1. Markov chains

Section 1. What is a Markov chain? How to simulate one.
Section 2. The Markov property.
Section 3. How matrix multiplication gets into the picture.
Section 4. Statement of the Basic Limit Theorem about convergence to stationarity. A motivating example shows how complicated random objects can be generated using Markov chains.
Section 5. Stationary distributions, with examples. Probability flux.
Section 6. Other concepts from the Basic Limit Theorem: irreducibility, periodicity, and recurrence. An interesting classical example: recurrence or transience of random walks.
Section 7. Introduces the idea of coupling.
Section 8. Uses coupling to prove the Basic Limit Theorem.
Section 9. A Strong Law of Large Numbers for Markov chains.

Markov chains are a relatively simple but very interesting and useful class of random processes. A Markov chain describes a system whose state changes over time. The changes are not completely predictable, but rather are governed by probability distributions. These probability distributions incorporate a simple sort of dependence structure, where the conditional distribution of future states of the system, given some information about past states, depends only on the most recent piece of information. That is, what matters in predicting the future of the system is its present state, and not the path by which the system got to its present state. Markov chains illustrate many of the important ideas of stochastic processes in an elementary setting. This classical subject is still very much alive, with important developments in both theory and applications coming at an accelerating pace in recent decades.

1.1 Specifying and simulating a Markov chain

What is a Markov chain*? One answer is to say that it is a sequence $\{X_0, X_1, X_2, \ldots\}$ of random variables that has the "Markov property"; we will discuss this in the next section. For now, to get a feeling for what a Markov chain is, let's think about how to *simulate* one, that is, how to use a computer or a table of random numbers to generate a typical "sample

* Unless stated otherwise, when we use the term "Markov chain," we will be restricting our attention to the subclass of *time-homogeneous* Markov chains. We'll do this to avoid monotonous repetition of the phrase "time-homogeneous." I'll point out below the place at which the assumption of time-homogeneity enters.

path." To start, how do I tell you which particular Markov chain I want you to simulate? There are three items involved: to specify a Markov chain, I need to tell you its

- State space $\mathcal{S}$.

  $\mathcal{S}$ is a finite or countable set of *states*, that is, values that the random variables X_i may take on. For definiteness, and without loss of generality, let us label the states as follows: either $\mathcal{S} = \{1, 2, \ldots, N\}$ for some finite N, or $\mathcal{S} = \{1, 2, \ldots\}$, which we may think of as the case "$N = \infty$".

- Initial distribution π_0.

  This is the probability distribution of the Markov chain at time 0. For each state $i \in \mathcal{S}$, we denote by $\pi_0(i)$ the probability $\mathbb{P}\{X_0 = i\}$ that the Markov chain starts out in state i. Formally, π_0 is a function taking $\mathcal{S}$ into the interval [0,1] such that

  $$\pi_0(i) \geq 0 \text{ for all } i \in \mathcal{S}$$

  and

  $$\sum_{i \in \mathcal{S}} \pi_0(i) = 1.$$

  Equivalently, instead of thinking of π_0 as a function from $\mathcal{S}$ to [0,1], we could think of π_0 as the vector whose ith entry is $\pi_0(i) = \mathbb{P}\{X_0 = i\}$.

- Probability transition rule.

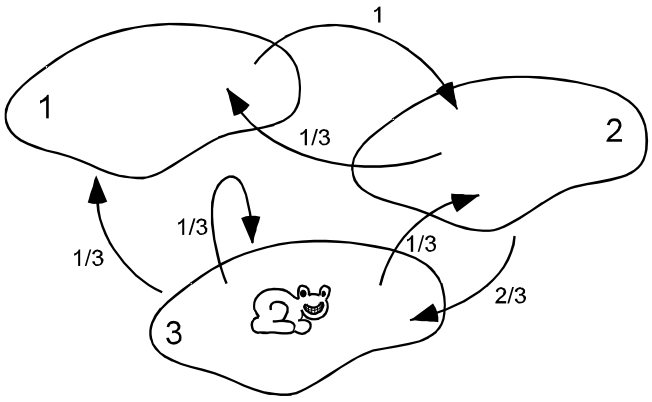
  This is specified by giving a matrix $P = (P_{ij})$. If $\mathcal{S}$ contains N states, then P is an $N \times N$ matrix. The interpretation of the number P_{ij} is the conditional probability, given that the chain is in state i at time n, say, that the chain jumps to the state j at time $n + 1$. That is,

  $$P_{ij} = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}.$$

  We will also use the notation $P(i, j)$ for the same thing. Note that we have written this probability as a function of just i and j, but of course it could depend on n as well. The ***time homogeneity*** restriction mentioned in the previous footnote is just the assumption that this probability does not depend on the time n, but rather remains constant over time.

  Formally, a ***probability transition matrix*** is an $N \times N$ matrix whose entries are all nonnegative and whose rows sum to 1.

  Finally, you may be wondering why we bother to arrange these conditional probabilities into a matrix. That is a good question, and will be answered soon.



(1.1) FIGURE. *The Markov frog.*

We can now get to the question of how to simulate a Markov chain, now that we know how to specify what Markov chain we wish to simulate. Let's do an example: suppose the state space is $\mathcal{S} = \{1, 2, 3\}$, the initial distribution is $\pi_0 = (1/2, 1/4, 1/4)$, and the probability transition matrix is

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 0 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}. \tag{1.2}$$

Think of a frog hopping among lily pads as in Figure 1.1. How does the Markov frog choose a path? To start, he chooses his initial position X_0 according to the specified initial distribution π_0. He could do this by going to his computer to generate a uniformly distributed random number $U_0 \sim \text{Unif}(0, 1)$, and then taking[†]

$$X_0 = \begin{cases} 1 & \text{if } 0 \le U_0 \le 1/2 \\ 2 & \text{if } 1/2 < U_0 \le 3/4 \\ 3 & \text{if } 3/4 < U_0 \le 1 \end{cases}$$

For example, suppose that U_0 comes out to be 0.8419, so that $X_0 = 3$. Then the frog chooses X_1 according to the probability distribution in row 3 of P, namely, $(1/3, 1/3, 1/3)$; to do this, he paws his computer again to generate $U_1 \sim \text{Unif}(0, 1)$ independently of U_0, and takes

$$X_1 = \begin{cases} 1 & \text{if } 0 \le U_0 \le 1/3 \\ 2 & \text{if } 1/3 < U_0 \le 2/3 \\ 3 & \text{if } 2/3 < U_0 \le 1. \end{cases}$$

[†]Don't be distracted by the distinctions between "$<$" and "$\le$" below—for example, what we do if U_0 comes out be exactly $1/2$ or $3/4$—since the probability of U_0 taking on any particular precise value is 0.

Suppose he happens to get $U_1 = 0.1234$, so that $X_1 = 1$. Then he chooses X_2 according to row 1 of P, so that $X_2 = 2$; there's no choice this time. Next, he chooses X_3 according to row 2 of P. And so on....

1.2 The Markov property

Clearly, in the previous example, if I told you that we came up with the values $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$, then the conditional probability distribution for X_3 is

$$\mathbb{P}\{X_3 = j \mid X_0 = 3, X_1 = 1, X_2 = 2\} = \begin{cases} 1/3 & \text{for } j = 1 \\ 0 & \text{for } j = 2 \\ 2/3 & \text{for } j = 3, \end{cases}$$

which is also the conditional probability distribution for X_3 given only the information that $X_2 = 2$. In other words, given that $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$, the only information relevant to the distribution to X_3 is the information that $X_2 = 2$; we may ignore the information that $X_0 = 3$ and $X_1 = 1$. This is clear from the description of how to simulate the chain! Thus,

$$\mathbb{P}\{X_3 = j \mid X_2 = 2, X_1 = 1, X_0 = 3\} = \mathbb{P}\{X_3 = j \mid X_2 = 2\} \text{ for all } j.$$

This is an example of the Markov property.

(1.3) DEFINITION. *A process $X_0, X_1, \ldots$ satisfies the **Markov property** if*

$$\begin{aligned} &\mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} \\ &\quad = \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n\} \end{aligned}$$

for all n and all $i_0, \ldots, i_{n+1} \in \mathcal{S}$.

The issue addressed by the Markov property is the *dependence structure* among random variables. The simplest dependence structure for $X_0, X_1, \ldots$ is no dependence at all, that is, independence. The Markov property could be said to capture the next simplest sort of dependence: in generating the process $X_0, X_1, \ldots$ sequentially, the "next" state X_{n+1} depends only on the "current" value X_n, and not on the "past" values $X_0, \ldots, X_{n-1}$. The Markov property allows much more interesting and general processes to be considered than if we restricted ourselves to independent random variables X_i, without allowing so much generality that a mathematical treatment becomes intractable.

▷ *The idea of the Markov property might be expressed in a pithy phrase, "Conditional on the present, the future does not depend on the past." But there are subtleties. Exercise [1.1] shows the need to think carefully about what the Markov property does and does not say.* ⟦*The exercises are collected in the final section of the chapter.*⟧

The Markov property implies a simple expression for the probability of our Markov chain taking any specified path, as follows:

$$\begin{aligned}
&\mathbb{P}\{X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\} \\
&= \mathbb{P}\{X_0 = i_0\}\mathbb{P}\{X_1 = i_1 \mid X_0 = i_0\}\mathbb{P}\{X_2 = i_2 \mid X_1 = i_1, X_0 = i_0\} \\
&\qquad \cdots \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} \\
&= \mathbb{P}\{X_0 = i_0\}\mathbb{P}\{X_1 = i_1 \mid X_0 = i_0\}\mathbb{P}\{X_2 = i_2 \mid X_1 = i_1\} \\
&\qquad \cdots \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}\} \\
&= \pi_0(i_0)P(i_0, i_1)P(i_1, i_2)\cdots P(i_{n-1}, i_n).
\end{aligned}$$

So, to get the probability of a path, we start out with the initial probability of the first state and successively multiply by the matrix elements corresponding to the transitions along the path.

The Markov property of Markov chains can be generalized to allow dependence on the previous several values. The next definition makes this idea precise.

(1.4) Definition. *We say that a process $X_0, X_1, \ldots$ is **rth order Markov** if*

$$\begin{aligned}
&\mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} \\
&\qquad = \mathbb{P}\{X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_{n-r+1} = i_{n-r+1}\}
\end{aligned}$$

for all $n \geq r$ and all $i_0, \ldots, i_{n+1} \in \mathcal{S}$.

▷ *Is this generalization general enough to capture everything of interest? No; for example, Exercise [1.6] shows that an important type of stochastic process, the "moving average process," is generally not r^{th} order Markov for any r.*

1.3 "It's all just matrix theory"

Recall that the vector π_0 having components $\pi_0(i) = \mathbb{P}\{X_0 = i\}$ is the initial distribution of the chain. Let π_n denote the distribution of the chain at time n, that is, $\pi_n(i) = \mathbb{P}\{X_n = i\}$. Suppose for simplicity that the state space is finite: $\mathcal{S} = \{1, \ldots, N\}$, say. Then the Markov chain has an $N \times N$ probability transition matrix

$$P = (P_{ij}) = (P(i, j)),$$

where $P(i, j) = \mathbb{P}\{X_{n+1} = j \mid X_n = i\} = \mathbb{P}\{X_1 = j \mid X_0 = i\}$. The law of total probability gives

$$\begin{aligned}
\pi_{n+1}(j) &= \mathbb{P}\{X_{n+1} = j\} \\
&= \sum_{i=1}^{N} \mathbb{P}\{X_n = i\}\mathbb{P}\{X_{n+1} = j \mid X_n = i\} \\
&= \sum_{i=1}^{N} \pi_n(i)P(i, j),
\end{aligned}$$

which, in matrix notation, is just the equation

$$\pi_{n+1} = \pi_n P.$$

Note that here we are thinking of π_n and π_{n+1} as *row vectors*, so that, for example,

$$\pi_n = (\pi_n(1), \ldots, \pi_n(N)).$$

Thus, we have

$$\begin{aligned} \pi_1 &= \pi_0 P \\ \pi_2 &= \pi_1 P = \pi_0 P^2 \\ \pi_3 &= \pi_2 P = \pi_0 P^3, \end{aligned} \tag{1.5}$$

and so on, so that by induction

$$\pi_n = \pi_0 P^n. \tag{1.6}$$

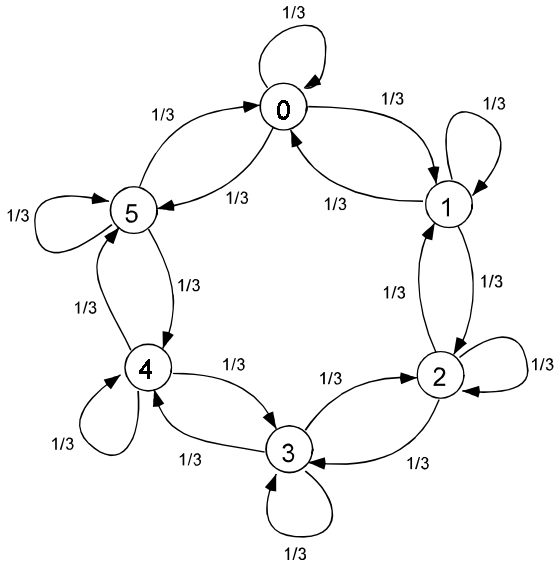
We will let $P^n(i,j)$ denote the (i,j) element in the matrix P^n.

