Discrete-time
Homogeneous Markov Chains
With Countable State Spaces

Supplementary handout for ECE 4271

prepared by

David F. Delchamps

Spring 2022

Table of Contents

Preface

Readers can find many standard references on random processes that include excellent treatments of discrete-time Markov chains with countable state spaces. My goal here is to present the essential theory in a largely self-contained fashion. An advanced undergraduate who has taken a good course in discrete probability should find it accessible. I've included proofs of all the usual limit theorems and other results on which those theorems depend, and I've tried to arrange the material in a way that makes both intuitive and logical sense to me. I don't offer the usual panoply of examples arising in applications, mainly due to a lack of space, but also because including them would disrupt the flow. Speaking of trying to avoid flow disruption, I've relegated to the Appendix several annoyingly complicated proofs that do little to enhance one's intuition about Markov chains. All the proofs are my own, but of course I've drawn inspiration from time-tested references, particularly Karlin and Taylor's *A First Course in Stochastic Processes* and Sid Resnick's *Adventures in Stochastic Processes.* In particular, I pretty much cribbed the proof of Theorem A2 in the Appendix from Karlin and Taylor's book. I'd like to acknowledge helpful conversations with Sid Resnick and Terry Fine and also express my gratitude to my army of proofreaders — the students who have taken ECE 4271 over the last few years — for their patience and keen-eyed close reading.

1. What is a Markov Chain? The lightwall model.

Most textbooks introduce Markov chains as special examples of random processes. First the books define what a random process is, and then they describe restrictions on a random process that make that process a Markov chain. Those restrictions guarantee mathematically that the randomness in the random process has certain features that make it especially easy to work with. Here, rather than take you down that standard route through general random processes to Markov chains, I'll try to get at Markov chains directly by essentially defining their building blocks and rules of operation in a somewhat informal way. I'll begin by describing a model that helps me visualize the key concepts.

Picture a huge wall of colored lights. The wall might feature finitely many lights or countably infinitely many of them. You are looking at the wall and can't see what's going on behind it. You observe that at each integer time $n \geq 0$, exactly one of the lights flashes. What light flashes at what time appears random. Sometimes one light will flash several times in a row. Some lights flash only rarely.

Now for the part you can't observe. Behind each light sits an operator. A light flashes if and only if its operator has been told to flash it. At time $n = 0$, a supervisor tells one of the operators to flash his light. The supervisor decides which operator goes first by (figuratively) rolling a die or flipping a coin, i.e., by some kind of random choice mechanism.

The lucky operator who flashes her light at time $n = 0$ (let's call her Arianna) then has the job of deciding which operator to direct to flash his light at time $n = 1$. She makes that decision by rolling her own figurative die. If you're straining to visualize many-sided dice or coins, think instead of each operator as possessing a spinner such as you might use in playing a game of Twister, say. Arianna's random-choice device has a certain probability of fingering each light operator, including Arianna herself. So Arianna spins her spinner and sees who the $n = 1$-operator will be, at which point she directs him to flash his light at $n = 1$.

The $n = 1$-operator, let's call him Brent, having flashed his light, spins his own spinner to decide who goes at time $n = 2$. Brent's spinner might be quite different from Arianna's. The most important thing about the two spinners is that they are completely independent of each other. The $n = 2$-operator to whom Brent gives the "Flash" order has his own independent spinner, which he will use to decide who goes at $n = 3$. And so on.

What you see from your side of the wall is some random sequence of lights flashing. The special way the randomness arises is what makes the lightwall a Markov chain. If, for example, the red light flashes at some specific time n, the probability that the blue light flashes at time $n + 1$ depends only on the structure of the spinner belonging to the operator behind the red light. What light actually flashes at time $n + 1$ depends only on the outcome of that spinner's spin. All past history, i.e., the record of what lights have flashed at times up through $n - 1$, is irrelevant to determining the probability that a given light will flash at time $n + 1$ given that the red light has flashed at time n. Furthermore, all the operators' spinners are independent and don't change over time.

More formally, every discrete-time homogeneous Markov chain has a compact collection of building blocks that determine it completely in a sense I'll make precise

in Section 2. I'll refer to the lightwall model to try to anchor things. The building blocks are

- A *state space* $\mathcal{S}$, which will generally be the set of positive integers $\{1, 2, 3, \ldots\}$ or the finite set $\{1, 2, \ldots, M\}$ for some integer $M > 0$. For each integer time $n \geq 0$, if you observe the Markov chain you will find that it is "in" one of its states. (Think of each light on the wall as representing a state; "being in a state" means the corresponding light is flashing.)
- For each $i \in \mathcal{S}$ a set of *one-step transition probabilities* $P(i, j)$ for $j \in \mathcal{S}$. $P(i, j)$ denotes the probability that the Markov chain will be "in" state j at the next time step given that it is "in" state i now. Note that $P(i, j)$ does not depend on what "now" is — that's what homogeneous means. For each i, $\{P(i, j) : j \in \mathcal{S}\}$ is a *conditional* probability distribution on the "next state" j, where the condition is that the Markov chain is currently in state i. Hence $P(i, j) \geq 0$ for all i and j, and $\sum_{j \in \mathcal{S}} P(i, j) = 1$ for every i. (In the lightwall model, the operator of light i has a time-invariant spinner divided into wedges — the area of j's wedge takes up a fraction $P(i, j)$ of the total area of i's spinner.)
- An *initial distribution* $\pi(0)$ over states. Think of $\pi(0)$ as an infinite (or M-dimensional) vector; $\pi(0)$ has one component $\pi_i(0)$ for each state $i \in \mathcal{S}$. $\pi_i(0)$ represents the probability that the Markov chain is in state i at time 0, so $\pi_i(0) \geq 0$ for all $i \in \mathcal{S}$ and $\sum_{i \in \mathcal{S}} \pi_i(0) = 1$. (Think of the supervisor's initial coin flip in the lightwall model as establishing $\pi(0)$.)

What makes homogeneous Markov chains particularly amenable to analysis is the following strong assumption.

Markov Property: For every $n \in \mathbb{N}$ and every i_0, i_1, $\ldots$, i_{n-1}, i, and j in $\mathcal{S}$, the probability that the Markov chain will be in state j at time $n + 1$ given that it is in state i at time n and state i_m at time m for every $0 \leq m < n$ is $P(i, j)$.

The Markov property is essentially about the irrelevance of past history to computing the probability of making an i to j transition at time n given that you know the chain is in state i at time n. The lightwall model exhibits the Markov property. If I'm a light operator and another operator tells me to flash my light, that operator's identity doesn't affect the outcome when I spin my spinner to pick the next operator — nor does the identity of any operator who has flashed his light before.

Here's one key consequence of the Markov property. Suppose we want to figure out the probability of starting in state 17 at time 0, then making a transition to state 3 at time 1, and from there to state 11 at time 2. The probability of starting at time 0 in state 17 is $\pi_{17}(0)$ and the probability of transitioning to 3 at time 1 given that we started in 17 at time 0 is $P(17, 3)$. Thus the probability that we'll see 17 at time 0 and 3 at time 1 is simply the product $\pi_{17}(0)P(17, 3)$. The probability that we'll see the sequence 17, 3, 11, by the rules of conditional probability, is the product of $\pi_{17}(0)P(17, 3)$ with

$$\text{Prob}\{\text{state at time 2 is 11 given state at time 0 is 17 and at time 1 is 3}\} .$$

By the Markov property, stipulating that the state at time 0 was 17 makes no difference — in other words, the conditional probability above is just $P(3, 11)$. Accordingly, the probability we set out to calculate is the product $\pi_{17}(0)P(17, 3)P(3, 11)$. More generally, for any sequence of states i_0 , i_1, ... , i_n, the probability of seeing the Markov chain pass through these states in order starting at time 0 is

$$\pi_{i_0}(0)P(i_0, i_1)P(i_1, i_2) \cdots P(i_{n-1}, i_n) \ .$$

It's convenient to represent the state-transition structure of homogeneous Markov chains pictorially as in Figure 1, which illustrates *transition diagrams* for various Markov chains. The absence of an arrow from i to j means $P(i, j) = 0$ — i.e., it's impossible to see state i at some time instant and state j at the next time instant. If $P(i, j) > 0$, there's an arrow from i to j, and the number next to the arrow is $P(i, j)$.

Define a *run* of a Markov chain as the particular infinite sequence of states it visits starting at time 0 as a result of the various random events that determine the starting state and subsequent state transitions. Possible runs of the Markov chain are in one-to-one correspondence with paths through the transition diagram, starting from some state, that always follow the arrows from state to state. In the lightwall example, the supervisor's initial coin flip begins a run by specifying the initial state for the run, namely, which operator flashes his light first. That operator spins his spinner to determine the next state in the run, and so on. In Figure 1(a), for example, $1, 1, 2, 3, 3$ is the possible start of a run, whereas $1, 1, 2, 1, 2, 3$ is not. The initial distribution $\pi(0)$ and the transition probabilities $P(i, j)$ induce a probability distribution over "run space." When I talk about probability distributions of various random quantities associated with the Markov chain, I'll be referring back to that probability distribution, at least implicitly. More on this point later.

The Markov property makes it easy to calculate the probability of seeing any finite sequence of possible state transitions. All you have to do is find the unique path through the transition diagram associated with the sequence and compute the product of the transition probabilities labeling the arrows along the path. In Figure 1(c), for example, given that you start in state 1, the probability that you'll see the sequence 1, 2, 2, 2, 1 is $p(1 - q)^2q$ — that is, the product of the probabilities of making a 1-to-2 transition, then two 2-to-2 transitions, and finally a 2-to-1 transition.

For any $m > 0$, define the m-step transition probabilities as follows: $P^{(m)}(i, j)$ is the probability that the Markov chain visits state j at time m given that it started in state i at time 0. Note that $P^{(m)}(i, j)$ is the same as the probability that, m time steps from now, the chain will be in state j given that it is in state i now — no matter what "now" is. This is, once again, a consequence of homogeneity.

You can generate the m-step transition probabilities recursively from the one-step transition probabilities. Consider first the case $m = 2$. The probability that the chain makes a two-step i-to-j transition is the same as the sum of the probabilities of all the two-step paths that begin in i and end up in j while passing through some state k in between. By the Markov property, each such i-to-k-to-j transition has probability $P(i, k)P(k, j)$, so

$$P^{(2)}(i, j) = \sum_{k \in \mathcal{S}} P(i, k)P(k, j)$$

for every i and j in $\mathcal{S}$. What about $m = 3$? To make a three-step transition from i to j, the chain must make a two-step transition from i to some state k followed by a one-step transition from k to j. Summing up the probabilities of all those possible three-step transitions from i to j yields

$$P^{(3)}(i,j) = \sum_{k \in \mathcal{S}} P^{(2)}(i,k)P(k,j)$$

for every i and j in $\mathcal{S}$. Plowing along inductively yields the recursive formula

$$P^{(m+1)}(i,j) = \sum_{k \in \mathcal{S}} P^{(m)}(i,k)P(k,j)$$

for every i and j in $\mathcal{S}$ and every $m > 0$.

Things are especially easy when the state space is finite. In that case, we can form a matrix P from the transition probabilities in a natural way. Suppose the Markov chain has M states. Define the $(M \times M)$ matrix P using

$$[P]_{ij} = P(i,j)$$

for $1 \leq i, j \leq M$. Raising P to powers gives the matrices of higher-order transition probabilities. For example,

$$P^{(2)}(i,j) = \sum_{k \in \mathcal{S}} P(i,k)P(k,j) = \sum_{k=1}^{M} [P]_{ik}[P]_{kj} = \left[P^2\right]_{ij}$$

for $1 \leq i, j \leq M$. By an easy induction based on the calculations in the preceding paragraph,

$$P^{(m)}(i,j) = \left[P^m\right]_{ij}$$

for $1 \leq i, j \leq M$ and every $m > 0$. As a reality check, convince yourself that the matrix P corresponding to the Markov chain in Figure 1(c) is

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

We'll be interested in answering questions about the behavior of Markov chains over time, by which I mean questions about properties of the various runs we might see. Many of our answers to these questions will be about what happens, on the average, over many runs of the chain, but some questions have sharper answers with strong implications regarding what we might ever expect to see happen in a particular run. Thinking in terms of the lightwall, we might ask, for example, how long do we have to wait, on the average, for the red light to flash after we see the blue light flash? If the green light flashes at some time, are we guaranteed to see it flash at some later time, or is there a nonzero probability that we'll never see it flash again? If we watch the chain for a very long time, what's the long-term average fraction of times that the magenta light flashes? And how do the answers to questions such as these depend, if at all, on the Markov chain's initial distribution $\pi(0)$?

2. Random and non-random quantities; transient and recurrent states

The building blocks of a Markov chain that I described in Section 1 are non-random quantities They're just parameters of the Markov chain that summarize its rules

of operation. In the lightwall model, the state space is the lightwall itself. The operators' spinners determine the $P(i,j)$, and the supervisor's spinner is laid out according to $\pi(0)$. Only when the operators and supervisor actually spin their spinners does randomness arise.

Random quantities associated with the Markov chain have values that depend on the outcomes of random events. One can view any such random quantity as a function whose domain is the set of all paths through the state space that the Markov chain might follow – "path space" for short. If Z is any random quantity arising from the Markov chain, then to determine the value of Z we have to look at the path the chain follows over time. The path the chain follows is influenced by random events, and that's what makes Z a random quantity. With the notable exception of the transition probabilities, I'll label random quantities with upper-case letters and non-random quantities with lower-case letters.

The most fundamental random quantity associated with a Markov chain is X_n, the state the Markov chain is in at time n. Whether $X_n = j$ for a given state j depends on the state the chain started in at time 0 along with all the subsequent state transitions up to time n, all of which arise from random events. The *probability* that $X_n = j$ depends on the initial distribution $\pi(0)$ and on all the *probabilities* of all the random outcomes that determine the state transitions. Like any random quantity associated with the Markov chain, you may regard X_n as a function on path space; X_n takes values in the state space $\mathcal{S}$, and $X_n = j$ precisely for those paths that hit state j at time n.

I'll elaborate now on what I meant when I asserted in Section 1 that a Markov chain's building blocks "determine it completely." You'll have to take my word for it when I say that you know everything about the Markov chain if you know all the joint probabilities of the form

$$\text{Prob}\{X_{n_1} = i_1, X_{n_2} = i_2, \ldots, X_{n_k} = i_k\}$$

for every increasing sequence of natural numbers n_1, n_2, ... , n_k and every i_1, i_2, ... , i_k in the state space $\mathcal{S}$. If you've ever studied random processes, you've heard some version of the mantra that the finite-dimensional joint distributions determine the process completely — but I'd prefer not to dwell on that. In any event, it's easy to show using the Markov property and the rules of conditional probability that the joint probability above is equal to the product of the following probabilities: the probability that $X_{n_1} = i_1$; the probability that the chain makes an $(n_2 - n_1)$-step transition from i_1 to i_2; the probability that the chain makes an $(n_3 - n_2)$-step transition from i_2 to i_3; ... ; and finally the probability that the chain ends up in state i_k by making an $(n_k - n_{k-1})$-step transition from i_{k-1}. In terms of the parameters of the Markov chain, that product is

$$\text{Prob}\{X_{n_1} = i_1\} P^{(n_2-n_1)}(i_1, i_2) P^{(n_3-n_2)}(i_2, i_3) \cdots P^{(n_k-n_{k-1})}(i_{k-1}, i_k) \; .$$

Meanwhile,

$$\text{Prob}\{X_{n_1} = i_1\} = \begin{cases} \sum_{j \in \mathcal{S}} \pi_j(0) P^{(n_1)}(j, i_1) & \text{when } n_1 > 0 \\ \pi_{i_1}(0) & \text{when } n_1 = 0. \; . \end{cases}$$

In this fashion, the initial distribution and one-step transition probabilities determine the Markov chain completely.

For any two states i and j in $\mathcal{S}$ and any $k > 0$, define $f_{ij}^{(k)}$ as follows: $f_{ij}^{(k)}$ is the probability that the first time after time 0 that the chain visits state j is time k,

given that the chain starts in state i at time 0. The $f_{ij}^{(k)}$, like the $P^{(m)}(i,j)$, are conditional probabilities. The letter f is supposed to represent the word "first:" $f_{ij}^{(k)}$ is the probability that the *first* time you see j starting from i is k steps later. This probability is the same no matter when you start from i because the chain is homogeneous. It's possible that $f_{ij}^{(k)} = 0$ for every $k > 0$, which is what happens when it is impossible to reach state j from state i by following the arrows in the Markov chain's transition diagram. If we take $i = j$, we get $f_{jj}^{(k)}$, the probability that the first time the chain *returns* to state j starting from state j is k time steps in the future.

