Lecture notes on Markov chains

for a graduate course given in the fall of 2004[1]

Olle Häggström

1 Monday, Sep 6

The reader is assumed to have has some prior exposure to Markov chains. If the following introduction feels too abrupt, then he or she is advised to spend some time with one of the basic introductory texts. Some examples of such texts, arranged approximately in order of increasing difficulty, are Enger and Grandell (2000), Häggström (2002), Norris (1997), Brémaud (1998), Grimmett and Stirzaker (1992), Durrett (1991), and Asmussen (2003).

*

We will soon get around to defining what a **Markov chain** is; for the moment, let us be content with saying that it is a special case of a **Markov process**. A Markov process is any stochastic process that is indexed by time (which may be discrete[2] or continuous) and that satisfies the so-called **Markov property**: at any time t the conditional distribution of what happens in the future (after time t) given everything that has happend until time t, depends only on the state of the process precisely *at* time t.

Every stochastic process indexed by time can be represented as a Markov process, by encoding it in a bigger state space that keeps track of "the full trajectory until now".

Example 1.1. Let $(\ldots, X_{-1}, X_0, X_1, X_2, \ldots)$ be any $\{0,1\}$-valued stochastic process indexed by $\mathbf{Z}$. By enlarging the state space to $\{0,1\}^{\mathbf{N}}$, we can obtain another process $(\ldots, Y_{-1}, Y_0, Y_1, Y_2, \ldots)$ by defining

$$Y_n = (\ldots, X_{n-2}, X_{n-1}, X_n)\,.$$

This new process has the properties that it

(a) captures all the interesting properties of the original process (because knowing the Y_i's tells us everything about the X_i's), and

[1] See also `http://www.math.chalmers.se/~olleh/Markovkedjor.html`

[2] By discrete time, I will always mean that the times are the *integers*, although of course other choices are possible. Usually, getting the set of times to be integers is simply a matter of choosing the right unit for measuring time.

(b) satisfies the Markov property (because knowing Y_n gives complete information about all the earlier Y_i's). $\diamond$

Therefore, very little interesting can be said about Markov processes *in complete generality*: we need to make some additional assumptions besides Markovianness in order to be able to conclude much interesting. But doing so leads to an extremely rich and fascinating theory: Markov theory, in its various guises, occupies perhaps 30% of all of probability theory today.[3]

Note, in the above example, that the state space $\{0, 1\}$ of the X-process is finite and therefore discrete, whereas the state space $\{0, 1\}^{\mathbf{N}}$ of the Y-process is equivalent to the unit interval $[0, 1]$ (because it can be thought of as the set of all binary expansions of numbers in $[0, 1]$) and therefore continuous. (Here I use the common terminology that a state space is discrete if is finite or countably infinite, and continuous otherwise.) In general, this type of artificial encodings to make a Markov process out of a non-Markovian one, leads to an uncountable state space. Therfore, requiring a Markov process to have discrete state space, might be enough to obtain some interesting (and genuinely Markovian) theory. This brings us closer to the definition of a Markov chain.

*

What is a Markov chain? There are in fact two conflicting definitions which both appear frequently in the literature. Consider the 2×2 matrix in Figure 1.1, which puts a Markov process in one of four classes depending on whether time is discrete or continuous, and whether the state space is discrete or continuous.

According to one definition, a Markov chain is a Markov process sitting in the top row of this matrix, i.e., one running in discrete *time*. According to the rival definition, a Markov chain is a Markov process sitting in the left column of the matrix, i.e., one having discrete *state space*.

Personally, I am in favor of a rather liberal definition of Markov chains, where having *either* discrete time *or* discrete state space is enough. Thus, I am prepared to call anything sitting in the upper-left, upper-right, or lower-left blocks in Figure 1.1 a Markov chain. And for good measure, I am in fact willing to throw in some of the processes in the lower-right block as well! Such as the following.

[3]There is a famous quote – it may have been due to Bernstein – from the early 1930's, that "by now, Markov theory is essentially complete". But things turned out rather differently...

Example 1.2. Let $\{X_t\}_{t\geq 0}$ be a continuous time Markov process with state space $\mathbf{R}$ and transition mechanism as follows. If $X_t = x$, then the chain stays in state x for an exponential amount of time with mean 1, and then it jumps to a new state chosen at random (i.e., according to Lebesgue measure) from the interval

$$\left[x - \frac{1}{2}, x + \frac{1}{2}\right] .$$

$\diamond$

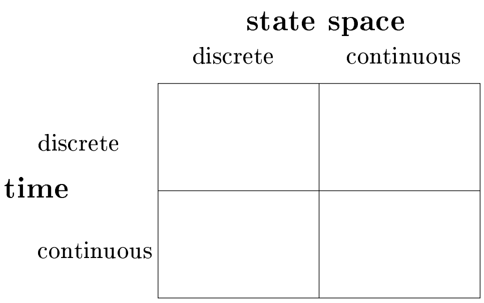


Figure 1.1. We may classify Markov processes according to whether time is discrete or continuous, and whether the state space is discrete or continuous; this leads to four classes of Markov processes.

There are various Markov processes, such as diffusions and certain infinite particle systems[4], in the lower-right block that nobody in his or her right mind would call a Markov chain. But the one in Example 1.2, while sitting in the lower-right block, has "in spirit" much more in common with the discrete time or discrete state-space Markov chains in the other three blocks than with, say, diffusions. In order to include it in the class of Markov chains, I propose the following definition.

Definition 1.1. *A Markov process is said to be a* **Markov chain** *if either*

(a) *it runs in discrete time, or*

(b) *it runs in continuous time and is piecewise constant, meaning that when the process arrives at a new state, it will a.s. stay there for a nonzero amount of time.*

[4]For the latter, see, for instance, Liggett (1999) or Jeff Steif's course from last spring: `http://www.math.chalmers.se/~steif/ips.html`

This definition is in the spirit of the term "chain", because for any process adhering to Definition 1.1, one can write down the sequence (a "chain") of states visited after time 0 say, one after the other. Obviously, this cannot be done for a diffusion.

*

To define a (discrete or continuous time) Markov chain with a given state space $\mathcal{S}$, say for positive times, we need to specify (a) the initial distribution, and (b) the transition mechanism. Let us focus, for now, on the latter.

If time is discrete and the state space $\mathcal{S} = (s_1, \ldots, s_k)$ is finite, then the transition mechanism is described by the **transition matrix**

$$P = \{P_{i,j}\}_{\substack{1 \le i \le k \\ 1 \le j \le k}}$$

where $P_{i,j} = \mathbf{P}(X_{n+1} = s_j \,|\, X_n = s_i)$ is the transition probability from state s_i to state s_j.[5] The same description of the transition mechanism works when the finite state space is replaced by a countably infinite one: the only difference is that the transition matrix P becomes infinite, something that one needn't worry too much about.

Staying within discrete time, things already become rather different when we move from discrete to continuous state space, such as in the following example.

Example 1.3. (A discrete time variant of Example 1.2.) Let $\{X_n\}_{n=0,1,\ldots}$ be a discrete time Markov chain with state space $\mathbf{R}$ and transition mechanism as follows. If $X_n = x$, then X_{n+1} is chosen at random (i.e., according to Lebesgue measure) from the interval

$$\left[x - \frac{1}{2}, x + \frac{1}{2}\right] .$$

◇

Clearly, we have for the chain in Example 1.3 that the transition probability $\mathbf{P}(X_{n+1} = y \,|\, X_n = x)$ is zero for all $x, y \in \mathbf{R}$. A description of the transition mechanism in terms of state-to-state transition probabilities is therefore entirely uninformative, and this will typically be the case also for other chains

[5] This transition probability may also depend on n, which yields a (time-)inhomogeneous Markov chain. Throughout this course, however, I will deal exclusively with (time-)homogeneous Markov chains. This is good enough for most applications, although an example where it is not is **simulated annealing**, see, e.g., Brémaud (1998) or Häggström (2002).

with continuous state space. Instead, the appropriate thing to do is to specify the state-to-(set of states) transition probabilities. The transition matrix is then replaced by the so-called **transition kernel**, defined by

$$P = \{P(x,A)\}_{\substack{x\in\mathcal{S}\\ A\in\Sigma}}$$

where Σ is the set of measurable subsets of the state space $\mathcal{S}$. Here $P(x,A) = \mathbf{P}(X_{n+1} \in A \,|\, X_n = x)$. For instance, in Example 1.3, $P(x,A)$ is the Lebesgue measure of A intersected with $[x-\frac{1}{2},\, x+\frac{1}{2}]$.

Next, how should we describe the transition mechanism of continuous time Markov chains? This is usually done using so-called generators. However, I will not discuss these, as the restriction made in Definition 1.1 (b) allows a more elementary description.

A continuous time Markov chain sitting in a state $x \in \mathcal{S}$ waits for an exponential amount of time with parameter λ_x (i.e., with mean $1/\lambda_x$) before jumping to another state.[6] When it jumps, it picks the new state according to the transition matrix or transition kernel of some discrete time Markov chain, as described above.

The discrete time Markov chain that is used in this way to generate the jumps in a continuous time Markov chain is called the **embedded chain** of the continuous time chain. For many questions concerning continuous time Markov chains, it is sufficient (and often simpler) to analyze their embedded discrete time chains. As an example, take the question *"What is the probability that the chain ever returns to its starting point?"*. The answer to this question is obviously the same regardless of whether we ask it for the original continuous time chain or for the embedded discrete time chain. Partly for this reason, I will in this lecture series focus mostly on discrete time chains.

However, before leaving the continuous time chains (for now!), I would like to mention a phenomenon which is inherently about continuous time chains, and whose analysis cannot be reduced to the study of a discrete time chain. What I have in mind is **explosions** of continuous time Markov chains.

Example 1.4. Let $\{X_t\}_{t\geq 0}$ be the continuous time Markov chain with state space equal to the set $\{1,2,3,\ldots\}$ of positive integers, initial value $X_0 = 1$, and transition mechanism as follows. When the chain arrives at state k, it stays there for an exponential amount of time with mean 2^{-k}, and then jumps to state $k+1$.

[6] The Markov property forces the waiting time distribution to exhibit the memoryless property, and, as we know, the exponential distribution is the only one that has this property. Modifying the cosntruction by allowing more general waiting time distributions leads to **semi-Markov chains** or **Markov renewal theory**; see, e.g., Asmussen (2003).

How will this chain behave? Clearly, it will visit each of the positive integers $1, 2, 3, \ldots$, and it will tend to move faster and faster upwards along these integers. But how fast? For any N, define the random variable T_N as the first time the chain reaches state N. Clearly, the expected value of T_N is

$$\mathbf{E}[T_N] = \sum_{k=1}^{N-1} 2^{-k}$$

which is less than 1. Consider the value X_2 of the chain at time 2. We get

$$\begin{aligned}
\mathbf{P}(X_2 \geq N) &= \mathbf{P}(T_N \leq 2) \\
&= 1 - \mathbf{P}(T_N > 2) \\
&\geq 1 - \frac{\mathbf{E}[T_N]}{2} \\
&\geq 1 - \frac{1}{2} = \frac{1}{2},
\end{aligned}$$

where the first inequality is an application of Markov's Inequality. Now, since this holds for any N, we have with probability at least $\frac{1}{2}$ that X_2 is greater than *every integer*. But no such integer exists! We therefore have, with probability at least $\frac{1}{2}$, that the value of the chain at time $t = 2$ fails to be well-defined.

What happens is that the chain runs so fast through the integers that after some finite time, it has travelled infinitely far. The moment this happens is called the **explosion time** of the chain; after the explosion the value of the chain is no longer well-defined. $\diamond$

It is clear that no such thing as an explosion can happen for a discrete time Markov chain.

2 Thursday, Sep 9

Let us consider another continuous time Markov chain, for which it needs to be carefully checked whether or not the unpleasant phenomenon of an explosion takes place.

Example 2.1. The Richardson model on a square lattice.[7] Let $\mathbf{Z}^2$ denote the set of integer points in the Euclidean plane. A stochastic growth model on this lattice can be informally described as follows. Each site $x \in \mathbf{Z}^2$

[7] I am presonally very fond of this model; see Häggström and Pemantle (1998) for a research contribution.

is in state 0 or 1, where 0 means "healthy", and 1 means "infected". An infected site stays infected forever, while a healthy site can become infected if at least one of its four neighbors (up, down, left, right) is infected, something that more precisely works as follows. Call a pair $\langle x, y\rangle$ of nearest neighbors **hot** is one of the sites is infected and the other is healthy. For each hot pair $\langle x, y\rangle$ of nearest neighbors, there is an exponential clock with rate 1 which, when it rings, causes the healthy site to become infected. Different hot pairs have independent exponential clocks. Since the minimum of k independent exponentials with rate 1, is in itself an exponential with rate k, the following description of the process is valid:

If at time t the number of nearest-neighbor pairs $\langle x, y\rangle$ is k, then the system waits for an exponential amount of time with rate k, after which it changes a healthy site y to become infected; the site y is chosen randomly among the healthy sites having at least one infected neighbor, with probabilities proportional to the number of infected neighbors.

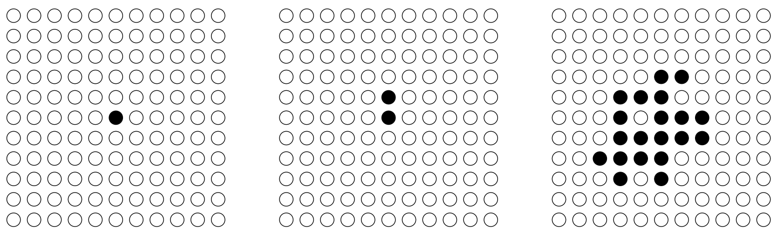


Figure 2.1. The Richardson model starting from a single infected site at three different stages: the initial state, the state after a single infection has taken place, and a state seen a while later.

This sounds very much like a continuous time Markov chain with state space $\mathcal{S} = \{0,1\}^{\mathbf{Z}^2}$, but the description only makes sense if the number of hot pairs at any time is finite.[8] One way to make sure this number is finite is to consider only configurations with a finite number of infected sites. Thus, we replace the state space $\mathcal{S}$ above with the more restrictive state space

$$\mathcal{S}^* = \{\xi \in \{0,1\}^{\mathbf{Z}^2} : \text{ at most finitely many sites in } \mathbf{Z}^2 \text{ take value 1 in } \xi\}.$$

If we take the initial state at time 0 to have just a single infected site, say at the origin, then the above description of the transition mechanism makes

[8]In fact, it *is* possible to make sense of the process even when this number is infinite, but only if we give up the idea of thinking of it as a Markov chain, and instead view it as a so-called interacting particle system (see Liggett, 1999).

perfect sense for the first transition, and also for all future transitions since obviously there will always be only finitely many infected sites. $\diamond$

But wait a minute! Is it really so obvious that the number of infected sites always remains finite, so that we stay within the required state space $\mathcal{S}^*$? Analogously to what happened in Example 1.4, mightn't it be possible that at some point the Markov chain explodes, by infecting infinitely many sites during a finite time interval? This needs to be carefully investigated before we can ligitimately claim that the Markov chain will be well-defined at all times.

Let us for simplicity insist that the process starts at time 0 with a single infected site (the generalization to arbitrary initical states in $\mathcal{S}^*$ will then be easy). Write T_k for the first time at which (at least) k sites are infected (plus the one at the origin). Also write D_k for the waiting time between the $(k-1)$th and the kth infection, with the convention that $D_1 = T_1$. Thus,

$$T_k \;=\; D_1 + D_2 + \cdots + D_k \,. \tag{1}$$

What we need in order to show that no explosion takes place is that

$$\mathbf{P}\left(\lim_{k\to\infty} T_k = \infty\right) \;=\; 1\,. \tag{2}$$

The first waiting time D_1 is exponential with parameter 4 (because the single initially infected sites has 4 neighbors), so that

$$\mathbf{E}[D_1] \;=\; \frac{1}{4}\,.$$

More generally, when $k-1$ sites are infected (plus the one at the origin), each one can contribute to at most 4 hot pairs, so there may be a total of at most $4k$ hot pairs. Hence, conditionally on the first $k-1$ infections, D_k will be exponentially distributed with rate at most $4k$, so that

$$\mathbf{E}[D_k] \geq \frac{1}{4k}\,.$$

Together with (1), this implies that the expected time $\mathbf{E}[T_k]$ until at least k sites are infected, satisfies

$$\begin{aligned} \mathbf{E}[T_k] \;&=\; \mathbf{E}[D_1] + \mathbf{E}[D_2] + \cdots + \mathbf{E}[D_k] \\ &\geq\; \sum_{i=1}^{k} \frac{1}{4i} \;=\; \frac{1}{4}\sum_{i=1}^{k}\frac{1}{i}\,. \end{aligned}$$

Now, we all know that the infinite harmonic series $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots$ diverges, so we may conclude that the expression in (3) tends to ∞ as $k \to \infty$. Hence,

$$\lim_{k\to\infty} \mathbf{E}[T_k] \;=\; \infty\,, \tag{3}$$

and it is very tempting to conclude that the desired formula (2) holds. But some more arguments need to be invoked in order to reach such a conclusion, since, for instance, there exist random variables which are always finite but which nevertheless have infinite expectation.[9]

Here we can proceed in various ways; let us opt for a very concrete and direct calculation. Let H_k denote the number of hot pairs during waiting time D_k. We know that conditionally on the first k infections, D_k is exponentially distributed with parameter H_k. This means in particular that given the full previous evolution, the conditional probability that $D_k \geq \frac{\log(2)}{H_k}$ is precisely $\frac{1}{2}$. Defining the indicator random variables $I_1, I_2, \ldots$ as

$$I_k = \begin{cases} 1 & \text{if } D_k \geq \frac{\log(2)}{H_k} \\ 0 & \text{otherwise,} \end{cases}$$

we thus get that

$$I_1, I_2, \ldots \text{ are i.i.d., each taking value 0 or 1 with probability } \tfrac{1}{2}. \tag{4}$$

Now we partition the sequence of waiting times into blocks of the type (D_1), (D_2, D_3), (D_4, D_5, D_6, D_7), ..., $(D_{2^{m-1}}, D_{2^{m-1}+1}, \ldots, D_{2^m-1})$, ... For fixed m, consider in particular the block $(D_{2^{m-1}}, D_{2^{m-1}+1}, \ldots, D_{2^m-1})$. Call the block **good** if at least half of the waiting times in the block satisfy $D_k \geq \frac{\log(2)}{H_k}$ (or in other words that $I_k = 1$). From (4), we may deduce that each block has probability at least $\frac{1}{2}$ of being good, and that the goodness of different blocks are independent of each other. By the Second Borel–Cantelli Lemma, we therefore get that

$$\mathbf{P}(\text{infinitely many blocks are good}) \;=\; 1\,. \tag{5}$$

Finally, let us see how much a good block contributes in terms of waiting time. For $k \in \{2^{m-1}, \ldots, 2^m - 1\}$, we have

$$\begin{aligned} H_k &\leq 4k \\ &\leq 4 \cdot 2^m \end{aligned}$$

[9] See Problem 2.2.

so that if $I_k = 1$, then $D_k \geq \frac{\log(2)}{4 \cdot 2^m}$. Thus, if a block is good, then at least half of its D_k's contribute a waiting time of at least $\frac{\log(2)}{4 \cdot 2^m}$ each. And in absolute terms, half of the D_k's means that there are 2^{m-2} of them, so that the sum of the D_k's along a good block will be at least

$$2^{m-2} \frac{\log(2)}{4 \cdot 2^m} = \frac{\log(2)}{16} .$$

We saw in (5) that there will a.s. be infinitely many good blocks, and we have now demonstrated that each of these will contribute at least $\frac{\log(2)}{16}$ to the sum $D_1 + D_2 + \cdots$ of all waiting times; hence

$$\mathbf{P}(D_1 + D_2 + \cdots = \infty) = 1 .$$

But this is the same as the desired relation (2), so we have – at last – shown that there is no explosion.

*

And now for something completely different: so-called **stationary distributions**[10] of discrete time Markov chains. For a discrete time Markov chain $(X_0, X_1, \ldots)$ with state space $\mathcal{S}$ and a given transition mechanism, we say that a probability distribution π on $\mathcal{S}$ is a stationary distribution for the chain if it has the following property:

> If the initial state X_0 of the Markov chain is chosen according to π, then the state X_1 will also have distribution π.

Clearly, by induction, if X_0 is chosen according to a stationary distribution π, then X_n will have distribution π for any n.

Intuitively, we may think of stationarity as a kind of statistical equilibrium: if the distribution of X_n is a stationary distribution, then what we observe at later times will have the same distribution.

Suppose that $\mathcal{S} = \{s_1, \ldots, s_k\}$ is finite and the transition matrix is P. A probability distribution ρ on $\mathcal{S}$ can be thought of as a vector $\rho = (\rho_1, \ldots, \rho_k)$ in the obvious way. If the initival value X_0 of the Markov chain is chosen according to ρ, then the distribution of X_1 can be calculated via

$$\mathbf{P}(X_1 = s_j) = \sum_{i=1}^{k} \mathbf{P}(X_0 = i, X_1 = j)$$

[10]The notion of stationarity should not be confused with the notion of time-homogeneity discussed in Footnote 5.

$$
\begin{aligned}
&= \sum_{i=1}^{k} \mathbf{P}(X_0 = i)\mathbf{P}(X_1 = j \mid X_0 = j) \\
&= \sum_{i=1}^{k} \rho_i P_{i,j} \, .
\end{aligned}
$$

Hence, ρ is a stationary distribution for the chain if and only if

$$
\rho_j = \sum_{i=1}^{k} \rho_i P_{i,j} \;\text{ for } j = 1, \ldots, k \, . \tag{6}
$$

In matrix notation, the criterion (6) for statonarity is the same as saying that

$$
\rho \;= \rho P \, . \tag{7}
$$

For a countable-state discrete time Markov chain with state space $\mathcal{S} = \{s_1, s_2, \ldots\}$, the same argument shows that a probability distribution $\rho = (\rho_1, \rho_2, \ldots)$ on $\mathcal{S}$ is stationary if and only if

$$
\rho_j = \sum_{i=1}^{\infty} \rho_i P_{i,j} \;\text{ for } j = 1, 2, \ldots
$$

which in matrix notation again translates into (7).

It is fundamental to understand when **existence** and **uniqueness** of stationary distributions of Markov chains hold. Let us begin with the issue of existence.

