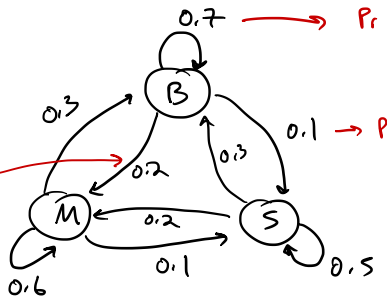
Markov Chains

B: Bertucci's
M: Margaritas
S: Sato



3 reds ⇒ must add to 1

Stationary Distribution:
long-term fraction of time spent visiting each state.

B, M, S

$\vec{\pi} = \langle \pi_B, \pi_M, \pi_S \rangle$

$\pi_B + \pi_M + \pi_S = 1$

state
Transition Matrix

$$P = \begin{array}{c} \\ B \\ M \\ S \end{array} \begin{array}{c} \begin{array}{ccc} B & M & S \end{array} \\ \begin{pmatrix} .7 & .2 & .1 \\ .3 & .6 & .1 \\ .3 & .2 & .5 \end{pmatrix} \end{array}$$

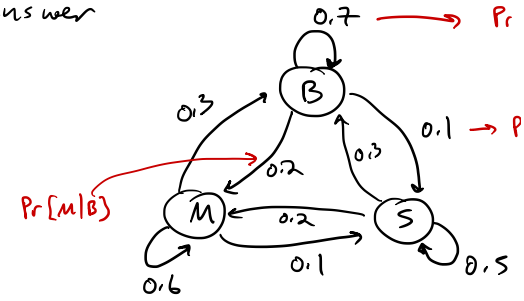
Stochastic matrix ⇒ all rows sum to 1

$B = 0.7 \cdot B + 0.3 \cdot M + 0.3 \cdot S$

Finding the s.d.:

① Simulation
- highly inefficient

② Create equations & solve for exact answer



① $B = .7B + .3M + .3S$

② $M = .2B + .6M + .2S$

③ $S = .1B + .1M + .5S$

But I have one more equation:

④ $B + M + S = 1$

① $.3B - .3M - .3S = 0$

②+③ $.3B - .3M - .3S = 0$

$B + M + S = 1$
$M = .2B + .6M + .2S$
$S = .1B + .1M + .5S$

$\Rightarrow$

$B + M + S = 1$ ①
$-.2B + .4M - .2S = 0$ ②
$-.1B - .1M + .5S = 0$ ③

① $B + M + S = 1$

5×② $-B + 2M - S = 0$ $\Rightarrow$

10×③ $-B - M + 5S = 0$

$B + M + S = 1$

$3M = 1 \Rightarrow M = 1/3$

$6S = 1 \Rightarrow S = 1/6$

$\Rightarrow B = 1/2$

2 MARKOV CHAINS

Let us begin with a simple example. We consider a "random walker" in a very small town consisting of four streets, and four street-corners v_1, v_2, v_3 and v_4 arranged as in Figure 1. At time 0, the random walker stands in corner v_1. At time 1, he flips a fair coin and moves immediately to v_2 or v_4 according to whether the coin comes up heads or tails. At time 2, he flips the coin again to decide which of the two adjacent corners to move to, with the decision rule that if the coin comes up heads, then he moves one step clockwise in Figure 1, while if it comes up tails, he moves one step counterclockwise. This procedure is then iterated at times 3, 4, ...

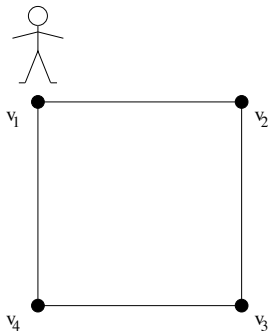


Figure 1: A random walker in a very small town.

For each n, let X_n denote the index of the street-corner at which the walker stands at time n. Hence, $(X_0, X_1, \ldots)$ is a random process taking values in $\{1, 2, 3, 4\}$. Since the walker starts at time 0 in v_1, we have

$$\mathbf{P}(X_0 = 1) = 1\,. \tag{6}$$

Next, he will move to v_2 or v_4 with probability $\frac{1}{2}$ each, so that

$$\mathbf{P}(X_1 = 2) = \frac{1}{2} \tag{7}$$

and

$$\mathbf{P}(X_1 = 4) = \frac{1}{2}\,. \tag{8}$$

To compute the distribution of X_n for $n \geq 2$ requires a little more thought; you will be asked to do this in Problem 2.1 below. To this end, it is useful to consider conditional probabilities. Suppose that at time n, the walker stands at, say, v_2. Then we get the conditional probabilities

$$\mathbf{P}(X_{n+1} = v_1 \,|\, X_n = v_2) = \frac{1}{2}$$

and

$$\mathbf{P}(X_{n+1} = v_3 \,|\, X_n = v_2) = \frac{1}{2}\,,$$

because of the coin-flipping mechanism for deciding where to go next. In fact, we get the same conditional probabilities if we condition further on the full history of the process up to time n, i.e.,

$$\mathbf{P}(X_{n+1} = v_1 \,|\, X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = v_2) = \frac{1}{2}$$

and

$$\mathbf{P}(X_{n+1} = v_3 \,|\, X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = v_2) = \frac{1}{2}$$

for any choice of $i_0, \ldots, i_{n-1}$. (This is because the coin-flip at time $n+1$ is independent of all previous coin-flips, and hence also independent of $X_0, \ldots, X_n$.) This phenomenon is called the **memoryless property**, also known as the **Markov property**: the conditional distribution of X_{n+1} given $(X_0, \ldots, X_n)$ depends only on X_n. Or in other words: to make the best possible prediction of what happens "tomorrow" (time $n+1$), we only need to consider what happens "today" (time n), as the "past" (times $0, \ldots, n-1$) gives no additional useful information[2].

Another interesting feature of this random process is that the conditional distribution of X_{n+1} given that $X_n = v_2$ (say) is the same for all n. (This is because the mechanism that the walker uses to decide where to go next is the same at all times.) This property is known as **time homogeneity**, or simply **homogeneity**.

These observations call for a general definition:

Definition 2.1 *Let P be a $(k \times k)$-matrix with elements $\{P_{i,j} : i, j = 1, \ldots, k\}$. A random process $(X_0, X_1, \ldots)$ with finite state space $S = \{s_1, \ldots, s_k\}$ is said to be a* **(homogeneous) Markov chain with transition matrix** *P, if for all n, all $i, j \in \{1, \ldots, k\}$ and all $i_0, \ldots, i_{n-1} \in \{1, \ldots k\}$ we have*

$$\begin{aligned} &\mathbf{P}(X_{n+1} = s_j \,|\, X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_{n-1} = s_{i_{n-1}}, X_n = s_i) \\ &\qquad = \mathbf{P}(X_{n+1} = s_j \,|\, X_n = s_i) \\ &\qquad = P_{i,j}\,. \end{aligned}$$

The elements of the transition matrix P are called transition probabilities. The transition probability $P_{i,j}$ is the conditional probability of being in state s_j "tomorrow" given that we are in state s_i "today". The term "homogeneous" is often dropped, and taken for granted when talking about "Markov chains".

[2]Please note that this is just a property of this particular mathematical model. It is *not* intended as a general advice that we should "never worry about the past". Of course, we have every reason, in daily life as well as in politics, to try to learn as much as we can from history in order to make better decisions for the future!