It's easy to understand the $f_{ij}^{(k)}$ in small examples. Consider again the Markov chain in Figure 1(c). Let's figure out $f_{11}^{(k)}$ for various values of $k > 0$. It's pretty clear that $f_{11}^{(1)} = 1 - p$, since the only way the chain can start in 1 and return to 1 in one step is by following the arrow from 1 to 1. As for $k = 2$, the chain can reach 1 at time 2 by following the sequence 1, 1, 1 or the sequence 1, 2, 1. The 1, 1, 1 sequence won't do, however, since we're interested in the probability that $k = 2$ is the first time the chain returns to 1 after starting in 1. Accordingly,

$$f_{11}^{(2)} = pq \; .$$

What about $k = 3$? The possible three-step paths starting and ending in state 1 are 1, 1, 1, 1; 1, 2, 1, 1; and 1, 2, 2, 1, only the last of which has first return to state 1 at time $k = 3$. Hence

$$f_{11}^{(3)} = p(1-q)q \; .$$

For any states i and j, set

$$r_{ij} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \; .$$

r_{ij} is the probability that the chain will reach state j in finite time given that it starts in state i. r_{ij} is, again, a conditional probability. Setting $i = j$ gives r_{jj}, the probability that, starting from state j, the chain will return to state j in finite time. For the Markov chain in Figure 1(a), it's clear that $f_{11}^{(1)} = 1 - p$, since the only way for the chain to start in state 1 and return to state 1 in one time step is by following the $(1-p)$-branch from 1 to 1. It's also clear that $f_{11}^{(k)} = 0$ for every $k > 1$, since the chain can never return to state 1 at time k unless it has been in state 1 for every time $m \leq k$ — the *first* return to 1 can never be at any time $k > 1$. Accordingly, $r_{11} = 1 - p$ for the chain in Figure 1(a). Like the transition probabilities and initial distribution, the $f_{ij}^{(k)}$ and the r_{ij} non- random quantities.

One important example of a random quantity are the first hitting times T_j, defined for any $j \in \mathcal{S}$. T_j is the first positive time that the chain visits state j. If the chain happens to follow a path that never hits state j, then $T_j = \infty$. If the chain follows a path that first hits state j at time k, then $T_j = k$. Think of T_j, like X_n, as a function on the set of all possible paths. The value of T_j can be either finite or infinite depending on the path.

We'll be interested mainly in studying T_j and related quantities under the condition that the chain starts in some given state i at time 0. To that end, I'll be employing some special notation. Given any state i, any random quantity Z associated with the Markov chain, and any possible value z for Z, define

$$\text{Prob}_i\{Z = z\}$$

as the probability that $Z = z$ given that the chain started in state i at time 0. Define

$$E_i(Z)$$

as the expected value of Z given that the chain started in state i at time 0.

So, for any states i and j and any $k > 0$, $\text{Prob}_i\{T_j = k\}$ is the probability that $T_j = k$ given that the Markov chain started in state i at time 0. In other words,

$$\text{Prob}_i\{T_j = k\} = f_{ij}^{(k)} .$$

Observe also that

- If it is a probability-zero event to get to state j in finite time starting from state i — i.e., if $r_{ij} = 0$ — then $\text{Prob}_i\{T_j = \infty\} = 1$.
- If the chain starting from state i will hit state j in finite time with probability 1 — i.e., if $r_{ij} = 1$ — then $\text{Prob}_i\{T_j < \infty\} = 1$.
- If the chain starting from state i has positive probability r_{ij} of hitting state j in finite time but $r_{ij} < 1$, then $\text{Prob}_i\{T_j = \infty\} = 1 - r_{ij}$ and $\text{Prob}_i\{T_j < \infty\} = r_{ij}$.

The T_j are called *first hitting times* for obvious reasons. If we start the chain off in state j, then we call T_j the *first return time* for state j.

To get a grip on hitting times, consider the Markov Chain in Figure 1(a). Suppose the chain starts in state 1; what is T_2? The answer, of course, depends on what path the chain follows. $T_2 = 1$ if and only if the chain moves immediately to state 2 along the p-branch; this occurs with probability p. $T_2 = 2$ if and only if the chain returns once to state 1 and then moves to state 2, which means following the $(1-p)$-branch and then the p-branch, an event with probability $(1-p)p$. It's easy to extend this argument to show that

$$\text{Prob}_1\{T_2 = m\} = (1-p)^{m-1}p$$

for every $m > 0$. From these probabilities we can compute the expected value of T_2 given that the chain starts in state 1. Our notation for that quantity is $E_1(T_2)$, and

$$E_1(T_2) = \sum_{m=1}^{\infty} m \times \text{Prob}_1\{T_2 = m\} = p \sum_{m=1}^{\infty} m(1-p)^{m-1} = \frac{1}{p} .$$

The last equality holds because

$$\sum_{m=1}^{\infty} m(1-p)^{m-1} = -\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1}{1-(1-p)}\right) = \frac{1}{p^2} ,$$

where I used the standard formula for the sum of a geometric series.

Another random quantity of great interest to us is $N_j(n)$, which for any $n > 0$ and $j \in \mathcal{S}$ is the number of visits the Markov chain makes to state j during the time interval $1 \le m \le n$. $N_j(n)$ is a random quantity for every j and n since its value depends on the path the Markov chain follows. I'll use N_j to denote the total number of times the Markov chain hits state j after time 0. Note that

$$N_j = \lim_{n\to\infty} N_j(n) \text{ for all } j \in \mathcal{S} .$$

It's possible that $N_j = 0$; this happens when the chain follows a path that never hits j. It's also possible that $N_j = \infty$. If the chain starts off in state j at time 0, N_j is the number of times the chain *returns* to state j after time 0. In particular, we don't count being in state j at time 0 as contributing to $N_j(n)$ or N_j.

As with the T_j, we'll be concerned mostly with the behavior of the $N_j(n)$ and N_j given that the chain starts off in some particular state i at time 0. An important identity we'll be using is

$$E_i(N_j(n)) = \sum_{m=1}^{n} P^{(m)}(i,j) \ . \ \text{for all} \ \ i,j \in \mathcal{S} \ . \tag{1}$$

To understand equation (1), regard $N_j(n)$ as the sum of a bunch of zero-one random variables $Z_j^{(m)}$ defined by

$$Z_j^{(m)} = \begin{cases} 1 & \text{if } X_m = j \\ 0 & \text{if } X_m \neq j \ , \end{cases}$$

where X_m is the state of the Markov chain at time m. Consider now the expected value of $Z_j^{(m)}$ given that the chain starts in state i at time 0. For each $m > 0$,

$$E_i(Z_j^{(m)}) = 1 \times P^{(m)}(i,j) + 0 \times (1 - P^{(m)}(i,j)) = P^{(m)}(i,j) \ ,$$

from which (1) follows because $N_j(n) = \sum_{m=1}^{n} Z_j^{(m)}$. Setting $i = j$ yields the following special case of (1):

$$E_j(N_j(n)) = \sum_{m=1}^{n} P^{(m)}(j,j) \ \ \text{for all} \ \ j \in \mathcal{S} \ . \tag{2}$$

Taking the limit of (1) as $n \to \infty$ gives

$$E_i(N_j) = \sum_{m=1}^{\infty} P^{(m)}(i,j) \ \ \text{for all} \ \ i,j \in \mathcal{S} \ . \tag{3}$$

Let's compute $E_i(N_j)$ for a couple of the Markov chains in Figure 1. Just to remind you, the notation $\text{Prob}_i\{N_j = k\}$ denotes the probability that $N_j = k$ given that the chain starts off in state i at time 0. It's clear that, in Figure 1(b),

$$\text{Prob}_3\{N_3 = \infty\} = 1$$

since there's no escape from state 3. Now consider starting the chain in Figure 1(a) in state 1. It might stay in state 1 for a while, but once it has left state 1 it can't return there. Theoretically, it could stay in state 1 forever. Such an eventuality would require that it follow the $1 - p$ arrow infinitely many times. Following that arrow m consecutive times has probability $(1-p)^m$; as $m \to \infty$, that probability goes to zero. Thus we can conclude informally (note: I could make the argument more rigorous) that if the chain starts in state 1, it will leave state 1 eventually (and therefore never return) with probability 1. Alternatively,

$$\text{Prob}_1\{N_1 < \infty\} = 1$$

and

$$\text{Prob}_1\{N_1 = \infty\} = 0 \ .$$

What about probabilities for the various finite values of N_1 in Figure 1(a) given that we start the chain in state 1? $N_1 = 0$ when the chain transitions immediately from state 1 to state 2, which happens with probability p. $N_1 = 1$ if the chain stays in state 1 for one step and then moves to 2; this happens with probability $(1-p)p$. The analysis is similar to that for T_2, and it leads to

$$\text{Prob}_1\{N_1 = k\} = (1-p)^k p$$

for every $k \geq 0$. Accordingly,

$$E_1(N_1) = \sum_{k=0}^{\infty} k(1-p)^k p = \frac{1-p}{p} \ .$$

Again, $E_1(N_1)$ is the expected value of N_1 given that the chain starts off in state 1 at time 0.

Definition 1: A state j is *recurrent* if and only if $r_{jj} = 1$. A state j is *transient* if and only if $r_{jj} < 1$. □

In other words, j is a recurrent state if and only the Markov chain starting in state j will return to j in finite time with probability 1. A state j is transient if and only there's positive probability that the chain, having started in j, will never return to j. Consider the Markov chain in Figure 1(a). It's clear that state 1 is transient because, with probability p, the chain starting in state 1 will leave 1 immediately (and therefore never return). In fact, we determined already that $r_{11} = 1 - p < 1$, which makes state 1 transient in the light of Definition 1. State 3 in the chain in Figure 1(b) is trivially recurrent, since there's no escape from it, so $r_{33} = 1$. What about states 1 and 2 in Figure 1(b)? If the chain starts in 2 it has positive probability of moving to the no-escape state 3 and therefore never returning to 2. Similarly, the chain starting in state 1 has positive probability of making a transition from 1 to 2 and then to 3. Accordingly, the chain starting in state 1 has positive probability of never returning to state 1, and likewise for state 2, which means both $r_{11} < 1$ and $r_{22} < 1$, so states 1 and 2 are transient.

Let S_T be the set of transient states and S_R be the set of recurrent states. A first decomposition of the state space S of the Markov chain is

$$S = S_T \cup S_R \ .$$

The following result summarizes and extends the discussion thus far.

Theorem 1: State j is recurrent if and only if

$$\text{Prob}_j\{N_j = \infty\} = 1 \ .$$

If j is recurrent, then for any state i

$$E_i(N_j) = \begin{cases} 0 & \text{if } r_{ij} = 0 \\ \infty & \text{if } r_{ij} > 0 \ . \end{cases}$$

State j is transient if and only if

$$\text{Prob}_j\{N_j < \infty\} = 1 \ ,$$

If j is transient, then for any state i

$$E_i(N_j) = \frac{r_{ij}}{1 - r_{jj}} \ .$$

Proof: N_j has positive probability of being finite if and only if the chain has a positive probability of leaving j and never coming back to j, i.e. if and only if

j is transient. Thus j is recurrent if and only if $\text{Prob}_j\{N_j = \infty\} = 1$. It follows that j is transient if and only if $\text{Prob}_j\{N_j < \infty\}$ is positive; in fact, we'll see that j is transient if and only if $N_j < \infty$ with probability 1. First let's address the expressions for $E_i(N_j)$.

Suppose j is recurrent and i is any state ($i = j$ is allowed). If $r_{ij} = 0$, the chain can never reach state j starting from state i, so $N_j = 0$ with probability 1 starting from state i, making $E_i(N_j) = 0$. If, on the other hand, $r_{ij} > 0$, the chain starting from i reaches j with positive probability and, by recurrence of j, returns to j infinitely often. Thus $N_j = \infty$ with probability r_{ij}, which means in particular that $E_i(N_j) = \infty$. If, on the other hand, j is transient, then $\text{Prob}_i\{N_j = 1\}$ is the probability that the Markov chain starting in state i first arrives in state j after some finite time k and then follows a path that never returns to j. The probability that it follows such a path after arriving in j is $1 - r_{jj}$, so

$$\text{Prob}_i\{N_j = 1\} = \sum_{k=1}^{\infty} f_{ij}^{(k)}(1 - r_{jj}) = r_{ij}(1 - r_{jj}) \; .$$

Similarly, $\text{Prob}_i\{N_j = 2\}$ is the probability that the chain, starting in i, first hits j after some finite time k, then first returns to j after some finite time l, then leaves j and never returns to j. Hence

$$\text{Prob}_i\{N_j = 2\} = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} f_{ij}^{(k)} f_{jj}^{(l)}(1 - r_{jj}) = r_{ij} r_{jj}(1 - r_{jj}) \; .$$

It's not hard to carry on with this reasoning and show that

$$\text{Prob}_i\{N_j = n\} = r_{ij} r_{jj}^{n-1}(1 - r_{jj})$$

for every $n > 0$. Taking $i = j$ reveals that

$$\text{Prob}_j\{N_j < \infty\} = (1 - r_{jj}) \sum_{n=0}^{\infty} r_{jj}^n = 1 \; .$$

Thus if j is transient, then $\text{Prob}_j\{N_j < \infty\} = 1$. Conversely, recall that that if N_j is finite with positive probability, then j is transient; in particular, if $N_j < \infty$ with probability 1, then j is transient. Finally, for arbitrary i,

$$E_i(N_j) = \sum_{n=1}^{\infty} n r_{ij} r_{jj}^{n-1}(1 - r_{jj}) = \frac{r_{ij}}{1 - r_{jj}} \; ,$$

when j is transient, where the last equality follows from geometric-series reasoning. □

For any two states i and j, let's use $i \to j$ to mean that the chain, starting in state i, has positive probability of hitting state j in finite time after leaving state i. That $i \to j$ if and only if $f_{ij}^{(k)} > 0$ for some $k > 0$, so $i \to j$ if and only if $r_{ij} > 0$. Another condition equivalent to $i \to j$ is that $P^{(m)}(i, j) > 0$ for some $m > 0$. Observe also that if $i \to j$ and $j \to k$, then $i \to k$.

Fact 1: If i is recurrent and $i \to j$, then j is also recurrent and $j \to i$. Furthermore, $r_{ij} = r_{ji} = 1$.

Proof: Since i is recurrent, the chain starting in state i will return to state i with probability 1. $i \to j$ means that, with positive probability, the chain starting in state i will hit state j in finite time. If it were not the case that $j \to i$, this would mean that, with positive probability, the chain could leave i and never return (all it would have to do is go to j). It follows that $j \to i$. Furthermore, $r_{ij} = r_{ji} = 1$. To see this, note that if $r_{ji} < 1$, then the chain could do the following with positive probability $1 - r_{ji}$: start in state i, reach state j in finite time, and then follow a path that never returns to state i. That would contradict recurrence of i. A similar argument proves that $r_{ij} = 1$ if we can show that j is recurrent.

To do that, I'll apply identity (3). Since $r_{ji} = 1$ and $r_{ij} > 0$, we can find k and l so that $P^{(k)}(j,i) > 0$ and $P^{(l)}(i,j) > 0$. For any $m > 0$, it's clear that

$$P^{(m+k+l)}(j,j) \geq P^{(k)}(j,i)P^{(m)}(i,i)P^{(l)}(i,j) \; .$$

The idea is that the right-hand side gives the probability of making a j-to-i transition in exactly k steps, followed by an i-to-i transition of exactly m steps, followed by an i-to-j transition of exactly l steps, and that's not necessarily the only way of making a j-to-j transition in exactly $m+k+l$ steps. Now sum over m and use (3) to obtain

$$E_j(N_j) \geq \sum_{m=1}^{\infty} P^{(m+k+l)}(j,j) = P^{(k)}(j,i)E_i(N_i)P^{(l)}(i,j) \; .$$

Since i is recurrent, $E_i(N_i) = \infty$ by Theorem 1. It follows that $E_j(N_j) = \infty$. If j were transient, then $E_j(N_j)$ would equal the finite value $r_{jj}/(1 - r_{jj})$ by Theorem 1, so j must be recurrent. and, again by Theorem 1, j is recurrent. □

3. Recurrence classes and limits of expected time averages

Thus the state space of a Markov chain decomposes into transient and recurrent states. We've learned so far that regardless of the initial distribution, the Markov chain will never visit a transient state infinitely often and that if the chain ever visits a recurrent state j, it will revisit j infinitely often. It's time now to take a closer look at the Markov chain's dynamics.

Definition 2: A set of states C is *closed* if and only if for every $i \in C$, $r_{ik} = 0$ when $k \notin C$. A closed set C of states is said to be *irreducible* if and only if $i \to j$ for every i and j in C. If the entire state space $\mathcal{S}$ is irreducible, we say that the Markov chain is irreducible.

The chain can't escape from a closed set C of states once it is in C. In particular, if C is a closed set of states, then for every $i \in C$, $P(i,k) = 0$ when $k \notin C$. As a consequence,

$$\sum_{j \in C} P(i,j) = \sum_{j \in \mathcal{S}} P(i,j) = 1$$

for every $i \in C$. You can think of any closed set of states C as constituting its own private little Markov chain with state space C and transition probabilities $\{P(i,j) : i,j \in C\}$.

Definition 3: For any recurrent state i, define the *recurrence class* of i as the set of all $j \in S$ such that $i \to j$.

If i is a state and C is the recurrence class of i, then C is closed. That's because if $j \in C$ and $j \to k$, then $i \to k$ because $i \to j$, so $k \in C$ as well. Furthermore, C is irreducible. That's because if j and k are in C, then $i \to j$ and $i \to k$, so $k \to i$ and $j \to i$ by Fact 1, from which it follows that $j \to k$ and $k \to j$.