Next week, we will see that every finite-state discrete time Markov chain has a stationary distribution. Before that, and in order to see that the issue of existence is not entirely trivial, let us look at a couple of countable-state Markov chains where one does have a stationary distribution and the other does not.

Example 2.2. Let $X_0, X_1, \ldots$ be simple symmetric random walk on the integers $\mathbf{Z}$, meaning that if $X_n = k$, then the next state X_{n+1} is taken to be $k-1$ or $k+1$ with probability $\frac{1}{2}$ each.

Let us see what happens if we try to construct a stationary distribution $\pi = (\ldots, \pi_{-1}, \pi_0, \pi_1, \ldots)$ for this Markov chain. If the Markov chain is started according to π, then we will have for any k that

$$
\mathbf{P}(X_1 = k) \;=\; \mathbf{P}(X_0 = k)
$$

Here, the right-hand side is π_k whereas the left-hand side equals

$$\begin{aligned}
\mathbf{P}(X_1 = k) &= \mathbf{P}(X_0 = k-1, X_1 = k) + \mathbf{P}(X_0 = k+1, X_1 = k) \\
&= \tfrac{1}{2}\mathbf{P}(X_0 = k-1) + \tfrac{1}{2}\mathbf{P}(X_0 = k+1) \\
&= \frac{\pi_{k-1} + \pi_{k+1}}{2}
\end{aligned}$$

whence π_k is the average of π_{k-1} and π_{k+1}:

$$\pi_k = \frac{\pi_{k-1} + \pi_{k+1}}{2}.$$

This may be rewritten as

$$\pi_{k+1} = 2\pi_k - \pi_{k-1}. \tag{8}$$

Now fix m, and suppose that we happen to know π_m and π_{m+1}. Setting $k = m+1$ in (8) yields

$$\pi_{m+2} = \pi_{m+1} + (\pi_{m+1} - \pi_m).$$

Now using (8) again, this time with $k = m+2$, and then iteratively with $k = m+3, m+4, \ldots$, yields

$$\begin{aligned}
\pi_{m+3} &= \pi_{m+1} + 2(\pi_{m+1} - \pi_m) \\
\pi_{m+4} &= \pi_{m+1} + 3(\pi_{m+1} - \pi_m) \\
&\vdots
\end{aligned}$$

Going off in the other direction, we also get

$$\begin{aligned}
\pi_m &= \pi_{m+1} - (\pi_{m+1} - \pi_m) \\
\pi_{m-1} &= \pi_{m+1} - 2(\pi_{m+1} - \pi_m) \\
\pi_{m-2} &= \pi_{m+1} - 3(\pi_{m+1} - \pi_m) \\
&\vdots
\end{aligned}$$

To see that this leads to unsurmountable problems, consider the following three cases.

Case 1: $\pi_{m+1} - \pi_m > 0$. This is impossible, since it will imply $\pi_{m+k} > 1$ for k large enough.

Case 2: $\pi_{m+1} - \pi_m < 0$. This is equally impossible, as it will imply $\pi_{m+k} < 0$ for k large enough.

CASE 3: $\pi_{m+1} - \pi_m = 0$. This implies that π_k is the same for *all* k. Now, if $\pi_m = 0$, then the total probability mass $\sum_{i=-\infty}^{\infty} \pi_i$ will equal 0, while if $\pi_m > 0$, then this sum will equal ∞. Both possibilities contradict the assumption that π is a probability distribution.

So we wee that in all cases, we arrive at untenable conclusions. All we assumed was the existence of a stationary distribution π, and we may therefore conclude that no stationary distribution exists. $\diamond$

The other example will have to be postponed to next week.

3 Monday, Sep 13

In the previous lecture I demonstrated (in Example 2.2) an example of a countable-state discrete time Markov chain that fails to have a stationary distribution. The following is a nice example of one that does have a stationary distribution.

Example 3.1. Let the discrete time Markov chain $(X_0, X_1, \ldots)$ have state space $\mathcal{S} = \{0, 1, 2, \ldots\}$ and the following transition mechanism. If $X_n = k > 0$, then it takes a step down or a step up with respective probabilities $\frac{2}{3}$ and $\frac{1}{3}$, i.e.,

$$X_{n+1} = \begin{cases} k-1 & \text{with probability } \frac{2}{3} \\ k+1 & \text{with probability } \frac{1}{3}\,. \end{cases}$$

For $k = 0$ this will not do since the state space $\mathcal{S}$ does not allow the chain to step down from state 0. Instead, if $X_n = 0$, we take X_{n+1} to be 0 with probability $\frac{2}{3}$ and 1 with probability $\frac{1}{3}$.

I claim that $\pi = (\pi_0, \pi_1, \pi_2, \ldots) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots)$ is a stationary distribution for this Markov chain. To verify this, let us assume that X_0 is chosen according to π, and calculate the distribution of X_1. We get

$$\begin{aligned} \mathbf{P}(X_1) = 0 &= \mathbf{P}(X_0 = 0, X_1 = 0) + \mathbf{P}(X_0 = 1, X_1 = 0) \\ &= \frac{1}{2} \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{2}{3} \\ &= \frac{1}{2} = \pi_0\,, \end{aligned}$$

and, for $k \geq 1$,

$$\begin{aligned} \mathbf{P}(X_1) = k &= \mathbf{P}(X_0 = k-1, X_1 = 0) + \mathbf{P}(X_0 = k+1, X_1 = 0) \\ &= \frac{1}{2^k} \cdot \frac{1}{3} + \frac{1}{2^{k+2}} \cdot \frac{2}{3} \\ &= \frac{1}{2^{k+1}} = \pi_k\,. \end{aligned}$$

Hence, X_1 has distribution π, and we have shown that π is a stationary distribution for the chain. $\diamond$

To summarize Examples 2.2 and 3.1, we thus see that a countable-state discrete time Markov chain may or may not have a stationary distribution. Our next task will be to show that if we reduce the size of the state space from countable to finite, then there *always* exists a stationary distribution:

Theorem 3.1. *Every finite-state discrete time Markov chain has at least one stationary distribution.*

In what follows, a way to prove this result will be indicated, but some of the details will be left to the reader.

An important notion in this context is that of **recurrent states** of a Markov chain. For a Markov chain with state space $\mathcal{S}$ and a given transition matrix, a state $s \in \mathcal{S}$ is said to be **recurrent** if it is the case that starting the chain in state s guarantees (with probability one) that the chain will come back to s at some later time. Otherwise, s is called **transient**.

For example, it is easy to see that in the Markov chain in Figure 3.1, state s_1 is transient while state s_3 is recurrent.[11]

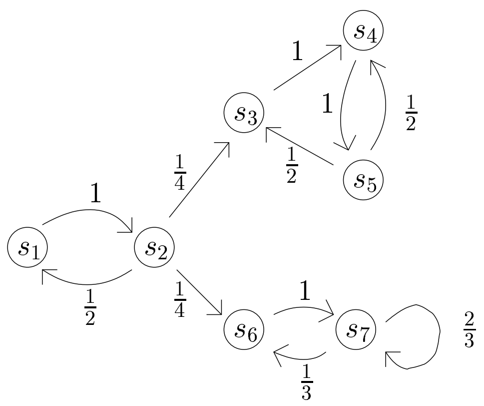


Figure 3.1. Transition probabilities of a Markov chain having recurrent states as well as transient states.

Here are two important facts about finite-state Markov chains and recurrent states:

Lemma 3.1. *Every finite-state discrete time Markov chain has at least one recurrent state.*

[11] See Problem 3.1.

Lemma 3.2. *Suppose that $X_0, X_1, \dots$ is a Markov chain with finite state space $\mathcal{S} = \{s_1, \dots, s_k\}$ and transition matrix P. Suppose that s_1 is a recurrent state, start the chain in $X_0 = s_1$, and let T_{s_1} denote the time of the first return to s_1. Then the probability $\mathbf{P}(T_{s_1} > n)$ decays exponentially in n, in the sense that there exist constants $C > 0$ and $c > 0$ such that for all n we have*

$$\mathbf{P}(T_{s_1} > n) \leq Ce^{-cn} . \tag{9}$$

Proof of Lemma 3.1: Consider a Markov chain with state space $\mathcal{S} = \{s_1, \dots, s_k\}$ and transition matrix P. Write q_i for the return probability of state s_i, i.e., the probability for ever returning to s_1 that we get if the Markov chain starting in s_i. Thus,

$$\text{a state } s_i \text{ is } \begin{cases} \text{recurrent} & \text{if } q_i = 1 \\ \text{transient} & \text{if } q_i < 1 . \end{cases}$$

If s_i is transient, then each time that s_i is visited, there is conditional probability q_i that it will ever come back again, so that

$$\mathbf{P}(\text{at least } n \text{ returns to } s_i) \leq q_i^n$$

which tends to 0 as $n \to \infty$, so that

$$\mathbf{P}(\text{infinitely many visits to } s_i) = 0 . \tag{10}$$

But if we keep running the chain forever, then at least one state must be visited infinitely often, and due to (10) this state cannot be transient, and is therefore recurrent. $\diamond$

For the next proof, the following piece of notation is convenient. For $s_i, s_j \in S$, take $s_i \Rightarrow s_j$ to represent the property of the transition matrix that if the chain starts in s_i, then the probability of sooner or later hitting state s_j is 1. (For instance, recurrence of s_i is the same as $s_i \Rightarrow s_i$.) Furthermore, write $s_i \to s_j$ to represent the property that if the chain starts in s_i, then the probability of ever hitting s_j is nonzero. Note that $s_i \Rightarrow s_j$ implies $s_i \to s_j$, but not vice versa.

Proof of Lemma 3.2: Write $\mathcal{S}^*$ for the set of states $s \in \mathcal{S}$ that have the property that $s_1 \to s$. Since s_1 is assumed to be recurrent, we have $s_1 \in S^*$. Assume that $\mathcal{S}^* = \{s_1, \dots, s_l\}$ for some l; by relabelling of the states if necessary, there is no loss of generality in making such an assumption.

Clearly, for $i = 1, \dots l$, we have $s_i \to s_1$, because otherwise a chain starting in s_1 could go to s_i and would then never return to s_1, contradicting the assumed recurrence of s_1. Hence, for $i = 1, \dots, l$ there exists an N_i such that whenever the chain sits in s_i, it has some positive probability $\delta_i > 0$ of reaching s_1 exactly N_i time units later.

Now set

$$N = \max\{N_1, \dots, N_l\}$$

and

$$\delta = \min\{\delta_1, \dots, \delta_l\}\,.$$

Recall that the chain $(X_0, X_1 \dots)$ starts in s_1, and that T_{s_1} denotes the time of the first return to s_1. By choice of N_1 and δ_1, we have

$$\begin{aligned}\mathbf{P}(T_{s_1} > N) &\leq \mathbf{P}(T_{s_1} > N_1) \\ &\leq 1 - \delta_1 \\ &\leq 1 - \delta\,.\end{aligned}$$

Conditional on $T_{s_1} > N$, we know that X_N stands in one of the states $s_2, \dots, s_l$, so that the conditional probability of avoiding s_1 for *another* N time units is at most $1 - \delta$, whence

$$\begin{aligned}\mathbf{P}(T_{s_1} > 2N) &= \mathbf{P}(T_{s_1} > N)\mathbf{P}(T_{s_1} > 2N \,|\, T_{s_1} > N) \\ &\leq (1 - \delta)^2\,.\end{aligned}$$

Iterating this reasoning gives

$$\mathbf{P}(T_{s_1} > kN) \;\leq\; (1 - \delta)^k$$

for $k = 3, 4, \dots$. Translating this into the language of eq. (9), we have for $n = N, 2N, 3N, \dots$ that

$$\begin{aligned}\mathbf{P}(T_{s_1} > n) &\leq (1 - \delta)^{\frac{n}{N}} \\ &= \exp\left(\frac{\log(1 - \delta)}{N} n\right)\,.\end{aligned}$$

And since $\mathbf{P}(T_{s_1} > n)$ is nonincreasing in n, we get for arbitrary $n \geq 0$ that

$$\mathbf{P}(T_{s_1} > n) \;\leq\; \frac{1}{1 - \delta} \exp\left(\frac{\log(1 - \delta)}{N} n\right)\,,$$

so (9) holds with $C = \frac{1}{1-\delta}$ and $c = -\frac{\log(1-\delta)}{N}$. $\diamond$

Let us move on towards a proof of Theorem 3.1 concerning the existence of a stationary distribution. Consider a discrete time Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and transition matrix P. By Lemma 3.1, there exists at least one recurrent state, say s_1. Taking the chain to start in initial state $X_0 = s_1$, we learned in Lemma 3.2 that the distribution of the first return time T_{s_1} has an exponential tail. This implies[12] that

$$\mathbf{E}[T_{s_1}] < \infty ,$$

and let us write τ_1 for this expectation. Then the expected number of visits to other states between two consecutive visits to s_1 is, obviously, $\tau_1 - 1$. For $i = 2, 3, \ldots, k$, let ρ_{1i} denote the expected number of visits to state s_i between two consecutive visits to s_1, so that in particular

$$\rho_{12} + \rho_{13} + \cdots + \rho_{1k} = \tau_1 - 1 .$$

Adding 1 and dividing by τ_1 gives

$$\frac{1}{\tau_1} + \frac{\rho_{12}}{\tau_1} + \frac{\rho_{13}}{\tau_1} + \cdots + \frac{\rho_{1k}}{\tau_1} = 1 .$$

Let us take the left-hand side of this equation as inspiration for constructing a candidate for stationary distribution.

Proposition 3.1. *For a discrete time Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$, with s_1 being a recurrent state, we have that*

$$\begin{aligned} \pi &= (\pi_1, \pi_2, \ldots, \pi_k) \\ &= \left(\frac{1}{\tau_1}, \frac{\rho_{12}}{\tau_1}, \frac{\rho_{13}}{\tau_1}, \ldots, \frac{\rho_{1k}}{\tau_1} \right) \end{aligned} \tag{11}$$

is a stationary distribution for the chain.

Sketch proof: First note that we can decompose ρ_{1i} as

$$\rho_{1i} = \sum_{n=1}^{\infty} \rho_{1i}^{(n)}$$

where $\rho_{1i}^{(n)}$ is the probability (given that the chain starts in $X_0 = s_1$) of the event $\{X_n = s_i, T_{s_1} > n\}$. Using this decomposition, it is a straightforward (but slightly tedious) calculation to verify that π is a stationary distribution (*Ansatz:* Assume that X_0 is chosen according to π. Then $\mathbf{P}(X_1 = s_i) = \ldots$).

[12] See Problem 3.2.

This calculation is left to the reader. (In case of severe laziness, he or she may consult, e.g., Grimmett and Stirzaker (1992) or Häggström (2002).) $\diamond$

Proof of Theorem 3.1: From Lemma 3.1, we know that any finite-state discrete-time Markov chain has some recurrent state s_1. Once such a state is identified, eq. (11) of Proposition 3.1 provides a stationary distribution. $\diamond$

*

The issue of *existence* of stationary distributions for finite-state Markov chains has thus been settled, but it remains to investigate the complementary issue of *uniqueness*. In fact, uniqueness does not alway hold: see Problem 3.3 for an example of a finite-state Markov chain with more than one stationary distribution. We will soon familiarize ourselves with a natural condition – irreducibility – that guarantees uniqueness, but before that, let us state and prove our first **limit theorem** for Markov chains. The following is a kind of Markov chain analogue of the Strong Law of Large Numbers.

Theorem 3.2. *Let $X_0, X_1, \ldots$ be a Markov chain with finite state space $\mathcal{S} = \{s_1, \ldots, s_k\}$, starting in state $X_0 = s_1$ which is assumed to be recurrent. Define the random variable $K(n, s_i)$ as the number of visits to state s_i during the time interval $\{1, 2, \ldots, n\}$. We then have that*

$$\lim_{n \to \infty} \frac{K(n, s_1)}{n} = \frac{1}{\tau_1} \quad a.s., \tag{12}$$

and, for $i = 2, 3, \ldots, k$, that

$$\lim_{n \to \infty} \frac{K(n, s_i)}{n} = \frac{\rho_{1i}}{\tau_1} \quad a.s. \tag{13}$$

The proof will be done first thing on Thursday.

4 Thursday, Sep 16

Proof of Theorem 3.2: For $m = 1, 2, \ldots$, let $T_m^{s_1}$ denote the time of the mth return to state s_1. Also let

$$D_m^{s_1} = T_m^{s_1} - T_{m-1}^{s_1}$$

with the convention that $D_1^{s_1} = T_1^{s_1}$. Thus, $D_m^{s_1}$ is the time between the $(m-1)$th visit to s_1 and the mth such visit, and

$$T_m^{s_1} = \sum_{l=1}^{m} D_m^{s_1} \,.$$

Note that $D_1^{s_1}, D_2^{s_1}, \ldots$ are i.i.d. random variables with mean τ_1. We may therefore apply the Strong Law of Large Numbers to deduce that, with probability 1,

$$\begin{aligned}\lim_{m\to\infty} \frac{T_m^{s_1}}{m} &= \lim_{m\to\infty} \frac{1}{m} \sum_{l=1}^{m} D_m^{s_1} \\ &= \tau_1 \,.\end{aligned}$$

For any m, we have $K(T_m^{s_1}, s_1) = m$. Hence,

$$\lim_{m\to\infty} \frac{K(T_m^{s_1}, s_1)}{T_m^{s_1}} = \frac{1}{\tau_1} \text{ a.s.,} \tag{14}$$

which is almost the desired asymptotics (12), but along the subsequence $n = T_1^{s_1}, T_2^{s_1}, \ldots$ rather than the full sequence $n = 1, 2, \ldots$. To go from (14) to the full limit, we proceed as follows. Fix $\varepsilon > 0$ and note that (14) tells us that there exists some (random) $N < \infty$ such that

$$\frac{1}{\tau_1} - \varepsilon \le \frac{K(T_m^{s_1}, s_1)}{T_m^{s_1}} \le \frac{1}{\tau_1} + \varepsilon$$

whenever $T_m^{s_1} \ge N$. Now take (for a fixed realization of the Markov chain such that the event in (14) happens) an n between $T_m^{s_1}$ and $T_{m+1}^{s_1}$, where m is large enough so that $T_m^{s_1} \ge N$. For such an n we have

$$\begin{aligned}\frac{K(n, s_1)}{n} &= \frac{K(T_m^{s_1}, s_1)}{n} \\ &\le \frac{K(T_m^{s_1}, s_1)}{T_m^{s_1}} \\ &\le \frac{1}{\tau_1} + \varepsilon \,,\end{aligned}$$

while, in the other direction,

$$\begin{aligned}\frac{K(n, s_1)}{n} &= \frac{K(T_{m+1}^{s_1}, s_1) - 1}{n} \\ &\ge \frac{K(T_{m+1}^{s_1}, s_1)}{T_{m+1}^{s_1}} - \frac{1}{n} \\ &\ge \frac{1}{\tau_1} - \varepsilon - \frac{1}{n} \,.\end{aligned}$$

Combining these inequalities into a single sandwiching relation, we thus have

$$\frac{1}{\tau_1} - \varepsilon - \frac{1}{n} \le \frac{K(n, s_1)}{n} \le \frac{1}{\tau_1} + \varepsilon$$

for all sufficiently large n, and since ε was arbitrary, (12) follows.

It remains to prove (13). For $i = 2, 3, \ldots, k$, let $D_m^{s_1,s_i}$ denote the number of visits to state s_i between the $(m-1)$th and the mth return to state s_1. Note that

$$K(T_m^{s_1}, s_i) = D_1^{s_1,s_i} + \ldots + D_m^{s_1,s_i} .$$

Furthermore, $D_1^{s_1,s_i}, D_2^{s_1,s_i}, \ldots$ are i.i.d. random variables with mean ρ_{1i}, whence another application of the Strong Law of Large Numbers yields that, with probability 1,

$$\begin{aligned}
\lim_{m\to\infty} \frac{K(T_m^{s_1}, s_i)}{K(T_m^{s_1}, s_1)} &= \lim_{m\to\infty} \frac{K(T_m^{s_1}, s_i)}{m} \\
&= \lim_{m\to\infty} \frac{1}{m} \sum_{l=1}^{m} D_l^{s_1,s_i} \\
&= \rho_{1i} .
\end{aligned}$$

Thus, the limit

$$\lim_{n\to\infty} \frac{K(n, s_i)}{K(n, s_1)}$$

exists along the subsequence $n = T_1^{s_1}, T_2^{s_1}, \ldots$, where it equals ρ_{1i}. To extend it to the full sequence $n = 1, 2, \ldots$, we employ the same epsilon-delta type argument as when going from (14) to (12) – the details are left to the reader – which yields

$$\lim_{n\to\infty} \frac{K(n, s_i)}{K(n, s_1)} = \rho_{1i} \text{ a.s.} \tag{15}$$

Finally, combining (15) and (12) yields (13), and the proof is complete. $\diamond$

*

While Theorem 3.2 on the limiting behavior of the Markov chain is of course highly interesting in its own right, it also has a nice implication for the problem of uniqueness of stationary distributions for so-called **irreducible** finite-state Markov chains.

For a discrete time Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and a given transition matrix, recall that $s_i \to s_j$ is used to denote the fact that if the chain sits in state s_i, it has positive probability of reaching s_j at some later time. The Markov chain is said to be **irreducible** if $s_i \to s_j$ for all $s_i, s_j \in \mathcal{S}$.

Theorem 4.1. *Any irreducible finite-state discrete time Markov chain has a unique stationary distribution.*

To prove this, we begin with a simple lemma.

Lemma 4.1. *Let $(X_0, X_1, \ldots)$ be a Markov chain with finite state space $\mathcal{S} = \{s_1, \ldots, s_k\}$. If the chain is irreducible, then for any $s_i \in \mathcal{S}$ and no matter how the chain is started, it will a.s. sooner or later reach state s_i.*

Proof: Follows by the same arguments as in the proof of Lemma 3.2. $\diamond$

Proof of Theorem 4.1: Write $\mathcal{S} = \{s_1, \ldots, s_k\}$ for the state space of the Markov chain. By Lemma 4.1, state s_1 is recurrent, and a stationary distribution π is therefore provided by Proposition 3.1:

$$\pi = \left(\frac{1}{\tau_1}, \frac{\rho_{12}}{\tau_1}, \frac{\rho_{13}}{\tau_1}, \cdots, \frac{\rho_{1k}}{\tau_1} \right) .$$

Next let π' be an arbitrary stationary distribution for this Markov chain; if we then manage to show that $\pi' = \pi$, then we know that *every* stationary distribution equals π, and the desired uniqueness conclusion is established.