For instance, the random walk example above is a Markov chain, with state space $\{1,\ldots,4\}$ and transition matrix

$$P = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix} . \tag{9}$$

Every transition matrix satisfies

$$P_{i,j} \geq 0 \text{ for all } i,j \in \{1,\ldots,k\}, \tag{10}$$

and

$$\sum_{j=1}^{k} P_{i,j} = 1 \text{ for all } i \in \{1,\ldots,k\}. \tag{11}$$

Property (10) is just the fact that conditional probabilities are always nonnegative, and property (11) is that they sum to 1, i.e.,

$$\mathbf{P}(X_{n+1} = s_1 \,|\, X_n = s_i) + \mathbf{P}(X_{n+1} = s_2 \,|\, X_n = s_i) + \cdots + \mathbf{P}(X_{n+1} = s_k \,|\, X_n = s_i) = 1 .$$

We next consider another important characteristic (besides the transition matrix) of a Markov chain $(X_0, X_1, \ldots)$, namely the **initial distribution**, which tells us how the Markov chain starts. The initial distribution is represented as a row vector $\mu^{(0)}$ given by

$$\begin{aligned} \mu^{(0)} &= (\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}) \\ &= (\mathbf{P}(X_0 = s_1), \mathbf{P}(X_0 = s_2), \ldots, \mathbf{P}(X_0 = s_k)) . \end{aligned}$$

Since $\mu^{(0)}$ represents a probability distribution, we have

$$\sum_{i=1}^{k} \mu_i^{(0)} = 1 .$$

In the random walk example above, we have

$$\mu^{(0)} = (1, 0, 0, 0) \tag{12}$$

because of (6).

Similarly, we let the row vectors $\mu^{(1)}, \mu^{(2)}, \ldots$ denote the distributions of the Markov chain at times $1, 2, \ldots$, so that

$$\begin{aligned} \mu^{(n)} &= (\mu_1^{(n)}, \mu_2^{(n)}, \ldots, \mu_k^{(n)}) \\ &= (\mathbf{P}(X_n = s_1), \mathbf{P}(X_n = s_2), \ldots, \mathbf{P}(X_n = s_k)) . \end{aligned}$$

For the random walk example, equations (7) and (8) tell us that

$$\mu^{(1)} = (0, \tfrac{1}{2}, 0, \tfrac{1}{2})\,.$$

It turns out that once we know the initial distribution $\mu^{(0)}$ and the transition matrix P, we can compute all the distributions $\mu^{(1)}, \mu^{(2)}, \ldots$ of the Markov chain. The following result tells us that this is simply a matter of matrix multiplication. We write P^n for the n^{th} power of the matrix P.

Theorem 2.1 *For a Markov chain $(X_0, X_1, \ldots)$ with state space $\{s_1, \ldots, s_k\}$, initial distribution $\mu^{(0)}$ and transition matrix P, we have for any n that the distribution $\mu^{(n)}$ at time n satisfies*

$$\mu^{(n)} = \mu^{(0)} P^n\,. \tag{13}$$

Proof: Consider first the case $n = 1$. We get, for $j = 1, \ldots, k$, that

$$\begin{aligned}
\mu_j^{(1)} &= \mathbf{P}(X_1 = s_j) = \sum_{i=1}^{k} \mathbf{P}(X_0 = s_i, X_1 = s_j) \\
&= \sum_{i=1}^{k} \mathbf{P}(X_0 = s_i)\mathbf{P}(X_1 = s_j \,|\, X_0 = s_i) \\
&= \sum_{i=1}^{k} \mu_i^{(0)} P_{i,j} = (\mu^{(0)} P)_j
\end{aligned}$$

where $(\mu^{(0)}P)_j$ denotes the j^{th} element of the row vector $\mu^{(0)}P$. Hence $\mu^{(1)} = \mu^{(0)}P$.

To prove (13) for the general case, we use induction. Fix m, and suppose that (13) holds for $n = m$. For $n = m + 1$, we get

$$\begin{aligned}
\mu_j^{(m+1)} &= \mathbf{P}(X_{m+1} = s_j) = \sum_{i=1}^{k} \mathbf{P}(X_m = s_i, X_{m+1} = s_j) \\
&= \sum_{i=1}^{k} \mathbf{P}(X_m = s_i)\mathbf{P}(X_{m+1} = s_j \,|\, X_m = s_i) \\
&= \sum_{i=1}^{k} \mu_i^{(m)} P_{i,j} = (\mu^{(m)} P)_j
\end{aligned}$$

so that $\mu^{(m+1)} = \mu^{(m)}P$. But $\mu^{(m)} = \mu^{(0)}P^m$ by the induction hypothesis, so that

$$\mu^{(m+1)} = \mu^{(m)}P = \mu^{(0)}P^m P = \mu^{(0)}P^{m+1}$$

and the proof is complete. □

Let us consider some more examples:

Example 2.1: The Gothenburg weather. It is sometimes claimed that the best way to predict tomorrow's weather[3] is simply to guess that it will be the same tomorrow as it is today. If we assume that this claim is correct[4], then it is natural to model the weather as a Markov chain. For simplicity, we assume that there are only two kinds of weather: rain and sunshine. If the above predictor is correct 75% of the time (regardless of whether today's weather is rain or sunshine), then the weather forms a Markov chain with state space $S = \{s_1, s_2\}$ (with $s_1 =$"rain" and $s_2 =$"sunshine") and transition matrix

$$P = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}.$$

Example 2.2: The Los Angeles weather. Note that in Example 2.1, there is a perfect symmetry between "rain" and "sunshine", in the sense that the probability that today's weather will persist tomorrow, is the same regardless of today's weather. This may be reasonably realistic in Gothenburg, but not in Los Angeles where sunshine is much more common than rain. A more reasonable transition matrix for the Los Angeles weather might therefore be (still with $s_1 =$"rain" and $s_2 =$"sunshine")

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{bmatrix}. \tag{14}$$

A useful way to picture a Markov chain is its so-called **transition graph**. The transition graph consists of nodes representing the states of the Markov chain, and arrows between the nodes, representing transition probabilities. This is explained easiest by just showing the transition graphs of the examples considered so far.

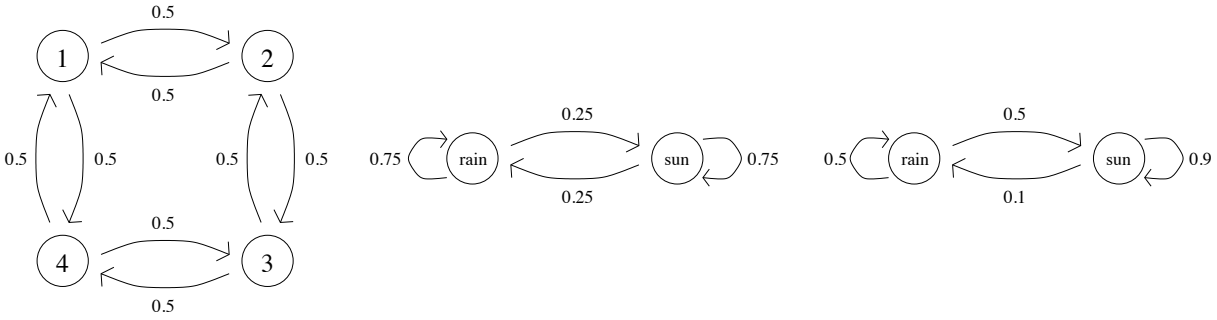


Figure 2: Transition graphs for the random walker in Figure 1, and for Examples 2.1 and 2.2.