Two different recurrence classes cannot intersect. Alternatively, if two recurrence classes intersect then they coincide. To see this, suppose that C_i is the recurrence class of state i and C_j is the recurrence class of state j. If $k \in C_i \cap C_j$, then $i \to k$ and $j \to k$ by definition of recurrence class. Furthermore, $k \to i$ and $k \to j$ by Fact 1. Thus $i \to j$ and $j \to i$, so C_i and C_j must be the same set of states.

Putting these ideas together leads to a refinement of our original decomposition of the Markov chain's state space S. The set S_R of recurrent states splits into the disjoint union of one or more recurrence classes. Each of these is a closed, irreducible set of states.

Consider the Markov chains in Figure 1. The chain in Figure 1(a) has $S_T = \{1\}$ and $S_R = \{2,3\}$. Since $2 \to 3$ and $3 \to 2$, $\{2,3\}$ is the one and only recurrence class for the chain. The chain in Figure 1(b) has $S_T = \{1,2\}$ and $S_R = \{3\}$, so state 3 is the sole member of its own recurrence class. The chain in Figure 1(c), provided p and q are both positive, has one recurrence class $S_R = \{1,2\}$, and $S_T = \phi$. The chain in Figure 1(d) has $S_T = \{1\}$, and its recurrent states divide up into the two recurrence classes $\{2,3\}$ and $\{4,5\}$. The Markov chain in Figure 1(e) is somewhat different because it has no recurrent states. For each j, state j is transient because the chain exits j in one step with probability $1/2$, and then never returns to j, and $r_{jj} = 1/2$.

The chain in Figure 1(f), on the other hand, has either only recurrent states or only transient states depending on the values of the $P(i,i+1)$. To see why, suppose first that $P(i,i+1) = i/i+1$ for all i. In that case, state 1 is recurrent because the only way the chain can start in 1 and never return to 1 is to march out through all the the states along the right-facing arrows, and following such a path has probability

$$\lim_{k\to\infty} (1/2)(2/3)(3/4)\cdots((k-1)/k) = \lim_{k\to\infty} \frac{1}{k} = 0 \, .$$

By Fact 1, since $1 \to j$ for every state j, all states are recurrent and belong to the same recurrence class. On the other hand, suppose that

$$P(i,i+1) = 2^{-\frac{1}{2^i}} \, .$$

In that case, the probability of following all right-facing arrows is

$$2^{-(1/2+1/4+1/8+\cdots)} = \frac{1}{2} \, ,$$

So $r_{11} = 1/2$ and state 1 is transient. By Fact 1, every other state j must also be transient since $j \to 1$ for all j.

We have almost enough machinery to prove a couple of theorems about the limiting behavior of Markov chains. The theorems address the following question: what fraction of the time do we expect the Markov chain, over the long haul, to spend in each state j? The answer depends on the value of $E_j(T_j)$, the so-called *mean first return time* for state j. Since

$$\text{Prob}_j\{T_j = k\} = f_{jj}^{(k)} \text{ for all } k > 0 ,$$

we have

$$E_j(T_j) = \sum_{k=1}^{\infty} k f_{jj}^{(k)}$$

for every state j. Standard notation for $E_j(T_j)$ is m_j. If j is a transient state, then by definition

$$\text{Prob}_j\{T_j = \infty\} = 1 - r_{jj} > 0 ,$$

so $m_j = \infty$ for every transient state j. If j is recurrent, it is still possible that $m_j = \infty$.

Definition 3: A recurrent state j is *positively recurrent* if and only if $m_j < \infty$ and *null-recurrent* if and only if $m_j = \infty$.

Null-recurrent states might seem a bit counter-intuitive. If state j is recurrent, you know that the Markov chain, starting from state j, will return to state j in finite time with probability 1. How could the average length of time you have to wait for such a return be infinite? Examining the infinite-series expression above for $E_j(T_j)$ reveals that $m_j = \infty$ occurs, roughly speaking, when $f_{jj}^{(k)}$ doesn't go to zero fast enough as $k \to \infty$. In other words, if with significant probability you have to wait a very long time to return to state j, then it's possible that, on average, you'll wait essentially forever.

As it happens, the chain in Figure 1(f) with transition probabilities $P(i, i+1) = i/i+1$ has all null-recurrent states. For example, consider state 1. Observe that

$$f_{11}^{(k)} = \frac{1}{k(k+1)} ,$$

so

$$m_1 = E_1(T_1) = \sum_{k=1}^{\infty} k f_{11}^{(k)} = \sum_{k=1}^{\infty} \frac{1}{k+1} = \infty .$$

So state 1 is null-recurrent. The fact that all the other states are null-recurrent is a consequence of the following result, which asserts that no recurrence class can contain a mixture of positively recurrent states and null-recurrent states.

Fact 2: For any recurrence class C, either every $j \in C$ is positively recurrent or every $j \in C$ is null-recurrent.

Proof: Suppose i and j are in the same recurrence class C. Then since $i \to j$ and $j \to i$, there exist l and m such that $P^{(l)}(i,j) > 0$ and $P^{(m)}(j,i) > 0$. For any $k > 0$, the probability that the chain, starting in state i, first returns to state i after $k+l+m$ time steps is bounded above by the probability that the state makes a transition from i to j in l steps, then first returns to j after k more steps, then makes a transition from j to i in m steps. That is, for every $k > 0$,

$$f_{ii}^{(k+l+m)} \leq P^{(l)}(i,j) f_{jj}^{(k)} P^{(m)}(j,i) \ .$$

Thus

$$\sum_{k=1}^{\infty} k f_{ii}^{(k+l+m)} \leq P^{(l)}(i,j) P^{(m)}(j,i) \sum_{k=1}^{\infty} k f_{jj}^{(k)} = P^{(l)}(i,j) P^{(m)}(j,i) m_j \ .$$

If $m_j < \infty$ — that is, if j is positively recurrent — then the right-hand side is finite, so the sum on the left-hand side converges.

Now, convergence of the sum on the left-hand side is equivalent to $m_i < \infty$. You can reason as follows: $m_i < \infty$ if and only if $\sum_{q=1}^{\infty} q f_{ii}^{(q)}$ converges, which happens if and only if $\sum_{q=l+m+1}^{\infty} q f_{ii}^{(q)}$ converges, which happens if and only if

$$\sum_{q=l+m+1}^{\infty} (q-l-m) f_{ii}^{(q)}$$

converges. The last equivalence holds because the $\sum_{q=1}^{\infty} f_{ii}^{(q)} = 1$. Setting $k = q-l-m$ reveals that $m_i < \infty$ if and only if the sum on the left-hand side converges.

The bottom line is that if $m_j < \infty$ then $m_i < \infty$. The argument is symmetric in i and j, which means that $m_i < \infty$ if and only if $m_j < \infty$. In other words, i and j are either both positively recurrent or both null-recurrent. Since i and j were arbitrary states in the recurrence class C, it follows that either every state in C is positively recurrent or every state in C is null-recurrent. □

It's time now to sneak up on the first convergence theorem. Roughly speaking, the theorem says that the long-term expected fraction of the time a Markov chain spends in a state j is zero if j is transient and inversely proportional to m_j if j is recurrent. This makes intuitive sense. Transient states are those that the chain stops visiting eventually, so, in the limit, the fraction of the time the chain spends visiting them will approach zero. Recurrent states are those that the chain visits infinitely often. How frequently the chain visits a given recurrent state j ought to depend on how long it takes, on the average, to return to j starting from j. In particular, the longer the average wait, the less frequent the visits ought to be.

To reason more quantitatively, suppose that $m_j = 17$ for some recurrent state j. Thus we "expect" the whenever the chain visits state j it will first re-visit j 17 time steps later on the average. This means that the chain will spend 1/17th of its time in state j, at least on the average. For any state j, recall that $N_j(n)$ denotes the number of visits the chain makes to state j during the time interval $1 \leq m \leq n$. Here is the central result. I'll relegate the proof to the Appendix since it's somewhat technical and doesn't contribute much to the exposition. Note that the expressions in the theorem statement for expected time-averages in terms of the $P^{(m)}(j,j)$ follow from equation (2).

Theorem 2: With notation as in the foregoing,

- If j is a transient state, then

$$\lim_{n\to\infty} \frac{E_j(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) = 0 \; .$$

- If j is a recurrent state, then

$$\lim_{n\to\infty} \frac{E_j(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) = \frac{1}{m_j} \; .$$

Theorem 2 addresses the convergence of a bunch of expected values, which are non-random quantities that depend only on the parameters of the Markov chain. One of the reasons I've decided not to include the proof here is that, in my view, it neither appeals to nor offers enhancement of your Markov-chain intuition. One particular consequence of Theorem 2 is worth mentioning. If j is null-recurrent, then the long-term expected fraction of the time the chain spends in state j is zero, just as it would have been if j were transient. That's because $m_j = \infty$ when j is null-recurrent. If you have to wait an infinite amount of time, on the average, to return to state j once you've landed there, then you don't spend much time in state j on the average.

Theorem 2 tells us what fraction of the time, on the average, a Markov chain spends revisiting a state j given that it starts in state j at time 0. A similar result holds if the chain starts in an arbitrary state i. The proof appears in the Appendix. Note that equation (1) accounts for the expressions in terms of the $P^{(m)}(i,j)$ for expected time-averages.

Theorem 3: With notation as in the foregoing,

- If j is a transient state, then for any state i

$$\lim_{n\to\infty} \frac{E_i(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = 0 \; .$$

- If j is a recurrent state, then for any state i

$$\lim_{n\to\infty} \frac{E_i(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = \frac{r_{ij}}{m_j} \; .$$

Observe that Theorem 2 is a special case of Theorem 3: set $i = j$ in Theorem 3 and you get Theorem 2. A simple corollary of Theorem 3 is that a finite-state Markov chain has no null-recurrent states and at least one recurrent state.

Corollary to Theorem 3: If the number of states in a Markov chain is finite, then the chain has at least one recurrent state and no null-recurrent states.

Proof: Suppose the chain has state space $\{1, 2, \ldots, M\}$. Start the chain off in any state i. Notice that, for any $n > 0$,

$$\sum_{j=1}^{M} N_j(n) = n \ .$$

This is because the Markov chain has to visit some state at each time instant. Now divide by n and take expected values to obtain

$$\sum_{j=1}^{M} \frac{E_i(N_j(n))}{n} = 1$$

for every $n > 0$. If every state were transient, then all the terms in the sum would go to zero as $n \to \infty$, which is impossible since they sum to 1 for every n. Hence at least one of the j's must be recurrent.

Now suppose some i is a null-recurrent state and let C be the recurrence class of i. Suppose C contains K states. All states in C must be null-recurrent by Fact 2. Start the chain off in state i. Since C is closed, the chain can't escape C, so for every $n > 0$ we have

$$\sum_{j \in C} N_j(n) = n$$

and therefore

$$\sum_{j \in C} \frac{E_i(N_j(n))}{n} = 1$$

The sum contains K terms, all of which go to zero as $n \to \infty$ by Theorem 3 because every $j \in C$ is null-recurrent. This is impossible because they must add up to 1 for every n, and thus we have a contradiction of our original supposition that i is null-recurrent. □

The argument in the proof of the Corollary also shows that in any Markov chain, a recurrence class consisting of null-recurrent states must be an infinite set of states. It also shows that any closed set of transient states in any Markov chain must be an infinite set of states. The chain in Figure 1(e) is an example of a chain with a (necessarily infinite) closed set of transient states, as is the chain in Figure 1(f) with transition probabilities $P(i, i+1) = 2^{-1/2^i}$. The chain in Figure 1(f) with transition probabilities $P(i, i+1) = i/i + 1$ has a (necessarily infinite) recurrence class consisting of null-recurrent states.

The qualitative implications of Theorems 2 and 3 are clear: overall, one expects to spend a negligible fraction of the time, over the long haul, visiting transient or null-recurrent states. On the other hand, over the medium term, transient and null-recurrent states might be hugely important. Suppose, for example, that p in Figure 1(a) is small. We calculated earlier that $E_1(T_2) = 1/p$. In particular, the expected time it takes to exit the transient state 1 given that the chain starts there will be 10,000 time steps if $p = .0001$.

Consider now the Markov chain in Figure 1(c). As we noted earlier, the chain is irreducible, and both states are positively recurrent. It's not hard to see that

$$f_{11}^{(k)} = \begin{cases} 1-p & \text{when } k=1 \\ pq(1-q)^{k-2} & \text{when } k>1 \end{cases}$$

Thus

$$\begin{aligned} m_1 &= E_1(T_1) \\ &= \sum_{k=1}^{\infty} k f_{11}^{(k)} \\ &= (1-p) + pq \sum_{k=2}^{\infty} k(1-q)^{k-2} \\ &= 1 + \frac{p}{q}\,, \end{aligned}$$

where the last equality follows from a geometric-series argument. Similarly, $m_2 = 1 + q/p$. These expressions make sense. If p is large relative to q, it's easy for the Markov chain to escape from state 1 and hard for it to return. The opposite holds when q is large relative to p. Theorem 3 implies, since $r_{12} = r_{21} = 1$, that

$$\lim_{n\to\infty} \frac{E_1(N_1)}{n} = \lim_{n\to\infty} \frac{E_2(N_1)}{n} = \frac{q}{p+q}$$

and

$$\lim_{n\to\infty} \frac{E_2(N_2)}{n} = \lim_{n\to\infty} \frac{E_1(N_2)}{n} = \frac{p}{p+q}\,,$$

confirming our intuition that the chain spends more time on average in state 1 when q is large relative to p and more time on average in state 2 when p is large relative to q.

Theorems 2 and 3 are relatively weak theorems in the sense that they make statements about the convergence of expected values rather than the "probability-1" convergence of random quantities. As it happens, Theorem 3 has just such a stronger version whose proof depends on the Strong Law of Large Numbers. Weak though they may be, Theorems 2 and 3 provide what's needed to characterize stationary distributions for Markov chains, which are the subject of the next section.

4. Stationary distributions and the marble-bins model

The formal definition of a Markov chain endows the chain with an initial distribution $\pi(0)$. In the framework of the lightwall analogy, $\pi(0)$ arises from the supervisor's initial spin deciding which operator is to flash his light first. The initial distribution has played essentially no role in the exposition so far. We've been concerned mainly with answering questions about what happens over time *given* that the Markov chain starts in a particular state i at time 0. To understand the role of the initial distribution, it pays to keep in mind a different but equivalent analogy for a Markov chain.

Imagine a large set $\mathcal{S}$ of bins and an even larger set of marbles (or pennies, or grains of rice or sand, or poker chips, or whatever). $\mathcal{S}$ is the state space of the Markov chain, and S is either finite or countably infinite. At time 0, a supervisor

allocates all the marbles among the various bins according to $\pi(0)$. If $\pi_i(0) = \alpha$, a fraction α of the marbles goes into bin i at time 0. Each bin has a caretaker. The caretaker's job is to re-allocate the marbles in his bin, once every time step. After the supervisor has distributed the marbles according to $\pi(0)$, the caretaker for bin i divides up the marbles in bin i according to the probability distribution $\{P(i,j) : j \in \mathcal{S}\}$.

It helps to think of each caretaker as having some kind of workbench off to the side on which he can divide up his marbles. After all the caretakers have parsed their marbles, they simultaneously allocate the various piles to the bins for which they're destined. For example, if $P(i,1) = 1/6$, $P(i,2) = 2/3$, $P(i,i) = 1/6$, and $P(i,j) = 0$ for all other j, the caretaker of bin i will end up keeping one sixth of his marbles, depositing one sixth of them in bin 1, depositing 2/3 of them in bin 2, and not putting any of them in any other bins. After all the deposits are in place, the new allocation of marbles among the bins is $\pi(1)$ — for all i, $\pi_i(1)$ is the fraction of all the marbles in bin i after the first re-allocation has taken place.

This process goes on. The caretakers do the same kind of re-allocation at every time step, leading to distributions $\pi(2)$, $\pi(3)$, and so on. For each m, $\pi(m)$ describes the allocation of the marbles among the bins in the sense that $\pi_i(m)$, for each $i \in \mathcal{S}$, is the fraction the marbles that bin i contains at time m (i.e., after m re-allocations). The distributions $\pi(m)$ follow a simple recursion in m. To figure out $\pi_j(m+1)$, remember that marbles end up in bin j at time $m+1$ due to re-allocations by the caretakers of all the bins. The caretaker of bin i allocates to bin j a fraction $P(i,j)$ of the marbles in bin i at time m. At time m, bin i contains a fraction $\pi_i(m)$ of all the marbles, so the fraction of all the marbles that the i-caretaker re-allocates to bin j at time $m+1$ is $\pi_i(m)P(i,j)$. As a result, the total fraction of all the marbles that sit in bin j at time $m+1$ is the sum over all i of these individual contributions, i.e.,

$$\pi_j(m+1) = \sum_{i \in \mathcal{S}} \pi_i(m) P(i,j) \tag{4}$$

for every $j \in \mathcal{S}$ and every $m > 0$. From (4) it follows easily by induction that

$$\pi_j(m) = \sum_{i \in \mathcal{S}} \pi_i(0) P^{(m)}(i,j)$$

for every $j \in \mathcal{S}$ and $m > 0$.

On the face of it, probability might seem to be absent from this picture. After all, $\pi(0)$ is just a vector of numbers and the $P(i,j)$ are just numbers as well. The caretakers appear to re-allocate in a purely deterministic manner — no coin flips, no spinner spins. How could the marble-bins model capture the evident randomness of the lightwall model?