Suppose now that we start the chain by choosing X_0 according to π'. Then X_n has distribution π' for any n. Furthermore, if N is a $\{0, 1, 2, \ldots\}$-valued random variable which is independent of the chain $(X_0, X_1, \ldots)$, then X_N has again distribution π'.

For $s_i \in \mathcal{S}$ and positive integers n_1 and n_2 with $n_1 \leq n_2$, write

$$K(\{n_1, \ldots, n_2\}, s_i)$$

for the number of visits by the chain to state s_i during the time interval $\{n_1, \ldots, n_2\}$.

By another application of Lemma 4.1, the chain started according to π' will a.s. sooner or later hit state s_1. Write T_{s_1} for this first encounter with state s_1. If we look at the chain from time T_{s_1} and on, the distribution of what we see is exactly the same as if we had started the chain in state s_1 and looked from time 0 and onwards. It therefore follows from Theorem 3.2 that, for $i = 2, 3, \ldots, k$,

$$\lim_{n \to \infty} \frac{K(\{T_{s_1} + 1, \ldots, n\}, s_i)}{n - T_{s_1}} = \frac{\rho_{1i}}{\tau_1} \text{ a.s.} \tag{16}$$

But the ratios $\frac{K(n, s_i)}{K(\{T_{s_1}+1, \ldots, n\}, s_i)}$ and $\frac{n}{n - T_{s_1}}$ both tend to 1 as $n \to \infty$, so it follows from (16) that

$$\lim_{n \to \infty} \frac{K(n, s_i)}{n} = \frac{\rho_{1i}}{\tau_1} \text{ a.s.}$$

Given $\varepsilon > 0$, we can thus find an n large enough so that

$$\mathbf{P}\left(\frac{\rho_{1i}}{\tau_1} - \varepsilon \leq \frac{K(n, s_i)}{n} \leq \frac{\rho_{1i}}{\tau_1} + \varepsilon\right) \geq 1 - \varepsilon\,. \tag{17}$$

Fix such an n, and let the random variable N be uniformly distributed on $\{1, \ldots, n\}$ and independent of the Markov chain. Then X_N has distribution π', but we also know from (17) that

$$\frac{\rho_{1i}}{\tau_1} - 2\varepsilon \leq \mathbf{P}(X_N = s_i) \leq \frac{\rho_{1i}}{\tau_1} + 2\varepsilon\,.$$

Since $\varepsilon > 0$ was arbitrary, we thus have

$$\begin{aligned} \pi'_i &= \frac{\rho_{1i}}{\tau_1} \\ &= \pi_i\,. \end{aligned}$$

This argument we did for $i = 2, 3, \ldots, k$, but since $\sum_{i=1}^{k} \pi_i$ and $\sum_{i=1}^{k} \pi'_i$ both equal 1, we may conclude that $\pi'_1 = \pi_1$ as well. Thus $\pi' = \pi$, as desired. $\diamond$

5 Monday, Sep 20

So now we know that an irreducible finite-state Markov chain has a unique stationary distribution π. From this we would like to conclude that no matter how the chain starts, if we run the chain for a sufficiently long time n, then X_n will have distribution π. This, however, is too much to ask for, as for instance the following simple example suggests.

Example 5.1. Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\mathcal{S} = \{s_1, s_2\}$, with initial value $X_0 = s_1$, and with transition matrix

$$P = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}.$$

This Markov chain is irreducible, and by Theorem 4.1 it therefore has a unique stationary distribution π. Write $\mu^{(n)}$ for the distribution of X_n. One can check (see Problem 5.1) that $\pi = (\frac{1}{2}, \frac{1}{2})$, while on the other hand, for each n,

$$\mu^{(n)} = \left(\frac{1}{2} + \frac{1}{2^{n+1}}, \frac{1}{2} - \frac{1}{2^{n+1}}\right).$$

$\diamond$

We see in this example that $\mu^{(n)} \neq \pi$ for every n. On the other hand, it is clear that in some sense that it comes "arbitrarily close" to π. In order to make this precise, we need to introduce some metric on the space of probability distributions on $\mathcal{S}$. A suitable metric is the so-called **total variation distance**.[13]

Definition 5.1. *Let $\mathcal{S}$ be an arbitrary set equipped with a σ-field Σ. For two probability measures μ and ν on $\mathcal{S}$, we define their* **total variation distance**, *denoted $d_{\mathrm{TV}}(\mu,\nu)$, as*

$$d_{\mathrm{TV}}(\mu,\nu) \;=\; \sup_{E\in\Sigma} |\mu(E)-\nu(E)|\,. \tag{18}$$

Thus, $d_{\mathrm{TV}}(\mu,\nu)$ is the maximum discrepancy between the μ-probability and the ν-probability among all measurable subsets of $\mathcal{S}$. When $\mathcal{S}$ is finite or countable, we need not worry about particular σ-fields. Or rather, we can then take *all* subsets of $\mathcal{S}$ to be measurable.[14] When $\mathcal{S}$ is finite, the following device for computing total variation distance if often useful.[15]

Lemma 5.1. *If $\mathcal{S} = \{s_1,\ldots,s_k\}$ is finite, then the total variation distance between two probability measures $\mu = (\mu_1,\ldots,\mu_k)$ and $\nu = (\nu_1,\ldots,\nu_k)$ satisfies*

$$d_{\mathrm{TV}}(\mu,\nu) \;=\; \frac{1}{2}\sum_{i=1}^{k} |\mu_i-\nu_i|\,. \tag{19}$$

Proof: See Problem 5.2. $\diamond$

For instance, we get for the Markov chain in Example 5.1 that

$$d_{\mathrm{TV}}(\mu^{(n)},\pi) \;=\; \frac{1}{2^{n+1}}\,.$$

Thus, $d_{\mathrm{TV}}(\mu^{(n)},\pi)$ tends to 0 as $n\to\infty$, and we therefore say that $\mu^{(n)}$ **converges to π in total variation as $n\to\infty$.**

A reasonable candidate for a theorem on the long-term behavior of irreducible finite-state Markov chains might now be that the distribution $\mu^{(n)}$

[13] The reader should be aware that two conflicting definitions of total variation distance between probability measures appear in the literature. Most authors go for the definition employed here, but some – notably, Lindvall (1992) and Thorisson (2000) – define it as twice the quantity we take it to be here. This has the advantage of removing the factor $\frac{1}{2}$ from the right-hand side of the formula (19), but of course at the cost of introducing a factor 2 in the right-hand side of (18).

[14] This is what we implicitly do all the time.

[15] The lemma also extends to the case where $\mathcal{S}$ is countable; see Problem 5.2.

of the chain at time n, converges in total variation to the unique stationary distribution π. This is too strong, however, as the following example shows:

Example 5.2. Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\mathcal{S} = \{s_1, s_2\}$, with initial value $X_0 = s_1$, and with transition matrix

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

It is easy to check that this Markov chain is irreducible with unique stationary distribution $\pi = (\frac{1}{2}, \frac{1}{2})$. However, the distribution $\mu^{(n)}$ of the chain at time n is

$$\mu^{(n)} = \begin{cases} (1,0) & \text{for } n \text{ even} \\ (0,1) & \text{for } n \text{ odd.} \end{cases}$$

Thus, $d_{\mathrm{TV}}(\mu^{(n)}, \pi) = \frac{1}{2}$ for all n, so $\mu^{(n)}$ does *not* converge to π in total variation as $n \to \infty$. $\diamond$

Some more condition is therefore needed to ensure convergence towards the unique stationary distribution. What goes wrong in Example 5.2 is a kind of deterministic oscillation phenomenon. The most general condition which avoids such phenomena in irreducible finite-state Markov chains is so-called **aperiodicity**, defined as follows.

Definition 5.2. *Consider a discrete time Markov chain with finite or countable state space $\mathcal{S}$ and transition matrix P. For $s \in S$, write Δ_s for the set of times n at which $\mathbf{P}(X_n = s) > 0$ provided that the chain starts with $X_0 = s$. Then let*

$$\Delta = \bigcup_{s \in \mathcal{S}} \Delta_s$$

and define the **period** *of the Markov chain to be*

$$\gcd(\Delta)\,,$$

the greatest common divisor of the numbers in Δ. If the period is 1, then the Markov chain is said to be **aperiodic**.

For example, the Markov chain in Example 5.2 has $\Delta = \{2, 4, 6, 8, \ldots\}$ and therefore period 2. A simple condition guaranteeing aperiodicity is to have some nonzero element on the diagonal of the transition matrix, i.e., to have some state where the chain has nonzero probability of staying for more than one time unit.[16]

[16] The reverse implication does not hold; see Problem 5.3 (b).

It turns out that in order to guarantee convergence of $\mu^{(n)}$ to the unique stationary distribution π, adding periodicity to the previously proposed conditions is enough:

Theorem 5.1. *Suppose that $(X_0, X_1, \ldots)$ is a finite-state irreducible and aperiodic Markov chain whose (unique) stationary distribution is π. Then, no matter how the distribution of the initial value X_0 is chosen, the distribution of X_n converges in total variation to π as $n \to \infty$.*

The best way to prove this result uses **coupling**[17]. I have demonstrated this proof on various occasions, for instance in propagandizing for the coupling method, and in order not to repeat myself too much I will here just give a sketch. For the details, the reader may consult pretty much any modern introductory book on Markov chains, or alternatively Lindvall (1992) or Thorisson (2000).

Sketch proof of Theorem 5.1: The basic idea is to construct the chain $(X_0, X_1, \ldots)$ whose initial value is chosen in some arbitrary fashion, together with another chain $(X'_0, X'_1, \ldots)$ which has the same transition matrix but is started acording to the stationary distribution π. The chains run independently until the first time T that they meet, and after that, they run together, i.e.,

$$X_n \;=\; X'_n \text{ for all } n \geq T\,.$$

It turns out, using irreducibility and aperiodicity, that the chains eventually do meet with probability 1, i.e., that

$$\mathbf{P}(T < \infty) \;=\; 1\,.$$

So if we pick n sufficiently large, then $T \leq n$ with probability close to 1, or in other words $X_n = X'_n$ with probability close to 1. But then the distribution of X_n cannot differ by much from the distribution of X'_n,[18] and since the distribution of X'_n is π for any n, the theorem follows. $\diamond$

*

Let us now go back to the Law of Large Numbers-type (LLN) result – Theorem 3.2 – that we looked at and proved last week, and see how easily we

[17] The modern meaning of **coupling** in probability theory is, roughly speaking, the simultaneous construction on the same probability space of two or more random objects (variables, processes, etc) for the purpose of drawing conclusions about their respective distributions.

[18] This uses the so-called **coupling inequality**, see Problem 5.4.

can obtain some variants of it. First of all, if the Markov chain considered in Theorem 3.2 is irreducible, then the choice of X_0 has no influence on the LLN behavior, as seen in Corollary 5.1 below. Recall that $K(n, s_i)$ denotes the number of visits by the chain to state s_i during the time interval $\{1, 2, \ldots, n\}$.

Corollary 5.1. *Let $(X_0, X_1, \ldots)$ be an irreducible aperiodic Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and (unique) stationary distribution $\pi = (\pi_1, \ldots, \pi_k)$. Then, for any $s_i \in \mathcal{S}$,*

$$\lim_{n\to\infty} \frac{K(n, s_i)}{n} = \pi_i \ \ a.s., \tag{20}$$

regardless of the distribution of X_0.

Proof: Since the chain is irreducible, s_1 is a recurrent state, so if the chain is started deterministically in state s_1, then Theorem 3.2 tells us that for $i = 1, \ldots, k$, $\frac{K(n,s_i)}{n}$ converges a.s. to the values predicted by the stationary distribution obtained in Proposition 3.1. But by uniqueness (Theorem 4.1) this distribution equals π, and (20) follows.

The cases where the chain starts deterministically in one of the states $s_2, \ldots, s_k$ follow immediately, since state s_1 plays no special role in the statement of the corollary.

Finally, the cases where X_0 is chosen randomly (non-deterministically) are taken care of by conditioning on the value of X_0. $\diamond$

Corollary 5.2. *Let the Markov chain $(X_0, X_1, \ldots)$ be as in Corollary 5.1, and let $f : \mathcal{S} \to \mathbf{R}$ be an arbitrary real-valued function of the state space. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n} f(X_l) = \mu_f \ \ a.s.,$$

where

$$\mu_f = \sum_{i=1}^{k} \pi_i f(s_i)\,. \tag{21}$$

Proof: First note that

$$\sum_{l=1}^{n} f(X_l) = \sum_{i=1}^{k} K(n, s_i) f(s_i)\,. \tag{22}$$

We get

$$\begin{aligned}
\mu_f &= \sum_{i=1}^k f(s_i)\pi_i \\
&= \sum_{i=1}^k f(s_i) \lim_{n\to\infty} \frac{K(n,s_i)}{n} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^k K(n,s_i) f(s_i) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^n f(X_l)
\end{aligned}$$

where the second equality uses Corollary 5.1, and the fourth uses (22). $\diamond$

When a Law of Large Numbers-type result such as Theorem 3.2 or Corollary 5.2 is established, then the almost automatic reply by a classically trained probabilist is to ask whether a corresponding central limit theorem (CLT) also holds. The answer, in this case, is yes, as we shall see in the next lecture.

6 Monday, Sep 27

Consider, as before, a Markov chain $(X_0, X_1, \ldots)$ with finite state space $\mathcal{S} = \{s_1, \ldots, s_k\}$, starting in state $X_0 = s_1$ which is assumed to be a recurrent state. Let

$$\begin{aligned}
\pi &= (\pi_1, \pi_2, \ldots, \pi_k) \\
&= \left(\frac{1}{\tau_1}, \frac{\rho_{12}}{\tau_1}, \frac{\rho_{13}}{\tau_1}, \cdots, \frac{\rho_{1k}}{\tau_1} \right)
\end{aligned} \tag{23}$$

denote the stationary distribution established in Proposition 3.1, and let f be some arbitrary function from $\mathcal{S}$ to $\mathbf{R}$. Define, as in (21),

$$\mu_f = \sum_{i=1}^k \pi_i f(s_i)\,, \tag{24}$$

so that in other words μ_f is the expected value of an $\mathcal{S}$-valued random variable chosen according to π. Furthermore, take as usual T_{s_1} to denote the first return time of the chain to state s_1, and define

$$\sigma_f^2 = \frac{1}{\tau_1} \mathbf{E} \left(\sum_{i=1}^{T_{s_1}} (f(X_i) - \mu_f) \right)^2 . \tag{25}$$

Since T_{s_1} is an unbounded random variable, it is not a priori clear that $\sigma_f^2 < \infty$, but this can in fact be proved using Lemma 3.2 concerning the exponential tail of the distribution of T_{s_i}; see Problem 6.1.

Corollary 5.2 tells us that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n} f(X_l) \;=\; \mu_f \;\text{ a.s.}$$

The following CLT-type result says that the deviations from μ_f that do appear even for large n, are typically of the order $\frac{1}{\sqrt{n}}$ and asymptotically normal.

Theorem 6.1.[19] *Consider a Markov chain $(X_0, X_1, \ldots)$ with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ starting in state $X_0 = s_1$ which is assumed to be recurrent, and a function $f : \mathcal{S} \to \mathbf{R}$. Define μ_f and σ_f^2 as in (24) and (25), and furthermore define*

$$Z_n \;=\; \frac{(\sum_{l=1}^{n} f(X_l)) - n\mu_f}{\sqrt{\sigma_f^2 n}} \,.$$

We then have that Z_n converges in distribution to a normal distribution with mean 0 *and variance* 1*. That is, for any $x \in \mathbf{R}$,*

$$\lim_{n\to\infty} \mathbf{P}(Z_n \le x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) dy \,.$$

We shall have a look both at a sketch proof of this result, giving a rough intuition for why it is true, and then a rigorous proof with all the details. Before that, however, a couple of remarks are in order.

1. Suppose we are interested in the CLT fluctuations of the number of visits to a given state, say s_1. Then a suitable choice of f is to take

   $$f(s_i) = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

   This gives $\mu = \frac{1}{\tau_1}$, while the asymptotic variance σ_f^2 simplifies to

   $$\sigma_f^2 \;=\; \frac{1}{\tau_1}\mathbf{E}\left(\sum_{i=1}^{T_{s_1}} (f(X_i) - \mu_f)\right)^2$$

[19] An implicit assumption in the theorem is that $\sigma_f^2 > 0$; otherwise, of course, its statement does not make sense. An example where $\sigma_f^2 = 0$ is of course when f is a constant, but there are other, slightly less trivial, examples. See Problem 6.2.

$$
\begin{aligned}
&= \frac{1}{\tau_1}\mathbf{E}\left(1-\frac{T_{s_1}}{\tau_1}\right)^2 \\
&= \frac{1}{\tau_1}\mathbf{Var}\left[\frac{T_{s_1}}{\tau_1}\right] \\
&= \frac{\mathbf{Var}[T_{s_1}]}{\tau_1^3}\,.
\end{aligned}
$$

2. Suppose that the Markov chain is irreducible. Then the CLT behavior predicted in Theorem 6.1 holds regardless of how the chain was started; this follows by the same argument that we used to go from Theorem 3.2 to Corollary 5.1. On the other hand, if the chain is *not* irreducible, then asymptotic normality may fail (see Problem 6.3).

The following lemma will be useful in our pursuit of a proof of Theorem 6.1.

Lemma 6.1. *For the Markov chain in Theorem 6.1, we have*

$$\mathbf{E}\left[\sum_{l=1}^{T_{s_1}} f(X_l)\right] = \mu_f \tau_1\,.$$

Proof: We have

$$\sum_{l=1}^{T_{s_1}} f(X_l) = f(s_1) + \sum_{i=2}^{k} K(T_{s_1}, s_i) f(s_i)$$

and taking expectations yields

$$
\begin{aligned}
\mathbf{E}\left[\sum_{l=1}^{T_{s_1}} f(X_l)\right] &= f(s_1) + \sum_{i=2}^{k} \mathbf{E}[K(T_{s_1}, s_i)] f(s_i) \\
&= f(s_1) + \sum_{i=2}^{k} \rho_{1i} f(s_i) \\
&= \sum_{i=1}^{k} \tau_1 \pi_i f(s_i) \\
&= \tau_1 \sum_{i=1}^{k} \pi_i f(s_i) \\
&= \tau_1 \mu_f\,,
\end{aligned}
$$

where the third inequality uses (23). $\diamond$

In the following, we shall write $N(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2.

Both in the sketch proof and in the actual rigorous proof of Theorem 6.1, it simplifies to consider the special case $\mu_f = 0$. There is no loss of generality in doing so, because if $\mu_f \neq 0$, we may instead of f consider the function $\tilde{f} = f - \mu_f$, and once we have proved that $\sum_{l=1}^{n} \tilde{f}(X_l)$ is approximately $N(0, n\sigma_f^2)$, it follows immediately that $\sum_{l=1}^{n} f(X_l)$ is approximately $N(n\mu_f, n\sigma_f^2)$.

Sketch proof of Theorem 6.1: Assume for simplicity (and without loss of generality) that $\mu_f = 0$. Write $\bar{\sigma}_{s_1}^2$ for the variance $\mathbf{Var}[T_{s_1}]$ of the time of first return to s_1. We have that $\bar{\sigma}_{s_1}^2 < \infty$ by the same argument as for the finiteness of σ_f^2 (recall Problem 6.1). Retrieving some notation from the proof of Theorem 3.2, write $T_m^{s_1}$ for the time of the mth return to state s_1, and

$$D_m^{s_1} = T_m^{s_1} - T_{m-1}^{s_1}$$

for the interarrival times, so that

$$T_m^{s_1} = \sum_{j=1}^{m} D_j^{s_1} .$$

Fixing a large n, let $m = \frac{n}{\tau_1}$ (rounding off to the nearest integer if needed). Then, since $D_1^{s_1}, D_2^{s_1}, \ldots$ are i.i.d. with mean τ_1 and variance $\bar{\sigma}_{s_1}^2$, we have from the classical CLT for sums of i.i.d. random variables that $T_m^{s_1}$ is approximately $N(m\tau_1, m\bar{\sigma}_{s_1}^2) \approx N(n, m\bar{\sigma}_{s_1}^2)$. Furthermore, defining

$$S_j^{s_1} = \sum_{l=T_{mj1}^{s_1}+1}^{T_j^{s_1}} f(X_l) ,$$

we may note that $S_1^{s_1}, S_2^{s_1}, \ldots$ are i.i.d. with mean $\mu_f \tau_1 = 0$ (due to Lemma 6.1 and the assumption that $\mu_f = 0$) and variance

$$\begin{aligned}
\mathbf{Var}[S_i^{s_i}] &= \mathbf{Var}\left[\sum_{l=1}^{T_{s_1}} f(X_l)\right] \\
&= \mathbf{E}\left[\left(\left(\sum_{l=1}^{T_{s_1}} f(X_l)\right) - \tau_1 \mu_f\right)^2\right] \\
&= \mathbf{E}\left[\left(\sum_{l=1}^{T_{s_1}} f(X_l)\right)^2\right]
\end{aligned}$$

$$\begin{aligned} &= \mathbf{E}\left[\left(\sum_{l=1}^{T_{s_1}}(f(X_l)-\mu_f)\right)^2\right] \\ &= \tau_1\sigma_f^2\,, \end{aligned}$$

where the third and fourth equalities exploit the assumption that $\mu_f = 0$, while the last one uses the definition (25) of σ_f^2. Also note that

$$\sum_{l=1}^{T_m^{s_1}} f(X_l) = \sum_{j=1}^{m} S_j^{s_1}$$

Another application of the classical CLT thus gives that

$$\sum_{l=1}^{T_m^{s_1}} \text{ is approximately } N(0, m\tau_1\sigma_f^2)$$

which by choice of m is (very close to) $N(0, n\sigma_f^2)$. Or in other words, if we set

$$Z_n^* = \frac{\sum_{l=1}^{T_m^{s_1}} f(X_l)}{\sqrt{n\sigma_f^2}}\,, \tag{26}$$

we get that Z_n^* is approximately $N(0,1)$.