In all examples above, as well as in Definition 2.1, the "rule" for obtaining X_{n+1} from X_n did not change with time. In some situations, it is more realistic, or for other reasons more

[3]Better than watching the weather forecast on TV.

[4]I doubt it.

desirable[5], to let this rule change with time. This brings us to the topic of **inhomogeneous Markov chains**, and the following definition, which generalizes Definition 2.1.

Definition 2.2 *Let $P^{(1)}, P^{(2)}, \ldots$ be a sequence $(k \times k)$-matrices, each of which satisfies* (10) *and* (11). *A random process $(X_0, X_1, \ldots)$ with finite state space $S = \{s_1, \ldots, s_k\}$ is said to be an* **inhomogeneous Markov chain with transition matrices** $P^{(1)}, P^{(2)}, \ldots$, *if for all n, all $i, j \in \{1, \ldots, k\}$ and all $i_0, \ldots, i_{n-1} \in \{1, \ldots k\}$ we have*

$$\begin{aligned} \mathbf{P}(X_{n+1} = s_j \,|\, X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_{n-1} = s_{i_{n-1}}, X_n = s_i) \\ &= \mathbf{P}(X_{n+1} = s_j \,|\, X_n = s_i) \\ &= P_{i,j}^{(n+1)}. \end{aligned}$$

Example 2.3: A refined model for the Gothenburg weather. There are of course many ways in which the crude model in Example 2.1 can be made more realistic. One way is to take into account seasonal changes: it does not seem reasonable to disregard whether the calendar says "January" or "July" when predicting tomorrow's weather. To this end, we extend the state space to $\{s_1, s_2, s_3\}$, where s_1 ="rain" and s_2 ="sunshine" as before, and s_3 ="snow". Let

$$P_{summer} = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0.25 & 0.75 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix} \quad \text{and} \quad P_{winter} = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.15 & 0.7 & 0.15 \\ 0.2 & 0.3 & 0.5 \end{bmatrix},$$

and assume that the weather evolves according to P_{summer} in May–September, and according to P_{winter} in October–April. This is an inhomogeneous Markov chain model for the Gothenburg weather. Note that in May–September, the model behaves exactly as the one in Example 2.1, except for some possible residual snowy weather on May 1.

The following result, which is a generalization of Theorem 2.1, tells us how to compute the distributions $\mu^{(1)}, \mu^{(2)}, \ldots$ at times $1, 2, \ldots$ of an inhomogeneous Markov chain with initial distribution $\mu^{(0)}$ and transition matrices $P^{(1)}, P^{(2)}, \ldots$.

Theorem 2.2 *Suppose that $(X_0, X_1, \ldots)$ is an inhomogeneous Markov chain with state space $\{s_1, \ldots, s_k\}$, initial distribution $\mu^{(0)}$ and transition matrices $P^{(1)}, P^{(2)}, \ldots$. For any n, we then have that*

$$\mu^{(n)} = \mu^{(0)} P^{(1)} P^{(2)} \cdots P^{(n)}.$$

Proof: Follows by a similar calculation as in the proof of Theorem 2.1. □

[5]Such as in the simulated annealing algorithms of Chapter 12.

Problems

2.1 (5) Consider the Markov chain corresponding to the random walker in Figure 1, with transition matrix P and initial distribution $\mu^{(0)}$ given by (9) and (12).

(a) Compute the square P^2 of the transition matrix P. How can we interpret P^2 ? (See Theorem 2.1, or glance ahead at Problem 2.5.)

(b) Prove by induction that

$$\mu^{(n)} = \begin{cases} (0, \frac{1}{2}, 0, \frac{1}{2}) & \text{for } n = 1, 3, 5, \ldots \\ (\frac{1}{2}, 0, \frac{1}{2}, 0) & \text{for } n = 2, 4, 6, \ldots \end{cases}$$

2.2 (2) Suppose that we modify the random walk example in Figure 1 as follows. At each integer time, the random walker tosses *two* coins. The first coin is to decide whether to stay or go. If it comes up heads, he stays where he is, whereas if it comes up tails, he lets the second coin decide whether he should move one step clockwise, or one step counterclockwise. Write down the transition matrix, and draw the transition graph, for this new Markov chain.

2.3 (5) Consider Example 2.1 (the Gothenburg weather), and suppose that the Markov chain starts on a rainy day, so that $\mu^{(0)} = (1, 0)$.

(a) Prove by induction that

$$\mu^{(n)} = (\tfrac{1}{2}(1 + 2^{-n}), \tfrac{1}{2}(1 - 2^{-n}))$$

for every n.

(b) What happens with $\mu^{(n)}$ in the limit as n tends to infinity?

2.4 (6)

(a) Consider Example 2.2 (the Los Angeles weather), and suppose that the Markov chain starts with initial distribution $(\frac{1}{6}, \frac{5}{6})$. Show that $\mu^{(n)} = \mu^{(0)}$ for any n, so that in other words the distribution remains the same at all times[6].

(b) Can you find an initial distribution for the Markov chain in Example 2.1 for which we get similar behavior as in (a)? Compare this result to the one in Problem 2.3 (b).

2.5 (6) Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\{s_1, \ldots, s_k\}$ and transition matrix P. Show, by arguing as in the proof of Theorem 2.1, that for any $m, n \geq 0$ we have

$$\mathbf{P}(X_{m+n} = s_j \,|\, X_m = s_i) = (P^n)_{i,j} \,.$$

[6]Such a Markov chain is said to be in **equilibrium**, and its distribution is said to be **stationary**. This is a very important topic, which will be treated carefully in Chapter 5.

4 IRREDUCIBLE AND APERIODIC MARKOV CHAINS

For several of the most interesting results in Markov theory, we need to put certain assumptions on the Markov chains we are considering. It is an important task, in Markov theory just like in all other branches of mathematics, to find conditions that on one hand are strong enough to have useful consequences, but on the other hand are weak enough to hold (and be easy to check) for many interesting examples. In this chapter, we will discuss two such conditions on Markov chains: **irreducibility** and **aperiodicity**. These conditions are of central importance in Markov theory, and in particular they play a key role in the study of stationary distributions, which is the topic of Chapter 5. We shall, for simplicity, discuss these notions in the setting of homogeneous Markov chains, although they do have natural extensions to the more general setting of inhomogeneous Markov chains.

We begin with irreducibility, which, loosely speaking, is the property that "all states of the Markov chain can be reached from all others". To make this more precise, consider a Markov chain $(X_0, X_1, \ldots)$ with state space $S = \{s_1, \ldots, s_k\}$ and transition matrix P. We say that a state s_i **communicates** with another state s_j, writing $s_i \to s_j$, if the chain has positive probability[8] of ever reaching s_j when we start from s_i. In other words, s_i communicates with s_j if there exists an n such that

$$\mathbf{P}(X_{m+n} = s_j \,|\, X_m = s_i) > 0\,.$$

By Problem 2.5, this probability is independent of m (due to the homogeneity of the Markov chain), and equals $(P^n)_{i,j}$.

If $s_i \to s_j$ and $s_j \to s_i$, then we say that the states s_i and s_j **intercommunicate**, and write $s_i \leftrightarrow s_j$. This takes us directly to the definition of irreducibility.

Definition 4.1 *A Markov chain $(X_0, X_1, \ldots)$ with state space $S = \{s_1, \ldots, s_k\}$ and transition matrix P is said to be* **irreducible** *if for all $s_i, s_j \in S$ we have that $s_i \leftrightarrow s_j$. Otherwise the chain is said to be* **reducible**.