▷ *Exercise* [1.7] *gives some basic practice with the definitions.*

So, in principle, we can find the answer to any question about the probabilistic behavior of a Markov chain by doing matrix algebra, finding powers of matrices, etc. However, what is viable in practice may be another story. For example, the state space for a Markov chain that describes repeated shuffling of a deck of cards contains 52! elements—the permutations of the 52 cards of the deck. This number 52! is large: about 80 million million million million million million million million million million million million. The probability transition matrix that describes the effect of a single shuffle is a 52! by 52! matrix. So, "all we have to do" to answer questions about shuffling is to take powers of such a matrix, find its eigenvalues, and so on! In a practical sense, simply reformulating probability questions as matrix calculations often provides only minimal illumination in concrete questions like "how many shuffles are required in order to mix the deck well?" Probabilistic reasoning can lead to insights and results that would be hard to come by from thinking of these problems as "just" matrix theory problems.

1.4 The basic limit theorem of Markov chains

As indicated by its name, the theorem we will discuss in this section occupies a fundamental and important role in Markov chain theory. What is it all about? Let's start with an example in which we can all see intuitively what is going on.



(1.7) FIGURE. *A random walk on a clock.*

(1.8) EXAMPLE [RANDOM WALK ON A CLOCK]. For ease of writing and drawing, consider a clock with 6 numbers on it: 0,1,2,3,4,5. Suppose we perform a random walk by moving clockwise, moving counterclockwise, and staying in place with probabilities 1/3 each at every time n. That is,

$$P(i,j) = \begin{cases} 1/3 & \text{if } j = i-1 \bmod 6 \\ 1/3 & \text{if } j = i \\ 1/3 & \text{if } j = i+1 \bmod 6. \end{cases}$$

Suppose we start out at $X_0 = 2$, say. That is,

$$\pi_0 = (\pi_0(0), \pi_0(1), \ldots, \pi_0(5)) = (0,0,1,0,0,0).$$

Then of course

$$\pi_1 = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0),$$

and it is easy to calculate

$$\pi_2 = (\frac{1}{9}, \frac{2}{9}, \frac{1}{3}, \frac{2}{9}, \frac{1}{9}, 0)$$

and

$$\pi_3 = (\frac{3}{27}, \frac{6}{27}, \frac{7}{27}, \frac{6}{27}, \frac{3}{27}, \frac{2}{27}).$$

Notice how the probability is spreading out away from its initial concentration on the state 2. We could keep calculating π_n for more values of n, but it is intuitively clear what will happen: the probability will continue to spread out, and π_n will approach the uniform distribution:

$$\pi_n \to (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$$

as $n \to \infty$. Just imagine: if the chain starts out in state 2 at time 0, then we close our eyes while the random walk takes 10,000 steps, and then we are asked to guess what state the random walk is in at time 10,000, what would we think the probabilities of the various states are? I would say: "$X_{10,000}$ is for all practical purposes uniformly distributed over the 6 states." By time 10,000, the random walk has essentially "forgotten" that it started out in state 2 at time 0, and it is nearly equally likely to be anywhere.

Now observe that the starting state 2 was not special; we could have started from anywhere, and over time the probabilities would spread out away from the initial point, and approach the same limiting distribution. Thus, π_n approaches a limit that does not depend upon the initial distribution π_0. □

The following "Basic Limit Theorem" says that the phenomenon discussed in the previous example happens quite generally. We will start with a statement and discussion of the theorem, and then prove the theorem later.

(1.9) THEOREM [BASIC LIMIT THEOREM]. *Let $X_0, X_1, \ldots$ be an irreducible, aperiodic Markov chain having a stationary distribution $\pi(\cdot)$. Let X_0 have the distribution π_0, an arbitrary initial distribution. Then $\lim_{n\to\infty} \pi_n(i) = \pi(i)$ for all states i.*

We need to define the words "irreducible," "aperiodic," and "stationary distribution." Let's start with "stationary distribution."

1.5 Stationary distributions

Suppose a distribution π on $\mathcal{S}$ is such that, if our Markov chain starts out with initial distribution $\pi_0 = \pi$, then we also have $\pi_1 = \pi$. That is, if the distribution at time 0 is π, then the distribution at time 1 is still π. Then π is called a ***stationary distribution*** for the Markov chain. From (1.5) we see that the definition of stationary distribution amounts to saying that π satisfies the equation

$$\pi = \pi P, \tag{1.10}$$

that is,

$$\pi(j) = \sum_{i\in\mathcal{S}} \pi(i)P(i,j) \quad \text{for all } j \in \mathcal{S}.$$

⟦In the case of an infinite state space, (1.10) is an infinite system of equations.⟧ Also from equations (1.5) we can see that if the Markov chain has initial distribution $\pi_0 = \pi$, then we have not only $\pi_1 = \pi$, but also $\pi_n = \pi$ for all n. That is, a Markov chain started out in a stationary distribution π stays in the distribution π forever; that's why the distribution π is called "stationary."

(1.11) EXAMPLE. If the $N \times N$ probability transition matrix P is symmetric, then the uniform distribution ⟦$\pi(i) = 1/N$ for all i⟧ is stationary. More generally, the uniform distribution is stationary if the matrix P is *doubly stochastic*, that is, the column-sums of P are 1 (we already know the row-sums of P are all 1). □

It should not be surprising that π appears as the limit in Theorem (1.9). It is easy to see that if π_n approaches a limiting distribution as $n \to \infty$, then that limiting distribution must be stationary. To see this, suppose that $\lim_{n\to\infty} \pi_n = \tilde{\pi}$, and let $n \to \infty$ in the equation $\pi_{n+1} = \pi_n P$ to obtain $\tilde{\pi} = \tilde{\pi} P$, which says that $\tilde{\pi}$ is stationary.

▷ *The argument just stated goes through clearly and easily when the state space is finite—there are no issues of mathematical analysis that arise in taking the limits. I'll leave it as Exercise [1.10] for the mathematically inclined among you to worry about the details of carrying through the above argument in the case of a countably infinite state space.*

Computing stationary distributions is an algebra problem.

(1.12) EXAMPLE. Let's find the stationary distribution for the frog chain, whose probability transition matrix was given in (1.2). Since most people are accustomed to solving linear systems of the form $Ax = b$, let us take the transpose of the equation $\pi(P - I) = 0$, obtaining the equation $(P^T - I)\pi^T = 0$. In our example, this becomes

$$\begin{pmatrix} -1 & 1/3 & 1/3 \\ 1 & -1 & 1/3 \\ 0 & 2/3 & -2/3 \end{pmatrix} \begin{pmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{pmatrix} = 0,$$

or

$$\begin{pmatrix} -1 & 1/3 & 1/3 \\ 0 & -2/3 & 2/3 \\ 0 & 2/3 & -2/3 \end{pmatrix} \begin{pmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{pmatrix} = 0,$$

which has solutions of the form $\pi = \text{const}(2/3, 1, 1)$. For the unique solution that satisfies the constraint $\sum \pi(i) = 1$, take the constant to be $3/8$, so that $\pi = (1/4, 3/8, 3/8)$.

As an alternative approach, here is another way, aside from solving the linear equations, to address the problem of finding a stationary distribution; this idea can work particularly well with computers. If we believe the Basic Limit Theorem, we should see the stationary distribution in the limit as we run the chain for a long time. Let's try it: Here are some calculations of powers of the transition matrix P from (1.2):

$$P^5 = \begin{pmatrix} 0.246914 & 0.407407 & 0.345679 \\ 0.251029 & 0.36214 & 0.386831 \\ 0.251029 & 0.366255 & 0.382716 \end{pmatrix},$$

$$P^{10} = \begin{pmatrix} 0.250013 & 0.37474 & 0.375248 \\ 0.249996 & 0.375095 & 0.374909 \\ 0.249996 & 0.375078 & 0.374926 \end{pmatrix},$$

$$P^{20} = \begin{pmatrix} 0.2500000002 & 0.3749999913 & 0.3750000085 \\ 0.2499999999 & 0.375000003 & 0.374999997 \\ 0.2499999999 & 0.3750000028 & 0.3749999973 \end{pmatrix}.$$

So we don't really have to solve equations; in this example, any of the rows of the matrix P^{20} provides a very accurate approximation for π. No matter what state we start from, the

distribution after 20 steps of the chain is very close to $(.25, .375, .375)$. This is the Basic Limit Theorem in action. □

(1.13) EXAMPLE [EHRENFEST CHAIN]. The Ehrenfest chain is a simple model of "mixing" processes. This chain can shed light on perplexing questions like "Why aren't people dying all the time due to the air molecules bunching up in some odd corner of their bedrooms while they sleep?" The model considers d balls distributed among two urns, and results in a Markov chain $\{X_0, X_1, \ldots\}$ having state space $\{0, 1, \ldots, d\}$, with the state X_n of the chain at time n being the number of balls in urn #1 at time n. At each time, we choose a ball at random uniformly from the d possibilities, take that ball out of its current urn, and drop it into the other urn. Thus, $P(i, i-1) = i/d$ and $P(i, i+1) = (d-i)/d$ for all i.

▷ *What is the stationary distribution of the Ehrenfest chain? Exercise [1.9] asks you to discover and explain the answer, which turns out to be a distribution that is one of your old friends.*

□

A Markov chain might have no stationary distribution, one stationary distribution, or infinitely many stationary distributions. We just saw examples with one stationary distribution. A trivial example with infinitely many is when P is the identity matrix, in which case all distributions are stationary. To find an example without any stationary distribution, we need to consider an infinite state space. ⟦We will see later that any finite-state Markov chain has at least one stationary distribution.⟧ An easy example of this has $\mathcal{S} = \{1, 2, \ldots\}$ and $P(i, i+1) = 1$ for all i, which corresponds to a Markov chain that moves deterministically "to the right." In this case, the equation $\pi(j) = \sum_{i \in \mathcal{S}} \pi(i) P(i, j)$ reduces to $\pi(j) = \pi(j-1)$, which clearly has no solution satisfying $\sum \pi(j) = 1$. Another interesting example is the *simple, symmetric random walk on the integers*: $P(i, i-1) = 1/2 = P(i, i+1)$. Here the equations for stationarity become

$$\pi(j) = \frac{1}{2}\pi(j-1) + \frac{1}{2}\pi(j+1).$$

Again it is easy to see ⟦how?⟧ that these equations have no solution π that is a probability mass function.

Intuitively, notice the qualitative difference: in the examples without a stationary distribution, the probability doesn't settle down to a limit probability distribution—in the first example the probability moves off to infinity, and in the second example it spreads out in both directions. In both cases, the probability on any fixed state converges to 0; one might say the probability escapes off to infinity (or $-\infty$).

▷ *Exercise [1.8] analyzes an example of a Markov chain that moves around on all of the integers, while no probability escapes to infinity, and the chain has a stationary distribution.*

A Markov chain in its stationary distribution π is at peace with itself; its distribution stays constant, with no desire to change into anything else. This property is explored further in terms of the idea of "probability flux."