Note now that (again using the assuption $\mu_f = 0$)

$$Z_n = \frac{\sum_{l=1}^{n} f(X_l)}{\sqrt{n\sigma_f^2}}$$

which is what we need to show is approximately $N(0,1)$. Comparing this to (26), we see that we are close to this goal, because we know that $T_m^{s_1}$ is (most likely) fairly close to n, and therefore Z_n^* and Z_n are (most likely) close to each other. More precisely, the time discrepancy $T_m^{s_1} - n$ is approximately $N(0, m\bar{\sigma}_{s_1}^2)$. We thus have that with overwhelming probability, $T_m^{s_1} - n$ is at most of the order $\sqrt{m\bar{\sigma}_{s_1}^2} = \sqrt{\frac{n}{\tau_1}\bar{\sigma}_{s_1}^2}$, corresponding to of the order $\frac{1}{\tau_1}\sqrt{\frac{n}{\tau_1}\bar{\sigma}_{s_1}^2}$ revisits to state s_1. The difference between $\sum_{l=1}^{T_m^{s_1}} f(X_l)$ and $\sum_{l=1}^{n} f(X_l)$ is therefore a sum of of the order $\frac{1}{\tau_1}\sqrt{\frac{n}{\tau_1}\bar{\sigma}_{s_1}^2}$ $S_j^{s_1}$-variables, and another application of the classical CLT gives that this sum is at most of the order $\sqrt{\frac{1}{\tau_1}\sqrt{\frac{n}{\tau_1}\bar{\sigma}_{s_1}^2}}$, i.e. of the order $n^{1/4}$. Dividing this by $\sqrt{n\sigma_f^2}$ yields the difference between Z_n^* and Z_n, which we see is of the order $n^{-1/4}$. This tends to 0 as $n \to \infty$, so if Z_n^* is approximately $N(0,1)$, then so is Z_n. $\diamond$

The last paragraph of the sketch proof makes such ample and free use of terms like "approximately" and "of the order" that sensitive readers may become a bit nervous, and a stricter treatment is called for. The final sentence of the sketch proof also makes implicit use of the following (intuitively obvious) lemma, which will be used again in the real proof.

Lemma 6.2. *Let $Y_1, Y_2, \ldots$ and $W_1, W_2, \ldots$ be real-valued random variables defined on the same probability space, and satisfying*

$$Y_n \textit{ converges in distribution to } N(0,1) \textit{ as } n \to \infty$$

and

$$W_n \textit{ converges in distribution to a point mass at } 0 \textit{ as } n \to \infty.$$

Then (regardless of any dependencies between the random variables)

$$(Y_n + W_n) \textit{ converges in distribution to } N(0,1) \textit{ as } n \to \infty.$$

Proof: See Problem 6.5.[20]

In the following, $\lceil x \rceil$ denotes "x rounded up to the nearest integer".

Proof of Theorem 6.1: As before, we may assume without loss of generality that $\mu_f = 0$. In the sketch proof, we used Z_n^* as a preliminary approximation of Z_n. Here, we shall make similar use of two other such approximations Z_n' and Z_n'', defined as follows. We take

$$Z_n' = \frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=1}^{T_m^{s_1}} f(X_l)$$

where[21] $m = \left\lceil \frac{n - n^{3/4}}{\tau_1} \right\rceil$, and

$$Z_n'' = \frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=1}^{N} f(X_l)$$

[20] In fact, there is nothing special about the choice of $N(0,1)$ here; the lemma extends easily to the setting where $N(0,1)$ is replaced by an arbitrary distribution.

[21] The point of the choice $m = \lceil \frac{n-n^{3/4}}{\tau_1} \rceil$ (rather than simply $\lceil \frac{n}{\tau_1} \rceil$) is to get $T_m^{s_1} < n$ with overwhelming probability as $n \to \infty$; why this is convenient will be clear later on.

where

$$N = \min\{l : l > n, X_l = s_1\},$$

i.e., N is the (random) time at which the chain first encounters state s_1 after time n.

We trivially have

$$Z_n = Z'_n + (Z''_n - Z'_n) + (Z_n - Z''_n)$$

and we will proceed to show that, as $n \to \infty$,

(a) Z'_n converges in distribution to $N(0,1)$,

(b) $(Z''_n - Z'_n)$ converges in distribution to 0, and

(c) $(Z_n - Z''_n)$ converges in distribution to 0.

Once (a), (b) and (c) have been established, we can use Lemma 6.2 twice – first to Z'_n and $(Z''_n - Z'_n)$, and then to Z''_n and $(Z_n - Z''_n)$ – to conclude that Z_n converges in distribution to $N(0,1)$. Thus, our only remaining task is to prove (a), (b) and (c).

To show (a), we follow the approach of the sketch proof by using the decomposition

$$\sum_{l=1}^{T_m^{s_1}} f(X_l) = \sum_{j=1}^{m} S_j^{s_1}.$$

Since $S_1^{s_1}, S_2^{s_1}, \ldots$ are i.i.d. with mean 0 and variance $\tau_1 \sigma_f^2$ we get from the classical CLT that

$$\frac{1}{\sqrt{\left\lceil \frac{n - n^{3/4}}{\tau_1} \right\rceil \tau_1 \sigma_f^2}} \sum_{l=1}^{T_m^{s_1}} f(X_l) \text{ converges in distribution to } N(0,1) \text{ as } n \to \infty. \tag{27}$$

Noting next that

$$\lim_{n \to \infty} \frac{\left\lceil \frac{n - n^{3/4}}{\tau_1} \right\rceil \tau_1}{n} = 1,$$

we see that we can replace $\left\lceil \frac{n - n^{3/4}}{\tau_1} \right\rceil \tau_1$ by n in (27), so (a) is established.

We go on to prove (c) (leaving the trickiest part (b) for last). By definitions of Z_n and Z''_n, we have

$$\begin{aligned} Z''_n - Z_n &= \frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=n+1}^{N} f(X_l) \\ &\geq 0. \end{aligned} \tag{28}$$

Furthermore, given any $\varepsilon > 0$, there exists a $t > \infty$ such that no matter what value X_n takes, the conditional probability that $N - l \leq t$ is at least $1 - \varepsilon$. Writing $f_{\max}$ for $\max\{f(s_1), \ldots, f(s_k)\}$, we have on that event that $\sum_{l=n+1}^{N} f(X_l) \leq t f_{\max}$. Thus,

$$\mathbf{P}\left(Z_n'' - Z_n \leq \frac{t f_{\max}}{\sqrt{n\sigma_f^2}} \right) \geq 1 - \varepsilon \,, \tag{29}$$

where we may note that $\frac{t f_{\max}}{\sqrt{n\sigma_f^2}} \to 0$ as $n \to \infty$. Since $\varepsilon > 0$ was arbitrary, we may conclude from (28) and (29) that $(Z_n'' - Z_n)$ converges in distribution to 0, and (c) follows.

Finally, we move on to proving (b). With interarrival times $D_1^{s_1}, D_2^{s_1}, \ldots$ defined as in the sketch proof, we can invoke the classical CLT to conclude that

$$\frac{T_m^{s_1} - m\tau_1}{\sqrt{m\bar{\sigma}_{s_1}^2}} \text{ converges in distribution to } N(0,1) \text{ as } n \to \infty.$$

In particular, since $\frac{n - m\tau}{\sqrt{m\bar{\sigma}_{s_1}^2}}$ tends to ∞ as $n \to \infty$, we get with probability approaching 1 as $n \to \infty$ that

$$T_m^{s_1} < n \,. \tag{30}$$

Now imaginge running the chain until time $T_m^{s_1}$, pausing, and then running it for another $m^* = \left\lceil \frac{2n^{3/4}}{\tau_1} \right\rceil$ visits to s_1, i.e., until time $T_{m+m^*}^{s_1}$. Similarly as we concluded that (30) holds with overwhelming probability, we also get $T_{m+m^*}^{s_1} > N$ with overwhelming probability. Yet another application of the classical CLT (this time applied to $S_{m+1}^{s_1}, S_{m+2}^{s_1}, \ldots, S_{m+m^*}^{s_1}$) yields that

$$\frac{1}{\sqrt{m^* \tau_1 \sigma_f^2}} \sum_{l=T_m^{s_1}+1}^{T_{m+m^*}^{s_1}} f(X_l) \text{ converges in distribution to } N(0,1) \text{ as } n \to \infty.$$

In particular, since $\frac{m^* \tau_1}{n} \to 0$ as $n \to \infty$, we have that

$$\frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=T_m^{s_1}+1}^{T_{m+m^*}^{s_1}} f(X_l) \text{ converges in distribution to 0 as } n \to \infty. \tag{31}$$

Let us next imagine that when running the chain from time $T_m^{s_1}$ until time $T_{m+m^*}^{s_1}$, we pause for a moment at time N. At that time, the probability that there are at least $\frac{n^{3/4}}{2\tau_1}$ visits to s_1 still to do until time $T^{s_1}_{m+m^*}$, tends to 1 as $n \to \infty$. And since N is a stopping time, and we can go on as if the number of such visits left to carry out were a fixed constant, and thus apply the classical CLT once more to conclude that (similarly as for (31))

$$\frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=N+1}^{T^{s_1}_{m+m^*}} f(X_l) \text{ converges in distribution to 0 as } n \to \infty. \tag{32}$$

Now fix $\varepsilon > 0$, and suppose for contradiction that

$$\limsup_{n\to\infty} \mathbf{P}\left(\left| \frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=T^{s_1}_m+1}^{N} \right| > \varepsilon \right) \geq \varepsilon\,. \tag{33}$$

Using (32), this would imply that

$$\limsup_{n\to\infty} \mathbf{P}\left(\left| \frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=T^{s_1}_m+1}^{T^{s_1}_{m+m^*}} \right| > \frac{\varepsilon}{2} \right) \geq \frac{\varepsilon}{2}$$

contradicting (31). Thus, (33) is false for any $\varepsilon > 0$, so that in other words

$$\frac{1}{\sqrt{n\sigma_f^2}} \sum_{l=T^{s_1}_m+1}^{N} \quad \text{converges in distribution to 0 as } n \to \infty. \tag{34}$$

But what we have in (34) is precisely $Z''_n - Z'_n$, so (b) is established, and the proof is complete. $\diamond$

7 Thursday, Sep 30

It is time for the next major topic of this course: reversibility, and reversible Markov chains. We jump right on to a definition:

Definition 7.1. *Consider a discrete time Markov chain $(X_0, X_1, \ldots)$ with finite state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and a given transition matrix P. A probability measure $\pi = (\pi_1, \ldots, \pi_k)$ on $\mathcal{S}$ is said to be* **reversible** *(for this Markov chain) if*

$$\pi_i P_{i,j} = \pi_j P_{j,i} \tag{35}$$

for all $i, j \in \{1, \ldots, k\}$.

What (35) says is that if X_0 is chosen according to π, then $\mathbf{P}(X_0 = s_i, X_1 = s_j) = \mathbf{P}(X_0 = s_j, X_1 = s_i)$: the probability of seeing a jump from state s_i to state s_j is the same as that of the reverse jump. The left-hand side of (35) may be seen as a flow of probability mass from state s_i to state s_j as we move from time 0 to time 1, and (35) is therefore often called a **local equilibrium** condition. Our first result today is that reversibility (in the sense of Definition 7.1) implies stationarity:

Lemma 7.1. *Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and transition matrix P. If the probability distribution $\pi = (\pi_1, \ldots, \pi_k)$ on $\mathcal{S}$ is reversible for this chain, then it is also a stationary distribution for the same chain.*

Proof: Suppose that X_0 is chosen according to π; our task is then to show that X_1 also gets distribution π. For any $s_i \in \mathcal{S}$, we get

$$\begin{aligned} \mathbf{P}(X_1 = s_i) &= \sum_{j=1}^{k} \mathbf{P}(X_0 = s_j, X_1 = s_i) \\ &= \sum_{j=1}^{k} \pi_j P_{j,i} \\ &= \sum_{j=1}^{k} \pi_i P_{i,j} \\ &= \pi_i \sum_{j=1}^{k} P_{i,j} \\ &= \pi_i \, , \end{aligned}$$

where the third equality follows from the local equilibrium condition (35). Since s_i was an arbitrary state, we can conclude that X_1 has distribution π. $\diamond$

From the notion of reversibility of distributions, it is natural to proceed to a notion of reversibility of the Markov chain *itself*, as follows:

Definition 7.2. *A discrete time Markov chain with finite state space is said to be reversible if all its stationary distributions are reversible.*

In particular, if we have a finite-state Markov chain for which we have found a reversible distribution π, and if we know that the chain is irreducible, then we know that π is the *only* stationary distribution, and it follows that the chain is reversible.

*

Before moving to deep into the *theory* of reversible Markov chains, we deserve to see some *examples*. An interesting class of examples is **random walks on graphs**.[22]

A **graph** $G = (V, E)$ is a set V of **vertices** together with a set E of **edges**, where each edge $e \in E$ is associated with exactly two vertices $x, y \in V$; we then write $e = \langle x, y \rangle$, and say that e **links** the vertices x and y. Multiple edges linking the same pair of vertices is not allowed. See Figure 7.1 for an example. Two vertices x and y are said to be **neighbors** if there exists an edge e linking x and y. The graph is said to be **finite** if V is finite, and **locally finite** if every vertex has only finitely many neighbors.

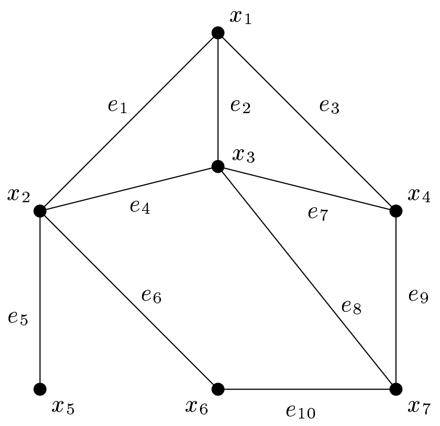


Figure 7.1.[23] *A graph $G = (V, E)$, with vertex set $V = \{x_1, x_2, \ldots, x_7\}$, and edge set $E = \{e_1, \ldots, e_{10}\}$, where $e_1 = \langle x_1, x_2 \rangle$, $e_2 = \langle x_1, x_3 \rangle$, etc.*

A random walk on a finite (or at least locally finite[24]) graph $G = (V, E)$ is a discrete time Markov chain $(X_0, X_1, \ldots)$ with state space V and the following transition mechanism: if $X_n = x$, and x has precisely d neighbors $y_1, \ldots, y_d$

[22] Random walks on graphs (and some generalizations) is the topic of the wonderful and easy-to-read book by Doyle and Snell (1984), which I strongly recommend as supplementary reading in connection with this chapter and some of those that follow.

[23] This example, like several others later on in this text, is borrowed from Häggström (2004).

[24] Without local finiteness, the following definition does not make sense, since it is in general not possible to choose among infinitely many neighbors with equal probabilities.

in the graph, then X_{n+1} is taken to be one of the vertices $y_1, \ldots, y_d$, each chosen with probability $\frac{1}{d}$. For instance, if we consider random walk on the graph G in Figure 7.1, and if $X_n = x_1$, then X_{n+1} is taken to be x_2, x_3 or x_4, each with probability $\frac{1}{3}$.

The choice of X_{n+1} in a random walk is taken to be as above even if we condition on the full past up until time n of the chain. Thus, a random walk is a Markov chain. The following result tells us more.

Theorem 7.1. *The Markov chain given by a random walk on a finite connected graph G with at least two vertices*[25]*, is reversible.*

Here, connected menas that for any two vertices x and y in the graph, it is possible to walk from x to y along edges of the graph. For the proof, it is convenient to define the degree of a vertex x, denoted $\deg(x)$, as the number of neighbors it has in G, and to define $\mathrm{sumdeg}(G)$ as the sum of $\deg(x)$ over all the vertices x in G.

Proof of Theorem 7.1: Since G is connected, it is clear that the Markov chain is irreducible. Thus, if we can find a reversible distribution π for the chain, then we know by irreducibility that π is the only stationary distribution for the chain, and we may then conclude that the chain is reversible.

Writing $V = \{x_1, \ldots, x_k\}$ for the vertex set of G, let us consider the distribution π on G given by

$$\begin{aligned} \pi &= (\pi_1, \ldots, \pi_k) \\ &= \left(\frac{\deg(x_1)}{\mathrm{sumdeg}(G)}, \frac{\deg(x_2)}{\mathrm{sumdeg}(G)}, \ldots, \frac{\deg(x_k)}{\mathrm{sumdeg}(G)} \right) . \end{aligned}$$

Clearly, π is a probability distribution on G. To show that π is reversible, we need to check that

$$\pi_i P_{i,j} = \pi_j P_{j,i} \tag{36}$$

for any $i, j \in \{1, \ldots, k\}$. For $i = j$, (36) holds trivially, so assume that $i \neq j$. This splits into two cases depending on whether or not x_i and x_j are neighbors of each other in G. If they are not neighbors, then both sides of (36) are 0, so it only remains to check (36) in the case where they are neighbors. Then $P_{i,j} = \frac{1}{\deg(x_i)}$ and $P_{j,i} = \frac{1}{\deg(x_j)}$, so that

$$\pi_i P_{i,j} = \frac{\deg(x_i)}{\mathrm{sumdeg}(G)} \cdot \frac{1}{\deg(x_i)}$$

[25] The condition of having at least two vertices is there just to ensure that the random walker always has somewhere to go: otherwise the random walk is not even well-defined.

$$\begin{aligned} &= \frac{1}{\mathrm{sumdeg}(G)} \\ &= \frac{\deg(x_j)}{\mathrm{sumdeg}(G)} \cdot \frac{1}{\deg(x_j)} \\ &= \pi_j P_{j,i}\,, \end{aligned}$$

and (36) is verified. ◇

Another class (beyond the random walks) of reversible Markov chains, are the so-called **weighted random walks**. When defining these, it is natural to extend the definition of a graph slightly by allowing so-called **self-edges**: a self-edge $e = \langle x, x\rangle$ is an edge leading from a vetex x to itself. A vertex having a self-edge counts as a neighbor of itself.

Assume that to the edges $e \in E$ of a graph $G = (V, E)$ we assign (strictly) positive **weights**, where the weight of e is denoted w_e. The weight w_x of a vertex $x \in V$ is taken to be the sum of w_e over all edges incident to e (where a self-edge at x is counted only once). Finally, if G is finite, we may define the weight w_G of the graph as $\sum_{x\in V} w_x$.

A weighted random walk on $G = (V, E)$ is the Markov chain with state space V and the following transition mechanism. If $X_n = x$, and x has neighbors $y_1, \ldots, y_d$, then X_{n+1} is chosen among $y_1, \ldots, y_d$ with respective probabilities

$$\frac{w_{\langle x,y_1\rangle}}{w_x}, \frac{w_{\langle x,y_2\rangle}}{w_x}, \ldots, \frac{w_{\langle x,y_d\rangle}}{w_x}\,.$$

Theorem 7.2. *Let $G = (V, E)$ be a graph (possibly with self-edges) with finite vertex set $V = \{x_1, \ldots, x_k\}$ and weights $\{W_e\}_{e\in E}$ attached to the edges. If G is connected, then the Markov chain given by weighted random walks on G with weights $\{W_e\}_{e\in E}$, is reversible, with stationary distribution*

$$\begin{aligned} \pi &= (\pi_1, \ldots, \pi_k) \\ &= \left(\frac{w_{x_1}}{w_G}, \frac{w_{x_2}}{w_G}, \ldots, \frac{w_{x_k}}{w_G}\right)\,. \end{aligned}$$

Proof: We need to check that the local equilibrium condition $\pi_i P_{i,j} = \pi_j P_{j,i}$ holds for every $i, j \in \{1, \ldots, k\}$. If x_i and x_j are not neighbors in the graph, then both sides of the local equilibrium condition is 0, while if they are neighbors with $e = \langle x_i, x_j\rangle$, we get

$$\pi_i P_{i,j} = \frac{w_{x_i}}{w_G} \cdot \frac{w_e}{w_{x_i}}$$

$$
\begin{aligned}
&= \frac{w_e}{w_G} \\
&= \frac{w_{x_j}}{w_G} \cdot \frac{w_e}{w_{x_j}} \\
&= \pi_j P_{i,j} \,.
\end{aligned}
$$

This shows that π is a reversible distribution for the chain. Finally, since G is connected, the chain is irreducible, so π is the only stationary distribution, whence the chain is reversible. $\diamond$

Having come this far, it is perhaps tempting to ask whether in fact *all* Markov chains are reversible. This, however, is far from the case, and Problem 7.1 is concerned with some counterexamples.

An interesting feature of the weighted random walks considered in Theorem 7.2 is that not only are they examples of reversible Markov chain, but they are in a sense the *only* examples of Markov chains.

Indeed, suppose that we have an irreducible Markov chain with finite state space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and transition matrix P, and suppose furthermore that the chain is reversible, so that its unique stationary distribution π satisfies the local equilibrium condition (35). Now let G be the weighted graph (possibly with self-edges) with vertex set $\{s_1, \ldots, s_k\}$, edges connecting precisely those pairs of vertices (states) between which jumps have positive probability in the Markov chain, and weights $w_e = w_{\langle s_i, s_j \rangle}$ given by the two (equal) sides of the local equilibrium relation (35). If $(X_0, X_1, \ldots)$ is a weighted random walk on this weighted graph, then we have

$$
\begin{aligned}
\mathbf{P}(X_{n+1} = s_j \mid X_n = s_i) &= \frac{w_{\langle s_i, s_j \rangle}}{w_{x_i}} \\
&= \frac{\pi_i P_{i,j}}{\sum \pi_i P_{i,i'}} \\
&= \frac{\pi_i P_{i,j}}{\pi_i} \\
&= P_{i,j}
\end{aligned}
$$

where the sum in the denominator of the second line is over all neighbors i' of i. Hence, the weighted random walk has transition matrix P, and we might therefore think of the original Markov chain as a weighted random walk with graph and weights as described here.

*

So far, we have only talked about, and defined, reversible Markov chains with *finite* state space. What would be a natural definition of reversibility of a *countable* state Markov chain?

To simply copy Definition 7.2 (a Markov chain is reversible if all its stationary distributions are reversible) to the countable state space setting turns out not to be a good idea. The reason is that many countable state Markov chains fail to have *any* stationary distribution (we saw an instance of that in Example 2.2), and for such Markov chains we trivially have that all their stationary distributions are reversible. This makes it much too easy for a Markov chain to qualify as reversible. It turns out to be a much better idea in this context to exploit the connection between reversibility and weighted random walks (recall that the definition of weighted random walk does not require the graph to be finite).