Another way of phrasing the definition would be to say that the chain is irreducible if for any $s_i, s_j \in S$ we can find an n such that $(P^n)_{i,j} > 0$.

An easy way to verify that a Markov chain is irreducible, is to look at its transition graph, and check that from each state there is a sequence of arrows leading to any other state. A

[8] Here and henceforth, by "positive probability", we always mean *strictly* positive probability.

quick glance at Figure 2 thus reveals that the Markov chains in Examples 2.1 and 2.2, as well as the random walk example in Figure 1, are all irreducible. Let us also have a look at an example which is *not* irreducible:

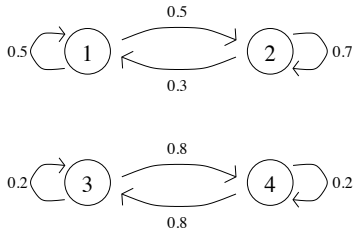


Figure 3: Transition graph for the Markov chain in Example 4.1.

Example 4.1: A reducible Markov chain. Consider a Markov chain $(X_0, X_1, \ldots)$ with state space $S = \{1, 2, 3, 4\}$ and transition matrix

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0.8 & 0.2 \end{bmatrix}.$$

By taking a look at its transition graph (see Figure 3), we immediately see that if the chain starts in state 1 or state 2, then it is restricted to states 1 and 2 forever. Similarly, if it starts in state 3 or state 4, then it can never leave the subset $\{3, 4\}$ of the state space. Hence, the chain is reducible.

Note that if the chain starts in state 1 or state 2, then it behaves exactly as if it were a Markov chain with state space $\{1, 2\}$ and transition matrix

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \end{bmatrix}.$$

If it starts in state 3 or state 4, then it behaves like a Markov chain with state space $\{3, 4\}$ and transition matrix

$$\begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}.$$

This illustrates a characteristic feature of reducible Markov chains, which also explains the term "reducible": If a Markov chain is reducible, then the analysis of its long-term behavior can be reduced to the analysis of the long-term behavior of one or more Markov chains with smaller state space.

We move on to consider the concept of aperiodicity. For a finite or infinite set $\{a_1, a_2, \ldots\}$ of positive integers, we write $\gcd\{a_1, a_2, \ldots\}$ for the greatest common divisor of $a_1, a_2, \ldots$. The **period** $\mathrm{d}(s_i)$ of a state $s_i \in S$ is defined as

$$\mathrm{d}(s_i) = \gcd\{n \geq 1 : (P^n)_{i,i} > 0\}.$$

In words, the period of s_i is the greatest common divisor of the set of times that the chain can return (i.e., has positive probability of returning) to s_i, given that we start with $X_0 = s_i$. If $\mathrm{d}(s_i) = 1$, then we say that the state s_i is **aperiodic**.

Definition 4.2 *A Markov chain is said to be* **aperiodic** *if all its states are aperiodic. Otherwise the chain is said to be* **periodic**.

Consider for instance Example 2.1 (the Gothenburg weather). It is easy to check that regardless whether the weather today is rain or sunshine, we have for any n that the probability of having the same weather n days later, is strictly positive. Or, expressed more compactly: $(P^n)_{i,i} > 0$ for all n and all states s_i.[9] This obviously implies that the Markov chain in Example 2.1 is aperiodic. Of course, the same reasoning applies to Example 2.2 (the Los Angeles weather).

On the other hand, let us consider the random walk example in Figure 1, where the random walker stands in corner v_1 at time 0. Clearly, he has to take an even number of steps in order to get back to v_1. This means that $(P^n)_{1,1} > 0$ only for $n = 2, 4, 6, \ldots$. Hence,

$$\gcd\{n \geq 1 : (P^n)_{i,i} > 0\} = \gcd\{2, 4, 6, \ldots\} = 2\,,$$

and the chain is therefore periodic.

One reason for the usefulness of aperiodicity is the following result.

Theorem 4.1 *Suppose that we have an aperiodic Markov chain $(X_0, X_1, \ldots)$ with state space $S = \{s_1, \ldots, s_k\}$ and transition matrix P. Then there exists an $N < \infty$ such that*

$$(P^n)_{i,i} > 0$$

for all $i \in \{1, \ldots, k\}$ and all $n \geq N$.

To prove this result, we shall borrow the following lemma from number theory.

Lemma 4.1 *Let $A = \{a_1, a_2, \ldots\}$ be a set of positive integers which is*

(i) *non-lattice, meaning that* $\gcd\{a_1, a_2, \ldots\} = 1$, *and*

(ii) *closed under addition, meaning that if $a \in A$ and $a' \in A$, then $a + a' \in A$.*

[9] By a variant of Problem 2.3 (a), we in fact have that $(P^n)_{i,i} = \frac{1}{2}(1 + 2^{-n})$.

Then there exists an integer $N < \infty$ such that $n \in A$ for all $n \geq N$.

Proof: See, e.g., the appendix of Brémaud [B]. □

Proof of Theorem 4.1: For $s_i \in S$, let $A_i = \{n \geq 1 : (P^n)_{i,i} > 0\}$, so that in other words A_i is the set of possible return times to state s_i starting from s_i. We assumed that the Markov chain is aperiodic, and therefore the state s_i is aperiodic, so that A_i is non-lattice. Furthermore, A_i is closed under addition, for the following reason: If $a, a' \in A_i$, then $\mathbf{P}(X_a = s_i \,|\, X_0 = s_i) > 0$ and $\mathbf{P}(X_{a+a'} = s_i \,|\, X_a = s_i) > 0$. This implies that

$$\begin{aligned} \mathbf{P}(X_{a+a'} = s_i \,|\, X_0 = s_i) &\geq \mathbf{P}(X_a = s_i, X_{a+a'} = s_i | \, X_0 = s_i) \\ &= \mathbf{P}(X_a = s_i \,|\, X_0 = s_i)\mathbf{P}(X_{a+a'} = s_i \,|\, X_a = s_i) \\ &> 0 \end{aligned}$$

so that $a + a' \in A_i$.

In summary, A_i satisfies assumptions (i) and (ii) of Lemma 4.1, which therefore implies that there exists an integer $N_i < \infty$ such that $(P^n)_{i,i} > 0$ for all $n \geq N_i$.

Theorem 4.1 now follows with $N = \max\{N_1, \ldots, N_k\}$. □

By combining aperiodicity and irreducibility, we get the following important result, which will be used in the next chapter to prove the so-called Markov chain convergence theorem (Theorem 5.2).