(1.14) DEFINITION. *For subsets A and B of the state space, define the **probability flux from the set A into the set B** to be*

$$\text{flux}(A, B) = \sum_{i \in A} \sum_{j \in B} \pi(i) P(i, j)$$

A fundamental balancing property occurs when we consider the probability flux between a set A and its complement A^c, in which case

$$\text{flux}(A, A^c) = \text{flux}(A^c, A). \tag{1.15}$$

▷ *Exercise [1.11] supplies some hints to help you prove this.*

The left side of (1.15) is the "probability flux flowing out of A into A^c." The equality says that this must be the same as the flux from A^c back into A. This has the suggestive interpretation that the stationary probabilities describe a stable system in which all the probability is happy where it is, and does not want to flow to anywhere else, so that the net flow from A to A^c must be zero. We can say this in a less mysterious way as follows. Think of $\pi(i)$ as the long run fraction of time that the chain is in state i. ⟦We will soon see a theorem ("a strong law of large numbers for Markov chains") that supports this interpretation.⟧ Then $\pi(i)P(i, j)$ is the long run fraction of times that a transition from i to j takes place. But clearly the long run fraction of times occupied by transitions going from a state in A to a state in A^c must equal the long run fraction of times occupied by transitions going the opposite way. ⟦In fact, along any sample path, the numbers of transitions that have occurred in the two directions up to any time n may differ by at most 1!⟧

1.6 Irreducibility, periodicity, and recurrence

We'll start by introducing some convenient notation to be used throughout the remainder of this chapter, then we'll define irreducibility and related terms.

(1.16) NOTATION. *We will use the shorthand "$\mathbb{P}_i$" to indicate a probability taken in a Markov chain started in state i at time 0. That is, "$\mathbb{P}_i(A)$" is shorthand for "$\mathbb{P}\{A \mid X_0 = i\}$." We'll also use the notation "$\mathbb{E}_i$" in an analogous way for expectation.*

(1.17) DEFINITION. *Let i and j be two states. We say that j **is accessible from** i if it is possible ⟦with positive probability⟧ for the chain ever to visit state j if the chain starts in state i, or, in other words,*

$$\mathbb{P}_i \left\{ \bigcup_{n=0}^{\infty} \{X_n = j\} \right\} > 0.$$

Clearly an equivalent condition is

$$\sum_{n=0}^{\infty} P^n(i,j) \triangleq \sum_{n=0}^{\infty} \mathbb{P}_i\{X_n = j\} > 0. \tag{1.18}$$

*We say i **communicates with** j if j is accessible from i and i is accessible from j. We say that the Markov chain is **irreducible** if all pairs of states communicate.*

▷ *In Exercise [1.15] you are asked to show that the relation "communicates with" is an equivalence relation. That is, you will show that the "communicates with" relation is reflexive, symmetric, and transitive.*

Recall that an equivalence relation on a set induces a partition of that set into equivalence classes. Thus, by Exercise [1.15], the state space $\mathcal{S}$ may be partitioned into what we will call "communicating classes," or simply "classes." The chain is irreducible if there is just one communicating class, that is, the whole state space $\mathcal{S}$. Note that whether or not a Markov chain is irreducible is determined by the state space $\mathcal{S}$ and the transition matrix $(P(i,j))$; the initial distribution π_0 is irrelevant. In fact, all that matters is the pattern of zeroes in the transition matrix.

Why do we require irreducibility in the "Basic Limit Theorem" (1.9)? Here is a trivial example of how the conclusion can fail if we do not assume irreducibility. Let $\mathcal{S} = \{0,1\}$ and let $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Clearly the resulting Markov chain is not irreducible. Also, clearly the conclusion of the Basic Limit Theorem does not hold; that is, π_n does not approach any limit that is independent of π_0. In fact, $\pi_n = \pi_0$ for all n.

Next, to discuss periodicity, let's begin with another trivial example: take $\mathcal{S} = \{0,1\}$ again, and let $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The conclusion of the Basic Limit Theorem does not hold here: for example, if $\pi_0 = (1,0)$, then $\pi_n = (1,0)$ if n is even and $\pi_n = (0,1)$ if n is odd. So in this case $\pi_n(1)$ alternates between the two values 0 and 1 as n increases, and hence does not converge to anything. The problem in this example is not lack of irreducibility; clearly this chain is irreducible. So, assuming the Basic Limit Theorem is true, the chain must not be aperiodic! That is, the chain is ***periodic***. The trouble stems from the fact that, starting from state 1 at time 0, the chain can visit state 1 only at even times. The same holds for state 2.

(1.19) DEFINITION. *Given a Markov chain $\{X_0, X_1, \ldots\}$, define the **period** of a state i to be the greatest common divisor (gcd)*

$$d_i = \gcd\{n : P^n(i,i) > 0\}.$$

Note that both states 1 and 2 in the example $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ have period 2. In fact, the next result shows that if two states i and j communicate, then they must have the same period.

(1.20) THEOREM. *If the states i and j communicate, then $d_i = d_j$.*

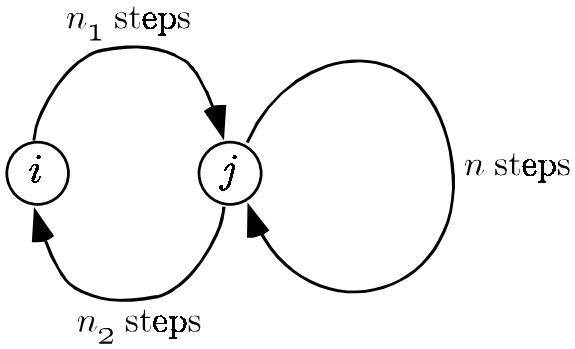
PROOF: Since j is accessible from i, by (1.18) there exists an n_1 such that $P^{n_1}(i,j) > 0$. Similarly, since i is accessible from j, there is an n_2 such that $P^{n_2}(j,i) > 0$. Noting that $P^{n_1+n_2}(i,i) > 0$, it follows that

$$d_i \mid n_1 + n_2,$$

that is, d_i divides $n_1 + n_2$, which means that $n_1 + n_2$ is an integer multiple of d_i. Now suppose that $P^n(j,j) > 0$. Then $P^{n_1+n+n_2}(i,i) > 0$, so that

$$d_i \mid n_1 + n + n_2.$$

Subtracting the last two displays gives $d_i \mid n$. Since n was an arbitrary integer satisfying $P^n(j,j) > 0$, we have found that d_i is a common divisor of the set $\{n : P^n(j,j) > 0\}$. Since d_j is defined to be the *greatest* common divisor of this set, we have shown that $d_j \geq d_i$. Interchanging the roles of i and j in the previous argument gives the opposite inequality $d_i \geq d_j$. This completes the proof. □



It follows from Theorem (1.20) that all states in a communicating class have the same period. We say that the period of a state is a "class property." In particular, all states in an irreducible Markov chain have the same period. Thus, we can speak of *the period of a Markov chain* if that Markov chain is irreducible: the period of an irreducible Markov chain is the period of any of its states.

(1.21) DEFINITION. *An irreducible Markov chain is said to be* ***aperiodic*** *if its period is 1, and* ***periodic*** *otherwise.*

▷ *A simple sufficient (but not necessary) condition for an irreducible chain to be aperiodic is that there exist a state i such that $P(i,i) > 0$. This is Exercise* [1.16].

We have now discussed all of the words we need in order to understand the statement of the Basic Limit Theorem (1.9). We will need another concept or two before we can

get to the proof, and the proof will then take some time beyond that. So I propose that we pause to discuss an interesting example of an application of the Basic Limit Theorem; this will help us build up some motivation to carry us through the proof, and will also give some practice that should be helpful in assimilating the concepts of irreducibility and aperiodicity. We'll also use the next example to introduce the important idea of using the Basic Limit Theorem, in a sense, *in reverse*, to generate random objects from specified distributions. This idea underlies many of the modern uses of Markov chains.

(1.22) Example [Generating a random table with fixed row and column sums]. Consider the 4×4 table of numbers that is enclosed within the rectangle below. The four numbers along the bottom of the table are the column sums, and those along the right edge of the table are the row sums.

68	119	26	7	220
20	84	17	94	215
15	54	14	10	93
5	29	14	16	64
108	286	71	127	

Suppose we want to generate a random, uniformly distributed, 4×4 table of nonnegative integers that has the same row and column sums as the table above. To make sure the goal is clear, define $\mathcal{S}$ to be the set of all nonnegative 4×4 tables that have the given row and column sums. Let $\#(\mathcal{S})$ denote the cardinality of $\mathcal{S}$, that is, the number of elements in $\mathcal{S}$. Remember, each element of $\mathcal{S}$ is a 4×4 table! We want to generate a random element, that is, a random 4×4 table, from $\mathcal{S}$, with each element having equal probability—that's the "uniform" part. That is, each of the $\#(\mathcal{S})$ tables in $\mathcal{S}$ should have probability $1/\#(\mathcal{S})$ of being the table actually generated.

In spirit, this problem is the same as the much simpler problem of drawing a uniformly distributed state from our random walk on a clock as described in Example (1.8). This much simpler problem is merely to generate a uniformly distributed random element X from the set $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, and we can do that without any fancy Markov chains. Just generate a random number $U \sim U[0, 1]$, and then take $X = i$ if U is between $(i-1)/6$ and $i/6$.

Although the two problems may be spiritually the same, there is a crucial practical difference. The set $\mathcal{S}$ for the clock problem has only 6 elements. The set $\mathcal{S}$ for the 4×4 tables is much larger, and in fact we don't know how many elements it has!

So an approach that works fine for $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$—generate a $U \sim U[0, 1]$ and chop up the interval $[0, 1]$ into the appropriate number of pieces—cannot be used to generate a random 4×4 table in our example. However, the Basic Limit Theorem suggests another general approach: start from any state in $\mathcal{S}$, and run an appropriate Markov chain ⟦such as the random walk on the clock⟧ for a sufficiently long time, and take whatever state the chain finds itself in. This approach is rather silly if $\mathcal{S}$ is very simple, like $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, but in many practical problems, it is the only approach that has a hope of working. In our 4×4 table problem, we can indeed generate an approximate solution, that is, a random

table having a distribution arbitrarily close to uniform, by running a Markov chain on $\mathcal{S}$, our set of tables.

Here is one way to do it. Start with any table having the correct row and column sums; so of course the 4×4 table given above will do. Denote the entries in that table by a_{ij}. Choose a pair $\{i_1, i_2\}$ of rows at random, that is, uniformly over the $\binom{4}{2} = 6$ possible pairs. Similarly, choose a random pair of columns $\{j_1, j_2\}$. Then flip a coin. If you get heads: add 1 to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtract 1 from $a_{i_1 j_2}$ and $a_{i_2 j_1}$ *if you can do so without producing any negative entries*—if you cannot do so, then do nothing. Similarly, if the coin flip comes up tails, then subtract 1 from $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and add 1 to $a_{i_1 j_2}$ and $a_{i_2 j_1}$, with the same nonnegativity proviso, and otherwise do nothing. This describes a random transformation of the original table that results in a new table in the desired set of tables $\mathcal{S}$. Now repeat the same random transformation on the new table, and so on.

▷ *In this example, a careful check that the conditions allowing application of the Basic Limit Theorem hold constitutes a challenging exercise, which you are asked to do in Exercise* [1.17]. *Exercise* [1.18] *suggests an alternative Markov chain for the same purpose, and Exercise* [1.19] *introduces a fascinating connection between two problems: generating an approximately uniformly distributed random element of a set, and approximately counting the number of elements in the set. My hope is that these interesting applications of the Basic Limit Theorem are stimulating enough to whet your appetite for digesting the proof of that theorem!*

□

For the proof of the Basic Limit Theorem, we will need one more concept: *recurrence*. Analogously to what we did with the notion of periodicity, we will begin by saying what a recurrent state is, and then show ⟦in Theorem (1.24) below⟧ that recurrence is actually a class property. In particular, in an irreducible Markov chain, either all states are recurrent or all states are *transient*, which means "not recurrent." Thus, if a chain is irreducible, we can speak of the chain being either recurrent or transient.

The idea of recurrence is this: a state i is recurrent if, starting from the state i at time 0, the chain is sure to return to i eventually. More precisely, define the *first hitting time* T_i of the state i by

$$T_i = \inf\{n > 0 : X_n = i\},$$

and make the following definition.

(1.23) DEFINITION. *The state i is* ***recurrent*** *if $\mathbb{P}_i\{T_i < \infty\} = 1$. If i is not recurrent, it is called* ***transient***.

The meaning of recurrence is this: state i is recurrent if, when the Markov chain is started out in state i, the chain is *certain* to return to i at some finite future time. Observe the difference in spirit between this and the definition of "accessible from" ⟦see the paragraph containing (1.18)⟧, which requires only that it be *possible* for the chain to hit a state j. In terms of the first hitting time notation, the definition of "accessible from" may be

restated as follows: for distinct states $i \neq j$, we say that j is accessible from i if and only if $\mathbb{P}_i\{T_j < \infty\} > 0$. ⟦Why did I bother to say "for distinct states $i \neq j$"?⟧

Here is the promised result that implies that recurrence is a class property.