Definition 7.3. *Consider a Markov chain $(X_0, X_1, \ldots)$ with countable state space $\mathcal{S}$ and transition matrix P. The chain is said to be* **reversible** *if it can be represented as a weighted random walk, i.e., if there exists a weighted random walk with the same state space $\mathcal{S}$ and the same transition matrix P.*

For examples, see Problem 7.2.

8 Thursday, Oct 7

We begin today's lecture with a theorem on reversible Markov chains which roughly speaking says the following. Suppose that a Markov chain is started in equilibrium, and that we have a camera that records the states of the chain at times $0, 1, \ldots, n$. Suppose furthermore that we show the movie of the chain's evolution, but that we have mixed up which is the beginning and which is the end, i.e., we do not know whether the movie is shown backwards or forwards. The theorem says that if the chain is reversible, then watching the movie gives no clue as to whether it is shown backwards or forwards. This gives a very nice intuitive support to the notion of reversibility.

Theorem 8.1. *Suppose that $(X_0, X_1, \ldots)$ is a reversible Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_k\}$, started with X_0 chosen according to a stationary distribution π. For any n, and any sequence $(i_0, i_1, \ldots, i_n)$ of numbers taking values in $\{1, \ldots, k\}$, we then have*

$$\begin{aligned} &\mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_1, X_n = i_n) \\ &\quad = \mathbf{P}(X_0 = i_n, X_1 = i_{n-1}, \ldots, X_{n-1} = i_1, X_n = i_0)\,. \end{aligned} \tag{37}$$

Proof: The left-hand side of (37) equals

$$\begin{aligned} &\mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n) \\ &= \pi_{i_0} P_{i_0,i_1} P_{i_1,i_2} \cdots P_{i_{n-2},i_{n-1}} P_{i_{n-1},i_n} \,. \end{aligned} \tag{38}$$

If we similarly rewrite the right-hand side of (37), and then exploit the local equlibrium condition $\pi_i P_{i,j} = \pi_j P_{j,i}$ to rewrite each of the $P_{i,j}$-factors (this we can do since the chain is reversible, and π is stationary and therefore reversible), we get

$$\begin{aligned} &\mathbf{P}(X_0 = i_n, X_1 = i_{n-1}, \ldots, X_{n-1} = i_1, X_n = i_0) \\ &= \pi_{i_n} P_{i_n,i_{n-1}} P_{i_{n-1},i_{n-2}} \cdots P_{i_2,i_1} P_{i_1,i_0} \\ &= \pi_{i_n} \cdot \frac{\pi_{i_{n-1}} P_{i_{n-1},i_n}}{\pi_{i_n}} \cdot \frac{\pi_{i_{n-2}} P_{i_{n-2},i_{n-1}}}{\pi_{i_{n-1}}} \cdots \frac{\pi_{i_1} P_{i_1,i_2}}{\pi_{i_2}} \cdot \frac{\pi_{i_0} P_{i_0,i_1}}{\pi_{i_1}} \\ &= P_{i_{n-1},i_n} P_{i_{n-2},i_{n-1}} \cdots P_{i_1,i_2} P_{i_0,i_1} \pi_{i_0} \,. \end{aligned}$$

But the last expression is just a rearrangement of the factors in (38), so (37) is established and the proof is complete. $\diamond$

It is worth remarking that even though we formulated Theorem 8.1 for the case of finite-state reversible Markov chains only, it is clear from the proof that it goes through for the case of a countable-state reversible Markov chain as well (provided it has a stationary distribution).

*

Next, let us discuss how to find the stationary distribution π of (say) a finite-state irreducible Markov chain in practice. In general one has to resort to some not-so-transparent linear algebra, but for the case of reversible Markov chains, a much more intuitively pleasing procedure is available.

Thus, suppose that we have an irreducible and reversible Markov chain $(X_0, X_1, \ldots)$ with finite or countable state space $\mathcal{S}$ and a given transition matrix P. Let $s_1 \in \mathcal{S}$ be some arbitrary state, and assign to s_1 some arbitrary number $a > 0$, which we take as a preliminary candidate for $\pi 1$. We write $\tilde{\pi} = (\tilde{\pi}_1, \tilde{\pi}_2, \ldots)$ for our preliminary choice of probability distribution; thus, $\tilde{\pi}_1 = a$. Then pick another state s_2 with the property that jumps between s_1 and s_2 are possible. The local equilibrium condition (35) applied to s_1 and s_2 forces us to choose $\tilde{\pi}_2$ in such a way that $\tilde{\pi}_1 P_{1,2} = \tilde{\pi}_2 P_{2,1}$, so that

$$\begin{aligned} \tilde{\pi}_2 &= \frac{\tilde{\pi}_1 P_{1,2}}{P_{2,1}} \\ &= \frac{a P_{1,2}}{P_{2,1}} \,. \end{aligned} \tag{39}$$

We may then inductively proceed as follows. Suppose that $\tilde{\pi}_1, \tilde{\pi}_2, \ldots, \tilde{\pi}_j$ have been decided. Then pick a state $s_{j+1} \in \mathcal{S}$ such that jumps between s_{j+1} and s_i are possible for some $i \in \{1, 2, \ldots, j\}$. Using the local equlibrium condition again, with states s_i and s_{j+1}, allows us to calculate $\tilde{\pi}_{j+1}$, analogously to (39). And so on.

This continues indefinitely or until (for the case where $\mathcal{S}$ is finite) there are no more states. Note that everything except the choice of a was forced, so *if* we were lucky and chose a to be precisely the correct value for π_1, then we have now found the full stationary distribution π. On the other hand, this argument shows that a stationary distribution π (if it exists) must be of the form

$$\begin{aligned} \pi &= b\tilde{\pi} \\ &= (b\tilde{\pi}_1, b\tilde{\pi}_2, \ldots) \end{aligned}$$

for some $b > 0$. Furthermore, the probabilities $\pi_1, \pi_2, \ldots$ must sum to 1, so the only possible choice of b is $1/\sum_i \tilde{\pi}_i$, so that

$$\begin{aligned} \pi &= (\pi_1, \pi_2, \ldots) \\ &= \left(\frac{\tilde{\pi}_1}{\sum_i \tilde{\pi}_i}, \frac{\tilde{\pi}_2}{\sum_i \tilde{\pi}_i}, \ldots \right) . \end{aligned} \tag{40}$$

When the state space $\mathcal{S}$ is finite, this is a fool-proof way of finding the stationary distribution π. When $\mathcal{S}$ is countably infinite, there are two distinct cases to consider, namely when $\sum_i \tilde{\pi}_i$ is finite, and when the sum is infinite. For the case $\sum_i \tilde{\pi}_i < \infty$, (40) yields a stationary distribution just as in the finite case. However, when $\sum_i \tilde{\pi}_i = \infty$ this obviously does not work, and we are forced to conclude that no stationary distribution exists.

*

The above recipe is nice, but a possible objection is that we might not know beforehand whether the Markov chain is reversible. Our next result gives a nice sufficient (but far from necessary) condition. Given a Markov chain with finite or countable state space $\mathcal{S} = \{s_1, s_2, \ldots\}$ and a given transition matrix P, define the **transition graph** $G = (V, E)$ corresponding to this Markov chain, to have vertex set $V = \mathcal{S}$, and edge sets consiting of those pairs of vertices s_i and s_j for which $P_{i,j} + P_{j,i} > 0$. (That is, two vertices are connected by an edge if jumps between the corresponding states are possible, one way or the other.)

A **path** in a graph is a sequence of edges of the form

$$(\langle x_1, x_2\rangle, \langle x_2, x_3\rangle, \ldots, \langle x_{n-2}, x_{n-1}\rangle, \langle x_{n-1}, x_n\rangle)\,.$$

A **circuit** is a path whose edges are all distinct and for which $x_n = x_1$, i.e., it ends where it begins. So a circuit can be seen as a way to walk from a vertex along the edges of the graph, back to the starting point, without ever walking more than once along an edge - a kind of sightseeing tour. A graph without circuits (other than possibly some self-edges) is called a **forest**, and the special case of a connected forest is called a **tree**.

Theorem 8.2. *Consider a Markov chain with finite or countable state space $\mathcal{S} = \{s_1, s_2, \ldots\}$ and transition matrix P. If the chain has the properties that*

(a) for each pair of states s_i and s_j, we have that $P_{i,j}$ and $P_{j,i}$ are both zero or both strictly positive, and

(b) the chain's transition graph is a forest,

then it is reversible.

Proof: It suffices to show that the Markov chain is a weighted random walk. As the graph for the weighted random walk to walk on, we take the chain's transition graph. Take an arbitrary edge of the graph, say $\langle s_1, s_2\rangle$ for concreteness, and assign it an arbitrary strictly positive weight $w_{\langle s_1, s_2\rangle}$. Then we can assign weights to all the other edges incident to s_1 in such a way that the jump probabilities out of vertex s_1 become the ones prescribed by row 1 of the transition matrix P. We then do the same thing with the edges incident to vertex s_2. And we keep doing the same thing with one vertex at a time, always making sure that the vertex we choose is one for which at least one of its edges has had its weight determined.

If the transition graph is connected, then eventually every vertex will be taken care of in this way. Otherwise, when the connected component containing $\langle s_1, s_2\rangle$ has been handled, we begin afresh with an arbitrary edge in a different connected component, and so on.

This defines the edge weights. The procedure obviously works fine if, whenever we decide to handle a vertex s_i, only one of its incident edges has had its weight determined. However, if, when we decide to handle a vertex s_i, two or more of its edges have already had their weights determined, then these weights may have proportions that make it impossible to make the jump probabilities from s_i agree with row i of the transition matrix P. The point now is that if the transition graph is a forest (i.e., if it has no circuits), then the described procedure for choosing the vertices will never result in a situation where more than one of the edges of the chosen vertex s_i has already been determined (right?). It is therefore clear that if the transition graph is free of circuits (other than possibly some self-edges), then this procedure yields a weighted random walk such that for every vertex s_i, the jump probabilities out of s_i agree with row i of the transition matrix P. Hence, the chain is a weighted random walk, and therefore reversible. $\diamond$

Example 8.1. Let the discrete time Markov chain $(X_0, X_1, \ldots)$ have state space $\mathcal{S} = \{0, 1, 2, \ldots\}$ and the following transition mechanism. Fix $p \in (0, 1)$. If $X_n = k > 0$, then it takes a step down or a step up with respective probabilities p and $1 - p$, i.e.,

$$X_{n+1} = \begin{cases} k-1 & \text{with probability } p \\ k+1 & \text{with probability } 1-p \,. \end{cases}$$

If $X_n = 0$, we instead take X_{n+1} to be 0 with probability p and 1 with probability $1 - p$. Note that taking $p = \frac{2}{3}$ yields Example 3.1.

The transition graph of this Markov chain consists just of the single infinite path $\langle 0,1\rangle, \langle 1,2\rangle, \langle 2,3\rangle, \ldots$ plus a self-edge at 0, so the transition graph is a forest and the chain is, due to Theorem 8.2, reversible.

Let us now see if we can find a stationary distribution π using the procedure outlined above.[26] Set, somewhat arbitrarily, $\tilde{\pi}_0 = 1$. We can then apply the local equilibrium condition to states 0 and 1, for which it says that $\tilde{\pi}_0 P_{0,1} = \tilde{\pi}_1 P_{1,0}$, so that

$$\begin{aligned} \tilde{\pi}_1 &= \frac{\tilde{\pi}_0 P_{0,1}}{P_{1,0}} \\ &= \frac{1-p}{p} \,. \end{aligned}$$

We can then calculate $\tilde{\pi}_2$ from $\tilde{\pi}_1$ in the same way, and then $\tilde{\pi}_3$ from $\tilde{\pi}_2$, and so on, getting

$$\begin{array}{lllll} \tilde{\pi}_2 & = & \frac{\tilde{\pi}_1 P_{1,2}}{P_{2,1}} = \tilde{\pi}_1 \cdot \frac{1-p}{p} & = & \left(\frac{1-p}{p}\right)^2 \\ \tilde{\pi}_3 & = & \frac{\tilde{\pi}_2 P_{2,3}}{P_{3,2}} = \tilde{\pi}_2 \cdot \frac{1-p}{p} & = & \left(\frac{1-p}{p}\right)^3 \\ \vdots & & & & \\ \tilde{\pi}_i & = & & & \left(\frac{1-p}{p}\right)^i \,. \end{array}$$

[26] This gives a much more satisfactory solution than that in Example 3.1, where the stationary distribution π was simply handed to us. Here we are able to find π through systematic thinking.

But clearly $\sum_{i=0}^{\infty} \tilde{\pi}_i > \tilde{\pi}_1 = 1$, so we need to normalize $\tilde{\pi}$ in order to get a probability measure. Note that

$$\sum_{i=0}^{\infty} \tilde{\pi}_i = \sum_{i=0}^{\infty} \left(\frac{1-p}{p}\right)^i$$

is a geometric series, summing to ∞ if $\frac{1-p}{p} \geq 1$ (i.e., if $p \leq \frac{1}{2}$), and summing to

$$\frac{1}{1 - \frac{1-p}{p}} = \frac{p}{2p-1}$$

otherwise. Hence, we conclude that if $p \leq \frac{1}{2}$, then no stationary distribution exists, while of $p > \frac{1}{2}$, then we have a stationary distribution π given by

$$\begin{aligned}
\pi &= (\pi_0, \pi_1, \ldots) \\
&= \left(\frac{2p-1}{p}, \frac{2p-1}{p}\left(\frac{1-p}{p}\right), \frac{2p-1}{p}\left(\frac{1-p}{p}\right)^2, \frac{2p-1}{p}\left(\frac{1-p}{p}\right)^3, \ldots\right).
\end{aligned}$$

The fact that a stationary distribution exists if and only if $p > \frac{1}{2}$ can be phrased as follows: Random walk on the positive integers has a stationary distribution if and only if it has a strictly positive drift back towards state 0. $\diamond$

9 Monday, Oct 11

Let us begin this lecture with a brief mentioning of an application of reversible Markov chains which today is extremely popular: Markov chain Monte Carlo (MCMC).[27] Often in areas such as image analysis, statistical mechanics, and Bayesian statistics, one is faced with the problem of simulating a random object with a given distribution π which is typically high-dimensional and very difficult to sample from directly. The idea of MCMC is the following. Instead of sampling from π directly, design a Markov chain which has π as its unique stationary distribution. If the Markov chain has a convergence property along the lines of Theorem 5.1, then we can run the chain from some arbitrary choice of X_0, and if the running time n is large enough, then the distribution of X_n will be close to π. In theory, the chain may or may not be reversible, but the choices of Markov chains that have become more or less standard in this field, are all reversible.

[27] For more on MCMC, see, e.g, Häggström (2002) or Gilks et al. (1996).

The whole idea may, at first sight, seem like a rather far-fetched detour: how in the world could it be easier to come up with, and run, such a Markov chain compared to sampling from π directly? But in fact the approach has turned out to be highly useful. Let us illustrate the potential for designing suitable Markov chains by describing one of the standard choices of MCMC chains: the so-called Metropolis–Hastings algorithm. We describe this in the case when π is a distribution on a finite set $\mathcal{S} = \{s_1, \ldots, s_k\}$ (there are analogues for more general state spaces that we do not describe here). We assume without loss of generality that π attaches positive probability to *every* $s_i \in \mathcal{S}$; otherwise, we can simply consider π to be a probability distribution on a smaller state space.

This first thing we do is to define a graph $G = (V, E)$, with vertex set $V = \mathcal{S}$. The edge set can be pretty much arbitrary. For the algorithm to work (in the sense that the distribution of X_n converges to π as $n \to \infty$), all we need is that the graph is connected.[28] However, for the purpose of making the algorithm work *efficiently* (in the sense that the distribution of X_n converges to π sufficiently *quickly*), the choice of edge set may matter a lot.

Given $G = (\mathcal{S}, E)$, the $\mathcal{S}$-valued Markov chain $(X_0, X_1, \ldots)$ is taken to have the following update mechanism. Recall that $\deg(s)$ denotes the number of neighbors in the graph of a vertex s. If $X_n = s_i$, then we first pick a vertex s_j among the $\deg(s_i)$ neighbors of s_i, each with probability $1/\deg(s_i)$. Conditional on the choice of s_j we set $X_{n+1} = s_j$ with probability $\min\{\frac{\pi_j \deg(s_i)}{\pi_i \deg(s_j)}, 1\}$, and $X_{n+1} = s_i$ with the complementary probability $1 - \min\{\frac{\pi_j \deg(s_i)}{\pi_i \deg(s_j)}, 1\}$.

It can now be checked (see Problem 9.1) that π satisfies local equlibrium for this chain. Thus, π is reversible and therefore stationary. Furthermore, if the edge set E is chosen in such a way that the graph is connected, then the chain is irreducible, so that π is the *only* stationary distribution. Hence, if we are prepared to run the chain for a long enough time, we have a Markov chain that serves our purposes for MCMC simulation. The topic of how long *long enough* is, and in particular whether this time span is short enough to admit practical usefulness of the algorithm, however, is a more subtle issue, to which a large literature has been devoted.[29]

[28] This is a slight lie, because we are ignoring the subtle issue that the chain that we are about to define may fail to be aperiodic. To ensure that it is aperiodic, it suffices to add a single self-edge to one of the vertices.

[29] See, e.g, Häggström (2002), Gilks et al. (1996), and the references therein.

*

And now for something completely different: hitting probabilities. Consider random walk on a finite connected graph $G = (V, E)$, and fix a vertex y_0. We know (e.g., from Lemma 4.1) that no matter where we start the random walk, the random walker will, a.s., sooner or later reach y_0. A drunk person walking around in his (finite) home town can be confident that he will find his way home.

The game becomes more interesting if we fix *two* vertices y_0 and y_1, and ask for the probability that the random walker reaches y_0 before he reaches y_1. What is the probability that the drunkard walking around in his home town finds his way home before encountering the pub? See Figure 9.1.

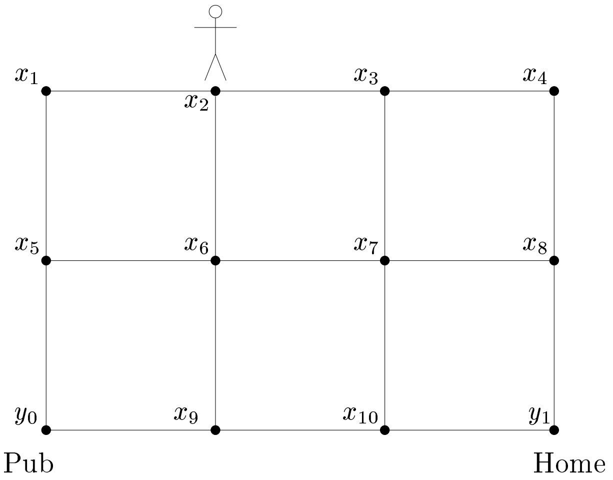


Figure 9.1. *A graph $G = (V, E)$ with $V = \{x_1, x_2, \ldots, x_{10}, y_0, y_1\}$. A random walker stands at x_2 and is about to begin his journey. What is the probability that he reaches y_0 before y_1?*

Consider a random walk $(X_0, X_1, \ldots)$ on the graph $G = (V, E)$ in Figure 9.1. For an arbitrary vertex $x \in V$, let $p(x)$ denote the probability that y_1 is encountered before y_0 provided the walk starts at $X_0 = x$. Here we use the convention that if X_0 equals y_0 or y_1, then the race is already over, so that $p(y_0) = 0$ and $p(y_1) = 1$. On the other hand, it is easy to see that for $i = 1, 2, \ldots, 10$, we have

$$0 < p(x_i) < 1 .$$

But how do we calculate $p(x_i)$ more precisely? In particular, in Figure 9.1, we are asked to calculate $p(x_2)$. Letting A denote the event that y_1 is encountered before y_0, we thus want to calculate $\mathbf{P}(A|X_0 = x_2)$. A first step is to condition on what happens at time 1. Since the random walker goes at time 1 to x_1, x_3 or x_6 with probability $\frac{1}{3}$ each, we get

$$\begin{aligned} p(x_2) &= \mathbf{P}(A \mid X_0 = x_2) \\ &= \tfrac{1}{3}\mathbf{P}(A \mid X_0 = x_2, X_1 = x_1) + \tfrac{1}{3}\mathbf{P}(A \mid X_0 = x_2, X_1 = x_3) \\ &\quad + \tfrac{1}{3}\mathbf{P}(A \mid X_0 = x_2, X_1 = x_6)\,. \end{aligned} \tag{41}$$

But of course

$$\begin{aligned} \mathbf{P}(A \mid X_0 = x_2, X_1 = x_1) &= \mathbf{P}(A \mid X_0 = x_1) \\ &= p(x_1) \end{aligned}$$

and similarly with x_1 replaced by x_3 and x_6. Hence, (41) becomes

$$p(x_2) = \tfrac{1}{3}(p(x_1) + p(x_3) + p(x_6))\,. \tag{42}$$

In other words, $p(x_2)$ is the average of $p(\cdot)$ taken over all the neighors of x_2 in the graph – a statement that we abbreviate $\Delta p(x_2) = 0$.[30]

This reduces the problem of calculating $p(x_2)$ to that of calculating $p(x_1)$, $p(x_3)$ and $p(x_6)$ – which may seem as not much of a simplification. If we try the same idea for calculating $p(x_1)$, we see that a random walk starting in x_1 goes at time 1 to either x_2 or x_5, each with probability $\frac{1}{2}$, so that

$$p(x_1) = \tfrac{1}{2}(p(x_2) + p(x_5))\,.$$

Similarly as in (42), we thus have that $p(x_1)$ is the average of $p(\cdot)$ among its neighbors, or in other words $\Delta p(x_1) = 0$.

Applying the same idea to $x_3, x_4, \ldots, x_{10}$, we get

$$\begin{aligned} \Delta p(x_1) &= 0 \\ \Delta p(x_2) &= 0 \\ &\vdots \\ \Delta p(x_{10}) &= 0 \\ p(y_0) &= 0 \\ p(y_1) &= 1\,. \end{aligned} \tag{43}$$

[30]We use this notation because Δ is a discrete variant of the Laplace operator, which is denoted in the same way.