Corollary 4.1 *Let $(X_0, X_1, \ldots)$ be an irreducible and aperiodic Markov chain with state space $S = \{s_1, \ldots, s_k\}$ and transition matrix P. Then there exists an $M < \infty$ such that $(P^n)_{i,j} > 0$ for all $i, j \in \{1, \ldots, k\}$ and all $n \geq M$.*

Proof: By the assumed aperiodicity and Theorem 4.1, there exists an integer $N < \infty$ such that $(P^n)_{i,i} > 0$ for all $i \in \{1, \ldots, k\}$ and all $n \geq N$. Fix two states $s_i, s_j \in S$. By the assumed irreducibility, we can find some $n_{i,j}$ such that $(P^{n_{i,j}})_{i,j} > 0$. Let $M_{i,j} = N + n_{i,j}$. For any $m \geq M_{i,j}$, we have

$$\begin{aligned} \mathbf{P}(X_m = s_j \,|\, X_0 = s_i) &\geq \mathbf{P}(X_{m-n_{i,j}} = s_i, X_m = s_j \,|\, X_0 = s_i) \\ &= \mathbf{P}(X_{m-n_{i,j}} = s_i \,|\, X_0 = s_i)\mathbf{P}(X_m = s_j \,|\, X_{m-n_{i,j}} = s_i) \quad (21) \\ &> 0 \end{aligned}$$

(the first factor in (21) is positive because $m - n_{i,j} \geq N$, and the second is positive by the choice of $n_{i,j}$). Hence, we have shown that $(P^m)_{i,j} > 0$ for all $m \geq M_{i,j}$. The corollary now follows with

$$M = \max\{M_{1,1}, M_{1,2} \ldots, M_{1,k}, M_{2,1}, \ldots, M_{k,k}\}.$$

□

Problems

4.1 (3) Show that if a Markov chain is irreducible and has a state s_i such that $P_{ii} > 0$, then it is also aperiodic.

4.2 (4) **Random chess moves.**

(a) Consider a chessboard with a lonely white king making random moves, meaning that at each move, he picks one of the possible squares to move to, uniformly at random. Is the corresponding Markov chain irreducible and/or aperiodic?

(b) Same question, but with the king replaced by a bishop.

(c) Same question, but instead with a knight.

4.3 (6) **Oriented random walk on a torus.** Let a and b be positive integers, and consider the Markov chain with state space

$$\{(x, y) : x \in \{0, \ldots, a-1\}, y \in \{0, \ldots, b-1\}\},$$

and the following transition mechanism: If the chain is in state (x, y) at time n, then at time $n+1$ it moves to $((x+1) \bmod a, y)$ or $(x, (y+1) \bmod b)$ with probability $\frac{1}{2}$ each.

(a) Show that this Markov chain is irreducible.

(b) Show that it is aperiodic if and only if $\gcd(a, b) = 1$.

5 STATIONARY DISTRIBUTIONS

In this chapter, we consider one of the central issues in Markov theory: asymptotics for the long-term behavior of Markov chains. What can we say about a Markov chain that has been running for a long time? Can we find interesting limit theorems?

If $(X_0, X_1, \ldots)$ is any nontrivial Markov chain, then the value of X_n will keep fluctuating infinitely many times as $n \to \infty$, and therefore we cannot hope to get results about X_n converging to a limit. However, we may hope that the *distribution* of X_n settles down to a limit. This is indeed the case if the Markov chain is irreducible and aperiodic, which is what the main result of this chapter, the so called Markov chain convergence theorem (Theorem 5.2), says.

Let us for a moment go back to the Markov chain in Example 2.2 (the Los Angeles weather), with state space $\{s_1, s_2\}$ and transition matrix given by (14). We saw in Problem 2.4 (a) that if we let the initial distribution $\mu^{(0)}$ be given by $\mu^{(0)} = (\frac{1}{6}, \frac{5}{6})$, then this distribution is preserved for all times, i.e., $\mu^{(n)} = \mu^{(0)}$ for all n. By some experimentation, we can easily convince ourselves that no other choice of initial distribution $\mu^{(0)}$ for this chain has the same property (try it!). Apparently, the distribution $(\frac{1}{6}, \frac{5}{6})$ plays a special role for this Markov chain, and we call it a **stationary distribution**[10]. The general definition is as follows.

Definition 5.1 *Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\{s_1, \ldots, s_k\}$ and transition matrix P. A row vector $\pi = (\pi_1, \ldots, \pi_k)$ is said to be a* **stationary distribution** *for the Markov chain, if it satisfies*

(i) $\pi_i \geq 0$ *for* $i = 1, \ldots, k$, *and* $\sum_{i=1}^k \pi_i = 1$, *and*

(ii) $\pi P = \pi$, *meaning that* $\sum_{i=1}^k \pi_i P_{i,j} = \pi_j$ *for* $j = 1, \ldots, k$.

Property (i) simply means that π should describe a probability distribution on $\{s_1, \ldots, s_k\}$. Property (ii) implies that if the initial distribution $\mu^{(0)}$ equals π, then the distribution $\mu^{(1)}$ of the chain at time 1 satisfies

$$\mu^{(1)} = \mu^{(0)} P = \pi P = \pi \,,$$

and by iterating we see that $\mu^{(n)} = \pi$ for every n.

[10] Another term which is used by many authors for the same thing, is **invariant distribution.** Yet another term is **equilibrium distribution**

Since the definition of a stationary distribution really only depends on the transition matrix P, we also sometimes say that a distribution π satisfying the assumptions (i) and (ii) in Definition 5.1, is **stationary for the matrix** P (rather than for the Markov chain).

The rest of this chapter will deal with three issues: the **existence** of stationary distributions, the **uniqueness** of stationary distributions, and the **convergence** to stationarity starting from any initial distribution. We shall work under the conditions introduced in the previous chapter (irreducibility and aperiodicity), although for some of the results these conditions can be relaxed somewhat[11]. We begin with the existence issue.

Theorem 5.1 (Existence of stationary distributions) *For any irreducible and aperiodic Markov chain, there exists at least one stationary distribution.*

To prove this existence theorem, we first need to prove a lemma concerning **hitting times** for Markov chains. If a Markov chain $(X_0, X_1, \ldots)$ with state space $\{s_1, \ldots, s_k\}$ and transition matrix P starts in state s_i, then we can define the hitting time

$$T_{i,j} = \min\{n \geq 1 : X_n = s_j\}$$

with the convention that $T_{i,j} = \infty$ if the Markov chain never visits s_j. We also define the **mean hitting time**

$$\tau_{i,j} = \mathbf{E}[T_{i,j}] .$$

This means that $\tau_{i,j}$ is the expected time taken until we come to state s_j, starting from state s_i. For the case $i = j$, we call $\tau_{i,i}$ the **mean return time** for state s_i. We emphasize that when delaing with the hitting time $T_{i,j}$, there is always the implicit assumption that $X_0 = s_i$.

Lemma 5.1 *For any irreducible aperiodic Markov chain with state space $S = \{s_1, \ldots, s_k\}$ and transition matrix P, we have for any two states $s_i, s_j \in S$ that if the chain starts in state s_i, then*

$$\mathbf{P}(T_{i,j} < \infty) = 1 . \tag{22}$$

Moreover, the mean hitting time $\tau_{i,j}$ is finite[12], i.e.,

$$\mathbf{E}[T_{i,j}] < \infty . \tag{23}$$

[11] By careful modification of our proofs, it is possible to show that Theorem 5.1 holds for arbitrary Markov chains, and that Theorem 5.3 holds without the aperiodicity assumption. That irreducibility and aperiodicity are needed for Theorem 5.2, and irreducibility is needed for Theorem 5.3, will be established by means of counterexamples in Problems 5.2 and 5.3.

[12] If you think that this should follow immediately from (22), then take a look at Example 1.1 to see that things are not always quite that simple.