To reconcile the two models, suppose that exactly one of the marbles is green and all the rest are black. At time 0, the supervisor allocates the marbles among the bins according to $\pi(0)$. We don't know which bin the green marble will end up in, but, at least intuitively, the probability that it will end up in bin i is $\pi_i(0)$, since that's the fraction of marbles allocated to bin i at time 0. As time goes on, the green marble will hop around from bin to bin as the marbles get re-allocated by the caretakers. It's helpful to think of the supervisor and caretakers as colorblind so that none of them knows where the green marble is at any time.

The bin containing the green marble at time m represents the state of the Markov chain at time m. In this way, the path the green marble follows through the bins represents a run of the Markov chain. We can phrase questions about Markov-chain runs in terms of the green marble's path. For example, "What is the first positive time that the the Markov chain is in state j given that it started in state i?" translates to, "Given that the green marble was in bin i at time 0, what is the first time after that when it's in bin j?"

At any time $m \geq 0$, let X_m be the bin the green marble is in at time m. For each $i \in \mathcal{S}$ and $m \geq 0$, the probability that $X_m = i$ is $\pi_i(m)$. The probability that $X_1 = j$ given that $X_0 = i$ is $P(i, j)$, so the probability that $X_1 = j$ is

$$\sum_{i \in \mathcal{S}} \pi_i(0) P(i, j) = \pi_j(1)$$

for every $j \in \mathcal{S}$. In other words, probability that the green marble is in bin j at time 1 is precisely the fraction of marbles in bin j after the first re-allocation. This is not surprising since the green marble is just one marble among many and is anonymous as far as the supervisor and caretakers are concerned, so the probability that it's sitting in some bin at any given time is simply the total fraction of marbles sitting in that bin at that time.

I don't want to push the analogies too far, but I think it's useful to keep both models in mind. The lightwall model is dynamic and its random nature sits center-stage. The marbles-bins model is somewhat less dynamic and less evidently "random." I might add that some folks prefer jars full of water to bins full of marbles. Although such an aesthetic choice is convenient when the transition probabilities $P(i, j)$ aren't rational numbers, the thought of following one specially marked water molecule from jar to jar gives me a quantum headache.

A stationary distribution for a Markov chain is an allocation of marbles to bins that doesn't change over time. The caretakers re-allocate and redistribute the marbles, but the fraction of all the marbles sitting in each bin after each such re-allocation is the same as the fraction sitting in that bin beforehand.

Definition 4: A vector $\overline{\pi} = (\overline{\pi}_i)_{i \in \mathcal{S}}$ is a *stationary distribution* for the Markov chain with transition probabilities $P(i, j)$ if and only if $\overline{\pi}_i \geq 0$ for every $i \in \mathcal{S}$, $\sum_{i \in \mathcal{S}} \overline{\pi}_i = 1$, and

$$\overline{\pi}_j = \sum_{i \in \mathcal{S}} \overline{\pi}_i P(i, j) \tag{5}$$

for every $j \in \mathcal{S}$.

If $\overline{\pi}$ is a stationary distribution for a Markov chain in the sense of Definition 4, and we set $\pi(0) = \overline{\pi}$, then clearly $\pi(m) = \overline{\pi}$ for every $m > 0$ because of recursion (4). So Definition 4 captures the notion of a marble distribution that persists after re-allocation. An easy induction (compare (4)) shows that

$$\overline{\pi}_j = \sum_{i \in \mathcal{S}} \overline{\pi}_i P^{(m)}(i, j)$$

for every $m > 0$ and every stationary distribution $\overline{\pi}$.

A Markov chain can have many stationary distributions, or exactly one, or none at all. Since a stationary distribution $\overline{\pi}$ is a fixed point for the recursion $\pi(m) \mapsto \pi(m+1)$, we'll want to determine under what circumstances $\overline{\pi}$ is "attracting" in the sense that $\pi(m) \to \overline{\pi}$ as $m \to \infty$ for a wide range of initial distributions $\pi(0)$. We'll also be interested in studying relationships between stationary distributions and the long-term behavior of runs of the Markov chain. These relationships constitute the nexus between the marbles-bins and lightwall models.

Think again about the meandering green marble. Suppose we initialize the Markov chain with a stationary distribution $\overline{\pi}$, assuming one exists. Then $\pi(m) = \overline{\pi}$ for every $m > 0$, and the probability that we'll find the green marble in bin j at time m, namely $\overline{\pi}_j$, doesn't change over time. Intuitively, one might expect this unchanging instantaneous probability of finding the green marble in bin j to correlate with the limiting frequency of visits that the green marble makes to bin j over time. If j is a transient or null-recurrent state, the limiting expected value of that frequency is zero. That intuition is sound, as the following result reveals.

Theorem 4: If $\overline{\pi}$ is a stationary distribution, then $\overline{\pi}_j = 0$ if j is a transient or null-recurrent state.

Proof: Since $\overline{\pi}$ is a stationary distribution,

$$\overline{\pi}_j = \sum_{i \in \mathcal{S}} \overline{\pi}_i P^{(m)}(i,j)$$

for every $m > 0$. Averaging over m yields

$$\overline{\pi}_j = \sum_{i \in \mathcal{S}} \overline{\pi}_i \left[\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) \right]$$

for every $n > 0$. Since $\sum_{i \in \mathcal{S}} \overline{\pi}_i = 1$, there exists for every $\delta > 0$ a finite set $\mathcal{S}_\delta$ of states such that $\sum_{i \notin \mathcal{S}_\delta} \overline{\pi}_i < \delta$. Given δ and such a $\mathcal{S}_\delta$, we have, since the bracketed terms in the last equation are bounded above by 1,

$$\overline{\pi}_j \leq \sum_{i \in \mathcal{S}_\delta} \overline{\pi}_i \left[\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) \right] + \delta \ .$$

The sum over $\mathcal{S}_\delta$ has finitely many terms, all of which, by Theorem 3, approach 0 as $n \to \infty$ if j is transient or null-recurrent. Taking the limit as $n \to \infty$ yields

$$\overline{\pi}_j \leq \delta$$

when j is transient or null-recurrent. Since δ was arbitrary, we conclude that $\overline{\pi}_j \leq \delta$ for all $\delta > 0$ – and therefore $\overline{\pi}_j = 0$ — when j is transient or null-recurrent. □

Theorem 4 asserts that any stationary distribution $\overline{\pi}$ for a Markov chain must be concentrated on the set of positively recurrent states in the sense that $\overline{\pi}_j > 0$ only if j is positively recurrent. So a Markov chain lacking positively recurrent states has no stationary distributions. The all-transient chain in Figure 1(e) and the null-recurrent version of the chain in Figure 1(f) are examples of chains that don't have stationary distributions.

Theorem 4, although it validates our intuition about the green marble, has a bit of a negative feel to it. On a more optimistic note, it turns out that any Markov chain with at least one positively recurrent state has at least one stationary distribution. The results we'll encounter presently are quite a bit sharper, but the existence of stationary distributions for "most" Markov chains is one of their consequences.

Consider, then, a Markov chain that has at least one positively recurrent state and therefore, by Fact 2, at least one recurrence class C consisting solely of positively recurrent states. In what follows, I'll refer to such a recurrence class as a *positively recurrent class.* Suppose i and j are in C. Since i and j are in the same recurrence class, $r_{ij} = 1$ by Fact 1, so, by Theorem 3,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = \lim_{n\to\infty} \frac{E_i(N_j(n))}{n} = \frac{1}{m_j} \ , \tag{6}$$

where $m_j = E_j(T_j)$ is the mean return time to state j. I'll use q_j to denote $1/m_j$ in the proofs of Theorem 5 below and its accompanying Lemma. The Lemma is easiest to understand when the state space is finite, so I'll leave for the Appendix the complete proof and present the proof for finite $\mathcal{S}$ here.

Lemma: Suppose a Markov chain has a positively recurrent class C. Define a distribution $\overline{\pi}$ for the Markov chain by setting

$$\overline{\pi}_j = \begin{cases} 1/m_j = q_j & \text{if } j \in C \\ 0 & \text{if} j \notin C \ , \end{cases}$$

Then $\overline{\pi}$ is a stationary distribution for the Markov chain. Furthermore, $\overline{\pi}$ is concentrated on the recurrence class C since $\overline{\pi}_j = 0$ for $j \notin C$.

Proof: Again, I'm assuming for the moment that the state space is finite, so C is finite, as well. Since $q_j > 0$ for every $j \in C$, all the elements in $\overline{\pi}$ are nonnegative. What we need to show is that they sum to 1 and satisfy the stationarity condition (5) in Definition 4. That the q_j sum to 1 follows from an argument embedded in the proof of the Corollary to Theorem 3. Since

$$\sum_{j\in\mathcal{S}} N_j(n) = n \ \text{ for all } \ n > 0 \ ,$$

we conclude that

$$\sum_{j\in\mathcal{S}} \frac{E_i(N_j(n))}{n} = 1 \ \text{ for all } \ n > 0 \ ,$$

and taking the limit as $n \to \infty$ yields

$$\sum_{j\in\mathcal{S}} \frac{1}{m_j} = \sum_{j\in\mathcal{S}} q_j = 1 \ .$$

Interchanging the limit with the sum is okay because $\mathcal{S}$ is finite.

Since C is a closed set of states, $P(i,j) = 0$ when $i \in C$ and $j \notin C$, so $\sum_{j\in C} P(i,j) = 1$ for every $i \in C$. Furthermore, when $j \notin C$,

$$\sum_{i\in\mathcal{S}} \overline{\pi}_i P(i,j) = \sum_{i\in C} q_i P(i,j) = 0 = \overline{\pi}_j \ .$$

So to prove that $\overline{\pi}$ satisfies (5), since $\overline{\pi}_j = q_j$ for $j \in C$, we need to show only that

$$q_j = \sum_{i \in C} q_i P(i,j) \text{ for all } j \in C \,. \tag{7}$$

Observe first that for every $m > 0$ and i, $j \in C$,

$$P^{(m+1)}(j,j) = \sum_{i \in C} P^{(m)}(j,i) P(i,j) \,.$$

Note that the sum on the right-hand side is over $i \in C$ because $P(j,i) = 0$ when $i \notin C$. Accordingly, for every $n > 0$,

$$\frac{1}{n} \sum_{m=1}^{n} P^{(m+1)}(j,j) = \sum_{i \in C} \left(\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,i) \right) P(i,j) \,.$$

The left-hand side is the same as

$$\frac{n+1}{n} \left(\frac{1}{n+1} \sum_{m=1}^{n+1} P^{(m)}(j,j) \right) - \frac{1}{n} P(j,j) \,.$$

Taking limits as $n \to \infty$ using (6) yields

$$q_j = \sum_{i \in C} q_i P(i,j) \text{ for all } j \in C \,,$$

which is what we needed to show. Again, interchanging limit with summation is okay because C is finite. □

Theorem 5: Suppose a Markov chain has at least one positively recurrent state. For each positively recurrent class C, the chain has a unique stationary distribution $\overline{\pi}^C$ concentrated on C. The distribution $\overline{\pi}^C$ is given by

$$\overline{\pi}^C_j = \begin{cases} \frac{1}{m_j} & \text{if } j \in C \\ 0 & \text{if } j \notin C \,, \end{cases}$$

where $m_j = E_j(T_j)$ is the expected return time to state j.

Proof: By the Lemma, the $\overline{\pi}^C$ in the theorem statement is a stationary distribution for the Markov chain concentrated on C. We need to make sure that it's the only one. Accordingly, let $\overline{\pi}$ be any stationary distribution concentrated on C. Since $\overline{\pi}$ is a stationary distribution, (5) holds, from which it follows easily as in the proof of the Lemma that

$$\overline{\pi}_j = \sum_{i \in \mathcal{S}} \overline{\pi}_i P^{(m)}(i,j) = \sum_{i \in C} \overline{\pi}_i P^{(m)}(i,j) \text{ for all } m > 0 \,.$$

Averaging over m yields

$$\overline{\pi}_j = \sum_{i \in C} \overline{\pi}_i \left(\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) \right) \text{ for all } n > 0 \text{ and } j \in C \,.$$

Since the $\overline{\pi}_j$ sum to 1, for any $\delta > 0$ we can find a finite subset C_δ of C such that $\sum_{i \notin C_\delta} \overline{\pi}_i < \delta$. Since the terms in the parentheses lie between 0 and 1 for every i

and j,

$$\sum_{i \in C_\delta} \overline{\pi}_i \left(\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) \right) \leq \overline{\pi}_j \leq \sum_{i \in C_\delta} \overline{\pi}_i \left(\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) \right) + \delta \; .$$

Taking the limit as $n \to \infty$ using (6) yields

$$\left(\sum_{i \in C_\delta} \overline{\pi}_i \right) \frac{1}{m_j} \leq \overline{\pi}_j \leq \left(\sum_{i \in C_\delta} \overline{\pi}_i \right) \frac{1}{m_j} + \delta \; .$$

Interchanging limit and summation is okay because C_δ is finite. Since $\sum_{i \in C} \overline{\pi}_i = 1$, $1/m_j \leq 1$, and δ is arbitrary, we can conclude that

$$\frac{1}{m_j} - \delta \leq \overline{\pi}_j \leq \frac{1}{m_j} + \delta \;\text{ for all }\; j \in C \text{ and } \delta > 0 \; ,$$

so $\overline{\pi}_j = 1/m_j$ for all $j \in C$. □

Theorem 5, in conjunction with Theorem 4 and the discussion that follows, asserts that a Markov chain that has any stationary distributions at all — i.e., a Markov chain with at least one positively recurrent state — has a unique stationary distribution concentrated on each positively recurrent class. If the chain is in addition irreducible, then the entire state space $\mathcal{S}$ constitutes a recurrence class, so the stationary distribution guaranteed by Theorem 5 is the only stationary distribution for the Markov chain. If the chain has some transient and/or null-recurrent states but has only one positively recurrent class, then, once again, the distribution from Theorem 5 is the Markov chain's only stationary distribution.

A Markov chain might, of course, have multiple positively recurrent classes. In that case, and only in that case, can the Markov chain have multiple stationary distributions. As it happens, the set of all stationary distributions has a neat characterization in terms of the distributions concentrated on individual recurrence classes.

Corollary to Theorem 5: If $\overline{\pi}$ is a stationary distribution for a Markov chain, then for any positively recurrent class C, either $\overline{\pi}_j = 0$ for every $j \in C$ or

$$\frac{\overline{\pi}_j}{\sum_{i \in C} \overline{\pi}_i} = \overline{\pi}_j^C \;\text{ for all }\; j \in C \; .$$

Proof: By Theorem 4, $\pi_i = 0$ if i is a transient or null-recurrent state. By Fact 1, if C is a positively recurrent class and i is a positively recurrent state not in C, then $P(i,j) = 0$ for every $j \in C$. Accordingly, the sum in (5) reduces to a sum over states in C:

$$\overline{\pi}_j = \sum_{i \in \mathcal{S}} \overline{\pi}_i P(i,j) = \sum_{i \in C} \overline{\pi}_i P(i,j) \tag{8}$$

for every $j \in C$. If $\overline{\pi}_j = 0$ for every $j \in C$, then we're done. If not, define another distribution π for the Markov chain by

$$\pi_j = \begin{cases} \frac{\overline{\pi}_j}{\sum_{i \in C} \overline{\pi}_i} & \text{if } \ j \in C \\ 0 & \text{if } \ j \notin C \ . \end{cases}$$

Then π is concentrated on C and

$$\sum_{i \in \mathcal{S}} \pi_i P(i,j) = \sum_{i \in C} \pi_i P(i,j) = \frac{1}{\sum_{i \in C} \overline{\pi}_i} \sum_{i \in C} \overline{\pi}_i P(i,j) \ .$$

By (8), the last sum yields $\overline{\pi}_j$ if $j \in C$. If $j \notin C$, the last sum is zero since $P(i,j) = 0$ when $j \notin C$ because C is closed. Thus

$$\sum_{i \in \mathcal{S}} \pi_i P(i,j) = \begin{cases} \frac{\pi_j}{\sum_{i \in C} \overline{\pi}_i} & \text{when } j \in C \\ 0 & \text{when } j \notin C \ , \end{cases}$$

so $\sum_{i \in \mathcal{S}} \pi_i P(i,j) = \pi_j$ for every $j \in \mathcal{S}$, and π is therefore a stationary distribution for the Markov chain concentrated on C. By Theorem 5, $\pi_j = \overline{\pi}_j^C$ for every $j \in C$, which is precisely what the Corollary asserts. □

One can re-cast the Corollary to Theorem 5 as a recipe for generating all the stationary distributions for a Markov chain. Consider a chain that has at least one positively recurrent class. Let Π be the set of all positively recurrent classes. The Corollary states that every stationary distribution π for the chain takes the form

$$\overline{\pi} = \sum_{C \in \Pi} \lambda_C \overline{\pi}^C \ , \tag{9}$$

where $\lambda_C \geq 0$ for each C and $\sum_{C \in \Pi} \lambda_C = 1$. For each C,

$$\lambda_C = \sum_{j \in C} \overline{\pi}_j \ .$$

Technically, every stationary distribution for the chain is a convex combination of the stationary distributions concentrated on the various recurrence classes.