(1.24) THEOREM. *Let i be a recurrent state, and suppose that j is accessible from i. Then in fact all of the following hold:*

(i) $\mathbb{P}_i\{T_j < \infty\} = 1$;

(ii) $\mathbb{P}_j\{T_i < \infty\} = 1$;

(iii) *The state j is recurrent.*

PROOF: The proof will be given somewhat informally; it can be rigorized. Suppose $i \neq j$, since the result is trivial otherwise.

Firstly, let us observe that (iii) follows from (i) and (ii): clearly if (ii) holds ⟦that is, starting from j the chain is certain to visit i eventually⟧ and (i) holds ⟦so that starting from i the chain is certain to visit j eventually⟧, then (iii) must also hold ⟦since starting from j the chain is certain to visit i, after which it will definitely get back to j⟧.

To prove (i), let us imagine starting the chain in state i, so that $X_0 = i$. With probability one, the chain returns at some time $T_i < \infty$ to i. For the same reason, continuing the chain after time T_i, the chain is sure to return to i for a second time. In fact, by continuing this argument we see that, with probability one, the chain returns to i infinitely many times. Thus, we may visualize the path followed by the Markov chain as a succession of infinitely many "cycles," where a cycle is a portion of the path between two successive visits to i. That is, we'll say that the first cycle is the segment $X_1, \ldots, X_{T_i}$ of the path, the second cycle starts with X_{T_i+1} and continues up to and including the second return to i, and so on. The behaviors of the chain in successive cycles are independent and have identical probabilistic characteristics. In particular, letting $I_n = 1$ if the chain visits j sometime during the nth cycle and $I_n = 0$ otherwise, we see that $I_1, I_2, \ldots$ is an *iid* sequence of Bernoulli trials. Let p denote the common "success probability"

$$p = \mathbb{P}\{\text{visit } j \text{ in a cycle}\} = \mathbb{P}_i\left[\bigcup_{k=1}^{T_i}\{X_k = j\}\right]$$

for these trials. Clearly if p were 0, then with probability one the chain would not visit j in any cycle, which would contradict the assumption that j is accessible from i. Therefore, $p > 0$. Now observe that in such a sequence of *iid* Bernoulli trials with a positive success probability, with probability one we will eventually observe a success. In fact,

$$\mathbb{P}_i\{\text{chain does not visit } j \text{ in the first } n \text{ cycles}\} = (1-p)^n \to 0$$

as $n \to \infty$. That is, with probability one, eventually there will be a cycle in which the chain does visit j, so that (i) holds.

It is also easy to see that (ii) must hold. In fact, suppose to the contrary that $\mathbb{P}_j\{T_i = \infty\} > 0$. Combining this with the hypothesis that j is accessible from i, we see that it is

possible with positive probability for the chain to go from i to j in some finite amount of time, and then, continuing from state j, never to return to i. But this contradicts the fact that starting from i the chain must return to i infinitely many times with probability one. Thus, (ii) holds, and we are done. □

The "cycle" idea used in the previous proof is powerful and important; we will be using it again.

The next theorem gives a useful equivalent condition for recurrence. The statement uses the notation N_i for the total number of visits of the Markov chain to the state i, that is,

$$N_i = \sum_{n=0}^{\infty} I\{X_n = i\}.$$

(1.25) THEOREM. *The state i is recurrent if and only if $\mathbb{E}_i(N_i) = \infty$.*

PROOF: We already know that if i is recurrent, then

$$\mathbb{P}_i\{N_i = \infty\} = 1,$$

that is, starting from i, the chain visits i infinitely many times with probability one. But of course the last display implies that $\mathbb{E}_i(N_i) = \infty$. To prove the converse, suppose that i is transient, so that $q := \mathbb{P}_i\{T_i = \infty\} > 0$. Considering the sample path of the Markov chain as a succession of "cycles" as in the proof of Theorem (1.24), we see that each cycle has probability q of never ending, so that there are no more cycles, and no more visits to i. In fact, a bit of thought shows that N_i, the total number of visits to i ⟦including the visit at time 0⟧, has a geometric distribution with "success probability" q, and hence expected value $1/q$, which is finite, since $q > 0$. □

(1.26) COROLLARY. *If j is transient, then $\lim_{n\to\infty} P^n(i,j) = 0$ for all states i.*

PROOF: Supposing j is transient, we know that $\mathbb{E}_j(N_j) < \infty$. Starting from an arbitrary state $i \neq j$, we have

$$\mathbb{E}_i(N_j) = \mathbb{P}_i\{T_j < \infty\}\mathbb{E}_i(N_j \mid T_j < \infty).$$

However, $\mathbb{E}_i(N_j \mid T_j < \infty) = \mathbb{E}_j(N_j)$; this is clear intuitively since, starting from i, if the Markov chain hits j at the finite time T_j, then it "probabilistically restarts" at time T_j. ⟦Exercise: give a formal argument.⟧ Thus, $\mathbb{E}_i(N_j) \leq \mathbb{E}_j(N_j) < \infty$, so that in fact we have $\mathbb{E}_i(N_j) = \sum_{n=1}^{\infty} P^n(i,j) < \infty$, which implies the conclusion of the Corollary. □

(1.27) EXAMPLE ["A DRUNK MAN WILL FIND HIS WAY HOME, BUT A DRUNK BIRD MAY GET LOST FOREVER," OR, RECURRENCE AND TRANSIENCE OF RANDOM WALKS]. The quotation is from Yale's own professor Kakutani, as told by R. Durrett in his probability book. We'll consider a certain model of a random walk in d dimensions, and show that the walk is recurrent if $d = 1$ or $d = 2$, and the walk is transient if $d \geq 3$.

In one dimension, our random walk is the "simple, symmetric" random walk on the integers, which takes steps of $+1$ and -1 with probability $1/2$ each. That is, letting $X_1, X_2, \ldots$ be *iid* taking the values ± 1 with probability $1/2$, we define the position of the random walk at time n to be $S_n = X_1 + \cdots + X_n$. What is a random walk in d dimensions? Here is what we will take it to be: the position of such a random walk at time n is

$$S_n = (S_n(1), \ldots, S_n(d)) \in \mathbb{Z}^d,$$

where the coordinates $S_n(1), \ldots, S_n(d)$ are independent simple, symmetric random walks in $\mathbb{Z}$. That is, to form a random walk in $\mathbb{Z}^d$, simply concatenate d independent one-dimensional random walks into a d-dimensional vector process.

Thus, our random walk S_n may be written as $S_n = X_1 + \cdots + X_n$, where $X_1, X_2, \ldots$ are *iid* taking on the 2^d values $(\pm 1, \ldots, \pm 1)$ with probability 2^{-d} each. This might not be the first model that would come to your mind. Another natural model would be to have the random walk take a step by choosing one of the d coordinate directions at random (probability $1/d$ each) and then taking a step of $+1$ or -1 with probability $1/2$. That is, the increments $X_1, X_2, \ldots$ would be *iid* taking the $2d$ values

$$(\pm 1, 0, \ldots, 0), (0, \pm 1, \ldots, 0), \ldots, (0, 0, \ldots, \pm 1)$$

with probability $1/2d$ each. This is indeed a popular model, and can be analyzed to reach the conclusion "recurrent in $d \leq 2$ and transient in $d \geq 3$" as well. But the "concatenation of d independent random walks" model we will consider is a bit simpler to analyze. Also, for all you Brownian motion fans out there, our model is the random walk analog of d-dimensional Brownian motion, which is a concatenation of d independent one-dimensional Brownian motions.

We'll start with $d = 1$. It is obvious that $S_0, S_1, \ldots$ is an irreducible Markov chain. Since recurrence is a class property, to show that every state is recurrent it suffices to show that the state 0 is recurrent. Thus, by Theorem (1.25) we want to show that

$$\mathbb{E}_0(N_0) = \sum_n P^n(0,0) = \infty. \tag{1.28}$$

But $P^n(0,0) = 0$ if n is odd, and for even $n = 2m$, say, $P^{2m}(0,0)$ is the probability that a Binomial$(2m, 1/2)$ takes the value m, or

$$P^{2m}(0,0) = \binom{2m}{m} 2^{-2m}.$$

This can be closely approximated in a convenient form by using Stirling's formula, which says that

$$k! \sim \sqrt{2\pi k}\,(k/e)^k,$$

where the notation "$a_k \sim b_k$" means that $a_k/b_k \to 1$ as $k \to \infty$. Applying Stirling's formula gives

$$P^{2m}(0,0) = \frac{(2m)!}{(m!)^2 2^{2m}} \sim \frac{\sqrt{2\pi(2m)}\,(2m/e)^{2m}}{2\pi m (m/e)^{2m} 2^{2m}} = \frac{1}{\sqrt{\pi m}}.$$

Thus, from the fact that $\sum(1/\sqrt{m}) = \infty$ it follows that (1.28) holds, so that the random walk is recurrent.

Now it's easy to see what happens in higher dimensions. In $d = 2$ dimensions, for example, again we have an irreducible Markov chain, so we may determine the recurrence or transience of chain by determining whether the sum

$$\sum_{n=0}^{\infty} \mathbb{P}_{(0,0)}\{S_{2n} = (0,0)\} \tag{1.29}$$

is infinite or finite, where S_{2n} is the vector (S^1_{2n}, S^2_{2n}), say. By the assumed independence of the two components of the random walk, we have

$$\mathbb{P}_{(0,0)}\{S_{2m} = (0,0)\} = \mathbb{P}_0\{S^1_{2m} = 0\}\mathbb{P}_0\{S^2_{2m} = 0\} \sim \left(\frac{1}{\sqrt{\pi m}}\right)\left(\frac{1}{\sqrt{\pi m}}\right) = \frac{1}{\pi m},$$

so that (1.29) is infinite, and the random walk is again recurrent. However, in $d = 3$ dimensions, the analogous sum

$$\sum_{n=0}^{\infty} \mathbb{P}_{(0,0,0)}\{S_{2n} = (0,0,0)\}$$

is finite, since

$$\mathbb{P}_{(0,0,0)}\{S_{2m} = (0,0,0)\} = \mathbb{P}_0\{S^1_{2m} = 0\}\mathbb{P}_0\{S^2_{2m} = 0\}\mathbb{P}_0\{S^3_{2m} = 0\} \sim \left(\frac{1}{\sqrt{\pi m}}\right)^3,$$

so that in three ⟦or more⟧ dimensions the random walk is transient.

The calculations are simple once we know that in one dimension $\mathbb{P}_0\{S_{2m} = 0\}$ is of order of magnitude $1/\sqrt{m}$. In a sense it is not very satisfactory to get this by using Stirling's formula and having huge exponentially large titans in the numerator and denominator fighting it out and killing each other off, leaving just a humble $\sqrt{m}$ standing in the denominator after the dust clears. In fact, it is easy to guess without any unnecessary violence or calculation that the order of magnitude is $1/\sqrt{m}$—note that the distribution of S_{2m}, having variance $2m$, is "spread out" over a range of order $\sqrt{m}$, so that the probabilities of points in that range should be of order $1/\sqrt{m}$. Another way to see the answer is to use a Normal approximation to the binomial distribution. We approximate the Binomial$(2m, 1/2)$ distribution by the Normal distribution $N(m, m/2)$, with the usual continuity correction:

$$\begin{aligned}\mathbb{P}\{\text{Binomial}(2m,1/2) = m\} &\sim \mathbb{P}\{m - 1/2 < N(m, m/2) < m + 1/2\} \\ &= \mathbb{P}\{-(1/2)\sqrt{2/m} < N(0,1) < (1/2)\sqrt{2/m}\} \\ &\sim \phi(0)\sqrt{2/m} = (1/\sqrt{2\pi})\sqrt{2/m} = 1/\sqrt{\pi m}.\end{aligned}$$

Although this calculation does not follow as a direct consequence of the usual Central Limit Theorem, it is an example of a "local Central Limit Theorem." □

▷ *Do you feel that the 3-dimensional random walk we have considered was not the one you would have naturally defined? Would you have considered a random walk that at each time moved either North or South, or East or West, or Up or Down? Exercise [1.20] shows that this random walk is also transient. The analysis is somewhat more complicated than that for the 3-dimensional random walk we have just considered.*

We'll end this section with a discussion of the relationship between recurrence and the existence of a stationary distribution. The results will be useful in the next section.