Proof: By Corollary 4.1, we can find an $M < \infty$ such that $(P^M)_{i,j} > 0$ for all $i, j \in \{1, \ldots, k\}$. Fix such an M, set $\alpha = \min\{(P^M)_{i,j} : i, j \in \{1, \ldots, k\}\}$, and note that $\alpha > 0$. Fix two states s_i and s_j as in the lemma, and suppose that the chain starts in s_i. Clearly,

$$\mathbf{P}(T_{i,j} > M) \leq \mathbf{P}(X_M \neq s_j) \leq 1 - \alpha\,.$$

Furthermore, given everything that has happened up to time M, we have conditional probability at least α of hitting state s_j at time $2M$, so that

$$\begin{aligned} \mathbf{P}(T_{i,j} > 2M) &= \mathbf{P}(T_{i,j} > M)\mathbf{P}(T_{i,j} > 2M \,|\, T_{i,j} > M) \\ &\leq \mathbf{P}(T_{i,j} > M)\mathbf{P}(X_{2M} \neq s_j \,|\, T_{i,j} > M) \\ &\leq (1-\alpha)^2\,. \end{aligned}$$

Iterating this argument, we get for any l that

$$\begin{aligned} \mathbf{P}(T_{i,j} > lM) &= \mathbf{P}(T_{i,j} > M)\mathbf{P}(T_{i,j} > 2M \,|\, T_{i,j} > M) \cdots \mathbf{P}(T_{i,j} > lM \,|\, T_{i,j} > (l-1)M) \\ &\leq (1-\alpha)^l\,, \end{aligned}$$

which tends to 0 as $l \to \infty$. Hence $\mathbf{P}(T_{i,j} = \infty) = 0$, so (22) is established.

To prove (23), we use the formula (1) for expectation, and get

$$\begin{aligned} \mathbf{E}[T_{i,j}] &= \sum_{n=1}^{\infty} \mathbf{P}(T_{i,j} \geq n) = \sum_{n=0}^{\infty} \mathbf{P}(T_{i,j} > n) \qquad (24) \\ &= \sum_{l=0}^{\infty} \sum_{n=lM}^{(l+1)M-1} \mathbf{P}(T_{i,j} > n) \\ &\leq \sum_{l=0}^{\infty} \sum_{n=lM}^{(l+1)M-1} \mathbf{P}(T_{i,j} > lM) = M \sum_{l=0}^{\infty} \mathbf{P}(T_{i,j} > lM) \\ &\leq M \sum_{l=0}^{\infty} (1-\alpha)^l = M \frac{1}{1-(1-\alpha)} = \frac{M}{\alpha} < \infty\,. \end{aligned}$$

□

Proof of Theorem 5.1: Write, as usual, $(X_0, X_1, \ldots)$ for the Markov chain, $S = \{s_1, \ldots, s_k\}$ for the state space, and P for the transition matrix. Suppose that the chain starts in state s_1, and define, for $i = 1, \ldots, k$,

$$\rho_i = \sum_{n=0}^{\infty} \mathbf{P}(X_n = s_i, T_{1,1} > n)$$

so that in other words, ρ_i is the expected number of visits to state i up to time $T_{1,1} - 1$. Since the mean return time $\mathbf{E}[T_{1,1}] = \tau_{1,1}$ is finite, and $\rho_i < \tau_{1,1}$, we get that ρ_i is finite as

well. Our candidate for a stationary distribution is

$$\pi = (\pi_1, \ldots, \pi_k) = \left(\frac{\rho_1}{\tau_{1,1}}, \frac{\rho_2}{\tau_{1,1}}, \ldots, \frac{\rho_k}{\tau_{1,1}} \right) .$$

We need to verify that this choice of π satisfies conditions (i) and (ii) of Definition 5.1.

We first show that the relation $\sum_{i=1}^{k} \pi_i P_{i,j} = \pi_j$ in condition (ii) holds for $j \neq 1$ (the case $j = 1$ will be treated separately). We get (hold on!)

$$\begin{aligned}
\pi_j = \frac{\rho_j}{\tau_{1,1}} &= \frac{1}{\tau_{1,1}} \sum_{n=0}^{\infty} \mathbf{P}(X_n = s_j, T_{1,1} > n) \\
&= \frac{1}{\tau_{1,1}} \sum_{n=1}^{\infty} \mathbf{P}(X_n = s_j, T_{1,1} > n) && (25) \\
&= \frac{1}{\tau_{1,1}} \sum_{n=1}^{\infty} \mathbf{P}(X_n = s_j, T_{1,1} > n-1) && (26) \\
&= \frac{1}{\tau_{1,1}} \sum_{n=1}^{\infty} \sum_{i=1}^{k} \mathbf{P}(X_{n-1} = s_i, X_n = s_j, T_{1,1} > n-1) \\
&= \frac{1}{\tau_{1,1}} \sum_{n=1}^{\infty} \sum_{i=1}^{k} \mathbf{P}(X_{n-1} = s_i, T_{1,1} > n-1) \mathbf{P}(X_n = s_j \,|\, X_{n-1} = s_i) && (27) \\
&= \frac{1}{\tau_{1,1}} \sum_{n=1}^{\infty} \sum_{i=1}^{k} P_{i,j} \mathbf{P}(X_{n-1} = s_i, T_{1,1} > n-1) \\
&= \frac{1}{\tau_{1,1}} \sum_{i=1}^{k} P_{i,j} \sum_{n=1}^{\infty} \mathbf{P}(X_{n-1} = s_i, T_{1,1} > n-1) \\
&= \frac{1}{\tau_{1,1}} \sum_{i=1}^{k} P_{i,j} \sum_{m=0}^{\infty} \mathbf{P}(X_m = s_i, T_{1,1} > m) \\
&= \frac{\sum_{i=1}^{k} \rho_i P_{i,j}}{\tau_{1,1}} = \sum_{i=1}^{k} \pi_i P_{i,j} && (28)
\end{aligned}$$

where in lines (25), (26) and (27) we used the assumption that $j \neq 1$.

Next, we verify condition (ii) also for the case $j = 1$. Note first that $\rho_1 = 1$; this is immediate from the definition of ρ_i. We get

$$\begin{aligned}
\rho_1 = 1 &= \mathbf{P}(T_{1,1} < \infty) = \sum_{n=1}^{\infty} \mathbf{P}(T_{1,1} = n) \\
&= \sum_{n=1}^{\infty} \sum_{i=1}^{k} \mathbf{P}(X_{n-1} = s_i, T_{1,1} = n)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{\infty}\sum_{i=1}^{k}\mathbf{P}(X_{n-1}=s_i, T_{1,1}>n-1)\mathbf{P}(X_n=s_1\,|\,X_{n-1}=s_i)\\
&= \sum_{n=1}^{\infty}\sum_{i=1}^{k}P_{i,1}\mathbf{P}(X_{n-1}=s_i, T_{1,1}>n-1)\\
&= \sum_{i=1}^{k}P_{i,1}\sum_{n=1}^{\infty}\mathbf{P}(X_{n-1}=s_i, T_{1,1}>n-1)\\
&= \sum_{i=1}^{k}P_{i,1}\sum_{m=0}^{\infty}\mathbf{P}(X_m=s_i, T_{1,1}>m)\\
&= \sum_{i=1}^{k}\rho_i P_{i,1}\,.
\end{aligned}
$$

Hence

$$\pi_1 = \frac{\rho_1}{\tau_{1,1}} = \sum_{i=1}^{k}\frac{\rho_i P_{i,1}}{\tau_{1,1}} = \sum_{i=1}^{k}\pi_i P_{i,1}\,.$$

By combining this with (28), we have established that condition (ii) holds for our choice of π.