As always, it's worth mentioning how the results specify to chains with finitely many states. By the Corollary to Theorem 3, any such chain has at least one positively recurrent state and and hence, by Theorem 4, at least one stationary distribution. If the chain has only one recurrence class — in particular, if it is irreducible — then it has a unique stationary distribution by Theorem 5.

5. Convergence of time averages with probability 1

Think back to the light-board model. Every run of the Markov chain gives rise to a sequence of lights flashing, or, alternatively, a path through the Markov chain's state space $\mathcal{S}$. Suppose you collect a certain amount of money every time a light flashes — say you collect $f(i)$ whenever light i flashes. Your accumulation grows as the run continues. The average amount of money you collect per unit time on a

given run up through time n is

$$S_n(f) = \frac{1}{n}\sum_{m=1}^{n} f(X_m) \ ,$$

where X_m is the state of the Markov chain at time m. (For ease of exposition, I'm assuming you collect no money from the flash at time 0.) It's natural to ask whether $S_n(f)$ approaches a limit and, if so, in what sense.

Theorems 3 addresses these convergence questions for specific choices of f. To see how this works, note that for any state j we can define

$$\chi_{\{j\}}(i) = \begin{cases} 1 & \text{if } \ i = j \\ 0 & \text{for all other } i \in \mathcal{S} \ . \end{cases}$$

Then

$$S_n(\chi_{\{j\}}) = \frac{1}{n} N_j(n) \ ,$$

the fraction of the time the Markov chain spends in state j during the time interval $1 \le m \le n$. Theorem 3 gives for each $i \in \mathcal{S}$ the limiting value as $n \to \infty$ of $E_i(S_n(\chi_{\{j\}}))$ for every j in $\mathcal{S}$. So Theorem 3 is about the convergence of expected values of n-fold time averages. You can augment the result of Theorem 3 by taking into account the initial distribution $\pi(0)$ for the Markov chain. The laws of conditional probability dictate that

$$E(S_n(\chi_{\{j\}})) = \sum_{i \in \mathcal{S}} \pi_i(0) E_i(S_n(\chi_{\{j\}})) \ \text{ for all } \ n > 0 \text{ and } j \in \mathcal{S} \ ,$$

and it follows from Theorem 3 and a bit of analysis that

$$\lim_{n\to\infty} E(S_n(\chi_{\{j\}})) = \sum_{i \in \mathcal{S}_R} \frac{\pi_i(0) r_{ij}}{m_j} \ ,$$

where $\mathcal{S}_R$ is the set of recurrent states. This formula, again, applies to expected values of time averages.

As I mentioned at the end of Section 3, one can prove a result far stronger than Theorem 3 that addresses the convergence of the time averages themselves rather than their expected values. Instead of asserting that the mean over many Markov-chain runs of the time average of $\chi_{\{j\}}$ will do such-and-such, Theorem 6 below states that the time average of $\chi_{\{j\}}$ in essentially any run — that is, in any run other than a set of runs that has zero probability — will converge nicely. After proving Theorem 6, we'll be able to generalize it to cover the convergence of $S_n(f)$ for a variety of state-dependent functions f other that $\chi_{\{j\}}$.

The key enabling tool, whose proof is way beyond the level of this handout, is the Strong Law of Large Numbers, known popularly as SLLN.

Strong Law of large Numbers (SLLN): Let $\{Z_m : m \ge 1\}$ be a sequence of independent real-valued random variables with common probability distribution ρ. Let $E_\rho(Z)$ be the common expected value of all the Z_m, and assume $E_\rho(Z)$ is finite Then, with probability 1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} Z_m = E_\rho(Z) \ .$$

Now for a quick word about terminology. The phrasing I used in stating SLLN can be confusing if you're not used to it. If you prefer, you can substitute the following equivalent formulation of the last sentence in SLLN:

$$\text{Prob}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}Z_m = E_\rho(Z)\right\} = 1\ .$$

At this point, I'll introduce a method of analysis particularly well suited to formulating arguments that apply SLLN to Markov chains. The method rests on the observation that the probabilistic behavior of a Markov chain between any two successive visits to any specific recurrent state is the same. It enables us to write various quantities of interest as sums of independent identically distributed random variables, which opens the door to applying SLLN.

Suppose j is a recurrent state and that we start the Markov chain off in state j at time 0. That's the same as setting the initial distribution $\pi(0)$ as

$$\pi_i(0) = \begin{cases} 1 & \text{if } \ i = j \\ 0 & \text{if } \ i \neq j\ . \end{cases}$$

Since j is recurrent, the first return time T_j is a random variable with probability distribution

$$\text{Prob}\{T_j = k\} = \text{Prob}_j\{T_j = k\} = f_{jj}^{(k)} \ \text{ for all } \ k > 0\ .$$

Note that total probability and Prob_j coincide because of the j-focused initial distribution. When j is positively recurrent, $E(T_j)E_j(T_j) = m_j < \infty$, whereas when j is null-recurrent then $E(T_j) = E_j(T_j) = \infty$.

For each $q > 0$, define $T_j^{(q)}$ as the qth time that the state returns to j after time 0. Set $T_j^{(0)} = 0$. Then

$$T_j^{(q)} = \sum_{l=1}^{q} T_j^{(l)} - T_j^{(l-1)} = \sum_{l=1}^{q} Z_l\ ,$$

where

$$Z_l = T_j^{(l)} - T_j^{(l-1)}$$

for every $l > 0$. For each $l > 0$, Z_l represents the waiting time from the $(l-1)$th visit to state j until the lth visit. The key observation is that the Z_l are independent random variables and are all distributed identically to T_j. Accordingly, $T_j^{(q)}$ is the sum of independent identically distributed random variables. Assume for the moment that j is positively recurrent, so the common mean m_j of all the Z_l is finite. Then SLLN yields

$$\lim_{q\to\infty}\frac{1}{q}T_j^{(q)} = \lim_{q\to\infty}\frac{1}{q}\sum_{l=1}^{q}Z_l = m_j \tag{10}$$

with probability 1.

A simple modification of the argument shows that (10) holds even for a null-recurrent state j, for which $m_j = \infty$. Here's how it goes. Let $\{Z_l : l > 0\}$ be the

sequence of independent identically distributed random variables defined above. Define for each integer $K > 0$ a new sequence $\{Y_l^K : l > 0\}$ as follows:

$$Y_l^K = \begin{cases} Z_l & \text{if } Z_l \le K \\ 0 & \text{if } Z_l > K \, . \end{cases}$$

The Y_l^K are again independent and identically distributed. Note that for every K and l we have

$$\text{Prob}\{Y_l^K = 0\} = \text{Prob}\{Z_l > K\} = \sum_{k=K+1}^{\infty} f_{jj}^{(k)}$$

and

$$\text{Prob}\{Y_l^K = k\} = \text{Prob}\{Z_l = k\} = f_{jj}^{(k)} \text{ if } 1 \le k \le K \, .$$

Hence the Y_l^K have a finite common expected value of

$$E(Y_l^K) = \sum_{k=1}^{K} k f_{jj}^{(k)} \, .$$

Because $Z_l \ge Y_l^K$ with probability 1 for every K and l, we have

$$\frac{1}{q} \sum_{l=1}^{q} Z_l \ge \frac{1}{q} \sum_{l=1}^{q} Y_l^K$$

with probability 1 for all positive K and q. SLLN implies that when $q \to \infty$ the right-hand side converges to $E(Y_1^K)$, which means that, with probability 1,

$$\liminf_{q \to \infty} \frac{1}{q} \sum_{l=1}^{q} Z_l \ge E(Y_1^K) = \sum_{k=1}^{K} k f_{jj}^{(k)} \text{ for all } K > 0 \, .$$

The right-hand side blows up as $K \to \infty$ because j is null-recurrent, and hence, with probability 1,

$$\lim_{q \to \infty} \frac{1}{q} T_j^{(q)} = \lim_{q \to \infty} \frac{1}{q} \sum_{l=1}^{q} Z_l = \infty \, ,$$

which is simply (10) in the null-recurrent case $m_j = \infty$.

Now, assume once again that we start the Markov chain in recurrent state j at time 0, and as usual for each $n > 0$ let $N_j(n)$ be the number of times the Markov chain is in state j during the time interval $1 \le m \le n$. Note that

$$N_j(n) = k \Longleftrightarrow T_j^{(k)} \le n < T_j^{(k+1)} \, .$$

This is equivalent to saying that

$$T_j^{(N_j(n))} \le n < T_j^{(N_j(n)+1)} \text{ for all } n > 0 \, ,$$

which is the same as

$$\frac{1}{\frac{1}{N_j(n)} T_j^{(N_j(n))}} \ge \frac{N_j(n)}{n} > \frac{1}{\frac{1}{N_j(n)} T_j^{(N_j(n)+1)}} \text{ for all } n > 0 \, .$$

Since j is recurrent, the outer terms in the inequality both approach $1/m_j$ with probability 1 by as $n \to \infty$ by (10). It follows that

$$\lim_{n \to \infty} \frac{N_j(n)}{n} = \frac{1}{m_j} \tag{11}$$

with probability 1 given that the chain starts in state j at time 0. Said another way, for any initial distribution $\pi(0)$ we have

$$\text{Prob}_j \left\{ \lim_{n\to\infty} \frac{N_j(n)}{n} = \frac{1}{m_j} \right\} = 1 \ . \tag{12}$$

Keep in mind that (12), which is the key to proving Theorem 6, holds for both positively recurrent and null-recurrent states j.

Theorem 6: For an arbitrary homogeneous Markov chain with countable state space, and with notation as in the foregoing, Markov chain,

- If j is a transient state, then for any state i,
$$\lim_{n\to\infty} \frac{N_j(n)}{n} = 0$$
for every path emanating from i with the possible exception of a set of paths of probability zero.
- If j is a recurrent state, then for any state i
$$\lim_{n\to\infty} \frac{N_j(n)}{n} = \frac{1}{m_j}$$
for any path emanating from i that hits j in finite time, and
$$\lim_{n\to\infty} \frac{N_j(n)}{n} = 0$$
for any path emanating from i that doesn't hit j — both assertions holding with the possible exception of a set of paths of probability zero.

Proof: When j is a transient state, $\text{Prob}_i\{N_j < \infty\} = 1$ for every state i by Theorem 1. This implies immediately that

$$\text{Prob}_i \left\{ \lim_{n\to\infty} \frac{N_j(n)}{n} = 0 \right\} = 1$$

for every state i, which is another way of writing the first bullet point of the Theorem.

Now suppose j is a recurrent state, and let i is an arbitrary state of the Markov chain. If we start the chain in state i at time 0, it will reach state j in finite time with probability r_{ij}. In other words, the set of all paths emanating from state i divides into two sets: the set of paths that hit j in finite time (call that set H) and the set that don't hit j in finite time (call that set $\overline{H}$). Given that the chain starts in state i, the probability that it follows a path in H is r_{ij} and the probability that it follows a path in $\overline{H}$ is $1 - r_{ij}$. $N_j(n)$ will be zero for all n along any path in $\overline{H}$, so, trivially, $\lim_{n\to\infty} N_j(n)/n = 0$ for every path in $\overline{H}$.

Every path in H, on the other hand, hits j first at some finite time $T_j^{(0)}$, which is a random quantity taking different values for different paths. Hitting state j at time $T_j^{(0)}$ is tantamount to re-starting the chain in state j at that time. As a result, the fraction of the time the chain spends in state j from then on will behave asymptotically according to (11). That proves the second bullet point of the Theorem modulo a bit of hand-waving.

To finish the proof carefully, note first that, for every $n > T_j^{(0)}$, $N_j(n) - 1$ is the number of times the chain visits j during the time interval $T_j^{(0)} < m \leq n$. So by (12), along any path in H except possibly for a set of probability zero,

$$\lim_{n\to\infty} \frac{1}{n - T_j^{(0)}}(N_j(n) - 1) = \frac{1}{m_j} \ .$$

Furthermore, along any path in H,

$$\lim_{n\to\infty} \frac{N_j(n)}{n} = \lim_{n\to\infty} \frac{1}{n - T_j^{(0)}}(N_j(n) - 1) \ .$$

To see this, note that

$$\frac{1}{n - T_j^{(0)}}(N_j(n) - 1) = \left(\frac{n}{n - T_j^{(0)}}\right)\left(\frac{N_j(n}{n}\right) - \frac{1}{n - T_j^{(0)}} \ .$$

The second term on the right-hand side goes to zero as $n \to \infty$ along any path in H. Similarly, the first factor in the first term on the right-hand side approaches 1 as $n \to \infty$ along any path in H. We conclude that along all paths in H, except for a set of zero probability,

$$\lim_{n\to\infty} \frac{N_j(n)}{n} = \frac{1}{m_j} \ ,$$

and that proves the second bullet point of the Theorem. □

Having dealt with Theorem 6, it's time to tackle the question posed at the start of this section, namely, what can we say about the convergence of time averages of the form

$$S_n(f) = \frac{1}{n}\sum_{m=1}^{n} f(X_m)$$

for functions $f : \mathcal{S} \to \mathbb{R}$? As I remarked previously, Theorem 6 takes care of the case $f = \chi_{\{j\}}$ because

$$S_n(\chi_{\{j\}}) = \frac{1}{n}\sum_{m=1}^{n} \chi_{\{j\}}(X_m) = \frac{1}{n}N_j(n)$$

for every $n > 0$. If j is transient, $S_n(\chi_{\{j\}})$ converges to 0 with probability 1 as $n \to \infty$. If j is recurrent, then $S_n(\chi_{\{j\}})$ always converges as $n \to \infty$, but what it converges to depends on the initial distribution for the Markov chain and can vary from run to run. If, for example, we start the chain off with an initial distribution concentrated on the recurrence class of j, then $S_n(\chi_{\{j\}})$ converges to $1/m_j$ with probability 1 as $n \to \infty$. If the initial distribution is concentrated instead on some other recurrence class, $S_n(\chi_{\{j\}})$ is zero for all $n > 0$. If the Markov chain has a transient state i for which $0 < r_{ij} < 1$, and we start the chain off with initial distribution concentrated on state i, then $S_n(\chi_{\{j\}})$ will converge either to $1/m_j$ or to 0 depending on the path Markov chain follows. If the path ever enters the recurrence class of j, it will stay there forever, and $S_n(\chi_{\{j\}})$ will converges to $1/m_j$. The probability that the chain follows such a path is r_{ij}.

When a Markov chain has finitely many states, a version of Theorem 6 applies to every $f : \mathcal{S} \to \mathbb{R}$. It turns out that $S_n(f)$ always converges as $n \to \infty$. What

it converges to is sometimes dependent on the initial distribution and sometimes not. Moreover, for a given initial distribution, $S_n(f)$ might converge to different limits on different runs with positive probability. The most straightforward kind of convergence applies to a Markov chain with only one recurrence class and hence only one stationary distribution.

Theorem 7: Given a Markov chain with finite state space $\mathcal{S}$, suppose the chain has only one recurrence class. Then for any $f : \mathcal{S} \to \mathbb{R}$ and for any initial distribution $\pi(0)$,

$$S_n(f) = \frac{1}{n} \sum_{m=1}^{n} f(X_m)$$

converges with probability 1 as $n \to \infty$ to

$$E_{\overline{\pi}^*}(f) = \sum_{j \in \mathcal{S}} f(j) \overline{\pi}_j^*$$

where $X(m)$ is the state of the Markov chain at time m and π^* is the unique stationary distribution for the Markov chain.

Proof: First of all, Theorem 5 guarantees the existence of π^*. Since the Markov chain has only one recurrence class, the entire set $\mathcal{S}_R$ of recurrent states must constitute that class, and π^* is given by

$$\pi_j^* = \begin{cases} 0 & \text{if } j \in \mathcal{S}_T \\ 1/m_j & \text{if } j \in \mathcal{S}_R \end{cases},$$

where $m_j = E_j(T_j)$ is the average first return time for state j. Note that since the state space is finite, the recurrent states are all positively recurrent. Furthermore, since the chain has only one recurrence class, $r_{ij} = 1$ for every state i and every recurrent state j. This implies, because of Theorem 6, that (11) holds with probability 1 for every recurrent state j, independently of the initial distribution $\pi(0)$.

Note that for any $n > 0$

$$S_n(f) = \frac{1}{n} \sum_{j \in \mathcal{S}} f(j) N_j(n) = \sum_{j \in \mathcal{S}} f(j) \left(\frac{N_j(n)}{n} \right) .$$

As $n \to \infty$, the jth term in large parentheses approaches $1/m_j$ for recurrent states j and zero for transient states j, both with probability 1. It follows that

$$\lim_{n \to \infty} S_n(f) = \sum_{j \in \mathcal{S}_R} \frac{f(j)}{m_j} = \sum_{j \in \mathcal{S}} f(j) \pi_j^* = E_{\pi^*}(f)$$

with probability 1. As always, interchanging the limit as $n \to \infty$ with the summation is permissible because $\mathcal{S}$ is finite. □

If a finite-state Markov chain has multiple recurrence classes, the limiting behavior of $S_n(f)$ depends on the initial distribution for the Markov chain. As in the

proof of Theorem 7,

$$S_n(f) = \frac{1}{n}\sum_{j\in\mathcal{S}} f(j)N_j(n) = \sum_{j\in\mathcal{S}} f(j)\left(\frac{N_j(n)}{n}\right)$$

for every $n > 0$. We know from Theorem 6 that, on every run of the Markov chain, $\lim_{n\to\infty} N_j(n)/n$ is either $1/m_j$ or 0. The limit is 0 if j is a transient state or if the run never enters the recurrence class of j. If the run enters the recurrence class of j, then the limit is $1/m_j$. The possible values of $\lim_{n\to\infty} S_n(f)$ are therefore $\sum_{j\in C} f(j)/m_j$, where C stands for any recurrence class. Alternatively, using the notation of Theorem 5, for any f and any run of the Markov chain, there exists a recurrence class C such that

$$\lim_{n\to\infty} S_n(f) = \sum_{j\in C} f(j)\overline{\pi}_j^C \ .$$

C is the recurrence class that the chain ends up in over the course of the run. The probability that the chain ends up in a particular C is simply $\sum_{i\in\mathcal{S}} \pi_i(0)r_{ij}$, where j is any state in C. Note that for any $i \in \mathcal{S}$, $r_{ik} = r_{ij}$ for any two states j and k in C because $r_{jk} = r_{kj} = 1$.