(1.30) PROPOSITION. *Suppose a Markov chain has a stationary distribution π. If the state j is transient, then $\pi(j) = 0$.*

PROOF: Since π is stationary, we have $\pi P^n = \pi$ for all n, so that

$$\sum_i \pi(i) P^n(i,j) = \pi(j) \quad \text{for all } n. \tag{1.31}$$

However, since j is transient, Corollary (1.26) says that $\lim_{n\to\infty} P^n(i,j) = 0$ for all i. Thus, the left side of (1.31) approaches 0 as n approaches ∞, which implies that $\pi(j)$ must be 0. □

The last bit of reasoning about equation (1.31) may look a little strange, but in fact $\pi(i)P^n(i,j) = 0$ for all i and n. In light of what we now know, this is easy to see. First, if i is transient, then $\pi(i) = 0$. Otherwise, if i is recurrent, then $P^n(i,j) = 0$ for all n, since if not, then j would be accessible from i, which would contradict the assumption that j is transient.

(1.32) COROLLARY. *If an irreducible Markov chain has a stationary distribution, then the chain is recurrent.*

PROOF: Being irreducible, the chain must be either recurrent or transient. However, if the chain were transient, then the previous Proposition would imply that $\pi(j) = 0$ for all j, which would contradict the assumption that π is a probability distribution, and so must sum to 1. □

The previous Corollary says that for an irreducible Markov chain, the existence of a stationary distribution implies recurrence. However, we know that the converse is not true. That is, there are irreducible, recurrent Markov chains that do not have stationary distributions. For example, we have seen that the simple symmetric random walk on the integers in one dimension is irreducible and recurrent but does not have a stationary distribution. This random walk is recurrent all right, but in a sense it is "just barely recurrent." That is, by recurrence we have $\mathbb{P}_0\{T_0 < \infty\} = 1$, for example, but we also have $\mathbb{E}_0(T_0) = \infty$. The name for this kind of recurrence is ***null recurrence***: the state i is null recurrent if it is recurrent and $\mathbb{E}_i(T_i) = \infty$. Otherwise, a recurrent state is called ***positive recurrent***: the state i is positive recurrent if $\mathbb{E}_i(T_i) < \infty$. A positive recurrent state i is not just barely recurrent, it is recurrent by a comfortable margin—when started at i, we have not only that T_i is finite almost surely, but also that T_i has finite expectation.

Positive recurrence is in a sense the right notion to relate to the existence of a stationary distibution. For now let me state just the facts, ma'am; these will be justified later. Positive recurrence is also a class property, so that if a chain is irreducible, the chain is either transient, null recurrent, or positive recurrent. It turns out that an irreducible chain has a stationary distribution if and only if it is positive recurrent. That is, strengthening "recurrence" to "positive recurrence" gives the converse to Corollary (1.32).

1.7 An aside on coupling

Coupling is a powerful technique in probability. It has a distinctly probabilistic flavor. That is, using the coupling idea entails thinking probabilistically, as opposed to simply applying analysis or algebra or some other area of mathematics. Many people like to prove assertions using coupling and feel happy when they have done so—a probabilisitic assertion deserves a probabilistic proof, and a good coupling proof can make obvious what might otherwise be a mysterious statement. For example, we will prove the Basic Limit Theorem of Markov chains using coupling. As I have said before, we could do it using matrix theory, but the probabilist tends to find the coupling proof much more appealing, and I hope you do too.

It is a little hard to give a crisp definition of coupling, and different people vary in how they use the word and what they feel it applies to. Let's start by discussing a very simple example of coupling, and then say something about what the common ideas are.

(1.33) Example [Connectivity of a random graph]. A graph is said to be *connected* if for each pair of distinct nodes i and j there is a path from i to j that consists of edges of the graph.

Consider a random graph on a given finite set of nodes, in which each pair of nodes is joined by an edge independently with probability p. We could simulate, or "construct," such a random graph as follows: for each pair of nodes $i < j$, generate a random number $U_{ij} \sim U[0,1]$, and join nodes i and j with an edge if $U_{ij} \leq p$. Here is a problem: show that the probability of the resulting graph being connected is nondecreasing in p. That is, for

$p_1 < p_2$, we want to show that

$$\mathbb{P}_{p_1}\{\text{graph connected}\} \leq \mathbb{P}_{p_2}\{\text{graph connected}\}.$$

I would say that this is intuitively obvious, but we want to give an actual *proof*. Again, the example is just meant to illustrate the idea of coupling, not to give an example that can be solved only with coupling!

One way that one might approach this problem is to try to find an explicit expression for the probability of being connected as a function of p. Then one would hope to show that that function is increasing, perhaps by differentiating with respect to p and showing that the derivative is nonnegative.

That is conceptually a straightforward approach, but you may become discouraged at the first step—I don't think there is an obvious way of writing down the probability the graph is connected. Anyway, doesn't it seem somehow very inefficient, or at least "overkill," to have to give a precise expression for the desired probability if all one desires is to show the inituitively obvious monotonicity property? Wouldn't you hope to give an argument that somehow simply formalizes the intuition that we all have?

One nice way to show that probabilities are ordered is to show that the corresponding events are ordered: if $A \subseteq B$ then $\mathbb{P}A \leq \mathbb{P}B$. So let's make two events by making two random graphs G_1 and G_2 on the same set of nodes. The graph G_1 is constructed by having each possible edge appear with probability p_1. Similarly, for G_2, each edge is present with probability p_2. We could do this by using two sets of $U[0,1]$ random variables: $\{U_{ij}\}$ for G_1 and $\{V_{ij}\}$ for G_2. OK, so now we ask: is it true that

$$\{G_1 \text{ connected}\} \subseteq \{G_2 \text{ connected}\}? \tag{1.34}$$

The answer is no; indeed, the random graphs G_1 and G_2 are independent, so that clearly

$$\mathbb{P}\{G_1 \text{ connected}, G_2 \text{ not connected}\} = \mathbb{P}\{G_1 \text{ connected}\}\mathbb{P}\{G_2 \text{ not connected}\} > 0.$$

The problem is that we have used different, independent random numbers in constructing the graphs G_1 and G_2, so that, for example, it is perfectly possible to have simultaneously $U_{ij} \leq p_1$ and $V_{ij} > p_2$ for all $i < j$, in which the graph G_1 would be completely connected and the graph G_2 would be completely disconnected.

Here is a simple way to fix the argument: use the *same random numbers* in defining the two graphs. That is, draw the edge (i,j) in graph G_1 if $U_{ij} \leq p_1$ and the edge (i,j) in graph G_2 if $U_{ij} \leq p_2$. Now notice how the picture has changed: with the modified definitions it is obvious that, if an edge (i,j) is in the graph G_1, then that edge is also in G_2. From this, it is equally obvious that (1.34) now holds. This establishes the desired monotonicity of the probability of being connected. Perfectly obvious, isn't it? □

So, what characterizes a coupling argument? In our example, we wanted to establish a statement about two distributions: the distributions of random graphs with edge probabilities p_1 and p_2. To do this, we showed how to "construct" ⟦i.e., *simulate* using uniform random numbers!⟧ random objects having the desired distributions in such a way that the desired conclusion became obvious. The trick was to make appropriate use of the same

uniform random variables in constructing the two objects. I think this is a general feature of coupling arguments: somewhere in there you will find the same set of random variables used to construct two different objects about which one wishes to make some probabilistic statement. The term "coupling" reflects the fact that the two objects are related in this way. □

▷ *Exercise [1.24] uses this type of coupling idea, proving a result for one process by comparing it with another process.*

1.8 Proof of the Basic Limit Theorem

The Basic Limit Theorem says that if an irreducible, aperiodic Markov chain has a stationary distribution π, then for each initial distribution π_0, as $n \to \infty$ we have $\pi_n(i) \to \pi(i)$ for all states i. Let me start by pointing something out, just in case the wording of the statement strikes you as a bit strange. Why does the statement read "... ***a*** stationary distribution"? For example, what if the chain has two stationary distributions? The answer is that this is impossible: the assumed conditions imply that a stationary distribution is in fact unique. In fact, once we prove the Basic Limit Theorem, we will know this to be the case. Clearly if the Basic Limit Theorem is true, an irreducible and aperodic Markov chain cannot have two different stationary distributions π and $\tilde{\pi}$, since obviously $\pi_n(i)$ cannot approach both $\pi(i)$ and $\tilde{\pi}(i)$ for all i.

An equivalent but conceptually useful reformulation is to define a distance between probability distributions, and then to show that as $n \to \infty$, the distance between the distribution π_n and the distribution π converges to 0. The notion of distance that we will use is called "total variation distance."

(1.35) Definition. *Let λ and μ be two probability distributions on the set $\mathcal{S}$. Then the* ***total variation distance*** *$\|\lambda - \mu\|$ between λ and μ is defined by*

$$\|\lambda - \mu\| = \sup_{A \subset \mathcal{S}} [\lambda(A) - \mu(A)].$$

(1.36) Proposition. *The total variation distance $\|\lambda - \mu\|$ may also be expressed in the alternative forms*

$$\|\lambda - \mu\| = \sup_{A \subset \mathcal{S}} |\lambda(A) - \mu(A)| = \frac{1}{2} \sum_{i \in \mathcal{S}} |\lambda(i) - \mu(i)| = 1 - \sum_{i \in \mathcal{S}} \min\{\lambda(i), \mu(i)\}.$$

▷ *The proof of this simple Proposition is Exercise [1.25].*

Two probability distributions λ and μ assign probabilites to all possible events. The total variation distance between λ and μ is the largest possible discrepancy between the

probabilities assigned by λ and μ to any event. For example, let π_7 denote the distribution of the ordering of a deck of cards after 7 shuffles, and let π denote the uniform distribution on all 52! permutations of the deck, which corresponds to the result of perfect shuffling (or "shuffling infinitely many times"). Suppose, for illustration, that the total variation distance $\|\pi_7 - \pi\|$ happens to be 0.17. This tells us that the probability of any event — for example, the probability of winning any specified card game — using a deck shuffled 7 times differs by at most 0.17 from the probability of the same event using a perfectly shuffled deck.

To introduce the coupling method, let $Y_0, Y_1, \ldots$ be a Markov chain with the same probability transition matrix as $X_0, X_1, \ldots$, but let Y_0 have the distribution π; that is, we start the Y chain off in the initial distribution π instead of the initial distribution π_0 of the X chain. Note that $\{Y_n\}$ is a stationary Markov chain, and, in particular, that Y_n has the distribution π for all n. Further let the Y chain be independent of the X chain.

Roughly speaking, we want to show that for large n, the probabilistic behavior of X_n is close to that of Y_n. The next result says that we can do this by showing that for large n, the X and Y chains will have met with high probability by time n. Define the *coupling time* T to be the first time at which X_n equals Y_n:

$$T = \inf\{n : X_n = Y_n\},$$

where of course we define $T = \infty$ if $X_n \neq Y_n$ for all n.