It remains to show that condition (i) holds as well. That $\pi_i \geq 0$ for $i = 1, \ldots, k$ is obvious. To see that $\sum_{i=1}^{k}\pi_i = 1$ holds as well, note that

$$
\begin{aligned}
\tau_{1,1} = \mathbf{E}[T_{1,1}] &= \sum_{n=0}^{\infty}\mathbf{P}(T_{1,1}>n) \qquad (29)\\
&= \sum_{n=0}^{\infty}\sum_{i=1}^{k}\mathbf{P}(X_n=s_i, T_{1,1}>n)\\
&= \sum_{i=1}^{k}\sum_{n=0}^{\infty}\mathbf{P}(X_n=s_i, T_{1,1}>n)\\
&= \sum_{i=1}^{k}\rho_i
\end{aligned}
$$

(where equation (29) uses (24)) so that

$$\sum_{i=1}^{k}\pi_i = \frac{1}{\tau_{1,1}}\sum_{i=1}^{k}\rho_i = 1\,,$$

and condition (i) is verified. □

We shall go on to consider the asymptotic behavior of the distribution $\mu^{(n)}$ of a Markov chain with arbitrary initial distribution $\mu^{(0)}$. To state the main result (Theorem 5.2), we

need to define what it means for a sequence of probability distributions $\nu^{(1)}, \nu^{(2)}, \ldots$ to converge to another probability distribution ν, and to this end it is useful to have a metric on probability distributions. There are various such metrics; one which is useful here is the so called **total variation distance**.

Definition 5.2 *If $\nu^{(1)} = (\nu_1^{(1)}, \ldots, \nu_k^{(1)})$ and $\nu^{(2)} = (\nu_1^{(2)}, \ldots, \nu_k^{(2)})$ are probability distributions on $S = \{s_1, \ldots, s_k\}$, then we define the* **total variation distance** *between $\nu^{(1)}$ and $\nu^{(2)}$ as*

$$\mathrm{d_{TV}}(\nu^{(1)}, \nu^{(2)}) = \frac{1}{2} \sum_{i=1}^{k} |\nu_i^{(1)} - \nu_i^{(2)}| \,. \tag{30}$$

If $\nu^{(1)}, \nu^{(2)}, \ldots$ and ν are probability distributions on S, then we say that $\nu^{(n)}$ **converges to ν in total variation as** *$n \to \infty$, writing $\nu^{(n)} \xrightarrow{\mathrm{TV}} \nu$, if*

$$\lim_{n \to \infty} \mathrm{d_{TV}}(\nu^{(n)}, \nu) = 0 \,.$$

The constant $\frac{1}{2}$ in (30) is designed to make the total variation distance $\mathrm{d_{TV}}$ take values between 0 and 1. If $\mathrm{d_{TV}}(\nu^{(1)}, \nu^{(2)}) = 0$, then $\nu^{(1)} = \nu^{(2)}$. In the other extreme case $\mathrm{d_{TV}}(\nu^{(1)}, \nu^{(2)}) = 1$, we have that $\nu^{(1)}$ and $\nu^{(2)}$ are "disjoint" in the sense that S can be partitioned into two disjoint subsets S' and S'' such that $\nu^{(1)}$ puts all of its probability mass in S', and $\nu^{(2)}$ puts all of its in S''. The total variation distance also has the natural interpretation

$$\mathrm{d_{TV}}(\nu^{(1)}, \nu^{(2)}) = \max_{A \subseteq S} |\nu^{(1)}(A) - \nu^{(2)}(A)| \,, \tag{31}$$

an identity that you will be asked to prove in Problem 5.1 below. In words, the total variation distance between $\nu^{(1)}$ and $\nu^{(2)}$ is the maximal difference between the probabilities that the two distributions assign to any one event.

We are now ready to state the main result about convergence to stationarity.

Theorem 5.2 (The Markov chain convergence theorem) *Let $(X_0, X_1, \ldots)$ be an irreducible aperiodic Markov chain with state space $S = \{s_1, \ldots, s_k\}$, transition matrix P, and arbitrary initial distribution $\mu^{(0)}$. Then, for any distribution π which is stationary for the transition matrix P, we have*

$$\mu^{(n)} \xrightarrow{\mathrm{TV}} \pi \,. \tag{32}$$

What the theorem says is that if we run a Markov chain for a sufficiently long time n, then, regardless of what the initial distribution was, the distribution at time n will be close to the stationary distribution π. This is often referred to as the Markov chain approaching **equilibrium** as $n \to \infty$.

For the proof, we will use a so-called **coupling** argument; coupling is one of the most useful and elegant techniques in contemporary probability. Before doing the proof, however, the reader is urged to glance ahead at Theorem 5.3 and its proof, to see how easily Theorem 5.2 implies that there cannot be more than one stationary distribution.

Proof of Theorem 5.2: When studying the behavior of $\mu^{(n)}$, we may assume that $(X_0, X_1, \ldots)$ has been obtained by the simulation method outlined in Chapter 3, i.e.,

$$\begin{aligned} X_0 &= \psi_{\mu^{(0)}}(U_0) \\ X_1 &= \phi(X_0, U_1) \\ X_2 &= \phi(X_1, U_2) \\ &\vdots \end{aligned}$$

where $\psi_{\mu^{(0)}}$ is a valid initiation function for $\mu^{(0)}$, ϕ is a valid update function for P, and $(U_0, U_1, \ldots)$ is an i.i.d. sequence of uniform $[0,1]$ random variables.

Next, we introduce a second Markov chain[13] $(X_0', X_1', \ldots)$ by letting ψ_π be a valid initiation function for the distribution π, letting $(U_0', U_1', \ldots)$ be another i.i.d. sequence (independent of $(U_0, U_1, \ldots)$) of uniform $[0,1]$ random variables, and setting

$$\begin{aligned} X_0' &= \psi_\pi(U_0) \\ X_1' &= \phi(X_0', U_1') \\ X_2' &= \phi(X_1', U_2') \\ &\vdots \end{aligned}$$

Since π is a stationary distribution, we have that X_n' has distribution π for any n. Also, the chains $(X_0, X_1, \ldots)$ and $X_0', X_1', \ldots)$ are independent of each other, by the assumption that the sequences $(U_0, U_1, \ldots)$ and $(U_0', U_1', \ldots)$ are independent of each other.

A key step in the proof is now to show that, with probability 1, the two chains will "meet", meaning that there exists an n such that $X_n = X_n'$. To show this, define the "first meeting time"

$$T = \min\{n : X_n = X_n'\}$$

with the convention that $T = \infty$ if the chains never meet. Since the Markov chain $(X_0, X_1, \ldots)$ is irreducible and aperiodic, we can find, using Corollary 4.1, an $M < \infty$ such that

$$(P^M)_{i,j} > 0 \text{ for all } i, j \in \{1, \ldots, k\}.$$

Set

$$\alpha = \min\{(P^M)_{i,j} : i \in \{1, \ldots, k\}\},$$

[13]This is what characterizes the coupling method: to construct two or more processes on the same probability space, in order to draw conclusions about their respective distributions.