Markov chains with infinitely many states can have null-recurrent states and/or closed subsets of transient states. Although it's possible to prove rather general versions of Theorem 7 that apply to such chains, I'll restrict attention here to irreducible chains with no transient or null-recurrent states. The state space $\mathcal{S}$ of any such chain is one big positively recurrent class. Theorem 5 guarantees the existence of a unique stationary distribution π^* with specification

$$\pi_j^* = 1/m_j \ \text{ for all } \ j \in \mathcal{S} \ .$$

The version of Theorem 7 that holds for such Markov chains applies only to functions $f : \mathcal{S} \to \mathbb{R}$ for which

$$E_{\pi^*}(f) = \sum_{j\in\mathcal{S}} f(j)\pi_j^*$$

is finite. Here is the result, whose proof appears in the Appendix.

Theorem 8: Suppose a Markov chain with state space $\mathcal{S}$ is irreducible and has only positively recurrent states. Let π^* be the unique stationary distribution guaranteed by Theorem 5. Let $f : \mathcal{S} \to \mathbb{R}$ be any function for which $E_{\pi^*}(f)$ is finite. Then, for any initial distribution $\pi(0)$,

$$S_n(f) = \frac{1}{n}\sum_{m=1}^{n} f(X_m)$$

converges with probability 1 as $n \to \infty$ to $E_{\pi^*}(f)$, where $X(m)$ is the state of the Markov chain at time m. □

6. Convergence of distributions

Consider again the two models for Markov chains — lightwall and marble-bins — in terms of which I've attempted to frame the discussion. The marble-bins model focuses our attention on distributions of marbles among the bins, particularly

stationary distributions. The lightwall model prompts us instead to ask questions about the behavior of typical runs of a Markov chain. Theorem 6 and the ensuing discussion characterize the possible stationary distributions. Theorems 7 and 8, in turn, demonstrate that these stationary distributions, at least in certain important generic cases, give answers to the questions about asymptotics of typical runs.

We haven't yet addressed the dynamics of the distribution of marbles among the bins. If we start a Markov chain off with initial distribution $\pi(0) = \overline{\pi}$, where $\overline{\pi}$ is a stationary distribution for the chain, the allocation of marbles among the bins stays fixed over time — in other words, $\pi(m) = \overline{\pi}$ for all $m > 0$ — because $\overline{\pi}$ is a fixed point of the recursion

$$\pi_j(m+1) = \sum_{i \in \mathcal{S}} \pi_i(m) P(i,j) \text{ for all } j \in \mathcal{S} .$$

How does $\pi(m)$ evolve over time if $\pi(0)$ is not a stationary distribution? Might $\pi(m)$ converge somehow to a stationary distribution $\overline{\pi}$ for the Markov chain? Such a result wouldn't be terribly surprising in the light of Theorem 8. If a chain has a unique stationary distribution π^*, the asymptotics of typical runs of the chain are embodied in π^*. Since π^* reflects the asymptotics of the lightwall model in this way, might one not expect π^* to reflect the asymptotics of the marbles-bins model similarly?

Suppose $C \subset \mathcal{S}$ is a recurrence class for the Markov chain. Once a run of the chain has entered C, it never leaves C thereafter. More prosaically, once the green marble has arrived in a bin corresponding to a state in C, it will spend the rest of its time bouncing between bins in C. The same is true of all other marbles that land in some bin in C because $P(i,j) = 0$ for every $i \in C$ when $j \notin C$. The caretaker of bin i follows marble-redistribution orders dictating that at each time step he reallocate the marbles in bin i among the bins in C only. Over time, new marbles might arrive in bins corresponding to states in C. Because

$$\pi_j(m+1) = \sum_{i \in \mathcal{S}} \pi_i(m) P(i,j) = \sum_{i \in C} \pi_i(m) P(i,j) + \sum_{i \notin C} \pi_i(m) P(i,j)$$

for all $j \in C$ and because $\sum_{j \in C} P(i,j) = 1$ for every $i \in C$, we have

$$\sum_{j \in C} \pi_j(m+1) = \sum_{i \in C} \pi_i(m) + \sum_{j \in C} \sum_{i \notin C} \pi_i(m) P(i,j) .$$

Thus the total probability allotted to states in C — if you will, the accumulation of marbles in bins corresponding to states in C — can only increase over time when C is a recurrence class.

Any additional marbles accruing to C's bins must come from bins corresponding to transient states because marbles in bins corresponding to states in any other recurrence class $\widehat{C}$ are stuck in $\widehat{C}$'s bins. Since the Markov chain never visits a transient state infinitely often, we know that the green marble will make only finitely many visits to any bin corresponding to a transient state. We might expect the same to be true of the other marbles. Perhaps, over time, all the marbles accumulate in the bins corresponding to recurrent states, leaving transient states' bins asymptotically empty. We'll discover in what follows that this is indeed the case.

Think again about the marbles in bins corresponding to states in a recurrence class C. Theorem 5 guarantees that if C is a positively recurrent class, the Markov

chain has a unique stationary distribution $\overline{\pi}^C$ concentrated on C. Might the marbles in C distribute themselves asymptotically among bins in C in a manner proportional to $\overline{\pi}^C$? A circumspect to approach to that question demands that we understand Markov chains such as the one in Figure 1(d). You can show easily that state 1 is transient and that the chain has two recurrence classes $C_1 = \{2, 3\}$ and $C_2 = \{4, 5\}$. Viewed as row vectors, the stationary distributions concentrated on the respective recurrence classes are

$$\overline{\pi}^{C_1} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

and

$$\overline{\pi}^{C_2} = \begin{bmatrix} 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix} .$$

Suppose that $\pi_2(0) = 1$ and $\pi_j(0) = 0$ for all other j, so all the marbles are in bin 2 at time 0. The bin-2 operator reallocates all the marbles to bin 3 at time 1, then they all move back to bin 2 at time 2, and so on. Thus the sequence $\{\pi(m) : m \geq 0\}$ alternates between two values. In particular, it doesn't converge to anything as $m \to \infty$

In terms of runs, the indicated specification of $\pi(0)$ mandates that the chain start in state 2 with probability 1. It then moves with probability 1 to state 3, then back to state 2, and so on. The run, like the distribution of marbles, follows a periodic sequence of states that alternates between 2 and 3. It spends half the time, on the average, in state 2 and half in state 3 — which is what we expect given $\overline{\pi}^{C_2}$ along with Theorems 5 and 6 — but it does so in a rigidly organized fashion. Other choices for $\pi(0)$ lead to similar nonconvergent asymptotics. For example, if $\pi_2(0) = p$, $\pi_3(0) = 1-p$, and $\pi_j(0) = 0$ for all other j, then when $p \neq 1/2$ the $\pi(m)$ alternates between two values as m increases. Thus a fraction p of the marbles sits in bin 2 and a fraction $1 - p$ in bin 3 at time 0, and at time 1 all the marbles that started in bin 3 end up in bin 2 (and vice versa), etc.

The Markov chain in Figure 1(d) exhibits a property that we'll need to rule out in order to have the distribution sequence $\{\pi(m) : m \geq 0\}$ behave nicely in the limit. The results that follow rest on some fundamental facts about integers that you can read about in the Appendix.

Definition 5: Let $d > 1$ be an integer. A state j of a Markov chain is *periodic of period d* if and only if d is the greatest common divisor of the set of integers $m > 0$ for which $P^{(m)}(j, j) > 0$. State j is *aperiodic* if and only if the greatest common divisor of the set of integers $m > 0$ for which $P^{(m)}(j, j) > 0$ is 1.

Here's the idea. Given a state $j \in \mathcal{S}$, list all $m > 0$ for which it's possible to make a j-to-j transition in exactly m steps. For these and only these m-values, $P^{(m)}(j, j) > 0$. If all are multiples of some integer $d > 1$, then state j is periodic with period at least d. On the other hand, state j is aperiodic if no such d exists. States $j = 2$, 3, 4, and 5 of the Markov chain in Figure 1(d) are all periodic with period 2 since, for each j, $P^{(m)}(j, j) > 0$ if and only if m is even.

Periodic states can be either transient or recurrent. If you reverse the arrow between states 1 and 2 in Figure 1(d), states 2, 3, 4, and 5 remain periodic with period 2, but states 2 and 3 become transient. Furthermore, recurrent periodic

states can be either positively recurrent or null-recurrent. Figure 1(g) depicts a chain whose every state is null-recurrent and periodic with period 2.

It turns out that aperiodicity is "generic" in the sense that a tiny adjustment in transition probabilities will eliminate periodic states. In Figure 1(d), for example, making $P(2,2) = \epsilon$ and $P(2,3) = 1 - \epsilon$ will render states 2 and 3 aperiodic for arbitrarily small $\epsilon > 0$. But periodicity is often an essential feature of a Markov chain modeling a particular real-world process. The state of the so-called random-walk Markov chain in Figure 1(h) "walks" one state to the right or one state to the left at each time step. One can imagine dynamical situations involving nearest-neighbor transitions that the random-walk chain might represent. Every state of the chain is periodic with period 2. Eliminating periodicity by tweaking transition probabilities would vitiate the Markov chain's modeling power.

If a recurrent state is periodic, so are all the other states in its recurrence class. In fact, all states in the class have the same period.

Fact 3: Let i be a recurrent state with period $d > 1$. If $i \to j$, then j is also periodic with period d.

Proof: By Fact 1, $i \to j$ implies that $j \to i$, j is in the same recurrence class as i, and $r_{ij} = r_{ji} = 1$. So there exist positive integers k and l so that $P^{(k)}(i,j) > 0$ and $P^{(l)}(j,i) > 0$. If $n > 0$ is such that $P^{(n)}(i,i) > 0$, then from

$$P^{(k+l+n)}(i,i) \geq P^{(k)}(i,j)P^{(l)}(j,i)P^{(n)}(i,i)$$

it follows that $P^{(k+l+m)}(i,i) > 0$, as well. Now let $q > 0$ be a common divisor of the set of all $m > 0$ such that $P^{(m)}(i,i) > 0$. Since q is a divisor of both n and $k+l+n$, q must also be a divisor of $k+l$. If $\hat{n} > 0$ is such that $P^{(\hat{n})}(j,j) > 0$, then from

$$P^{(k+l+\hat{n})}(i,i) \geq P^{(k)}(i,j)P^{(\hat{n})}(j,j)P^{(l)}(j,i)$$

it follows that $P^{(k+l+\hat{n})}(i,i) > 0$, as well, so q must also be a divisor of $k+l+\hat{n}$ and hence of $\hat{n}$, as well. We conclude that the common divisors of all $m > 0$ for which $P^{(m)}(i,i) > 0$ are also common divisors of all $m > 0$ for which $P^{(m)}(j,j) > 0$. A symmetric argument shows that the reverse inclusion holds. Since i is periodic, the largest element of this set of common divisors is $d > 1$, so j is also periodic with period d. □

Fact 3 implies that if i is an aperiodic recurrent state, then every j in the recurrence class of i is also aperiodic. It follows that one can speak about periodic and aperiodic recurrence classes similarly to how one can speak about positively recurrent classes and null-recurrent classes (compare Fact 2). Furthermore, every state in a periodic recurrence class has the same period, so we can talk about things like "d-periodic recurrence classes." The recurrence classes in the Markov chains of Figures 1(d), 1(g), and 1(h) are all 2-periodic recurrence classes.

The central result of this section is that the unique stationary distribution π^* for a Markov chain whose state space consists of one aperiodic positively recurrent class is also a limiting distribution in the sense that regardless of the chain's initial

distribution $\pi(0)$, $\pi(m)$ approaches π^* as $m \to \infty$. I'll prove a version of this result and mention how one can strengthen it. The argument rests on the following observation.

Fact 4: If j is an aperiodic state and i is any state for which $i \to j$, then there exists $M > 0$ such that $P^{(m)}(i,j) > 0$ for every $m \geq M$.

Proof: By Theorem A2 of the Appendix and definition of aperiodicity, we can find positive integers $m_1, \ldots, m_K$, and M_1, where $P^{(m_k)}(j,j) > 0$ for all k, such that every $m \geq M_1$ has an expansion of the form

$$m = \sum_{k=1}^{K} n_k m_k$$

for some nonnegative integers $n_1, \ldots, n_K$. For any $m \geq M_1$,

$$\begin{aligned} P^{(m)}(j,j) &= P^{(n_1 m_1 + \cdots + n_K m_K)}(j,j) \\ &\geq \left(P^{(m_1)}(j,j)\right)^{n_1} \left(P^{(m_2)}(j,j)\right)^{n_2} \cdots \left(P^{(m_K)}(j,j)\right)^{n_K} > 0 \,. \end{aligned}$$

The weak inequality holds because the product on the second line is the probability of making an m_1-step j-to-j transition n_1 times, followed by an m_2-step j-to-j transition n_2 times, etc., and that is but one way of making a j-to-j transition in $\sum_{k=1}^{K} n_k m_k$ steps. If $i \to j$, then $P^{(l)}(i,j) > 0$ for some $l > 0$. Hence for any $m \geq M = l + M_1$, $P^{(m)}(i,j) > 0$. □

Theorem 9: Suppose a Markov chain with state space $\mathcal{S}$ is irreducible and has only aperiodic positively recurrent states. Let π^* be the unique stationary distribution guaranteed by Theorem 6. Then

$$\lim_{n\to\infty} P^{(n)}(i,j) = \frac{1}{m_j} = \pi_j^* \text{ for all } i \text{ and } j \text{ in } \mathcal{S} \,.$$

It follows that for any initial distribution $\pi(0)$, $\pi(n)$ converges to π^* as $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \pi_j(n) = \pi_j^*$$

for every $j \in \mathcal{S}$.

Partial proof: The hard part is proving that the $P^{(n)}(i,j)$ converge, and I've relegated that to the Appendix. To see why the last sentence of the theorem holds, observe first that

$$\pi_j(n) = \sum_{i\in\mathcal{S}} \pi_i(0) P^{(n)}(i,j) \text{ for all } j \in \mathcal{S} \text{ and } n > 0 \,.$$

Given $\delta > 0$, pick a finite subset $\mathcal{S}_\delta$ of states so that $\sum_{i\notin\mathcal{S}_\delta} \pi_i(0) < \delta$. Then

$$\pi_j(n) = \sum_{i\in\mathcal{S}_\delta} \pi_i(0) P^{(n)}(i,j) + \sum_{i\notin\mathcal{S}_\delta} \pi_i(0) P^{(n)}(i,j)$$

for every $j \in \mathcal{S}$ and $n > 0$. The finite sum over states in $\mathcal{S}_\delta$ converges as $n \to \infty$ to $\left(\sum_{i\in\mathcal{S}_\delta} \pi_i(0)\right) \pi_j^*$, which is bounded from below by $\pi_j^* - \delta$ and bounded from above

by π_j^*. The infinite sum is bounded from below by zero and from above by δ for every $n > 0$ because $0 \le P^{(n)}(i,j) \le 1$ for all i. It follows that

$$\limsup_{n\to\infty} \pi_j(n) \le \pi_j^* + \delta \text{ for all } j \in \mathcal{S} \text{ and } \delta > 0$$

and

$$\liminf_{n\to\infty} \pi_j(n) \ge \pi_j^* - \delta \text{ for all } j \in \mathcal{S} \text{ and } \delta > 0 \,,$$

from which it follows that

$$\lim_{n\to\infty} \pi_j(n) = \pi_j^*$$

for every $j \in \mathcal{S}$. □

Consider the implications of Theorem 9 with regard to the marble-bins model. The first sentence of the theorem statement asserts that, for any state i, if we start the chain off with all the marbles in bin i — that is, with initial distribution $\pi_i(0) = 1$ and $\pi_j(0) = 0$ for all other j — then the marbles redistribute among the bins in such a way that for each bin j, the limiting fraction of marbles in bin j is π_j^*. The point is that the i-focused initial distribution makes $P^{(n)}(i,j)$ the fraction of marbles in bin j at time $n > 0$, and $P^{(n)}(i,j) \to \pi_j^*$ as $n \to \infty$. The second sentence of the theorem statement says that the same limiting behavior occurs when you start the chain off with an arbitrary initial distributions $\pi(0)$.