(1.37) Lemma ["The coupling inequality"]. *For all n we have*

$$\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}.$$

Proof: Define the process $\{Y_n^*\}$ by

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

It is easy to see that $\{Y_n^*\}$ is a Markov chain, and it has the same probability transition matrix $P(i,j)$ as $\{X_n\}$ has. ⟦To understand this, start by thinking of the X chain as a frog carrying a table of random numbers jumping around in the state space. The frog uses his table of *iid* uniform random numbers to generate his path as we described earlier in the section about specifying and simulating Markov chains. He uses the first number in his table together with his initial distribution π_0 to determine X_0, and then reads down successive numbers in the table to determine the successive transitions on his path. The Y frog does the same sort of thing, except he uses his own, different table of uniform random numbers so he will be independent of the X frog, and he starts out with the initial distribution π instead of π_0. How about the Y^* frog? Is he also doing a Markov chain? Well, is he choosing his transitions using uniform random numbers like the other frogs? Yes, he is; the only difference is that he starts by using Y's table of random numbers (and hence he follows Y) until the coupling time T, after which he stops reading numbers from Y's table and switches to X's table. But big deal; he is still generating his path by using

uniform random numbers in the way required to generate a Markov chain.⟧ The chain $\{Y_n^*\}$ is stationary: $Y_0^* \sim \pi$, since $Y_0^* = Y_0$ and $Y_0 \sim \pi$. Thus, $Y_n^* \sim \pi$ for all n. so that for $A \subseteq \mathcal{S}$ we have

$$\begin{aligned}\pi_n(A) - \pi(A) &= \mathbb{P}\{X_n \in A\} - \mathbb{P}\{Y_n^* \in A\} \\ &= \mathbb{P}\{X_n \in A, T \le n\} + \mathbb{P}\{X_n \in A, T > n\} \\ &\quad -\mathbb{P}\{Y_n^* \in A, T \le n\} - \mathbb{P}\{Y_n^* \in A, T > n\}.\end{aligned}$$

However, on the event $\{T \le n\}$, we have $Y_n^* = X_n$, so that the two events $\{X_n \in A, T \le n\}$ and $\{Y_n^* \in A, T \le n\}$ are the same, and hence they have the same probability. Therefore, the first and third terms in the last expression cancel, yielding

$$\pi_n(A) - \pi(A) = \mathbb{P}\{X_n \in A, T > n\} - \mathbb{P}\{Y_n^* \in A, T > n\}.$$

Since the last difference is obviously bounded by $\mathbb{P}\{T > n\}$, we are done. □

Note the significance of the coupling inequality: it reduces the problem of showing that $\|\pi_n - \pi\| \to 0$ to that of showing that $\mathbb{P}\{T > n\} \to 0$, or equivalently, that $\mathbb{P}\{T < \infty\} = 1$. To do this, we consider the "bivariate chain" $\{Z_n = (X_n, Y_n) : n \ge 0\}$. A bit of thought confirms that $Z_0, Z_1, \ldots$ is a Markov chain on the state space $\mathcal{S} \times \mathcal{S}$. Since the X and Y chains are independent, the probability transition matrix P_Z of the Z chain can be written as

$$P_Z(i_x i_y, j_x j_y) = P(i_x, j_x)P(i_y, j_y).$$

It is easy to check that the Z chain has stationary distribution

$$\pi_Z(i_x i_y) = \pi(i_x)\pi(i_y).$$

Watch closely now; we're about to make an important reduction of the problem. Recall that we want to show that $\mathbb{P}\{T < \infty\} = 1$. Stated in terms of the Z chain, we want to show that with probability one, the Z chain hits the "diagonal" $\{(j, j) : j \in \mathcal{S}\}$ in $\mathcal{S} \times \mathcal{S}$ in finite time. To do this, it is sufficient to show that the Z chain is irreducible and recurrent ⟦why?⟧. However, since we know that the Z chain has a stationary distribution, by Corollary (1.32), to prove the Basic Limit Theorem, it suffices to show that the Z chain is irreducible.

This is, strangely[‡], the hard part. This is where the aperiodicity assumption comes in. For example, consider a Markov chain $\{X_n\}$ having the "type A frog" transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ started out in the condition $X_0 = 0$. Then the stationary chain $\{Y_n\}$ starts out in the uniform distribution: probability 1/2 on each state 0,1. The bivariate chain $\{(X_n, Y_n)\}$ is not irreducible: for example, from the state $(0, 0)$, we clearly cannot reach the state $(0, 1)$. And this ruins everything. For example, if $Y_0 = 1$, which happens with probability 1/2, *the X and Y chains can never meet,* so that $T = \infty$. Thus, $\mathbb{P}\{T < \infty\} < 1$.

[‡]Or maybe not so strangely, in view of Exercise [1.17].

A little number-theoretic result will help us establish irreducibility of the Z chain.

(1.38) LEMMA. *Suppose A is a set of positive integers that is closed under addition and has greatest common divisor (gcd) one. Then there exists an integer N such that $n \in A$ for all $n \geq N$.*

PROOF: First we claim that A contains at least one pair of consecutive integers. To see this, suppose to the contrary that the minimal "spacing" between successive elements of A is $s > 1$. That is, any two distinct elements of A differ by at least s, and there exists an integer n_1 such that both $n_1 \in A$ and $n_1 + s \in A$. Let $m \in A$ be such that s does not divide m; we know that such an m exists because $\gcd(A) = 1$. Write $m = qs + r$, where $0 < r < s$. Now observe that, by the closure under addition assumption, the two numbers $a_1 = (q+1)(n_1+s)$ and $a_2 = (q+1)n_1 + m$ are both in A. However, $a_1 - a_2 = s - r \in (0, s)$, which contradicts the definition of s. This proves the claim.

Thus, A contains two consecutive integers, say, c and $c+1$. Now we will finish the proof by showing that $n \in A$ for all $n \geq c^2$. If $c = 0$ this is trivially true, so assume that $c > 0$. We have, by closure under addition,

$$\begin{aligned} c^2 &= (c)(c) \in A \\ c^2 + 1 &= (c-1)c + (c+1) \in A \\ &\vdots \\ c^2 + c - 1 &= c + (c-1)(c+1) \in A. \end{aligned}$$

Thus, $\{c^2, c^2+1, \ldots, c^2+c-1\}$, a set of c consecutive integers, is a subset of A. Now we can add c to all of these numbers to show that the next set $\{c^2+c, c^2+c+1, \ldots, c^2+2c-1\}$ of c integers is also a subset of A. Repeating this argument, clearly all integers c^2 or above are in A. □

Let $i \in \mathcal{S}$, and retain the assumption that the chain is aperiodic. Then since the set $\{n : P^n(i,i) > 0\}$ is clearly closed under addition, and, by the aperiodicity assumption, has greatest common divisor 1, the previous lemma applies to give that $P^n(i,i) > 0$ for all sufficiently large n. From this, for any $i, j \in \mathcal{S}$, since irreducibility implies that $P^m(i,j) > 0$ for some m, it follows that $P^n(i,j) > 0$ for all sufficiently large n.

Now we complete the proof of the Basic Limit Theorem by showing that the chain $\{Z_n\}$ is irreducible. Let $i_x, i_y, j_x, j_y \in \mathcal{S}$. It is sufficient to show, in the bivariate chain $\{Z_n\}$, that $(j_x j_y)$ is accessible from $(i_x i_y)$. To do this, it is sufficient to show that $P_Z^n(i_x i_y, j_x j_y) > 0$ for some n. However, by the assumed independence of $\{X_n\}$ and $\{Y_n\}$,

$$P_Z^n(i_x i_y, j_x j_y) = P^n(i_x, j_x) P^n(i_y, j_y),$$

which, by the previous paragraph, is positive for all sufficiently large n. Of course, this implies the desired result, and we are done.

▷ *Exercises [1.27] and [1.28] give you a chance to think about the coupling idea used in this proof.*

1.9 A SLLN for Markov chains

The usual Strong Law of Large Numbers for independent and identically distributed (*iid*) random variables says that if $X_1, X_2, \ldots$ are *iid* with mean μ, then the average $(1/n)\sum_{t=1}^{n} X_t$ converges to μ with probability 1 as $n \to \infty$.

> Some fine print: It is possible to have $\mu = +\infty$, and the SLLN still holds. For example, supposing that the random variables X_t take their values in the set of nonnegative integers $\{0, 1, 2, \ldots\}$, the mean is defined to be $\mu = \sum_{k=0}^{\infty} k\mathbb{P}\{X_0 = k\}$. This sum could diverge, in which case we define μ to be $+\infty$, and we have $(1/n)\sum_{t=1}^{n} X_t \to \infty$ with probability 1.

For example, if $X_0, X_1, \ldots$ are *iid* with values in the set $\mathcal{S}$, then the SLLN tells us that

$$(1/n)\sum_{t=1}^{n} I\{X_t = i\} \to \mathbb{P}\{X_0 = i\}$$

with probability 1 as $n \to \infty$. That is, the fraction of times that the *iid* process takes the value i in the first n observations converges to $\mathbb{P}\{X_0 = i\}$, the probability that any given observation is i.

We will do a generalization of this result for Markov chains. This law of large numbers will tell us that the fraction of times that a Markov chain occupies state i converges to a limit.

It is possible to view this result as a consequence of a more general and rather advanced *ergodic theorem* (see, for example, Durrett's *Probability: Theory and Examples*). However, I do not want to assume prior knowledge of ergodic theory. Also, the result for Markov chains is quite simple to derive as a consequence of the ordinary law of large numbers for *iid* random variables. Although the successive states of a Markov chain are not independent, of course, we have seen that certain features of a Markov chain are independent of each other. Here we will use the idea that the path of the chain consists of a succession of independent "cycles," the segments of the path between successive visits to a recurrent state. This independence makes the treatment of Markov chains simpler than the general treatment of stationary processes, and it allows us to apply the law of large numbers that we already know.

(1.39) THEOREM. *Let $X_0, X_1, \ldots$ be a Markov chain starting in the state $X_0 = i$, and suppose that the state i communicates with another state j. The limiting fraction of time that the chain spends in state j is $1/\mathbb{E}_j T_j$. That is,*

$$\mathbb{P}_i\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} I\{X_t = j\} = \frac{1}{\mathbb{E}_j T_j}\right\} = 1.$$

PROOF: The result is easy if the state j is transient, since in that case $\mathbb{E}_j T_j = \infty$ and (with probability 1) the chain visits j only finitely many times, so that

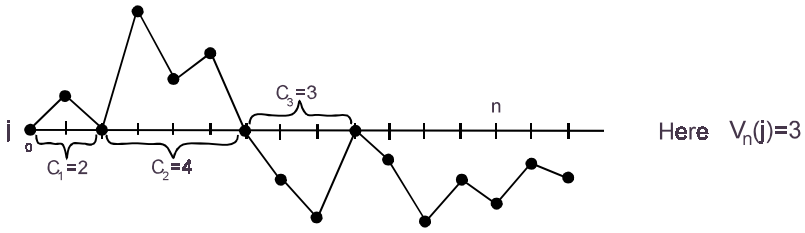
$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} I\{X_t = j\} = 0 = \frac{1}{\mathbb{E}_j T_j}$$

with probability 1. So we assume that j is recurrent. We will also begin by proving the result in the case $i = j$; the general case will be an easy consequence of this special case. Again we will think of the Markov chain path as a succession of *cycles*, where a cycle is a segment of the path that lies between successive visits to j. The cycle lengths $C_1, C_2, \ldots$ are *iid* and distributed as T_j; here we have already made use of the assumption that we are starting at the state $X_0 = j$. Define $S_k = C_1 + \cdots + C_k$ and let $V_n(j)$ denote the number of visits to state j made by $X_1, \ldots, X_n$, that is,

$$V_n(j) = \sum_{t=1}^{n} \{X_t = j\}.$$

A bit of thought [[see also the picture below]] shows that $V_n(j)$ is also the number of cycles completed up to time n, that is,

$$V_n(j) = \max\{k : S_k \le n\}.$$



To ease the notation, let V_n denote $V_n(j)$. Notice that

$$S_{V_n} \le n < S_{V_n+1},$$

and divide by V_n to obtain

$$\frac{S_{V_n}}{V_n} \le \frac{n}{V_n} < \frac{S_{V_n+1}}{V_n}.$$

Since j is recurrent, $V_n \to \infty$ with probability one as $n \to \infty$. Thus, by the ordinary Strong Law of Large Numbers for *iid* random variables, we have both

$$\frac{S_{V_n}}{V_n} \to \mathbb{E}_j(T_j)$$

and

$$\frac{S_{V_n+1}}{V_n} = \left(\frac{S_{V_n+1}}{V_n+1}\right)\left(\frac{V_n+1}{V_n}\right) \to \mathbb{E}_j(T_j) \times 1 = \mathbb{E}_j(T_j)$$

with probability one. Note that the last two displays hold whether $\mathbb{E}_j(T_j)$ is finite or infinite. Thus, $n/V_n \to \mathbb{E}_j(T_j)$ with probability one, so that

$$\frac{V_n}{n} \to \frac{1}{\mathbb{E}_j T_j}$$

with probability one, which is what we wanted to show.