This is a systems of equations with 12 unknowns and 12 equations, which sounds like good news.[31] It nevertheless needs to be checked whether we have existence and uniqueness of solution(s) to this system.

Existence is no problem at all, because we have shown that the hitting probabilities $p(x_1), p(x_2), \ldots, p(x_{10}), p(y_0), p(y_1)$ do satisfy these equations.[32] To handle the issue of uniqueness, let

$$\mathbf{p}' = (p'(x_1), p'(x_2), \ldots, p'(x_{10}), p'(y_0), p'(y_1))$$

and

$$\mathbf{p}'' = (p''(x_1), p''(x_2), \ldots, p''(x_{10}), p''(y_0), p''(y_1))$$

denote two (a priori possible different) solutions to system (43). Write

$$\mathbf{d} = (d(x_1), d(x_2), \ldots, d(x_{10}), d(y_0), d(y_1))$$

for the difference between these solutions, i.e., $d(x_1) = p'(x_1) - p''(x_1)$ and so on; if we can show that $\mathbf{d} = (0, 0, \ldots 0)$, then $\mathbf{p}' = \mathbf{p}''$, and since p' and $\mathbf{p}''$ were arbitrary solutions, the desired uniqueness will follow.

To analyze $\mathbf{d}$, we can begin by noting that

$$\begin{aligned} d(y_1) &= p'(y_1) - p''(y_1) \\ &= 1 - 1 \\ &= 0\,, \end{aligned}$$

and similarly

$$d(y_0) = 0\,.$$

Next, we consider the first equation in (43). Inserting solutions $\mathbf{p}'$ and $\mathbf{p}''$ gives

$$p'(x_1) = \tfrac{1}{2}(p'(x_2) + p'(x_5)) \tag{44}$$

and

$$p''(x_1) = \tfrac{1}{2}(p''(x_2) + p''(x_5))\,. \tag{45}$$

Subtracting (45) from (44) gives

$$\begin{aligned} d(x_1) &= p'(x_1) - p''(x_1) \\ &= (\tfrac{1}{2}(p'(x_2) + p'(x_5))) - (\tfrac{1}{2}(p''(x_2) + p''(x_5))) \\ &= \tfrac{1}{2}((p'(x_2) - p''(x_2)) + (p'(x_5) - p''(x_5))) \\ &= \tfrac{1}{2}(d(x_2) + d(x_5))\,, \end{aligned}$$

[31] Alternatively, since the last two equations are so trivial, we should perhaps view it as a system with 10 unknowns and 10 equations, which is of course just as good news.

[32] An extremely prudent reader might object that we need to show that these hitting probabilities are well-defined. But of course they are, and probabilists stopped spending time over such moronic measurability issues many decades ago.

or in other words $\Delta d(x_1) = 0$. Similarly, we get $\Delta d(x_i) = 0$ for $i = 2, 3, \ldots, 10$. To summarize, we have that

$$\begin{array}{rcl} \Delta d(x_1) & = & 0 \\ \Delta d(x_2) & = & 0 \\ & \vdots & \\ \Delta d(x_{10}) & = & 0 \\ d(y_0) & = & 0 \\ d(y_1) & = & 0\,. \end{array} \tag{46}$$

This system may at first sight look very similar to the one in (43), but the difference in the last equation ($p(y_1) = 1$ versus $d(y_1) = 0$) is crucial.

Lemma 9.1. *The only solution to system (46) is* $\mathbf{d} = (0, 0, \ldots, 0)$.

Proof: We are done if we can show that

$$d(x_i) \leq 0 \quad \text{for } i = 1, 2, \ldots, 10 \tag{47}$$

and that

$$d(x_i) \geq 0 \quad \text{for } i = 1, 2, \ldots, 10\,. \tag{48}$$

Let us show (47) only; (48) follows by an identical argument.

To this end, define

$$M \;=\; \max\{d(x_1), d(x_2), \ldots, d(x_{10}), d(y_0), d(y_1)\}\,.$$

The desired (47) follows if we can show that $M = 0$.

The maximum is achieved at some vertex x. If $x = y_0$ or y_1, then we are done, so suppose that $x = x_i$ for some i. Thus,

$$d(x_i) = M\,. \tag{49}$$

By definition of M, we have $d(x_j) \leq M$ for every neighbor x_j of x_i. But then we must have $d(x_j)$ *equal* to M for each such neighbor, because if some of them are *strictly* less than M, then their average would be strictly less than M, contradicting (49).

Thus, every neighbor of x_i must also take value M. If one of these neighbors is y_0 or y_1, then we know that $M = 0$, and we are done. Otherwise, we just apply the same argument to the neighbors of x_i to show that *their* neighbors are also equal to M. And so on. Sooner or later[33] we will encounter

[33] Pretty soon, in fact.

y_0 or y_1 and conclude that $M = 0$. Thus, (47), and since (48) follows by the same reasoning (just replace maxima by minima), the proof is complete. $\diamond$

From Lemma 9.1, we get that the two solutions $\mathbf{p}'$ and p'' to system (43) are identical, and since $\mathbf{p}'$ and p'' we can conclude that (43) has a unique solution. Thus, solving (43) gives a way of calculating the hitting probabilities $p(x_1), \ldots, p(x_{10})$. Feeding the equation system into *Matlab*, or *Mathematica*, or whatever, gives

$$\begin{aligned} \mathbf{p} &= (p(x_1), \ldots, p(x_{10}), p(y_0), p(y_1)) \\ &= (\tfrac{92}{267}, \tfrac{117}{267}, \tfrac{150}{267}, \tfrac{175}{267}, \tfrac{67}{267}, \tfrac{109}{267}, \tfrac{158}{267}, \tfrac{200}{267}, \tfrac{94}{267}, \tfrac{173}{267}, 0, 1)\,. \end{aligned} \tag{50}$$

Finally, it is worth reflecting over what would happen if we tried to apply the ideas in the proof of Lemma 9.1 directly to system (43) rather than to (46). What the arguments show in the more general setting is that any $\mathbf{p}$ satisfying the first 10 lines in (43) attains its maximum (and its minimum) at y_0 or y_1. Thus, $p(x_i)$ is between 0 and 1 for every $i \in \{1, \ldots, 10\}$. This is good (or perhaps, highly expected) news, since the $p(x_i)$'s are meant to represent probabilities.

10 Thursday, October 14

We will now consider something that may at first sight appear to be completely irrelevant for our study of Markov chains, but which will turn out to be intimately related to reversible Markov chains, namely electrical networks.

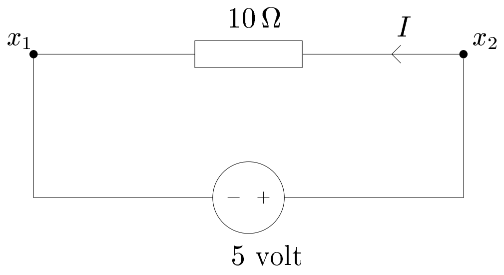


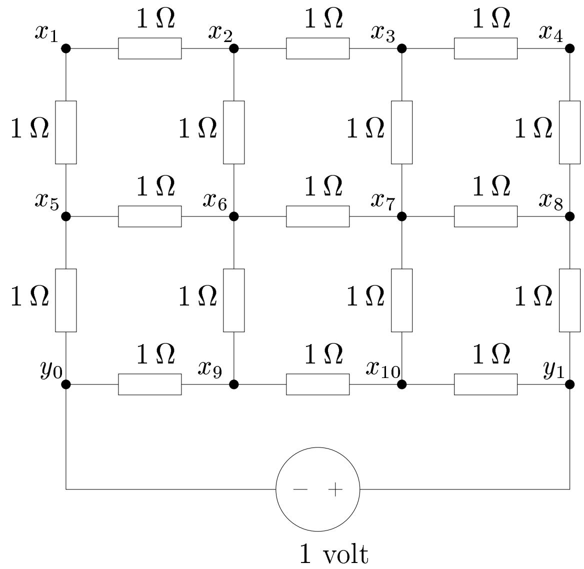
Figure 10.1. *An electrical network consisting of a 5 volt battery connected to a 10 ohm resistor. Using Ohm's law, the current I through the resistor can be calculated as $I = \frac{U}{R} = \frac{5}{10} = 0.5$ ampere.*

In Figure 10.1, we have one of the simplest possible electrical networks: a 5 volt battery hooked up to a pair of nodes, to which a 10 ohm resistor has also been hooked. We call x_2 the plus pole of the circuit (because this is the node to which the plus pole of the battery has been hooked), and x_1 the minus pole. Through the resistor, there will run a current I from the plus pole to the minus pole, and this current can (as we all learned in high school, but may or may not remember) be calculated using **Ohm's law**, which says that the current through a resistor equals the voltage V divided by the resistance R:

$$I = \frac{U}{R}. \tag{51}$$

Thus, in this case, the current I will be $\frac{5}{10} = 0.5$ ampere.

Ohm's law (51) is one of the two fundamental laws of electricity that we will be concerned with. The other is **Kirchhoff's current law**, which says that the total amount of current flowing into a node must be 0: what comes in, must also come out. Applied to the circuit in Figure 10.1, this means that the same current I is flowing everywhere in the circuit: 5 ampere in counterclockwise direction.



Figue 10.2. *An electrical network whose resistors all have resistance* 1 *ohm. (Note that the graph structure is the same as in Figure 9.1; this is not a coincidence.)*

A voltage across a resistor – say, the one along the edge $\langle x_1, x_2\rangle$ – can be described as the difference between the voltage **potentials** $V(x_1)$ and $V(x_2)$ of the two vertices. The potential at the nodes are always relative to some pre-specified node (which itself gets potential 0). As far as the voltages along the resistors, the choice of which nodes to measure against does not matter.[34] Here, we will only be considering circuits with just a single voltage source (a single battery), and we will use the convention that voltages are measured relative to the minus pole of the battery. Thus, in Figure 10.1, node x_1 has potential $V(x_1) = 0$ while node x_2 has potential $V(x_2) = 5$. And in the circuit in Figure 10.2, we have $V(y_0) = 0$ and $V(y_1) = 1$, while the remaining voltages $V(x_1), \ldots V(x_{10})$ remain to be found.

For two nodes x and y linked by a resistor, write $R_{x,y}$ for the resistance, and $I_{x,y}$ for the current flowing from x to y through the resistor. Let's say that we want to calculate the voltage V_{x_2} in the circuit in Figure 10.2. What can we do? We have Ohm's and Kirchhoff's laws at our disposal, so let us begin by applying Kirchhoff's law to the vertex x_2: the total amount of current flowing into x_2 is 0, so that

$$I_{x_1,x_2} + I_{x_3,x_2} + I_{x_6,x_2} = 0\,. \tag{52}$$

Each of the three currents can be related to the voltages across their respective resistors via Ohm's law:

$$I_{x_i,x_2} = \frac{V(x_i) - V(x_2)}{R_{i,2}}$$

for $i = 1, 3, 6$. Noting that $R_{i,2} = 1$ in each case, and inserting into (52), yields

$$(V(x_1) - V(x_2)) + (V(x_3) - V(x_2)) + (V(x_6) - V(x_2)) = 0$$

and solving for $V(x_2)$ gives

$$V(x_2) = \tfrac{1}{3}(V(x_1) + V(x_3) + V(x_6))\,. \tag{53}$$

Using the shorthand notation introduced in Section 9, we may write this as

$$\Delta V(x_2) = 0\,.$$

Proceeding analogously with the node x_2 replaced by x_i for $i = 1, \ldots, 10$ yields $\Delta V(x_i) = 0$ for each i. Hence, we arrive at the following equation

[34] This is vaguely analogous to the fact that a 10 degree drop in temperature means the same thing regardless of whether we use the Celsius scale or the Kelvin scale. Or to the fact that the primitive function $F(x)$ of a Riemann integrable function $f(x)$ is well-defined only up to an additive constant.

system for the $V(x_i)$'s:

$$\begin{array}{rcl} \Delta V(x_1) & = & 0 \\ \Delta V(x_2) & = & 0 \\ & \vdots & \\ \Delta V(x_{10}) & = & 0 \\ V(y_0) & = & 0 \\ V(y_1) & = & 1\,. \end{array} \tag{54}$$

Does it ring a bell? Yes, this is precisely the same system of equations as that for the hitting probabilities $(p(x_1), \ldots, p(x_{10}), p(y_0), p(y_1))$ in (43). And that system of equations was shown in Section 9 to have a unique solution, and we can therefore conclude that

$$V(x_i) \;=\; p(x_i)$$

for $i = 1, \ldots, 10$. The numerical values for $(p(x_1), \ldots, p(x_{10}))$ were given in (50), and we may conclude that $(V(x_1), \ldots, V(x_{10}))$ take the same values. The currents can then be read off from the voltages using Ohm's law.

*

Of course, the connection between random walks and electric networks goes further than just to the particular graph considered in Figures 9.1 and 10.2. The arguments we have gone through for this particular graph go through in a straightforward manner to prove the following result.

Theorem 10.1. *Let $G = (V, E)$ be a finite connected graph with vertex set $V = \{x_1, x_2, \ldots, x_k, y_0, y_1\}$, and consider random walk on G. For $i = 1, 2, \ldots, k$, let $p(x_i)$ denote the probability that the walk encounters the vertex y_1 before its first arrival at y_0, provided that it starts in x_i.*

Also, consider the electric network with unit resistors placed along the edges of G, and a unit voltage source with the minus pole at y_0 and the plus pole at y_1. For $i = 1, 2, \ldots, k$, let $V(x_i)$ denote the resulting voltage (relative to the minus pole) in node x_i.

Then, for $i = 1, 2, \ldots, k$, we have

$$p(x_i) \;=\; V(x_i)\,.$$

There are several things to say in praise of this result. First of all, it is a beautiful illustration of the power and generality of mathematics: two seemingly unrelated objects – random walks versus electrical networks – turn out to be governed by the very same mathematical equations. Furthermore, the result opens up interesting avenues for applying probabilistic techniques in the study of electrical networks, and, conversely, for using ideas from electrical network theory in the study of random walks. Doyle and Snell (1984) present a variety of such cross-fertilization ideas[35], and the in remaining lectures of this course we will see some of them.

Is Theorem 10.1 as far as the analogy goes, or can it be generalized even further? For instance, what happens if we replace the unit voltage source by one with a larger or smaller voltage? Or what happens if we replace all the unit resistors with ones with resistance R (the same R for all of them)? It turns out (see Problem 10.1) in both cases that this only leads to a trivial rescaling of the voltages and currents.[36]

But suppose that we replace the unit resistors in the network of Figure 10.2 by *different* resistors, i.e., each edge $\langle x_i, x_j \rangle$ is allowed to have its own resistance $R_{i,j}$. Then no such trivial rescaling of the entire systems exists. Let us see what happens if we try to copy the above analysis, using Ohm's and Kirchhoff's laws, to this modified network. Kirchhoff's current law applied to the node x_2 gives

$$I_{x_1,x_2} + I_{x_3,x_2} + I_{x_6,x_2} = 0 \,. \tag{55}$$

For $i = 1, 3, 6$ we get from Ohm's law that

$$I_{x_i,x_2} = \frac{V(x_i) - V(x_2)}{R_{i,2}}$$

and inserting into (55) yields

$$\frac{V(x_1) - V(x_2)}{R_{1,2}} + \frac{V(x_3) - V(x_2)}{R_{3,2}} + \frac{V(x_6) - V(x_2)}{R_{6,2}} = 0 \,.$$

Solving for $V(x_2)$ gives

$$V(x_2) = \frac{\frac{1}{R_{1,2}} V(x_1) + \frac{1}{R_{3,2}} V(x_3) + \frac{1}{R_{6,2}} V(x_6)}{\frac{1}{R_{1,2}} + \frac{1}{R_{3,2}} + \frac{1}{R_{6,2}}} \,. \tag{56}$$

What (56) says is not (as in the original network with unit resistors) that $V(x_2)$ is an *average* of $V(x_1)$, $V(x_3)$ and $V(x_6)$, but rather that it is a

[35] See also Häggström (2004) for a treatment in Swedish.

[36] Physically, we can think of both changes as being just a change in the choice of unit with which to measure voltages, resistances, and currents, while the actual physical system remains exactly the same. Mathematically, this sort of invariance can be seen as resulting from the fact that the equation system (54) is linear.

weighted avarege of those voltages, with respective weights $\frac{1}{R_{1,2}}$, $\frac{1}{R_{3,2}}$ and $\frac{1}{R_{6,2}}$.

At this point, it is natural to stop talking about resistances and switching to **conductances**. The conductance C of a resistor with resistance R is simply $C = \frac{1}{R}$.[37] For the graphs under consideration, write $C_{i,j}$ for the conductance of the resistor along the edge $\langle x_i, x_j \rangle$. The formula (56) then turns into the slightly more elegant

$$V(x_2) = \frac{C_{1,2}V(x_1) + C_{3,2}V(x_3) + C_{6,2}V(x_6)}{C_{1,2} + C_{3,2} + C_{6,2}} . \tag{57}$$

Thus $V(x_2)$ is a weighted average of $V(x_1)$, $V(x_3)$ and $V(x_6)$ with respective weights $C_{1,2}$, $C_{3,2}$ and $C_{6,2}$. We abbreviate this as

$$\Delta_C V(x_2) = 0$$

where we might call Δ_C the **weighted discrete Laplacian** with the prescribed set of weights $C = \{C_e\}_{e \in E}$ to the edges.

Applying Kirchoff's current law to x_i rather than to x_2, and then using Ohm's law as above, we get that $\Delta_C V(x_i) = 0$ for $i = 1, \ldots, 10$. Together with the relations $V(y_0) = 0$ and $V(y_1) = 1$ that are forced by the unit voltage source, we get the equation system

$$\begin{array}{rcl} \Delta_C V(x_1) & = & 0 \\ \Delta_C V(x_2) & = & 0 \\ & \vdots & \\ \Delta_C V(x_{10}) & = & 0 \\ V(y_0) & = & 0 \\ V(y_1) & = & 1 . \end{array} \tag{58}$$

By copying the same argument as for the case with averages rather than weighted averages, we get that this system of equations has a unique solution.

Now consider a weighted random walk on the graph G in Figure 9.1, with edge weights $\{C_e\}_{e \in E}$ equal to the conductances in the electrical network that we just had a look at. Let A denote the event that this weighted random walk visits y_1 before the first visit to y_0. As in the ordinary random walk, let $p(x_i)$ denote the probability of A given that we start from x_i, i.e., conditional on $X_0 = x_i$.

[37] Physically, conductances are at least as natural quantities as resistances. While the resistance is defined as the amount of voltage needed over a resistor in order to push a unit current through the resistor, the conductance is defined as the amount of current flowing through the resistor when it is attached to a unit voltage source.

If we begin by trying to get an alternative expression for $p(x_2)$, we get, analogously to (41),

$$\begin{aligned} p(x_2) &= \mathbf{P}(A \,|\, X_0 = x_2) \\ &= \tfrac{C_{1,2}}{C_{1,2}+C_{3,2}+C_{6,2}} \mathbf{P}(A \,|\, X_0 = x_2, X_1 = x_1) \\ &\quad + \tfrac{C_{3,2}}{C_{1,2}+C_{3,2}+C_{6,2}} \mathbf{P}(A \,|\, X_0 = x_2, X_1 = x_3) \\ &\quad + \tfrac{C_{6,2}}{C_{1,2}+C_{3,2}+C_{6,2}} \mathbf{P}(A \,|\, X_0 = x_2, X_1 = x_6) \\ &= \tfrac{C_{1,2}}{C_{1,2}+C_{3,2}+C_{6,2}} p(x_1) + \tfrac{C_{3,2}}{C_{1,2}+C_{3,2}+C_{6,2}} p(x_3) + \tfrac{C_{6,2}}{C_{1,2}+C_{3,2}+C_{6,2}} p(x_6)\,, \end{aligned}$$

or in other words

$$\Delta_C p(x_2) \;=\; 0\,.$$

By the same reasoning, we get $\Delta_C p(x_i) = 0$ for $i = 1, \ldots, 10$. Furthermore, it is immediate that $p(y_0) = 0$ and $p(y_1) = 1$. In summary, we have the equation system

$$\begin{array}{rcl} \Delta_C p(x_1) & = & 0 \\ \Delta_C p(x_2) & = & 0 \\ & \vdots & \\ \Delta_C p(x_{10}) & = & 0 \\ p(y_0) & = & 0 \\ p(y_1) & = & 1\,. \end{array} \tag{59}$$

We now see that the equation system (58) for $V(x_1), \ldots, V(y_1)$ and the system (59) for $p(x_1), \ldots, p(y_1)$ are identical. We have already noted that (58) has a unique solution, so it follows that $p(x_i) = V(x_i)$ for each i. This leads us to the following generalization of Theorem 10.1.

Theorem 10.2. *Let $G = (V, E)$ be a finite connected graph with vertex set $V = \{x_1, x_2, \ldots, x_k, y_0, y_1\}$, and consider weighted random walk on G with edge weights $\{C_e\}_{e \in E}$. For $i = 1, 2, \ldots, k$, let $p(x_i)$ denote the probability that this walk encounters the vertex y_1 before its first arrival at y_0, provided that it starts in x_i.*

Also, consider the electric network which has resistors with conductances $\{C_e\}_{e \in E}$ placed along the edges of G, and a unit voltage source with the minus pole at y_0 and the plus pole at y_1. For $i = 1, 2, \ldots, k$, let $V(x_i)$ denote the resulting voltage (relative to the minus pole) in node x_i.

Then, for $i = 1, 2, \ldots, k$, we have

$$p(x_i) \;=\; V(x_i)\,.$$

11 Monday, Oct 18

Consider random walks on the two graphs G and G^* in Figure 11.1, both of them starting in y_0. We are interested in the probability of reaching y_1 before the first return to y_0. The two graphs differ only by a single edge: the extra edge $\langle x_1, x_{18} \rangle$ in G^*.