Theorem 9 fails to address any kind of uniformity of convergence. Some bins might take a lot longer than others to get close to acquiring their limiting fractions of marbles. If we allocate all the marbles to bin 1 at time 0 in the chain of Figure 1(i), bin j will sit empty at least until time j, and j can be arbitrarily large. Nonetheless, the convergence of $\pi(n)$ to π^* is strongly uniform, as the following result, whose proof I'll omit, asserts.

Theorem 10: Suppose a Markov chain with state space $\mathcal{S}$ is irreducible and has only aperiodic positively recurrent states. Let π^* be the unique stationary distribution guaranteed by Theorem 6. Then, for any initial distribution $\pi(0)$, $\pi(n)$ converges to π^* as $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \left(\sup_{A\subset\mathcal{S}} \left| \sum_{j\in A} (\pi_j(n) - \pi_j^*) \right| \right) = 0 \,.$$

If the state space of a Markov chain satisfying the conditions in Theorems 9 and 10 is finite, we can say even more. Let P be the transition matrix for the Markov chain. Suppose the chain has state space $\mathcal{S} = \{1, 2, \ldots, M\}$, so P is an $(M \times M)$ stochastic matrix. In this case, we can think of $\pi(0)$, π^*, and $\pi(n)$ for $n > 0$ as row M-vectors. Since $\pi(n) = \pi(0)P^n$ for every $n > 0$ and since $\pi_j(n)$ converges to π_j^* for each $j \in \mathcal{S}$, the vector $\pi(n)$ converges to the vector π^* in the sense that

$$\lim_{n\to\infty} \|\pi(n) - \pi^*\| = 0$$

for any norm $\| \ \|$ on $\mathbb{R}^M$.

Furthermore, the sequence of matrices $\{P^n : n > 0\}$ converges as $n \to \infty$ to a matrix P^* all of whose rows are π^*. To see this, consider starting the chain in state i, which is the same as setting $\pi_i(0) = 1$ and $\pi_j(0) = 0$ for all other j. Then $\pi(n) = \pi(0)P^n$ is simply the ith row of P^n. Thus for every i, the ith row of P^n converges as $n \to \infty$ to π^*, so $P^n \to P^*$ as $n \to \infty$.

The first part of Theorem 9 extends readily to cover any Markov chain whose state contains no transient states and all of whose recurrence classes are positively recurrent and aperiodic. If the transition probabilities of such a chain are $P(i,j)$ for i and j in $\mathcal{S}$, then each recurrence class C, as I noted in Section 3, is the state space of a irreducible Markov chain with state space C and transition probabilities $P(i,j)$ for i and j in C. Theorem 9 applies to each of these Markov mini-chains separately, with $\overline{\pi}^C$ standing in for π^*. In other words, for any recurrence class C,

$$\lim_{n\to\infty} P^{(n)}(i,j) = \frac{1}{m_j} = \overline{\pi}^C_j \text{ for all } i \text{ and } j \text{ in } C\ .$$

If i and j lie in different recurrence classes, $P^{(n)}(i,j) = 0$ for all $n > 0$, so $\lim_{n\to\infty} P(i,j) = 0$ trivially. Extending the second part of Theorem 9 to cover chains with multiple recurrence classes requires a bit more work.

Theorem 11: Suppose every state of a Markov chain is positively recurrent and aperiodic. Then

$$\lim_{n\to\infty} P^{(n)}(i,j) = \begin{cases} \frac{1}{m_j} & \text{when } i \text{ and } j \text{ are in the same recurrence class} \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, for any initial distribution $\pi(0)$, $\pi(n)$ converges as $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \pi_j(n) = \sum_{C\in\Pi} \lambda_C \overline{\pi}^C_j \text{ for all } j \in \mathcal{S}\ ,$$

where Π is the set of all recurrence classes and

$$\lambda_C = \sum_{i\in C} \pi_i(0) \text{ for all } C \in \Pi\ .$$

Proof: We've seen already why the first assertion holds. As for the statement about convergence of $\pi(n)$, first recall from Theorem 5 that

$$\overline{\pi}^C_j = \begin{cases} \frac{1}{m_j} & \text{when } j \in C \\ 0 & \text{when } j \notin C\ , \end{cases}$$

so $\sum_{C\in\Pi} \lambda_C \overline{\pi}^C_j$ has at most one nonzero term for every $j \in \mathcal{S}$ — namely, the term corresponding to the one recurrence class C in which j lies, and that term is equal to $\lambda_C \overline{\pi}^C_j$. In any event, if C is any recurrence class and $j \in C$, then because $P^{(n)}(i,j) = 0$ when $i \notin C$, we have

$$\pi_j(n) = \sum_{i\in C} \pi_i(0) P^{(n)}(i,j)\ .$$

We can now apply essentially the same argument as in the partial proof of Theorem 9. To wit, given $\delta > 0$, pick a finite subset C_δ of states in C so that $\sum_{i\notin C_\delta} \pi_i(0) < \delta$.

Then

$$\pi_j(n) = \sum_{i \in C_\delta} \pi_i(0) P^{(n)}(i,j) + \sum_{i \in C - C_\delta} \pi_i(0) P^{(n)}(i,j)$$

for every $j \in C$ and $n > 0$. By the first part of the theorem and the fact that $\overline{\pi}_j^C = 1/m_j$, the finite sum over states in C_δ converges as $n \to \infty$ to $\left(\sum_{i \in C_\delta} \pi_i(0)\right) \overline{\pi}_j^C$, which satisfies the inequality

$$\left(\sum_{i \in C} \pi_i(0)\right) \overline{\pi}_j^C - \delta \leq \left(\sum_{i \in C_\delta} \pi_i(0)\right) \overline{\pi}_j^C \leq \left(\sum_{i \in C} \pi_i(0)\right) \overline{\pi}_j^C \ .$$

The infinite sum over states in $C - C_\delta$ is bounded from below by zero and from above by δ for every $n > 0$ because $0 \leq P^{(n)}(i,j) \leq 1$ for all i. It follows that

$$\limsup_{n \to \infty} \pi_j(n) \leq \left(\sum_{i \in C} \pi_i(0)\right) \overline{\pi}_j^C + \delta \ \text{ for all } \ j \in C \text{ and } \delta > 0$$

and

$$\liminf_{n \to \infty} \pi_j(n) \geq \left(\sum_{i \in C} \pi_i(0)\right) \overline{\pi}_j^C - \delta \ \text{ for all } \ j \in C \text{ and } \delta > 0 \ ,$$

from which it follows that

$$\lim_{n \to \infty} \pi_j(n) = \overline{\pi}_j^C \ \text{ for all } \ j \in C \ .$$

That completes the proof because C was an arbitrary recurrence class and every state in $\mathcal{S}$ lies in some recurrence class. □

Note that the limiting value of $\pi(n)$ in Theorem 11 is a stationary distribution that takes the general form (9). For any recurrence class C, λ_C is the total probability that the initial distribution $\pi(0)$ allots to C. The limiting distribution allots the same total probability to C, but in general it's spread differently among the states in C. Think about the marbles in the bins. The marbles meted out at time zero to bins in C remain among bins in C because C is a recurrence class. Over time, they redistribute themselves among those bins, and in the limit they're allocated proportionally to $\overline{\pi}^C$.

Theorem 11 applies only to Markov chains with no transient states. If a Markov chain all of whose recurrent states are positively recurrent and aperiodic has in addition some transient states, then $\pi(n)$ will still converge to a stationary distribution of the form (9) but with different values for the λ_C. The new values of λ_C take into account the "leakage" over time of marbles in bins corresponding to the transient states into the bins corresponding to recurrent states.

Suppose, then, that every state in $\mathcal{S}$ is either transient or positively recurrent and aperiodic. Again let Π be the set of all recurrence classes. If i is transient, then r_{ij} takes on the same value for every j in a single recurrence class $C \in \Pi$. This is because for any j and k in C, $r_{jk} = r_{kj} = 1$. In words, any path starting from i that reaches j in finite time will also reach k in finite time, and any path starting from i that reaches k in finite time will also reach j in finite time, so the paths emanating from i that reach j in finite time are the same as those that reach k in finite time — possibly modulo, as usual, a set of paths of probability zero. Accordingly, for each transient state i and recurrence class C, we can define r_i^C as

the common value of r_{ij} for $j \in C$. People call r_i^C the *absorption probability* of the transient state i into the recurrence class C. With this added piece of notation we can state the final result about convergence of distributions.

Theorem 12: Suppose every state of a Markov chain is either transient or positively recurrent and aperiodic. Let $\mathcal{S}_T$ be the set of transient states and $\mathcal{S}_R$ the set of recurrent states. Then for any $i \in \mathcal{S}_T$

$$\lim_{n\to\infty} P^{(n)}(i,j) = \begin{cases} 0 & \text{when } j \in \mathcal{S}_T \\ \frac{r_{ij}}{m_j} & \text{when } j \in \mathcal{S}_R \ . \end{cases}$$

Furthermore, for any initial distribution $\pi(0)$, $\pi(n)$ converges as $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \pi_j(n) = \sum_{C\in\Pi} \lambda_C \overline{\pi}_j^C \ \text{ for all } \ j \in \mathcal{S} \ ,$$

where Π is the set of all recurrence classes and

$$\lambda_C = \sum_{i\in C} \pi_i(0) + \sum_{i\in\mathcal{S}_T} \pi_i(0) r_i^C \ \text{ for all } \ C \in \Pi \ .$$

In particular, $\lim_{n\to\infty} \pi_j(n) = 0$ when $j \in \mathcal{S}_T$.

Partial proof: I'll leave for the Appendix the proof of the first assertion about convergence of the $P^{(n)}(i,j)$. To prove the second assertion, first observe that for every state j,

$$\pi_j(n) = \sum_{i\in\mathcal{S}_T} \pi_i(0) P^{(n)}(i,j) + \sum_{i\in\mathcal{S}_R} \pi_i(0) P^{(n)}(i,j) \ .$$

If j is transient, the second sum is term is zero for all $n > 0$ because $P(i,j) = 0$ for all $i \in \mathcal{S}_R$. The first term approaches zero as $n \to \infty$ by an argument similar to the one in the proof of Theorem 11. Thus $\pi_j(n) \to 0$ as $n \to \infty$ when j is transient.

If j is recurrent, then by Theorem 11 the second term converges as $n \to \infty$ to

$$\sum_{C\in\Pi} \left(\sum_{i\in C} \pi_i(0) \right) \overline{\pi}_j^C \ .$$

As for the first term when j is recurrent, another argument similar to the one in the proof of Theorem 11 shows that it converges as $n \to \infty$ to

$$\left(\sum_{i\in\mathcal{S}_T} \pi_i(0) \right) \lim_{n\to\infty} P^{(n)}(i,j) = \left(\sum_{i\in\mathcal{S}_T} \pi_i(0) \right) \frac{r_{ij}}{m_j} \ .$$

If C is any recurrence class, then $r_{ij} = r_i^C$ for all $i \in \mathcal{S}_T$ and $j \in C$, while

$$\overline{\pi}_j^C = \begin{cases} \frac{1}{m_j} & \text{when } j \in C \\ 0 & \text{when } j \notin C \ . \end{cases}$$

Accordingly,

$$\lim_{n\to\infty} \pi_j(n) = \sum_{C\in\Pi} \lambda_C \overline{\pi}_j^C \ ,$$

where

$$\lambda_C = \left(\sum_{i \in \mathcal{S}_T} \pi_i(0) \right) r_i^C + \sum_{i \in C} \pi_i(0) \text{ for all } C \in \Pi \,,$$

which completes the proof. □

I'll conclude with a brief description of the limiting behavior of the probability distributions on Markov chains that feature periodic states. To keep things relatively simple, consider an irreducible Markov chain whose state space $\mathcal{S}$ constitutes one positively recurrent class. Because of irreducibility and positive recurrence, the chain has a unique stationary distribution π^* by Theorem 5. Applying Theorem 3 to this chain, we see that

$$\pi_j^* = \frac{1}{m_j} = \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} E_i(N_j(n)) = \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j)$$

for every i and $j \in \mathcal{S}$. The second equality holds because $r_{ij} = 1$ for all i and j. The third equality follows from equation (1).

In other words, the long-term time average of the sequence $\{P^{(n)}(i,j) : n > 0\}$ converges as $n \to \infty$ even if the sequence itself does not. If the sequence itself converges, as it does in the case of an aperiodic chain by Theorem 9, its long-term time average converges to the same limit. But what if the chain has periodic states? If the chain has a state with period $d > 1$, Fact 3 implies that all the chain's states have that same period d. In that case, it is not hard to show using the Corollary to Theorem A2 in the Appendix that the following periodic version of Fact 4 holds.

Fact 5: If i and j are states with period d that lie in the same recurrence class, then there exists an integer ρ_{ij}, with $0 \leq \rho_{ij} < d$, such that when n is sufficiently large, $P^{(n)}(i,j) > 0$ if and only if $n = qd + \rho_{ij}$ for some positive integer q. □

Applied to an irreducible chain with period-d positively recurrent states, Fact 5 implies that for any states i and j, the sequence $\{P^{(n)}(i,j) : n > 0\}$, for n large enough, cycles as follows: $(d-1)$ zeroes followed by a positive value followed by $(d-1)$ zeroes followed by a positive value followed by $(d-1)$ zeroes and so on. Using techniques similar to those in the proof of Theorem 9, one can show that these positive values, which in the notation of Fact 5 constitute the sequence

$$\left\{ P^{(n)}(i,j) : n = qd + \rho_{ij} \text{ with } q \text{ large} \right\} \,,$$

converge to a limit $\widehat{\pi}_j$ as $q \to \infty$. But then

$$\pi_j^* = \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = \lim_{Q \to \infty} \frac{1}{Qd + \rho_{ij}} \sum_{q=1}^{Q} P^{(qd+\rho_{ij})}(i,j) = \frac{\widehat{\pi}_j}{d} \,,$$

which means that

$$\widehat{\pi}_j = d\pi_j^*$$

for every $j \in \mathcal{S}$.

Theorem 13: Suppose a Markov chain with state space $\mathcal{S}$ is irreducible and has only positively recurrent states with period $d > 1$. Let π^* be the unique stationary distribution guaranteed by Theorem 6. Then, for every i and j in $\mathcal{S}$, there exists an integer ρ_{ij}, with $0 \leq \rho_{ij} < d$, such that $P^{(n)}(i,j) > 0$ when n is sufficiently large if and only if $n = qd + \rho_{ij}$ for some integer $q > 0$. Furthermore, for every i and j,

$$\lim_{q \to \infty} P^{(qd+\rho_{ij})}(i,j) = d\pi_j^* \; .$$

Appendix

Some fundamental facts about integers

I think you all know what a prime number is, but let's start from the beginning. Let $\mathbb{N}$ denote the set of natural numbers $\{0, 1, 2, 3, \ldots\}$. Given two natural numbers d and m, we say that *d is a divisor of m* if and only if $m = dk$ for some natural number k. The standard notation for "d is a divisor of m" is $d|m$. Often we just say "d divides m" for short. Observe that 1 divides every $m \in \mathbb{N}$ and every $d \in \mathbb{N}$ divides 0. Note also that if $d|m$ and $m|n$, then $d|n$. A natural number p is *prime* if and only if the only natural-number divisors of p are 1 and p itself. By convention, 1 is not a prime number even though it satisfies the technical definition. The first few prime numbers are 2, 3, 5, 7, 11, and 13. Note that 2 is the only even prime number. Can you see why?

It's often convenient to use induction when proving things about natural numbers. It takes some practice to get the hang of inductive arguments, but they're quite useful. Here's an example of a typical such argument. I'll prove that every $m \in \mathbb{N}$, $m > 1$, has at least one prime divisor. You start with

- the base case $m = 2$: clearly, 2 has a prime divisor (2 itself).

Then you move on to

- the induction step: suppose we have shown that every $m \leq M$ has at least one prime divisor. Consider $m = M + 1$. If $M + 1$ is prime, we're done, since $M + 1$ is then a prime divisor of itself. If $M + 1$ is not prime, then we can write

  $$M + 1 = nq$$

  for some natural numbers n and q with $1 < n, q \leq M$. But since we've shown already that every such n and q has at least one prime divisor, and since n and q here are both divisors of $M + 1$, we conclude that $M + 1$ must have at least one prime divisor.

I hope you see how the induction works. We know that the theorem is true for $m = 2$ by the base case. What about $m = 3$? It's true for $m = 2$, and the induction step shows that if it's true for $m = 2$, then it's true for $m = 3$; hence it's also true for $m = 3$. What about $m = 4$? We know now that it's true for $m = 2$ and $m = 3$, and the induction step enables us to conclude that it's also true for $m = 4$. And so the dominoes fall.

Euclid used the result we just proved to demonstrate that there are infinitely many prime numbers. His argument proceeds as follows. Suppose that we have a list of K primes. Index them as $p_1, p_2, \ldots, p_K$. Consider the number

$$R = 1 + p_1 p_2 p_3 \cdots p_K \ .$$

Our "theorem" above guarantees that R has at least one prime factor p, and p could not possibly be among the p_j on our list. If it were on the list, then $p|R$ would imply that

$$p|(R - p_1 p_2 p_3 \cdots p_K) \text{ i.e. } p|1 \ ,$$

which is impossible. So p is not on our list. In particular, our list doesn't contain every prime. More trenchantly, nothing about our list is special, so no finite list of primes can be exhaustive, and infinitely many primes exist.