Next, to treat the general case where i may be different from j, note that $P_i\{T_j < \infty\} = 1$ by Theorem 1.24. Thus, with probability one, a path starting from i behaves as follows. It starts by going from i to j in some finite number T_j of steps, and then proceeds on from state j in such a way that the long run fraction of time that $X_t = j$ for $t \geq T_j$ approaches $1/\mathbb{E}_j(T_j)$. But clearly the long run fraction of time the chain is at j is not affected by the behavior of the chain on the finite segment $X_0, \ldots, X_{T_j-1}$. So with probability one, the long run fraction of time that $X_n = j$ for $n \geq 0$ must approach $1/\mathbb{E}_j(T_j)$. □

The following result follows directly from Theorem (1.39) by the Bounded Convergence Theorem from the Appendix. ⟦That is, we are using the following fact: if $Z_n \to c$ with probability one as $n \to \infty$ and the random variables Z_n all take values in the same bounded interval, then we also have $\mathbb{E}(Z_n) \to c$. To apply this in our situation, note that we have

$$Z_n := \frac{1}{n}\sum_{t=1}^{n} I\{X_t = j\} \to \frac{1}{\mathbb{E}_j T_j}$$

with probability one as $n \to \infty$, and also each Z_n lies in the interval [0,1]. Finally, use the fact that the expectation of an indicator random variable is just the probability of the corresponding event.⟧

(1.40) Corollary. *For an irreducible Markov chain, we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} P^t(i,j) = \frac{1}{\mathbb{E}_j(T_j)}$$

for all states i and j.

There's something suggestive here. Consider for the moment an irreducible, aperiodic Markov chain having a stationary distribution π. From the Basic Limit Theorem, we know that, $P^n(i,j) \to \pi(j)$ as $n \to \infty$. However, it is a simple fact that if a sequence of numbers converges to a limit, then the sequence of "Cesaro averages" converges to the same limit; that is, if $a_t \to a$ as $t \to \infty$, then $(1/n)\sum_{t=1}^{n} a_t \to a$ as $n \to \infty$. Thus, the Cesaro averages of $P^n(i,j)$ must converge to $\pi(j)$. However, the previous Corollary shows that the Cesaro averages converge to $1/\mathbb{E}_j(T_j)$. Thus, it follows that

$$\pi(j) = \frac{1}{\mathbb{E}_j(T_j)}.$$

It turns out that the aperiodicity assumption is not needed for this last conclusion; we'll see this in the next result. Incidentally, we could have proved this result much earlier; for example we don't need the Basic Limit Theorem in the development.

(1.41) THEOREM. *An irreducible, positive recurrent Markov chain has a unique stationary distribution π given by*

$$\pi(j) = \frac{1}{\mathbb{E}_j(T_j)}.$$

PROOF: For the uniqueness, let π be a stationary distribution. We start with the relation

$$\sum_i \pi(i) P^t(i,j) = \pi(j),$$

which holds for all t. Averaging this over values of t from 1 to n gives

$$\sum_i \pi(i) \frac{1}{n} \sum_{t=1}^{n} P^t(i,j) = \pi(j).$$

By Corollary 1.40 ⟦and the Dominated Convergence Theorem⟧, the left side of the last equation approaches

$$\sum_i \pi(i) \frac{1}{\mathbb{E}_j(T_j)} = \frac{1}{\mathbb{E}_j(T_j)}$$

as $n \to \infty$. Thus, $\pi(j) = 1/\mathbb{E}_j(T_j)$, which establishes the uniqueness assertion.

We begin the proof of existence by doing the proof in the special case where the state space is finite. The proof is simpler here than in the general case, which involves some distracting technicalities.

So assume for the moment that the state space is finite. We begin again with Corollary 1.40, which says that

$$\frac{1}{n} \sum_{t=1}^{n} P^t(i,j) \to \frac{1}{\mathbb{E}_j(T_j)}. \tag{1.42}$$

However, the sum over all j of the left side of (1.42) is 1, for all n. Therefore,

$$\sum_j \frac{1}{\mathbb{E}_j(T_j)} = 1.$$

That's good, since we want our claimed stationary distribution to be a probability distribution.

Next we write out the matrix equation $P^t P = P^{t+1}$ as follows:

$$\sum_k P^t(i,k) P(k,j) = P^{t+1}(i,j). \tag{1.43}$$

Averaging this over $t = 1, \ldots, n$ gives

$$\sum_k \left[\frac{1}{n} \sum_{t=1}^{n} P^t(i,k) \right] P(k,j) = \frac{1}{n} \sum_{t=1}^{n} P^{t+1}(i,j).$$

Taking the limit as $n \to \infty$ of the last equation and using (1.42) again gives

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right) P(k,j) = \frac{1}{\mathbb{E}_j T_j}.$$

Thus, our claimed stationary distribution is indeed stationary.

Finally, let's see how to handle the infinite state space case. Let $A \subset \mathcal{S}$ be a finite subset of the state space. Summing (1.42) over $j \in A$ gives the inequality

$$\sum_{j \in A} \frac{1}{\mathbb{E}_j(T_j)} \leq 1.$$

Therefore, since this is true for all subsets A, we get

$$\sum_{j \in \mathcal{S}} \frac{1}{\mathbb{E}_j(T_j)} =: C \leq 1.$$

By the assumption of positive recurrence, we have $C > 0$; in a moment we'll see that $C = 1$. The same sort of treatment of (1.43) ⟦i.e., sum over $k \in A$, average over $t = 1, \ldots, n$, let $n \to \infty$, and then take supremum over subsets A of $\mathcal{S}$⟧ gives the inequality

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right) P(k,j) \leq \frac{1}{\mathbb{E}_j T_j}. \tag{1.44}$$

However, the sum over all j of the left side of (1.44) is

$$\sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right) \sum_j P(k,j) = \sum_k \left(\frac{1}{\mathbb{E}_k T_k} \right),$$

which is the same as the sum of the right side of (1.44). Thus, the left and right sides of (1.44) must be the same for all j. From this we may conclude that the distribution

$$\tilde{\pi}(j) = \frac{1}{C} \left(\frac{1}{\mathbb{E}_j(T_j)} \right)$$

is stationary, so that, in particular, we know that our chain does have a stationary distribution. Thus, by the uniqueness assertion we proved above, we must have $C = 1$, and we are done. □

▷ *You might like to try Exercise [1.29] at this point. I hope you can play chess.*

1.10 Exercises

[1.1] Let $X_0, X_1, \ldots$ be a Markov chain, and let A and B be subsets of the state space.

(a) Is it true that $\mathbb{P}\{X_2 \in B \mid X_1 = x_1, X_0 \in A\} = \mathbb{P}\{X_2 \in B \mid X_1 = x_1\}$? Give a proof or counterexample.

(b) Is it true that $\mathbb{P}\{X_2 \in B \mid X_1 \in A, X_0 = x_0\} = \mathbb{P}\{X_2 \in B \mid X_1 \in A\}$? Give a proof or counterexample.

⟦The moral: be careful about what the Markov property says!⟧

[1.2] Let $X_0, X_1, \ldots$ be a Markov chain on the state space $\{-1, 0, 1\}$, and suppose that $P(i, j) > 0$ for all i, j. What is a necessary and sufficient condition for the sequence of absolute values $|X_0|, |X_1|, \ldots$ to be a Markov chain?

▷ *Exercise [1.3] uses a basic and important technique: conditioning on what happens in the first step of the chain. And then in Exercise [1.4] you get to use this to do something interesting.*

[1.3] Let $\{X_n\}$ be a finite-state Markov chain and let A be a subset of the state space. Suppose we want to determine the expected time until the chain enters the set A, starting from an arbitrary initial state. That is, letting $\tau_A = \inf\{n \geq 0 : X_n \in A\}$ denote the first time to hit A ⟦defined to be 0 if $X_0 \in A$⟧, we want to determine $\mathbb{E}_i(\tau_A)$. Show that

$$\mathbb{E}_i(\tau_A) = 1 + \sum_k P(i, k)\mathbb{E}_k(\tau_A)$$

for $i \notin A$.

[1.4] You are tossing a coin repeatedly. Which pattern would you expect to see faster: HH or HT? For example, if you get the sequence TTHHHTH..., then you see "HH" at the 4th toss and "HT" at the 6th. Letting N_1 and N_2 denote the times required to see "HH" and "HT", respectively, can you guess intuitively whether $\mathbb{E}(N_1)$ is smaller than, the same as, or larger than $\mathbb{E}(N_2)$? Go ahead, make a guess ⟦and my day⟧. Why don't you also simulate some to see how the answer looks; I recommend a computer, but if you like tossing real coins, enjoy yourself by all means. Finally, you can use the reasoning of the Exercise [1.3] to solve the problem and evaluate $\mathbb{E}(N_i)$. A hint is to set up a Markov chain having the 4 states HH, HT, TH, and TT.

[1.5] Here is a chance to practice formalizing some typical "intuitively obvious" statements. Let $X_0, X_1, \ldots$ be a finite-state Markov chain.

a. We start with an observation about conditional probabilities that will be a useful tool

throughout the rest of this problem. Let $F_1, \ldots, F_m$ be disjoint events. Show that if $\mathbb{P}(E|F_i) = p$ for all $i = 1, \ldots, m$ then $\mathbb{P}(E \mid \bigcup_{i=1}^m F_i) = p$.

b. Show that

$$\begin{aligned}&\mathbb{P}\{X_{n+1} \in A_1, \ldots, X_{n+r} \in A_r \mid X_n = j, X_{n-1} \in B_{n-1}, \ldots, X_0 \in B_0\}\\ &\quad = \mathbb{P}_j\{X_{n+1} \in A_1, \ldots, X_{n+r} \in A_r\}.\end{aligned}$$

c. Recall the definition of hitting times: $T_i = \inf\{n > 0 : X_n = i\}$. Show that $\mathbb{P}_i\{T_i = n + m \mid T_j = n, T_i > n\} = \mathbb{P}_j\{T_i = m\}$, and conclude that $\mathbb{P}_i\{T_i = T_j + m \mid T_j < \infty, T_i > T_j\} = \mathbb{P}_j\{T_i = m\}$. This is one manifestation of the statement that the Markov chain "probabilistically restarts" after it hits j.

d. Show that $\mathbb{P}_i\{T_i < \infty \mid T_j < \infty, T_i > T_j\} = \mathbb{P}_j\{T_i < \infty\}$. Use this to show that if $\mathbb{P}_i\{T_j < \infty\} = 1$ and $\mathbb{P}_j\{T_i < \infty\} = 1$, then $\mathbb{P}_i\{T_i < \infty\} = 1$.

e. Let i be a recurrent state and let $j \neq i$. Recall the idea of "cycles," the segments of the path between successive visits to i. For simplicity let's just look at the first two cycles. Formulate and prove an assertion to the effect that whether or not the chain visits state j during the first and second cycles can be described by *iid* Bernoulli random variables.

[1.6] ⟦A moving average process⟧ Moving average models are used frequently in time series analysis, economics and engineering. For these models, one assumes that there is an underlying, unobserved process $\ldots, Y_{-1}, Y_0, Y_1, \ldots$ of *iid* random variables. A ***moving average process*** takes an average (possibly a weighted average) of these *iid* random variables in a "sliding window." For example, suppose that at time n we simply take the average of the Y_n and Y_{n-1}, defining $X_n = (1/2)(Y_n + Y_{n-1})$. Our goal is to show that the process $X_0, X_1, \ldots$ defined in this way is not Markov. As a simple example, suppose that the distribution of the *iid* Y random variables is $\mathbb{P}\{Y_i = 1\} = 1/2 = \mathbb{P}\{Y_i = -1\}$.

(a) Show that $X_0, X_1, \ldots$ is not a Markov chain.

(b) Show that $X_0, X_1, \ldots$ is not an rth order Markov chain for any finite r.

[1.7] Let $P^n(i, j)$ denote the (i, j) element in the matrix P^n, the nth power of P. Show that $P^n(i, j) = \mathbb{P}\{X_n = j \mid X_0 = i\}$. Ideally, you should get quite confused about what is being asked, and then straighten it all out.

[1.8] Consider a Markov chain on the integers with

$$\begin{aligned}&P(i, i + 1) = .4 \text{ and } P(i, i - 1) = .6 \text{ for } i > 0,\\ &P(i, i + 1) = .6 \text{ and } P(i, i - 1) = .4 \text{ for } i < 0,\\ &P(0, 1) = P(0, -1) = 1/2.\end{aligned}$$

This is a chain with infinitely many states, but it has a sort of probabilistic "restoring force" that always pushes back toward 0. Find the stationary distribution.

[1.9] Recall the definition the Ehrenfest chain from Example (1.13).

(a) What is the stationary distribution? You might want to solve the problem for a few small values of d. You should notice a pattern, and come up with a familiar answer.

(b) Can you explain without calculation why this distribution is stationary? That is, supposing you start the Ehrenfest chain at time 0 by choosing a state according to the distribution that you claim is stationary, you should argue without calculation that the state at time 1 should also have this same distribution.