and note that $\alpha > 0$. We get that

$$\begin{aligned}
&\mathbf{P}(T \le M) \\
&\quad \ge \ \mathbf{P}(X_M = X'_M) \\
&\quad \ge \ \mathbf{P}(X_M = s_1, X'_M = s_1) \\
&\quad = \ \mathbf{P}(X_M = s_1)\mathbf{P}(X'_M = s_1) \\
&\quad = \ \left(\sum_{i=1}^{k} \mathbf{P}(X_0 = s_i, X_M = s_1)\right)\left(\sum_{i=1}^{k} \mathbf{P}(X'_0 = s_i, X'_M = s_1)\right) \\
&\quad = \ \left(\sum_{i=1}^{k} \mathbf{P}(X_0 = s_i)\mathbf{P}(X_M = s_1 \,|\, X_0 = s_i)\right)\left(\sum_{i=1}^{k} \mathbf{P}(X'_0 = s_i)\mathbf{P}(X'_M = s_1 \,|\, X'_0 = s_i)\right) \\
&\quad \ge \ \left(\alpha \sum_{i=1}^{k} \mathbf{P}(X_0 = s_i)\right)\left(\alpha \sum_{i=1}^{k} \mathbf{P}(X'_0 = s_i)\right) \ = \ \alpha^2
\end{aligned}$$

so that

$$\mathbf{P}(T > M) \le 1 - \alpha^2 .$$

Similarly, given everything that has happened up to time M, we have conditional probability at least α^2 of having $X_{2M} = X'_{2M} = s_1$, so that

$$\begin{aligned}
\mathbf{P}(T > 2M) &\le \mathbf{P}(T > M)\mathbf{P}(T > 2M \,|\, T > M) \\
&\le (1 - \alpha^2)\mathbf{P}(T > 2M \,|\, T > M) \\
&\le (1 - \alpha^2)\mathbf{P}(X_{2M} \ne X'_{2M} \,|\, T > M) \\
&= (1 - \alpha^2)(1 - \mathbf{P}(X_{2M} = X'_{2M} \,|\, T > M)) \\
&\le (1 - \alpha^2)^2 .
\end{aligned}$$

By iterating this argument, we get for any l that

$$\mathbf{P}(T > lM) \le (1 - \alpha^2)^l$$

which tends to 0 as $l \to \infty$. Hence,

$$\lim_{n \to \infty} \mathbf{P}(T > n) = 0 \tag{33}$$

so that in other words, we have shown that the two chains will meet with probability 1.

The next step of the proof is to construct a third Markov chain $(X''_0, X''_1, \ldots)$, by setting

$$X''_0 = X_0 \tag{34}$$

and, for each n,

$$X''_{n+1} = \begin{cases} \phi(X''_n, U_{n+1}) & \text{if } X''_n \ne X'_n \\ \phi(X''_n, U'_{n+1}) & \text{if } X''_n = X'_n \end{cases}$$

In other words, the chain $(X_0'', X_1'', \ldots)$ evolves exactly like the chain $(X_0, X_1, \ldots)$ until the time T when it first meets the chain $(X_0', X_1', \ldots)$. It then switches to evolving exactly like the chain $(X_0', X_1', \ldots)$. It is important to realize that $(X_0'', X_1'', \ldots)$ really is a Markov chain with transition matrix P; this may require a pause for thought, but the basic reason why it is true is that at each update, the update function is exposed to a "fresh" new uniform $[0, 1]$ variable, i.e., one which is independent of all previous random variables.

Because of (34), we have that X_0'' has distribution $\mu^{(0)}$. Hence, for any n, X_n'' has distribution $\mu^{(n)}$. Now, for any $i \in \{1, \ldots, k\}$ we get,

$$\begin{aligned} \mu_i^{(n)} - \pi_i &= \mathbf{P}(X_n'' = s_i) - \mathbf{P}(X_n' = s_i) \\ &\leq \mathbf{P}(X_n'' = s_i, X_n' \neq s_i) \\ &\leq \mathbf{P}(X_n'' \neq X_n') \\ &= \mathbf{P}(T > n) \end{aligned}$$

which tends to 0 as $n \to \infty$, due to (33). Using the same argument (with the roles of X_n'' and X_n' interchanged), we see that

$$\pi_i - \mu_i^{(n)} \leq \mathbf{P}(T > n)$$

as well, again tending to 0 as $n \to \infty$. Hence,

$$\lim_{n \to \infty} |\mu_i^{(n)} - \pi_i| = 0 \,.$$

This implies that

$$\begin{aligned} \lim_{n \to \infty} \mathrm{d_{TV}}(\mu^{(n)}, \pi) &= \lim_{n \to \infty} \left(\tfrac{1}{2} \textstyle\sum_{i=1}^{k} |\mu_i^{(n)} - \pi_i| \right) \\ &= 0 \end{aligned} \tag{35}$$

since each term in the right hand side of (35) tends to 0. Hence, (32) is established. □

Theorem 5.3 (Uniqueness of the stationary distribution) *Any irreducible and aperiodic Markov chain has exactly one stationary distribution.*

Proof: Let $(X_0, X_1, \ldots)$ be an irreducible and aperiodic Markov chain with transition matrix P. By Theorem 5.1, there exists *at least* one stationary distribution for P, so we only need to show that there is *at most* one stationary distribution. Let π and π' be two (a priori possibly different) stationary distributions for P; our task is to show that $\pi = \pi'$.

Suppose that the Markov chain starts with initial distribution $\mu^{(0)} = \pi'$. Then $\mu^{(n)} = \pi'$ for all n, by the assumption that π' is stationary. On the other hand, Theorem 5.2 tells us that $\mu^{(n)} \xrightarrow{\mathrm{TV}} \pi$, meaning that

$$\lim_{n \to \infty} \mathrm{d_{TV}}(\mu^{(n)}, \pi) = 0 \,.$$

Since $\mu^{(n)} = \pi'$, this is the same as

$$\lim_{n\to\infty} \mathrm{d_{TV}}(\pi', \pi) = 0\,.$$

But $\mathrm{d_{TV}}(\pi', \pi)$ does not depend on n, and hence equals 0. This implies that $\pi = \pi'$, so the proof is complete. □

To summarize Theorems 5.2 and 5.3: If a Markov chain is irreducible and aperiodic, then it has a unique stationary distribution π, and the distribution $\mu^{(n)}$ of the chain at time n approaches π as $n \to \infty$, regardless of the initial distribution $\mu^{(0)}$.

Problems

5.1 (7) Prove the formula (31) for total variation distance. Hint: consider the event

$$A = \{s \in S : \nu^{(1)}(s) \geq \nu^{(2)}(s)\}\,.$$

5.2 (4) **Theorems 5.2 and 5.3 fail for reducible Markov chains.** Consider the reducible Markov chain in Example 4.1.

(a) Show that both $\pi = (0.375, 0.625, 0, 0)$ and $\pi' = (0, 0, 0.5, 0.5)$ are stationary distributions for this Markov chain.

(b) Use (a) to show that the conclusions of Theorem 5.2 and 5.3 fail for this Markov chain.

5.3 (6) **Theorem 5.2 fails for periodic Markov chains.** Consider the Markov chain $(X_0, X_1, \ldots)$ describing a knight making random moves on a chessboard, as in Problem 4.2 (c). Show that $\mu^{(n)}$ does not converge in total variation, if the chain is started in a fixed state (such as the square `a1` of the chessboard).

5.4 (7) **If there are two different stationary distributions, then there are infinitely many.** Suppose that $(X_0, X_1, \ldots)$ is a reducible Markov chain with two different stationary distributions π and π'. Show that, for any $p \in (0, 1)$, we get yet another stationary distribution as $p\pi + (1-p)\pi'$.

5.5 (6) Show that the stationary distribution obtained in the proof of Theorem 5.1, can be written as

$$\pi = \left(\frac{1}{\tau_{1,1}}, \frac{1}{\tau_{2,2}}, \ldots, \frac{1}{\tau_{k,k}}\right)\,.$$