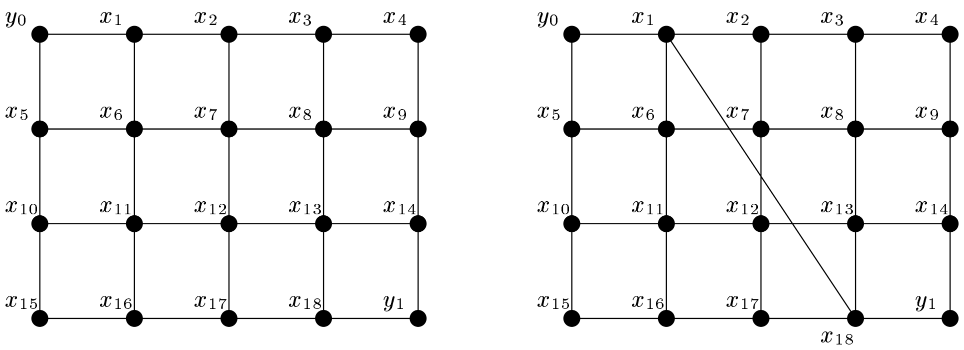


Figure 11.1. *Two graphs G and G^*, differing only in that G^* has an extra edge between the nodes x_1 and x_{18}. If in both graphs we start random walks at y_0, then which of them has the greatest probability of reaching y_0 before returning to y_1?*

Denote by A the event that the random walk on G reaches y_1 before returning to y_0, and write A^* for the corresponding event for the random walk on G^*; we call $\mathbf{P}(A)$ and $\mathbf{P}(A^*)$ the respective **escape probabilities** for the two graphs. Is $\mathbf{P}(A) < \mathbf{P}(A^*)$, or vice versa? A natural first guess is that $\mathbf{P}(A) < \mathbf{P}(A^*)$, because the extra edge $\langle x_1, x_{18} \rangle$ can help the walk to move from x_1 to x_{18} in one giant leap, something that definitely ought to improve the escape probability. But on second thought, things are not so clear: the nightmare scenario is that the random walk on G^* tediously works its way towards y_1 along the ordinary (short) edges of the graph, but then it reaches x_{18} and is on the next step thrown all the way back to x_1. So perhaps $\mathbf{P}(A) > \mathbf{P}(A^*)$ after all? The answer is, in any case, not obvious.

It turns out that the first guess is in fact correct: the extra edge $\langle x_1, x_{18} \rangle$ does improve the escape probability. A numerical calculation can be done to show that $\mathbf{P}(A) < \mathbf{P}(A^*)$. Here we shall be content with proving $\mathbf{P}(A) \leq \mathbf{P}(A^*)$. Although the result is weaker, the proof – which involves arguments with a highly electrical flavor – will in itself be much more rewarding, compared to the numerical calculation which is not particularly illuminating.

Focus for the time being on the random walk in G, and define probabilities $p(x_1), p(x_2), \ldots, p(x_{18})$ as in Sections 9 and 10, i.e., $p(x_i)$ is the probability that the random walk on G hits y_1 before y_0, starting from node x_i. For the random walk starting with $X_0 = y_0$, we get $X_1 = x_1$ or x_5 with probability $\frac{1}{2}$ each, so that

$$\begin{aligned} \mathbf{P}(A) &= \tfrac{1}{2}\mathbf{P}(A \,|\, X_1 = x_1) + \tfrac{1}{2}\mathbf{P}(A \,|\, X_1 = x_5) \\ &= \tfrac{1}{2}(p(x_1) + p(x_5))\,. \end{aligned} \tag{60}$$

Next consider the electrical network with unit resistors along the edges of G, and a unit voltage source with the minus pole at y_0 and the plus pole at y_1. Writing $V(x_i)$ for the resulting voltage at x_i, we get from Theorem 10.1 that $V(x_i) = p(x_i)$ for each i, and (60) can therefore be rewritten as

$$\mathbf{P}(A) \;=\; \tfrac{1}{2}(V(x_1) + V(x_5))\,. \tag{61}$$

Since the resistors along the edges $\langle y_0, x_1\rangle$ and $\langle y_0, x_5\rangle$ both have resistance 1, we get from Ohm's law that the two currents flowing into y_0 are $I_{x_1,y_0} = (V(x_1) - 0)/1 = V(x_1)$ and $I_{x_5,y_0} = (V(x_5) - 0)/1 = V(x_5)$, so we may further rewrite (61) as

$$\mathbf{P}(A) \;=\; \frac{I_{x_1,y_0} + I_{x_5,y_0}}{2}\,. \tag{62}$$

Note that by Kirchhoff's law applied to y_0, the sum $I_{x_1,y_0} + I_{x_5,y_0}$ in the numerator of (62) equals precisely the current that the unit voltage source receives from y_0. This is also the amount of current that the voltage source pushes through the circuit; let us denote it by I_{tot}.

The resistance R of a single resistor equals 1 divided by the amount of current flowing through the resistor when hooked up to a unit voltage source. Analogously, for a graph $G = (V, E)$ with resistances $\mathbf{R} = \{R_e\}_{e \in E}$ along the edges and a unit voltage source hooked up at y_0 and y_1, we define **the effective resistance** $R_{\mathrm{eff}} = R_{\mathrm{eff}}(G, \mathbf{R}, y_0, y_1)$ as 1 divided by the total current I_{tot} that the voltage source pushes through the circuit. It is easy to see (cf. Problem 10.1) that if the unit voltage source is replaced by one with voltage U, then the total current becomes

$$I_{\mathrm{tot}} = \frac{U}{R_{\mathrm{eff}}} \tag{63}$$

which is strongly reminiscent of Ohm's law, thereby justifying the definition of R_{eff}. Rewriting (62) once more gives

$$\mathbf{P}(A) \;=\; \frac{I_{x_1,y_0} + I_{x_5,y_0}}{2}$$

$$
\begin{aligned}
&= \frac{I_{\text{tot}}}{2} \\
&= \frac{1}{2R_{\text{eff}}(G, \mathbf{1}, y_0, y_1)} \qquad (64)
\end{aligned}
$$

where $\mathbf{1} = (1, 1, \ldots, 1)$ is for the assumption that all the resistors in the circuit have resistance 1.

Arguing in the exact same way for the graph G as for the graph G^*, we obtain that

$$
\mathbf{P}(A^*) = \frac{1}{2R_{\text{eff}}(G^*, \mathbf{1}, y_0, y_1)}
$$

analogously to (64). So to prove $\mathbf{P}(A) \leq \mathbf{P}(A^*)$, it suffices to show that

$$
R_{\text{eff}}(G, \mathbf{1}, y_0, y_1) \geq R_{\text{eff}}(G^*, \mathbf{1}, y_0, y_1) . \qquad (65)
$$

That (65) should hold seems highly plausible, because it would seem very strange if adding the edge $\langle x_1, x_{18} \rangle$ (which only gives the current more options) would *decrease* the amount of current that the unit voltage source manages to push through the circuit. This physical intuition can be justified mathematically by proving the following more general result.

Theorem 11.1. (Rayleigh's Monotonicity Principle) *If we decrease the resistance of one or more of the resistors in an electrical network, then the resulting effective resistance is at most as large as the effective resistance in the original circuit.*

The manoeuvre of adding the edge $e = \langle x_1, x_{18} \rangle$ with a unit resistor, to the unit resistor network corresponding to G, can alternatively be thought of as having had the edge e there all along, and decreasing its resistance from ∞ (allowing zero current to go through it) to 1. Thus, (65) follows once we have proved Theorem 11.1, and the desired inequality $\mathbf{P}(A) \leq \mathbf{P}(A^*)$ follows as well.

*

Do we need to prove Theorem 11.1? A physicist migh say no, and even the great James Clerk Maxwell claimed that "this principle is self-evident".[38] Readers with a taste for mathematical rigor will on the other hand agree with me that the answer is yes, and are invited to come along in the following search for the mathematical truth about Rayleigh's Monotonicity Principle. Along the way, we will learn some more about the physics of electrical networks.

[38] Maxwell (1891), quoted in Doyle and Snell (1984).

A key quantity to consider in this context is the so-called **energy dissipation** $E(e)$ in a resistor of resistance R_e along an edge $e = \langle x, y \rangle$ that is exposed to voltage $U_e = V(x) - V(y)$. The energy dissipation is defined as

$$E(e) \;=\; \frac{U_e^2}{R_e}\,.$$

By Ohm's law, the current in e from x to y is $I_e = U_e/R_e$, whence the energy dissipation may alternatively be expressed as $E(e) = U_e I_e$ or $E(e) = I_e^2 R_e$.[39] The **total energy dissipation** E_{tot} of an entire circuit, is simply $E(e)$ summed over all the resistors in the network.

The proof of Theorem 11.1 will be based on two lemmas whose proofs will be deferred until the end. The first one is as follows.

Lemma 11.1. *If to a graph $G = (V, E)$ with $V = \{x_0, x_1, \ldots, x_k\}$ and resistances $\mathbf{R} = \{R_e\}_{e \in E}$ assigned to the edges, we attach a voltage source with voltage U with the minus and plus poles at x_0 and x_1, then the total energy dissipation E_{tot} of the circuit equals*

$$E_{\mathrm{tot}} \;=\; \frac{U^2}{R_{\mathrm{eff}}(G, \mathbf{R}, x_0, x_1)}\,.$$

For the next Lemma, we need to introduce the notions of **unit current** and **unit flow** from x_1 to x_0 in a graph $G = (V, E)$ with $V = \{x_0, x_1, \ldots, x_k\}$ and resistances $\mathbf{R} = \{R_e\}_{e \in E}$ assigned to the edges. The unit current

$$\mathbf{I} = (I_{0,0}, I_{0,1}, \ldots, I_{0,k}, I_{1,0}, \ldots, I_{k,k}) \tag{66}$$

is defined as the currents in the circuit when a voltage source with U is attached with the minus and plus poles at x_0 and x_1 respectively, where U is chosen in such a way that $I_{\mathrm{tot}} = 1$. By (63), this means that U should equal the effective resistance R_{eff} of the circuit. In (66), $I_{i,j}$ is 0 whenever $i = j$ or when x_i and x_j do not share an edge, while otherwise it equals the current flowing from x_i to x_j through $\langle x_i, x_j \rangle$. Thus

$$I_{i,j} \;=\; -I_{j,i}$$

[39]Physically, the energy dissipation $E(e)$ is the amount of energy (per time unit) that is transformed from electrical energy (originating from the voltage source) to heat in the resistor.

for each i and j (the sign of the current depends on from which way we look at it). Also, using $I_{\text{tot}} = 1$ and Kirchhoff's law, we get

$$\sum_{j=0}^{k} I_{i,j} = \begin{cases} -1 & \text{for } i = 0 \\ 1 & \text{for } i = 1 \\ 0 & \text{for } i = 2, 3, \ldots, k\,. \end{cases} \tag{67}$$

Writing $R_{i,j}$ for $R_{\langle x_i, x_j \rangle}$, the energy dissipation of the unit current $\mathbf{I}$ equals

$$E(\mathbf{I}) = \tfrac{1}{2} \sum_{i=0}^{k} \sum_{j=0}^{k} I_{i,j}^2 R_{i,j} \tag{68}$$

with the convention that the summand $I_{i,j}^2 R_{i,j}$ is 0 whenever $i = j$ or when x_i and x_j do not share an edge. The factor $\frac{1}{2}$ in (68) is there to compensate for the fact that all edges appear twice in the double sum.

More generally, define a **unit flow** $\mathbf{J}$ from x_1 to x_0 in G as a vector

$$\mathbf{J} = (J_{0,0}, J_{0,1}, \ldots, J_{0,k}, J_{1,0}, \ldots, J_{k,k}) \tag{69}$$

with the property that

$$J_{i,j} = -J_{j,i}$$

for each i and j, and furthermore that $J_{i,j} = 0$ whenever $i = j$ or when x_i and x_j do not share an edge, and finally that

$$\sum_{j=0}^{k} J_{i,j} = \begin{cases} -1 & \text{for } i = 0 \\ 1 & \text{for } i = 1 \\ 0 & \text{for } i = 2, 3, \ldots, k\,. \end{cases} \tag{70}$$

The energy dissipation of $\mathbf{J}$ is defined as

$$E(\mathbf{J}) = \tfrac{1}{2} \sum_{i=0}^{k} \sum_{j=0}^{k} J_{i,j}^2 R_{i,j} \tag{71}$$

with the same convention as in (68).

The following result, known as Thomson's Principle, is an example of a kind of meta-principle that tends to show up in all sorts of areas of physics: that nature "wants to" minimize energy.

Lemma 11.2. (Thomson's Principle) *Among all unit flows* $\mathbf{J}$ *from* x_1 *to* x_0 *in* G*, the energy dissipation* $E(\mathbf{J})$ *is minimized when* $\mathbf{J}$ *is taken to be equal to the unit current* $\mathbf{I}$*.*

Assuming Lemmas 11.1 and 11.2, the proof of Rayleigh's Monotonicity Principle is now quite easy.

Proof of Theorem 11.1: Write R_{eff} and R^*_{eff} for the effective resistances in the original circuit and the modified circuit, respectively. Also write E and E^* for energies of flows in the original circuit and the modified circuit. Consider the unit current $\mathbf{I}$ through the original circuit. Since the voltage source has voltage $U = R_{\text{eff}}$, we get from Lemma 11.1 that $E(\mathbf{I}) = \frac{U^2}{R_{\text{eff}}} = R_{\text{eff}}$.

In the modified circuit, the energy dissipation $E^*(\mathbf{I})$ becomes smaller than $E(\mathbf{I})$ because each term in (68) will be smaller. But if we switch to the actual current $\mathbf{I}^*$ in the modified circuit, then by Lemma 11.2 the energy dissipation $E^*(\mathbf{I}^*)$ will be even smaller. Putting things together, we get

$$R_{\text{eff}} \;=\; E(\mathbf{I}) \;\geq\; E^*(\mathbf{I}) \;\geq\; E^*(\mathbf{I}^*) \;=\; R^*_{\text{eff}}$$

as desired. $\diamond$

It only remains to prove Lemmas 11.1 and 11.2.

Proof of Lemma 11.1: Write, as usual, $V(x_0), V(x_1), \ldots V(x_k)$ for the voltages in the nodes (relative to the minus pole x_0). The energy dissipation in (68) may be rewritten, using Ohm's law, as

$$\begin{aligned} E(\mathbf{I}) &= \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} I_{i,j}^2 R_{i,j} \\ &= \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} (V(x_i) - V(x_j)) I_{i,j} \\ &= \tfrac{1}{2}\left(\sum_{i=0}^{k} V(x_i)\sum_{j=0}^{k} I_{i,j} - \sum_{j=0}^{k} V(x_j)\sum_{i=0}^{k} I_{i,j}\right). \end{aligned} \tag{72}$$

Consider now the expression $V(x_i)\sum_{j=0}^{k} I_{i,j}$ in (72). For $i = 0$ we have $V(x_i) = 0$ while for $i = 2, \ldots, k$ we have $\sum_{j=0}^{k} I_{i,j} = 0$ by Kirchhoff's law. Thus, only in the case $i = 1$ can $V(x_i)\sum_{j=0}^{k} I_{i,j}$ be nonzero. Since $V(x_i) = U$ and $\sum_{j=0}^{k} I_{1,j} = I_{\text{tot}}$, we get

$$\begin{aligned} \sum_{i=0}^{k} V(x_i)\sum_{j=0}^{k} I_{i,j} &= V(x_1)\sum_{j=0}^{k} I_{1,j} \\ &= U I_{\text{tot}}\,. \end{aligned}$$

Next, the expression $\sum_{j=0}^{k} V(x_j) \sum_{i=0}^{k} I_{i,j}$ in (72) equals $-UI_{\mathrm{tot}}$ by the same reasoning. Inserting into (72) yields

$$\begin{aligned} E(\mathbf{I}) &= \tfrac{1}{2}(UI_{\mathrm{tot}} - (-UI_{\mathrm{tot}})) \\ &= UI_{\mathrm{tot}} \,. \end{aligned} \tag{73}$$

From (63) we know that $I_{\mathrm{tot}} = U/R_{\mathrm{eff}}$, which when plugged into (73) gives $E(\mathbf{I}) = U^2/R_{\mathrm{eff}}$. $\diamond$

Proof of Lemma 11.2: Let $\mathbf{I}$ be the unit current from x_1 to x_0, let $\mathbf{J}$ be an arbitrary unit flow from x_1 to x_0, and define their **difference flow** $\mathbf{D} = \mathbf{J} - \mathbf{I}$ by setting

$$D_{i,j} = J_{i,j} - I_{i,j}$$

for all i, j. Subtracting (67) from (70) yields

$$\sum_{j=0}^{k} D_{i,j} = 0 \tag{74}$$

for all i (including the cases $i = 0$ and 1). The energy dissipation $E(\mathbf{J})$ in (71) equals

$$\begin{aligned} E(\mathbf{J}) &= \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} J_{i,j}^2 R_{i,j} \\ &= \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} (I_{i,j} + D_{i,j})^2 R_{i,j} \\ &= \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} I_{i,j}^2 R_{i,j} + \sum_{i=0}^{k}\sum_{j=0}^{k} I_{i,j} D_{i,j} R_{i,j} + \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} D_{i,j}^2 R_{i,j} \,. \end{aligned} \tag{75}$$

The second double sum in (75) may be rewritten, using Ohm's law, as

$$\begin{aligned} \sum_{i=0}^{k}\sum_{j=0}^{k} I_{i,j} D_{i,j} R_{i,j} &= \sum_{i=0}^{k}\sum_{j=0}^{k} (V(x_i) - V(x_j)) D_{i,j} \\ &= \sum_{i=0}^{k} V(x_i) \sum_{j=0}^{k} D_{i,j} - \sum_{j=0}^{k} V(x_j) \sum_{i=0}^{k} D_{i,j} \\ &= 0 - 0 \end{aligned}$$

where the last equality uses (74). Continuing the calculation in (75), we get

$$\begin{aligned} E(\mathbf{J}) &= \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} I_{i,j}^2 R_{i,j} + \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} D_{i,j}^2 R_{i,j} \\ &\geq \tfrac{1}{2}\sum_{i=0}^{k}\sum_{j=0}^{k} I_{i,j}^2 R_{i,j} \\ &= E(\mathbf{I}) \end{aligned}$$

as desired. ◇

12 Thursday, Oct 21

Fix the dimension $d \in \{1, 2, \ldots\}$, and consider random walk on the graph $G = (V, E)$ whose vertex set $V = \mathbf{Z}^d$, where $\mathbf{Z}^d$ denotes the set of integer points in d-dimensional Euclidean space, and whose edge set consists of those pairs of vertices sitting at Euclidean distance precisely 1 from each other. With a slight abuse of notation, we write $\mathbf{Z}^d$ for this graph, which is called **the d-dimensional Euclidean lattice**; a small portion of the $\mathbf{Z}^2$ lattice is shown in Figure 12.1.

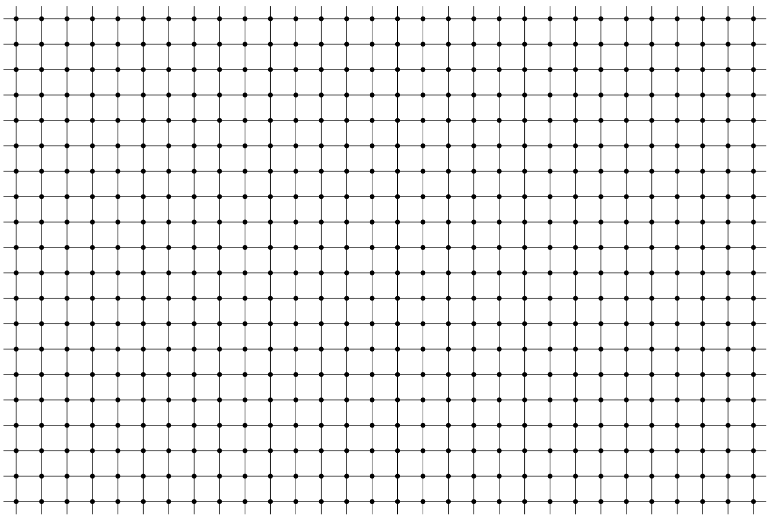


Figure 12.1. *A small part of the square lattice $\mathbf{Z}^2$.*

Starting the random walk at some $x \in \mathbf{Z}^d$, we are interested in the probability of ever returning to the starting point; in particular we are interested in whether this probability is 1 or not. Recall from Section 3 that a state x of a Markov chain is said to be **recurrent** if starting from x the chain will a.s. return to x, and **transient** otherwise.

Random walk on $\mathbf{Z}^d$ is clearly an irreducible Markov chain. For finite state irreducible Markov chains, we have seen that all states are recurrent, but here we are dealing with a countably infinite state space, and then the situation is not so clear. One thing that is not so difficult to show, however, is that for an irreducible countable state space Markov chain, either all states are transient or all states are recurrent (see Problem 12.1); this is especially obvious in lattices such as $\mathbf{Z}^d$ which "look the same from every vertex". Thus, it makes sense to talk about the entire chain being either recurrent or transient.

Is random walk on $\mathbf{Z}^d$ recurrent or transient? The answer is given in the following result, which dates back to Pólya (1921), and which I rank without hesitation to be among the ten most important and beautiful results in probability theory.

Theorem 12.1. *Random walk on* $\mathbf{Z}^d$ *is*

$$\begin{cases} \textit{recurrent} & \textit{for } d = 1 \textit{ and } 2 \\ \textit{transient} & \textit{for } d \geq 3. \end{cases}$$

The fact that the answer depends on the dimension d shows that there is something highly nontrivial going on. There are various possible approaches towards proving Pólya's theorem; here, following the direction set out in the last two lectures, we will focus on an electrical approach.

First, we make a simple observation concerning how the recurrence/transience property may depend on d. Consider random walk on $\mathbf{Z}^2$ starting from the origin $\mathbf{0} = (0, 0)$. If we look only at the x-coordinate of this walk, and only at the times at which the walk moves in the x-direction, then the successive x-coordinates that we observe behave precisely as a random walk on $\mathbf{Z}$. Suppose now that random walk on $\mathbf{Z}^2$ is recurrent, so that with probability 1 it returns to $\mathbf{0}$ infinitely often. That means in particular that it returns to the y-axis infinitely often, so that the embedded random walk on $\mathbf{Z}$ returns to 0 infinitely often, making it recurrent. In summary, recurrence of random walk on $\mathbf{Z}^2$ implies recurrence of random walk on $\mathbf{Z}$. Similarly, for any d, recurrence of random walk on $\mathbf{Z}^d$ implies recurrence on $\mathbf{Z}^{d-1}$. Hence,

> if for a given d we have that random walk on $\mathbf{Z}^d$ is transient, then this holds for all larger d as well. (76)

Alternatively, a more general result encompassing this statement can be proved using Rayleigh's Monotonicity Principle; see Problem 12.2. One useful consequence of (76) is that in order to prove Theorem 12.1, it is enough to consider the cases $d = 2$ and 3.