[1.10] On page 13 we argued that a limiting distribution must be stationary. This argument was clear in the case of a finite state space. For you fans of mathematical analysis, what happens in the case of a countably infinite state space? Can you still make the limiting argument work?

[1.11] Consider a partition of the state space $\mathcal{S}$ of a Markov chain into two complementary subsets A and A^c. Suppose the Markov chain has stationary distribution π. Show that $\text{flux}(A, A^c) = \text{flux}(A^c, A)$. As a hint, here is an outline of steps you might follow.

(i) Show that the flux function has the following sort of linearity properties: If B and C are disjoint,

$$\text{flux}(A, B \cup C) = \text{flux}(A, B) + \text{flux}(A, C)$$
$$\text{flux}(B \cup C, A) = \text{flux}(B, A) + \text{flux}(C, A)$$

(ii) Show that $\text{flux}(\mathcal{S}, \{k\}) = \text{flux}(\{k\}, \mathcal{S})$ for all singleton sets $\{k\}$.

(iii) Using the first two steps, show that $\text{flux}(\mathcal{S}, A) = \text{flux}(A, \mathcal{S})$.

(iv) By subtracting a certain flux quantity from both sides, conclude that $\text{flux}(A, A^c) = \text{flux}(A^c, A)$.

[1.12] Show by example that for general subsets A and B, the equality $\text{flux}(A, B) = \text{flux}(B, A)$ does not necessarily hold.

[1.13] Use Exercise [1.11] to re-do Exercise [1.9], by writing the equations produced by (1.15) with the choice $A = \{0, 1, \ldots, i\}$ for various i. The calculation should be easier.

[1.14] [[Renewal theory, the residual, and length-biased sampling]] Let $X_1, X_2, \ldots$ be *iid* taking values in $\{1, \ldots, d\}$. You might, for example, think of these random variables as lifetimes of light bulbs. Define $S_k = X_1 + \cdots + X_k$, $\tau(n) = \inf\{k : S_k \geq n\}$, and $R_n = S_{\tau(n)} - n$. Then R_n is called the *residual lifetime* at time n. This is the amount of lifetime remaining in the light bulb that is in operation at time n.

(a) The sequence $R_0, R_1, \ldots$ is a Markov chain. What is its transition matrix? What is the stationary distribution?

(b) Define the *total lifetime* L_n at time n by $L_n = X_{\tau(n)}$. This is the total lifetime of the light bulb in operation at time n. Show that $L_0, L_1, \ldots$ is not a Markov chain. But L_n still has a limiting distribution, and we'd like to find it. We'll do this by constructing a Markov chain by enlarging the state space and considering the sequence of random vectors $(R_0, L_0), (R_1, L_1), \ldots$. This sequence does form a Markov chain. What is its probability transition function and stationary distribution? Now, assuming the Basic Limit Theorem applies here, what is the limiting distribution of L_n as $n \to \infty$? This is the famous "length-biased sampling" distribution.

[1.15] Show that the relation "communicates with" is an equivalence relation. That is, show that the "communicates with" relation is reflexive, symmetric, and transitive.

[1.16] Show that if an irreducible Markov chain has a state i such that $P(i,i) > 0$, then the chain is aperiodic. Also show by example that this sufficient condition is not necessary.

[1.17] ⟦Generating a random 4×4 table of numbers satisfying given restrictions⟧ Show that if we run the process described in Example (1.22) for a sufficiently long time, then we will end up with a random table having probability distribution arbitrarily close to the desired distribution (that is, uniform on $\mathcal{S}$). In order to do this, you need to demonstrate that the conditions of the Basic Limit Theorem are satisfied in this example, by showing that

(a) The procedure generates a Markov chain whose state space is $\mathcal{S}$,
(b) that Markov chain is irreducible,
(c) that Markov chain is aperiodic, and
(d) that Markov chain has the desired distribution as its stationary distribution.

[1.18] ⟦More on 4×4 tables⟧ Refer to the description of the Markov chain in Example (1.22). Imagine that we have already chosen a random pair of rows $\{i_1, i_2\}$ and a random pair of columns $\{j_1, j_2\}$. The Markov chain described in Example (1.22) takes very small steps, adding ± 1 to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtracting ± 1 from $a_{i_1 j_2}$ and $a_{i_2 j_1}$, when doing so produces no negative entries. We could make larger changes by choosing uniformly from all possible modifications of the form: add m to $a_{i_1 j_1}$ and $a_{i_2 j_2}$, and subtract m from $a_{i_1 j_2}$ and $a_{i_2 j_1}$, where m is any integer that does not cause any table entries to become negative. Describe in a more explicit way (explicit enough to make it clear how to write a computer program to do this) how to run this Markov chain. Show that the Basic Limit Theorem applies here to guarantee convergence to the uniform distribution on $\mathcal{S}$. If you feel inspired and/or your instructor asks you to do so, simulate this chain in our example and show the world some random tables from $\mathcal{S}$.

[1.19] ⟦A computing project: Approximate counting⟧ In Example (1.22), we don't know the cardinality of the state space, $\#(\mathcal{S})$. How many such tables are there? About a million? A billion? A trillion? Hey, we don't even know approximately *how many digits* the cardinality has! In some problems there is a nice connection between being able to generate a nearly

uniformly distributed element of a set and the problem of approximating the number of elements in the set. You can try the idea out in the setting of Example (1.22). This is stated in a somewhat open-ended way; there are many variations in how you might approach this, some more or less efficient that the others, and there will be lots of details to work out. The basic idea of the connection between random generation and approximate counting is use the approximate uniform generation to reduce the original approximate counting problem recursively to smaller and smaller counting problems. For example, suppose we knew the fraction, f_{11}, of elements of $\mathcal{S}$ that have a "68" as their (1,1) ⟦upper left-hand corner⟧ entry. Then we have reduced the problem to counting a smaller set, namely, the subset $\mathcal{S}_{11} = \{a \in \mathcal{S} : a_{11} = 68\}$ of $\mathcal{S}$ meeting this additional restriction, because $\#(\mathcal{S}) = \#(\mathcal{S}_{11})/f_{11}$. How do we estimate f_{11}? Well, f_{11} is the probability of a uniformly distributed $A \in \mathcal{S}$ satisfying the extra restriction $A_{11} = 68$. Now you see where the uniform generation comes in: you can estimate f_{11} by generating many nearly uniformly distributed tables from $\mathcal{S}$ and taking the fraction of those that have "68" in their upper left corner. The same idea may be applied recursively in this example. Estimating $\#(\mathcal{S}_{11})$ involves adding an extra restriction, say on the $(1, 2)$ entry of the table, which defines a further subset $\mathcal{S}_{11,12}$ of $\mathcal{S}_{11}$. Estimating the fraction $\#(\mathcal{S}_{11,12})/\#(\mathcal{S}_{11})$ involves running a Markov chain in the smaller state space $\mathcal{S}_{11}$. And so on.

Note: as a practical matter and to preserve your sanity, before applying your methodology to the original large problem, it's a good idea to test it on some much smaller version of the problem (smaller than a 4×4 table) where you know the answer.

[1.20] ⟦The other 3-dimensional random walk⟧ Consider a random walk on the 3-dimensional integer lattice; at each time the random walk moves with equal probability to one of the 6 nearest neighbors, adding or subtracting 1 in just one of the three coordinates. Show that this random walk is transient.

Hint: You want to show that some series converges. An upper bound on the terms will be enough. How big is the largest probability in the Multinomial$(n; 1/3, 1/3, 1/3)$ distribution?

▷ *Here are three additional problems about a simple symmetric random walk $\{S_n\}$ in one dimension starting from $S_0 = 0$ at time 0.*

[1.21] Let a and b be integers with $a < 0 < b$. Defining the hitting times $\tau_c = \inf\{n \geq 0 : S_n = c\}$, show that the probability $\mathbb{P}\{\tau_b < \tau_a\}$ is given by $(0 - a)/(b - a)$.

[1.22] Let $S_0, S_1, \ldots$ be a simple, symmetric random walk in one dimension as we have discussed, with $S_0 = 0$. Show that

$$\mathbb{P}\{S_1 \neq 0, \ldots, S_{2n} \neq 0\} = \mathbb{P}\{S_{2n} = 0\}.$$

Now you can do a calculation that explains why the expected time to return to 0 is infinite.

[1.23] As in the previous exercise, consider a simple, symmetric random walk started out at 0. Letting $k \neq 0$ be any fixed state, show that the expected number of times the random walk visits state k before returning to state 0 is 1.

[1.24] Consider a Markov Chain on the nonnegative integers $\mathcal{S} = \{0, 1, 2, \ldots\}$. Defining $P(i, i+1) = p_i$ and $P(i, i-1) = q_i$, assume that $p_i + q_i = 1$ for all $i \in \mathcal{S}$, and also $p_0 = 1$, and $0 < p_i \leq 1/2$ for all $i \geq 1$. Use what you know about the simple, symmetric random walk to show that the given Markov chain is recurrent.

[1.25] Prove Proposition (1.36).

[1.26] Let π_0 and ρ_0 be probability mass functions on $\mathcal{S}$, and define $\pi_1 = \pi_0 P$ and $\rho_1 = \rho_0 P$, where P is a probability transition matrix. Show that $\|\pi_1 - \rho_1\| \leq \|\pi_0 - \rho_0\|$. That is, in terms of total variation distance, π_1 and ρ_1 are closer to each other than π_0 and ρ_0 were.

[1.27] Here is a little practice with the coupling idea as used in the proof of the Basic Limit Theorem. Consider a Markov chain $\{X_n\}$ having probability transition matrix

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Note that $\{X_n\}$ has stationary distribution $\pi = (1/3, 1/3, 1/3)$. Using the kind of coupling we did in the proof of the Basic Limit Theorem, show that, no matter what the initial distribution π_0 of X_0 is, we have

$$\|\pi_n - \pi\| \leq \frac{2}{3}\left(\frac{11}{16}\right)^n$$

for all n.

[1.28] Do you think the bound you just derived in Exercise [1.27] is a good one? In particular, is 11/16 the smallest we can get, or can we do better? What is the actual rate of geometric decrease of $\|\pi_n - \pi\|$? You could think about this in your head, investigate numerically by matrix multiplication, or both.

⟦Hint about coupling: Try to think of a more "aggressive" coupling to get a better bound What does this mean? The coupling we used in the proof of the Basic Limit Theorem was not very aggressive, in that it let the two chains evolve independently until they happened to meet, and only then started to use the same uniform random numbers to generate the paths. No attempt was made to get the chains together as fast as possible. A more aggressive coupling would somehow make use of some random numbers in common to both chains in generating their paths right from the beginning.⟧

[1.29] Consider a knight sitting on the lower left corner square of an ordinary 8×8 chess board. The knight has residual frog-like tendencies, left over from an old spell an older witch cast

upon him. So he performs a random walk on the chess board, at each time choosing a random move uniformly distributed over the set of his possible knight moves. What is the expected time until he first returns to the lower left corner square?

[1.30] Recall the definition of positive recurrence on page 24. Show that positive recurrence is a class property.

[1.31] Suppose a Markov chain has a stationary distribution π and the state j is null recurrent. Show that $\pi(j) = 0$.

[1.32] ⟦Birth-collapse chain⟧ Consider a Markov chain on $\mathcal{S} = \{0, 1, 2, \ldots\}$ having $P(i, i+1) = p_i$, $P(i, 0) = 1 - p_i$ for all i, with $p_0 = 1$ and $0 < p_i < 1$ for all $i > 0$. Show that

(i) The chain is recurrent if and only if $\lim_{n\to\infty} \prod_{i=1}^{n} p_i = 0$. ⟦This, in turn, is equivalent to the condition $\sum_{i=1}^{\infty}(1 - p_i) = \infty$. (This was just for interest; not a problem or a hint.)⟧

(ii) The chain is positive recurrent if and only if $\sum_{n=1}^{\infty} \prod_{i=1}^{n} p_i < \infty$.

(iii) What is the stationary distribution if $p_i = 1/(i + 1)$?

[1.33] Consider an irreducible Markov chain $\{X_0, X_1, \ldots\}$ on a state space having $n < \infty$ states. Let π denote the stationary distribution of the chain, and suppose X_0 is distributed according to π. Define τ to be the first return time to the initial state, that is, $\tau = \inf\{k > 0 : X_k = X_0\}$. What is the expectation of τ?