (1-p)^{k-1}p$, $k \in \mathbb{N}$.	Example 1.11
$\text{Unif}(a, b)$	The uniform distribution on the interval (a, b) for some $a < b$. $X \sim \text{Unif}(a, b)$ if X has pdf $f(x) = \frac{1}{b-a}$ for $x \in (a, b)$.	Review problem
$:=$	The symbol on the left is *defined* by what follows on the right. For example, $\mathbb{P}_{ij} := \mathbb{P}(X_1 = j \,\vert\, X_0 = i)$.	§1.1.2
$u \sim v$	Vertices u and v in a graph are adjacent.	§1.1.2
$P_{ij}, P(i, j)$	The probability of moving from state i to j.	
$\hat{P}_{ij}, \hat{P}(i, j)$	The probability of moving from state i to j in the time-reversed Markov chain.	(1.58)

τ_A, τ_x	The hitting time of A and x, respectively. This is the first non-negative time $k \geq 0$ at which the Markov chain satisfies $X_k \in A$ or $X_k = x$, respectively	(1.33)
τ_A^+, τ_x^+	The return time of A and x, respectively, or the first positive time $k \geq 1$ that the Markov chain is in the set A or state x.	(1.34)
$\mathbb{P}_x(A)$	The conditional probability $\mathbb{P}(A \,\vert\, X_0 = x)$.	(1.43)
$\mathbb{P}_\mu(A)$	The conditional probability of A, given $X_0 \sim \mu$, where μ is a probability distribution on Ω. ($X_0 \sim \mu$ means X_0 is randomly assigned to be one of the states in Ω, and X_0's pdf is μ. That is, $\mathbb{P}(X_0 = k) = \mu(k)$.)	(1.53)
$\mathbb{E}_x(A)$	The conditional expectation of A, given $X_0 = x$.	(1.44)
$\mathbb{E}_\mu(A)$	The conditional expectation of A, given $X_0 \sim \mu$, where μ is a probability distribution on Ω.	(1.54)
$\partial\Omega$	The boundary of the set Ω. If $\Omega = \{0, 1, \ldots, n\}$, then $\partial\Omega = \{0, n\}$.	(2.1)
$a \wedge b$	The minimum $\min\{a, b\}$ of a and b	(5.35)

Appendix 2: Suggested homework sets

The following are suggestions for (approximately) weekly homework sets. Problems numbered 6.x are from the review problems in Appendix 3.

Note that there also a number of exercises within the text of each chapter. Students should be encouraged to do *all* of these as they are reading to ensure they are internalizing the concepts. Although they vary in difficulty, most of these exercises are elementary and give an opportunity to immediately engage with the new ideas.

- Homework set 1: Problems 6.1, 6.3, 1.1, 1.3, 1.4, 1.6.
- Homework set 2: Problems 6.2, 6.5, 1.2, 1.5, 1.7, 1.8, 1.9.
- Homework set 3: Problems 1.10 to 1.14 and 1.25.
- Homework set 4: Problem 1.17 and 2.1 to 2.4.
- Homework set 5: Problems 1.20, 1.24, 2.5, 2.6 and 2.7
- Homework set 6: Problems 1.21, 1.22, 2.8, 3.7 and 3.8
- Homework set 7: Problems 3.1 to 3.6

Appendix 3: Review problems

The following problems help you brush up on familiar distributions, conditional expectations, and linear algebra.

Problem 6.1. Let N be a geometric random variable with parameter p. That is, the p.m.f. of N is given by

$$\mathbb{P}(N = n) = (1-p)^{n-1}p, \quad n = 1, 2, 3, \ldots.$$

Given $N = n$, generate n many i.i.d. exponentials with rate 1: $X_1, X_2, \ldots, X_n \sim$ Exp(1). Let $W = \min\{X_1, X_2, \ldots, X_n\}$ be their minimum.

(*a*) Find the conditional c.d.f. of W by computing $\mathbb{P}(W > t \mid N = n)$, n=1,2,....

(*b*) Use part (*a*) to compute $\mathbb{P}(W > t)$.

(*c*) Compute the conditional probability mass function of N, given $\{W > t\}$, and identify it as one of the named distributions (binomial, Poisson, geometric, exponential, normal etc.) and specify all the parameters.

Problem 6.2. Let N be a Poisson random variable with mean λ. Given $N = n$, generate $X_1, X_2, \ldots, X_n, X_{n+1}$ i.i.d. Exp(1) random variables.

(*a*) Let $Y = \min\{X_1, \ldots, X_{n+1}\}$ when $N = n$. Compute $P(Y > a)$ for all $a > 0$.

(*b*) Find $E(Y)$. (*Hint:* conditional expectation and the tower property.)

Problem 6.3. Let X be the number of rolls of a fair die until I see the first six. Given $X = x$, I choose a sample, with replacement, of size x from an urn with 5 red balls and 4 green balls. Let Y be the number of green balls in my sample.

(*a*) Find $\mathbb{E}(Y \mid X = x)$.

(*b*) Use part (*a*) to find $\mathbb{E}(Y)$.

Problem 6.4. Take a stick of length one and denote it by the unit interval $[0, 1]$. Pick two i.i.d. Unif$(0, 1)$ random variables U and V and mark them on the stick. Break the stick at those two marks so that you get three broken pieces of the stick. Choose one these three pieces uniformly at random. Let L be the length of the chosen piece. What is the expected value of L?

Problem 6.5. Start with an urn with a red ball and a black ball. At each turn, pick a ball at random from the urn, return it to the urn along with another ball of the same color.

(*a*) Find the conditional probability that the second ball picked is red, given that the third ball picked is red.

(*b*) Find the expected number of red balls in the urn after the fourth turn (i.e., when there are six balls in the urn).

Bibliography

[1] Lawler, G. F. *Introduction to stochastic processes*, 2nd ed. ed. Chapman & Hall/CRC, Boca Raton, 2006.

[2] Levin, D. A., and Peres, Y. *Markov chains and mixing times*, second ed. American Mathematical Society, Providence, Rhode Island, 2017.