Let us begin with focusing on the case $d = 2$. Let $(X_0, X_1, \ldots)$ be a random walk on $\mathbf{Z}^2$ starting at the origin: $X_0 = \mathbf{0}$. We are interested in the event that the walk returns to the origin, and in order to analyse this event it is useful to introduce a couple of auxiliary random variables T and Y, defined as follows. Take T to be the time of the of the first return to $\mathbf{0}$, i.e.,

$$T = \inf\{n \geq 1 : X_n = \mathbf{0}\}$$

with the convention that $T = \infty$ if the walk *never* returns to $\mathbf{0}$. Let B_k denote the box $\{-k, -k+1, \ldots, k\}^2$ of side-length $2k+1$ centered at the origin, let $\partial B_k = B_k \setminus B_{k-1}$ denote the "outermost layer" of the box B_k, and define Y as

$$Y = \sup\{k : X_n \in \partial B_k \text{ for some } n < T\},$$

which is thus a measure of how far the walk goes from $\mathbf{0}$ before its first return, with the convention that $Y = \infty$ if the walk goes arbitrarily far without ever returning to $\mathbf{0}$.

Lemma 12.1. *We have with probability* 1 *that either* T *is finite or* Y *is infinite.*

Proof: We need to show that

$$\mathbf{P}(T = \infty, Y < \infty) = 0. \tag{77}$$

The event in (77) means that there is some box B_k which the walk never leaves, yet it never returns to the origin. That this has probability 0 of happening follows by an easy modification of the argument in Lemma 4.1. (See Problem 12.3.) $\diamond$

To prove recurrence of random walk on $\mathbf{Z}^2$ is the same as showing that $\mathbf{P}(T = \infty) = 0$, and in view of Lemma 12.1 it is enough to show that

$$\mathbf{P}(Y = \infty) = 0.$$

This is the same as showing that

$$\lim_{k \to \infty} \mathbf{P}(Y \geq k) = 0. \tag{78}$$

An advantage with working with the quantity $\mathbf{P}(Y \geq k)$ in (78) is that its analysis can be reduced to a random walk on a finite graph: the event $\{Y \geq$

$k\}$ is the same as saying that the random walk hits ∂B_k before returning to $\mathbf{0}$, and to study the probability of this event we may as well think of the random walk as taking place on B_k rather than on the full $\mathbf{Z}^2$ lattice.

Thus, take G_k to be the graph with vertex set B_k and edges connecting pairs of vertices at Euclidean distance 1 from each other. The desired probability

$$\mathbf{P}(\text{the walk hits } \partial B_k \text{ before returning to } \mathbf{0})$$

looks much like the escape probability for the random walk in the example in the beginning of Section 11. There, the escape probability turned out to equal

$$\frac{1}{2R_{\text{eff}}(G, \mathbf{1}, x_0, x_1)}\,. \tag{79}$$

An examination of the derivation of (79) in Section 11 shows that the '2' in (79) comes from the fact that x_0 is incident to precisely 2 edges. In our setting, $\mathbf{0}$ is incident to 4 edges, and we get

$$\begin{aligned} \mathbf{P}(Y \geq k) &= \mathbf{P}(\text{the walk hits } \partial B_k \text{ before returning to } \mathbf{0}) \\ &= \frac{1}{4R_{\text{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k)} \end{aligned} \tag{80}$$

where the last two arguments of R_{eff} mean that the minus pole of the voltage source is hooked to $\mathbf{0}$ while the plus pole is hooked to *all* vertices in ∂B_k. The desired relation (78) follows if we can show that

$$\lim_{k\to\infty} R_{\text{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k) = \infty\,. \tag{81}$$

Now, $R_{\text{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k)$ equals the energy dissipation $E[\mathbf{I}]$ of the unit current from $\mathbf{0}$ to ∂B_k. For $i = 1, \ldots, k-1$, let F_i denote the set of edges connecting ∂B_{i-1} to ∂B_i, and let $E[\mathbf{I}(F_i)]$ denote the part of $E[\mathbf{I}]$ coming from edges in F_i. Note that F_i consists of precisely $8i - 4$ edges, and that

$$\sum_{e \in F_i} \mathbf{I}_e = 1 \tag{82}$$

where the current through e is measured as going from ∂B_{i-1} to ∂B_i. The following lemma now comes in handy.

Lemma 12.2. *Let $a_1, a_2, \ldots, a_n$ be arbitrary real numbers satisfying $a_1 + \cdots + a_n = 1$. We then have*[40]

$$\sum_{j=1}^n a_j^2 \geq \frac{1}{n}\,.$$

[40]Note that taking $a_1 = \cdots = a_n = \frac{1}{n}$ shows that the lemma is sharp.

Proof: For each j, set $b_j = a_j - \frac{1}{n}$, so that in particular

$$\sum_{j=1}^{n} b_j \;=\; 0\,. \tag{83}$$

We get

$$\begin{aligned}
\sum_{j=1}^{n} a_j^2 \;&=\; \sum_{j=1}^{n}\left(b_j + \frac{1}{n}\right)^2 \\
&=\; \sum_{j=1}^{n} b_j^2 + 2\sum_{j=1}^{n}\frac{b_j}{n} + \sum_{j=1}^{n}\frac{1}{n^2} \\
&=\; \sum_{j=1}^{n} b_j^2 + \sum_{j=1}^{n}\frac{1}{n^2} \\
&=\; \sum_{j=1}^{n} b_j^2 + \frac{1}{n} \\
&\geq\; \frac{1}{n}
\end{aligned}$$

where the third equality uses (83). ◇

From (82) and Lemma 12.2, we now get

$$\begin{aligned}
E[\mathbf{I}(F_i)] \;&=\; \sum_{e\in F_i} (\mathbf{I}_e)^2 \\
&\geq\; \frac{1}{8i-4}\,.
\end{aligned}$$

Summing over the layers $F_1, \ldots F_k$ we get

$$\begin{aligned}
R_{\text{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k) \;&=\; E[\mathbf{I}] \\
&\geq\; \sum_{i=1}^{k} E[\mathbf{I}(F_i)] \\
&\geq\; \sum_{i=1}^{k} \frac{1}{8i-4}
\end{aligned}$$

so that

$$\lim_{k\to\infty} R_{\text{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k) \;\;\geq\;\; \lim_{k\to\infty}\sum_{i=1}^{k}\frac{1}{8i-4}$$

$$\begin{aligned} &= \sum_{i=1}^{\infty} \frac{1}{8i-4} \\ &= \infty \end{aligned} \tag{84}$$

where the last step follows from the well-known divergence of the harmonic series: $\sum_{i=1}^{\infty} \frac{1}{i} = \infty$. So (81) is established, and plugging it into (80) yields $\lim_{k\to\infty} P(Y \geq k) = 0$, so that $\mathbf{P}(Y = \infty) = 0$, and recurrence of random walk on $\mathbf{Z}^2$ is established.

*

Having thus established Pólya's Theorem for $d = 2$, it remains to prove it for $d = 3$. It is instructive to consider what happens if we try apply the $d = 2$ argument to prove the (incorrect) statement that random walk on $\mathbf{Z}^3$ is recurrent. In three dimensions, the number of edges between B_{i-1} and B_i is $6(2i-1)^2$. Hence, the three-dimensional counterpart of (84) becomes

$$\lim_{k\to\infty} R_{\mathrm{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k) \geq \sum_{i=1}^{\infty} \frac{1}{6(2i-1)^2} \tag{85}$$

This sum is finite, due to the well-known finiteness of $\sum_{i=1}^{\infty} \frac{1}{i^2}$, and the argument for proving recurrence therefore collapses.

On the other hand, this strongly suggests that

$$\lim_{k\to\infty} R_{\mathrm{eff}}(G_k, \mathbf{1}, \mathbf{0}, \partial B_k) < \infty \tag{86}$$

so that $\mathbf{P}(Y = \infty) > 0$, giving the desired transience. The problem, however, with this simple-minded argument is that the inequlity in (85) goes in the wrong direction.

What can we do instead? The best approach is to construct, for any k, a unit flow from $\mathbf{0}$ to ∂B_k, in such a way that the energy dissipation is bounded by some constant C, uniformly in k. By Lemma 11.2 (Thomson's Principle) this shows that the unit currents are bounded by C, which implies (86), and transience of random walk on $\mathbf{Z}^3$ follows. The construction of such a flow is a non-trivial task, but the corageous reader is invited to work it out, using a series of hints provided in Problem 12.4.

And with this challenge, ladies and gentlemen, I close this lecture series.

Problems

Problem 1.1. Consider the Markov chain in Example 1.4, and show that

$$\mathbf{P}(\text{the chain sooner or later explodes}) = 1 .$$

Problem 2.1. For the Richardson model, we considered the state space $\mathcal{S} = \{0, 1\}^{\mathbf{Z}^d}$ and the more restrictive state space $\mathcal{S}^*$ consisting of only those elements of $\mathcal{S}$ that have only finitely many 1's. Show that $\mathcal{S}$ is uncountable while $\mathcal{S}^*$ is countable.

Problem 2.2. (This problem is meant to illustrate the nontrivialness of going from eq. (3) to eq. (2)). Give an example of an increasing sequence of positive random variables $X_1, X_2, \ldots$ such that

$$\lim_{k\to\infty} \mathbf{E}[X_k] = \infty$$

while nevertheless

$$\mathbf{P}\left(\lim_{k\to\infty} X_k < \infty\right) = 1 .$$

Problem 2.3. The Richardson model can be defined on other lattices than the square lattice, or even on an arbitrary graph structure where we allow infections between two nodes x and y if and only if they share an edge.

(a) Can the argument for non-explosion on $\mathbf{Z}^2$ be adapted to get the analogous result for the cubic $\mathbf{Z}^3$-lattice? If yes, then what would be the most far-reaching generalization of this result that you can prove using the same technique?

(b) Can you find an example of a graph structure where the Richardson model *does* explode?

Problem 3.1. Determine, for each state s_i in the Markov chain in Figure 3.1, whether s_i is recurrent or transient.

Problem 3.2. Let X be a positive random variable with the property that for some $C, c > 0$ and for any n, we have

$$\mathbf{P}(X > n) \leq Ce^{-cn} .$$

Prove that X has finite expectation.

Problem 3.3. Show that, for the Markov chain in Figure 3.1, both

$$\pi = (0, 0, \tfrac{1}{5}, \tfrac{2}{5}, \tfrac{2}{5}, 0, 0)$$

and

$$\pi' = (0, 0, 0, 0, 0, \tfrac{1}{4}, \tfrac{3}{4})$$

are stationary distributions. Then show that for any $p \in [0, 1]$, $p\pi + (1-p)\pi'$ is also a stationary distribution.

Problem 5.1. Consider the Markov chain in Example 5.1.

(a) Show that its unique stationary distribution equals $(\frac{1}{2}, \frac{1}{2})$.

(b) Prove by induction that for every n, $\mu^{(n)} = (\frac{1}{2} + \frac{1}{2^{n+1}}, \frac{1}{2} - \frac{1}{2^{n+1}})$.

Problem 5.2. Prove Lemma 5.1. Also formulate and prove a generalization to the case where $\mathcal{S}$ is countable.

Problem 5.3.

(a) Give an example of a Markov chain with period 7.

(b) Give an example of an aperiodic Markov chain in which jumps from a state to itself always has probability 0.

Problem 5.4. Prove the **coupling inequality**, which says the following. If X and Y are two random variables taking values in the same space $\mathcal{S}$, and having respective distributions μ and ν, then

$$d_{\mathrm{TV}}(\mu, \nu) \leq \mathbf{P}(X \neq Y)\,.$$

Problem 6.1. Show that $\sigma_f < \infty$, where σ_f^2 is defined as in (25). Hint: Note that, since $\mathcal{S}$ is finite, $(f(X_i) - \mu_i)$ is bounded by some absolute constant. Then exploit Lemma 3.2.

Problem 6.2. This problem is meant to illustrate that the asymptotic variance σ_f^2 can be zero also in other cases beyond the totally trivial one where f is constant.

(a) Consider the irreducible finite-state Markov chain with state space $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5, s_6\}$ and transition matrix

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

Furthermore, define $f : \mathcal{S} \to \mathbf{R}$ by taking $f(1) = 0$, $f(2) = f(4) = f(6) = -1$, $f(3) = 1$ and $f(5) = 2$, and show that $\mu_f = 0$ and $\sigma_f^2 = 0$.

(b) Show that for any irreducible Markov chain with finite state-space $\mathcal{S} = \{s_1, \ldots, s_k\}$ and any f such that $\sigma_f^2 = 0$, we have that the discrepancy

$$\left| \left(\sum_{i=1}^{n} f(X_i) \right) - n\mu_f \right|$$

is actually *bounded* (uniformly in n).

Problem 6.3. Consider the Markov chain in Figure 3.1, together with the function $f : \mathcal{S} \to \mathbf{R}$ given by $f(s_1) = f(s_2) = f(s_4) = f(s_5) = f(s_7) = 0$, $f(s_3) = 1$ and $f(s_6) = -1$. Then Theorem 6.1 applies to show that if the chain is started in state s_3, then $\sum_{i=1}^{n} f(X_i)$ is asymptotically normal. The same conclusion holds if we start the chain (deterministically) in one of the states s_4, s_5, s_6 or s_7. Show that, on the other hand, if the chain is started in s_1 or s_2, then the distribution of $\sum_{i=1}^{n} f(X_i)$ is *not* asymptotically normal.

Problem 6.4. Consider the Markov chain with state space $\mathcal{S} = \{s_1, s_2\}$ and transition matrix

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

with $p \in (0, 1)$. Also define $f : \mathcal{S} \to \mathbf{R}$ by taking $f(s_1) = 1$ and $f(s_2) = 0$. Clearly, the chain is irreducible with stationary distribution $\pi = (\frac{1}{2}, \frac{1}{2})$, and $\mu_f = \frac{1}{2}$. Compute the asymptotic variance σ_f^2 and comment on its dependence on p. (Does it depend monotonically on p? Is this intuitively plausible?)

Problem 6.5. Prove Lemma 6.2.

Problem 7.1.

(a) Show that the Markov chain with state space $\mathcal{S} = \{s_1, s_2, s_3\}$ and transition matrix

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

is not reversible.

(b) Give an example of a (necessarily non-irreducible) Markov chain which has both a reversible stationary distribution and a stationary distribution which is *not* reversible.

Problem 7.2.

(a) Show that the two Markov chains in Examples 2.2 and 3.1 are reversible.

(b) More generally, show that any Markov chain with state space $\mathbf{Z}$ and the property that jumps between states that differ by at least 2 have probability 0, are reversible.

(c) Give an example of a countable state Markov chain which is *not* reversible.

Problem 8.1. Consider the Markov chain with state space $\mathcal{S} = \{s_1, \ldots, s_7\}$ and transition matrix

$$\begin{bmatrix} 0 & p_1 & 0 & 0 & 0 & 0 & 1-p_1 \\ 1-p_2 & 0 & p_2 & 0 & 0 & 0 & 0 \\ 0 & 1-p_3 & 0 & p_3 & 0 & 0 & 0 \\ 0 & 0 & 1-p_4 & 0 & p_4 & 0 & 0 \\ 0 & 0 & 0 & 1-p_5 & 0 & p_5 & 0 \\ 0 & 0 & 0 & 0 & 1-p_6 & 0 & p_6 \\ p_7 & 0 & 0 & 0 & 0 & 1-p_7 & 0 \end{bmatrix}$$

for some $p_1, \ldots, p_7 \in (0, 1)$. Thinking of the states $s_1, \ldots, s_7$ as arranged in a circle, this can be thought of a random walk on a circle, where there is a drift towards clockwise (if $p_i > \frac{1}{2}$) or counterclockwise (if $p_i < \frac{1}{2}$) movement that may depend on the site. Find a necessary and sufficient condition on $p_1, \ldots, p_7$ for the chain to be reversible.

Problem 9.1. Consider the Metropolis–Hastings Markov chain described in the beginning of Section 9, together with the distribution π, and show that π satisfies the local equilibrium condition $\pi_i P_{i,j} = \pi_j P_{j,i}$ for all $s_i, s_j \in \mathcal{S}$.

Problem 9.2. Many probability distributions, particularly in statistical mechanics and related fields, arise in the following way. Let $\mathcal{S}$ be some set, which for concreteness we assume to be finite, and let $H : \mathcal{S} \to \mathbf{R}$ be some real-valued function of $\mathcal{S}$. For $s \in \mathcal{S}$, we call $H(s)$ the **energy** of s. Nature tends to assign small probabilities to states with large energy, and vice versa. More precisely, we want to set probability $e^{-H(s)}$ to state s. This might not lead to a proper probability distribution, since the "probabilities" might not sum to 1. We therefore introduce a so-called **normalizing constant** Z as

$$Z = \sum_{s \in \mathcal{S}} e^{-H(s)}$$

and define the **Boltzmann distribution** corresponding to the energy function H, as the probability distribution which to each $s \in \mathcal{S}$ assigns probability $\frac{1}{Z}e^{-H(s)}$.

A major obstacle in working with Boltzmann distributions is that when the state space is very large, the normalizing constant Z may be very difficult to compute or even find a reasonable approximation to. However, the Metropolis–Hastings algorithm is particularly suitable for simulating Boltzmann distributions, because it has the property that

$$\text{in order to implement the algorithm, there is no need to know the normalizing constant } Z. \quad (87)$$

Demonstrate that statement (87) is correct.

Problem 9.3. The vector $\mathbf{p}$ in (50) exhibits certain properties that may at first sight look like numerical coincidences, such as $p(x_1)+p(x_4) = 1$. Show that these are in fact not coincidences at all.

Problem 10.1. Consider an electrical network such as the one in Figure 10.2.

(a) Show that if all the unit resistors are replaced by resistors with resistance $R > 0$ (the same for all of them), then all the voltages $V(x_i)$ remain the same as in the original network.

(b) Show that if the unit voltage source is replaced by one with voltage $U > 0$, then all the voltages $V(x_i)$ get multiplied by U compared to the original network.

Problem 12.1. Consider an irreducible Markov chain with countable state space $\mathcal{S}$, and show that, for two arbitrary states $x, y \in \mathcal{S}$, if x is recurrent then so is y.

Problem 12.2. Let G and G' be two infinite locally finite connected graphs, such that G is a subgraph[41] of G'. Prove, using Rayleigh's Monotonicity Principle, that if random walk on G' is recurrent, then so is random walk on G.

Problem 12.3. Fill in the details of the proof of Lemma 12.1.

Problem 12.4. For each k, construct a unit flow in the $\mathbf{Z}^3$ lattice from the origin $\mathbf{0}$ to the layer ∂B_k, in such a way that the energy dissipations of

[41] A graph $G = (V, E)$ is a subgraph of another graph $G' = (V', E')$ if V is a subset of V' and E is a subset of E'.

these flows are bounded by a constant C (independent of k).[42] You may, for instance, use the following approach.[43]

(a) One way to define a unit flow $\mathbf{J}$ from $\mathbf{0}$ to ∂B_k is as follows. Define a random procedure for selecting a path ξ from $\mathbf{0}$ to ∂B_k in the graph, and for each edge $e = \langle x, y \rangle$ take $J_{\langle x,y \rangle}$ to be the expected number of times that the path passes through e, counting positive if it goes from x to y and negative otherswise. Prove that this makes $\mathbf{J}$ a unit flow from $\mathbf{0}$ to ∂B_k. In other words, prove the Kirchhoff's law holds for $\mathbf{J}$ at all vertices except $\mathbf{0}$ (which exports flow 1) and the vertices in ∂B_k (who import at total flow of 1).

(b) Show that the following way to choose the path π is possible. First pick a vertex v in ∂B_k unifrmly at random, and draw a straight line in $\mathbf{R}^3$ from $\mathbf{0}$ to v. Then construct a path from $\mathbf{0}$ to v in the lattice, sitting inside the cylinder of radius 1 around this line, and such that no edge is traversed at most 1.

(c) Show that when ξ is chosen as in (b), and $\mathbf{J}$ is then chosen as in (a), then the current J_e going through an edge e either sitting in layer ∂B_i or connecting ∂B_{i-1} to ∂B_i, is bounded by a universal constant D divided by i^2.

(d) Use (c) and the fact that $\sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$ to show that the flows thus constructed satisfy the desired boundedness of energy dissipation.

[42] As explained at the end of Section 12, this is precisely what needs to be done in order to prove transience of random walk on $\mathbf{Z}^3$.

[43] There are alternative ways to proceed; see, e.g., Doyle and Snell (1984).

Bibliography

Asmussen, S. (2003) *Applied Probability and Queues* (2nd ed), Springer, New York.

Brémaud, P. (1998) *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer, New York.

Doyle, P.G. and Snell, J.L. (1984) *Random Walks and Electric Networks*, Mathematical Association of America. The authors have made their manuscript available free of charge at `http://arxiv.org/abs/math.PR/0001057` .

Durrett, R. (1991) *Probability: Theory and Examples*, Wadsworth & Brooks/Cole, Pacific Grove.

Enger, J. and Grandell, J. (2000) *Markovprocesser och köteori*, Mathematical statistics, KTH, Stockholm.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*, Clarendon, Oxford.

Grimmett, G.R. and Stirzaker, D. (1992) *Probability and Random Processes*, Clarendon, Oxford.

Häggström, O. (2002) *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press.

Häggström, O. (2004) *Slumpens skördar*, Studentlitteratur, Lund.

Häggström, O. and Pemantle, R. (1998) First passage percolation and a model for competing spatial growth, *Journal of Applied Probability* **35**, 683–692.

Liggett, T.M. (1999) *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, Springer, New York.

Lindvall, T. (1992) *Lectures on the Coupling Method*, Wiley, New York.

Maxwell, J.C. (1891) *Treatise on Electricity and Magnetism* (3rd ed.), reprinted by Dover, New York, 1954.

Norris, J. (1997) *Markov Chains*, Cambridge University Press.

Pólya, G. (1921) Über eine Aufgabe betreffend die Irrfahrt im Strassennetz, *Mathematische Annalen* **84**, 149–160.

Thorisson, H. (2000) *Coupling, Stationarity, and Regeneration*, Springer, New York.