Natural numbers $m_1, \ldots, m_K$ are said to be *relatively prime* or *coprime* if and only if they have no common divisors except 1. One of the workhorses of number theory is the following result.

Theorem A1: If $m_1, \ldots, m_K$ are positive natural numbers and are coprime, then there exist integers $n_1, \ldots, n_K$ (note: negative integers allowed) such that $\sum_{k=1}^{K} n_k m_k = 1$.

Proof: Define a set I of integers as follows:

$$I = \{i \in \mathbb{Z} : i = \sum_{k=1}^{K} n_k m_k \text{ for some } n_1, \ldots, n_K \in \mathbb{Z}\} .$$

I obviously contains some positive elements. Let $d \in \mathbb{N}$ be the smallest positive element of I; suppose $d = \sum_{k=1}^{K} \overline{n}_k m_k$. I'll show that $d = 1$. First of all, since all the m_k are in I, $d \leq m_k$ for every k.

Given j, suppose d is not a divisor of m_j. Then since $d \leq m_j$ we get a positive remainder $r > 0$ when we divide d into m_j. In other words,

$$m_j = qd + r$$

for some positive $r \in \mathbb{N}$ with $r < d$. But this means that

$$m_j = q \sum_{k=1}^{K} n_k m_k + r$$

which in turn implies that

$$r = (1 - q\overline{n}_j)m_j - \sum_{k \neq j} \overline{n}_k m_k = \sum_{k=1}^{K} \hat{n}_k a_k ,$$

so $r \in I$, as well. This is a contradiction since $r < d$ and we defined d as the smallest positive element of I. It follows that $d | m_j$ after all. Since j was arbitrary, $d | m_k$ for all k. But the m_k are coprime, so $d = 1$. Conclude that $\sum_{k=1}^{K} n_k m_k = 1$. □

The *greatest common divisor* of a (possibly infinite) family $\{m_1, m_2, m_3, \ldots\}$ of natural numbers is the largest natural number that divides all the m_k. The standard notation for that greatest common divisor is $\gcd(m_1, m_2, m_3 \ldots)$. There always exists $K > 0$ such that

$$\gcd(m_1, m_2, m_3 \ldots) = \gcd(m_1, m_2, \ldots, m_K) .$$

To see this, observe that $\{d_k = \gcd(m_1, m_2, \ldots, m_k) : k > 0\}$ is a decreasing sequence of natural numbers bounded from below by 1, so it must "stop decreasing" after k reaches some $K > 0$. If you set

$$\overline{m}_k = \frac{m_k}{\gcd(m_1, m_2, m_3, \ldots)}$$

for all k, then $\gcd(\overline{m}_1, \overline{m}_2, \overline{m}_3, \ldots) = 1$. Dividing by the greatest common divisor cancels out any common divisors that the m_k might have. More rigorously, if d divides all the $\bar{m}_k$, then $d \gcd(m_1, m_2, m_3, \ldots)$ divides all the m_k, and it follows

that $d = 1$ because $\gcd(m_1, m_2, m_3, \ldots)$ is by definition the largest common divisor of the m_k.

Corollary to Theorem A1: If $m_1, \ldots, m_K$ are positive natural numbers whose greatest common divisor is d, then there exist integers $n_1, \ldots, n_K$ (note: negative integers allowed) such that $\sum_{k=1}^{K} n_k m_k = d$.

Proof: Form the coprime natural numbers $\overline{m}_1, \ldots, \overline{m}_K$ as above and conclude from Theorem A1 that there exist $n_1, \ldots, n_K$ such that $\sum_{k=1}^{k} n_k \overline{m}_k = 1$. Then multiply both sides of the last identity by d. □

The following results have important consequences for Markov chains.

Theorem A2: If $m_1, m_2, m_3, \ldots$ are positive natural numbers whose greatest common divisor is 1, then there exist positive integers $K > 0$ and $M > 0$ with the following property: for every integer $m \geq M$ there exist exist nonnegative integers $v_1, \ldots, v_K$ such that $\sum_{k=1}^{K} v_k m_k = m$.

Proof: As I noted earlier, you can find $K > 0$ so that $\gcd(m_1, m_2, \ldots, m_K) = 1$. Choose such a K and use Theorem A1 to find integers $n_1, \ldots, n_k$ so that $\sum_{k=1}^{K} n_k m_k = 1$. Re-write the left-hand side as a difference $N_1 - N_2$, where N_1 and N_2 are both positive. Let $M = N_2^2$. If $m \geq M$, then we can write $m = N_2^2 + l$ for some nonnegative integer l. We can also decompose l as $l = qN_2 + r$, where q and r are nonnegative integers with $r < N_2$. Since $N_1 - N_2 = 1$, it follows that

$$\begin{aligned} m &= N_2^2 + qN_2 + r(N_1 - N_2) \\ &= N_2(N_2 - r + q) + rN_1 \, . \end{aligned}$$

Because $N_2 > r$, the last line is of the form $\sum_{k=1}^{K} v_k m_k$ for nonnegative integers $v_1, \ldots, v_k$, and the proof is complete. □

Corollary to Theorem A2: If $m_1, m_2, m_3, \ldots$ are positive natural numbers whose greatest common divisor is d, then there exist positive integers $K > 0$ and $M > 0$ with the following property: for every integer $m \geq M$ there exist exist nonnegative integers $v_1, \ldots, v_K$ such that $\sum_{k=1}^{K} v_k m_k = md$.

Proof: Choose K so $\gcd(m_1, \ldots, m_K) = d$. Now apply Theorem A2 to the coprime integers $\overline{m}_1 = m_1/d, \ldots, \overline{m}_K = m_k/d$ to find $M > 0$ so that for every $m \geq M$ there exist nonnegative integers $v_1, \ldots, v_K$ so that $m = \sum_{k=1}^{K} v_k \overline{m}_k$. Finally, multiply through by d to obtain $\sum_{k=1}^{K} v_k m_k = md$. □

Proofs of Theorems 2 and 3

Here's a re-statement of Theorem 2.

Theorem 2: With notation as usual,

- If j is a transient state, then

$$\lim_{n\to\infty} \frac{E_j(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) = 0 \ .$$

- If j is a recurrent state, then

$$\lim_{n\to\infty} \frac{E_j(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) = \frac{1}{m_j} \ .$$

Now for the proof. Assume first that j is transient. Start the chain off in state j. With probability 1, the chain follows a path that returns to j only finitely often, which implies that

$$\mathrm{Prob}_j \left\{ \lim_{n\to\infty} \frac{N_j(n)}{n} = 0 \right\} = 1 \ ,$$

which implies in turn the weaker assertion that

$$\lim_{n\to\infty} \frac{E_j(N_j)}{n} = 0 \ ,$$

which is the first bullet point of the theorem.

The argument is far more difficult when j is recurrent. I'll be applying the following identity, which relates the transition probabilities $P^{(m)}(i,j)$ and the first hitting probabilities $f_{ij}^{(k)}$:

$$P^{(m)}(i,j) = f_{ij}^{(m)} + \sum_{k=1}^{m-1} f_{ij}^{(k)} P^{(m-k)}(j,j) \ \text{ for all } \ m > 0 \text{ and } i,j \in \mathcal{S} \ . \tag{13}$$

Identity (13) reflects the fact that the chain can make an m-step transition from i to j in two mutually exclusive ways. It can either make the first i-to-j transition in time exactly m or else make the first i-to-j transition some time k between time 1 and m and then make an $(m-k)$-step j-to-j transition after that. The right-hand side of (13) is the sum of the probabilities of those two occurrences.

Identity (13) looks convolution-ish, and it helps to think of it that way. Recall that if g_1 and g_2 are two functions defined on the integers $\mathbb{Z}$, then their convolution $g = g_1 * g_2$ is the following function defined on $\mathbb{Z}$:

$$g(m) = \sum_{k=-\infty}^{\infty} g_1(k) g_2(m-k) \ \text{ for all } \ m \in \mathbb{Z} \ .$$

The convolution of two arbitrary functions g_1 and g_2 might not exist because the defining sums might diverge. However, if $g_1(m) = g_2(m) = 0$ for $m < 0$, then $g = g_1 * g_2$ exists and is given by

$$g(m) = \begin{cases} \sum_{k=0}^{m} g_1(k) g_2(m-k) & \text{if } m \geq 0 \\ 0 & \text{if } m < 0 \ . \end{cases}$$

Returning to (13), regard $P^{(m)}(i,j)$ as a function of m defined for all integers m by setting $P^{(m)}(i,j) = 0$ for $m \leq 0$. Similarly, extend the definition of $f_{ij}^{(k)}$ so that

$f_{ij}^{(k)} = 0$ for $k \leq 0$, which allows you to think of $f_{ij}^{(k)}$ as a function of k defined for all integers k. The following compact schematic revision of (13) arises:

$$P = f + f * P \tag{14}$$

where $*$ denotes convolution.

A final identity that will play a central role arises from the following infinite table:

$$\begin{array}{ccccccc} f_{jj}^{(1)} & f_{jj}^{(2)} & f_{jj}^{(3)} & f_{jj}^{(4)} & f_{jj}^{(5)} & f_{jj}^{(6)} & \cdots \\ & f_{jj}^{(2)} & f_{jj}^{(3)} & f_{jj}^{(4)} & f_{jj}^{(5)} & f_{jj}^{(6)} & \cdots \\ & & f_{jj}^{(3)} & f_{jj}^{(4)} & f_{jj}^{(5)} & f_{jj}^{(6)} & \cdots \\ & & & f_{jj}^{(4)} & f_{jj}^{(5)} & f_{jj}^{(6)} & \cdots \\ & & & & f_{jj}^{(5)} & f_{jj}^{(6)} & \cdots \end{array}$$

By definition, $m_j = \sum_{k=1}^{\infty} k f_{jj}^{(k)}$. If you stare at the table above, it becomes clear that m_j is just the sum of all the table entries. You can group them row by row if you wish by setting d_k equal to the sum across the $(k+1)$th row, i.e.

$$d_k = \sum_{l=k+1}^{\infty} f_{jj}^{(l)} \text{ for all } k \geq 0 .$$

You discover that

$$m_j = \sum_{k=0}^{\infty} d_k . \tag{15}$$

It turns out that the d_k-sequence has a convolutional representation of its own. First extend the d_k-sequence by setting $d_k = 0$ for $k < 0$. Next, recall that $\sum_{k=1}^{\infty} f_{jj}^{(k)} = r_{jj}$. When j is recurrent, $r_{jj} = 1$, in which case it follows that

$$d_k = \begin{cases} 1 - \sum_{l=0}^{k} f_{jj}^{(l)} & \text{if } k \geq 0 \\ 0 & \text{if } k < 0 . \end{cases}$$

Alternatively, in terms of convolutions,

$$d = u - u * f , \tag{16}$$

where u is the unit step function

$$u(k) = \begin{cases} 1 & k \geq 0 \\ 0 & k < 0 . \end{cases}$$

Now let j be any state, set $i = j$ in (13)-(14), and convolve both sides with the unit step u to obtain

$$u * P = u * f + u * f * P$$

from which it follows, using (16), that

$$d * P = (u - u * f) * P = u * f = u - d .$$

Convolving both sides of this last equation with the unit step yields

$$d * (u * P) = u * u - u * d .$$

Divide both sides of this equation by n and use the fact that

$$u * u(n) = n + 1 \text{ for all } n > 0$$

to obtain

$$\sum_{k=1}^{n-1} d_k \left(\frac{1}{n} \sum_{m=1}^{n-k} P^{(m)}(j,j) \right) = 1 + \frac{1}{n}(1 - u * d(n)) \text{ for all } n > 0 . \tag{17}$$

Since all the terms on the left-hand side of (17) are nonnegative, we can peel off a few of them to obtain the inequality

$$\sum_{k=1}^{M} d_k \left(\frac{1}{n} \sum_{m=1}^{n-k} P^{(m)}(j,j) \right) \leq 1 + \frac{1}{n}(1 - u * d(n)) \text{ for all } M > 0 \text{ and } n > M .$$

All the terms in parentheses on the left-hand side of this last inequality are bounded from below by the k-independent term

$$\frac{1}{n} \sum_{m=1}^{n-M} P^{(m)}(j,j)$$

because all the P's are nonnegative. Thus

$$\left(\sum_{k=0}^{M} d_k \right) \left(\frac{1}{n} \sum_{m=1}^{n-M} P^{(m)}(j,j) \right) \leq 1 + \frac{1}{n}(1 - u * d(n)) \text{ for all } M > 0 \text{ and } n > M .$$

Since all the P's are bounded from above by 1, we have

$$\frac{1}{n} \sum_{m=1}^{n-M} P^{(m)}(j,j) \geq \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) - \frac{M}{n} \text{ for all } M > 0 \text{ and } n > M ,$$

whereby

$$\left(\sum_{k=0}^{M} d_k \right) \left(\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) - \frac{M}{n} \right) \leq 1 + \frac{1}{n}(1 - u * d(n)) \text{ for all } M > 0 \text{ and } n > M .$$

A straightforward manipulation leads to

$$\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{1}{\sum_{k=0}^{M} d_k} + \frac{1 - u * d(n)}{n \sum_{k=0}^{M} d_k} + \frac{M}{n} \text{ for all } M > 0 \text{ and } n > M . \tag{18}$$

If j is positively recurrent, then the numerator in the second term approaches $1 - m_j$ as $n \to \infty$. It follows that for any $M > 0$ and any $\epsilon > 0$ we can pick $N > M$ so large that

$$\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{1}{\sum_{k=0}^{M} d_k} + \epsilon \text{ for all } n > N ,$$

implying that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{1}{\sum_{k=0}^{M} d_k} \text{ for all } M > 0 .$$

Because this last inequality holds for all $M > 0$ and because the d_k are nonnegative and sum to m_j, we obtain

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{1}{m_j} \tag{19}$$

when j is positively recurrent.

Regardless of whether j is positively recurrent or null-recurrent, the numerator in the second term on the right-hand side of (18) satisfies the inequality

$$|1 - u * d(n)| \leq n - 1 \text{ for all } n > 1 ,$$

so in this case

$$\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{2}{\sum_{k=0}^{M} d_k} + \frac{1}{n\sum_{k=0}^{M} d_k} + \frac{M}{n} \text{ for all } M > 0 \text{ and } n > M .$$

Again, given M and ϵ you can find $N > M$ so large that

$$\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{2}{\sum_{k=0}^{M} d_k} + \epsilon \text{ for all } n > N ,$$

implying that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \leq \frac{2}{\sum_{k=0}^{M} d_k} \text{ for all } M > 0 .$$

Because this last inequality holds for all $M > 0$ and because the d_k are nonnegative and sum to ∞ when j is null-recurrent, we obtain

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \leq 0$$

for null-recurrent j, implying, since all the P's are nonnegative, that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) = 0 \tag{20}$$

when j is null-recurrent.

Returning now to (17), note that all the terms in parentheses on the left-hand side are bounded from above by

$$\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) ,$$

from which it follows that

$$\left(\sum_{k=0}^{n-1} d_k\right)\left(\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j)\right) \geq 1 + \frac{1}{n}(1 - u * d(n)) \text{ for all } n > 0 ,$$

whereby

$$\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \geq \frac{1}{\sum_{k=0}^{n-1} d_k} + \frac{1 - u * d(n)}{n\sum_{k=0}^{n-1} d_k} \text{ for all } n > 0 . \tag{21}$$

When j is positively recurrent, the first term on the left-hand side is decreasing in n and converges to $1/m_j$. The numerator of the second term is bounded. Thus given $\epsilon > 0$ we can choose N so large that for $n > N$ we have

$$\frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \geq \frac{1}{m_j} - \epsilon ,$$

implying that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(j,j) \geq \frac{1}{m_j} \tag{22}$$

when j is positively recurrent.

Taken together, the three equations (19), (20), and (22) imply that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) = \frac{1}{m_j}$$

when j is recurrent, where I'm using the fact that $m_j = \infty$ when j is null-recurrent. By (2), this is the same as the second bullet-point of the theorem. □

Proving Theorem 3 given Theorem 2 is a walk in the park compared to the analytical morass we slogged through en route to Theorem 2. First a re-statement.

Theorem 3: With notation as usual,

- If j is a transient state, then for any state i

$$\lim_{n\to\infty} \frac{E_i(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = 0 \; .$$

- If j is a recurrent state, then for any state i

$$\lim_{n\to\infty} \frac{E_i(N_j(n))}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = \frac{r_{ij}}{m_j} \; .$$

The proof starts with equation (13). Sum both sides from $m = 1$ to $m = n$ and divide by n and you obtain, after re-ordering and tidying up the sums on the right-hand side,

$$\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(i,j) = \sum_{k=1}^{n-1} f_{ij}^{(k)} \left[\frac{1}{n} \sum_{m=k+1}^{n} P^{(m)}(j,j) \right] + \frac{1}{n} \sum_{m=1}^{n} f_{ij}^{(m)} \; . \tag{23}$$

The left-hand side is $E_i(N_j(n))/n$ by (1). For each k, the bracketed term in the right-hand side is k terms short of

$$\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,j) \; ,$$

which by Theorem 2 converges as $n \to \infty$ to 0 if j is transient and to $1/m_j$ if j is recurrent. The missing k terms make no contribution in the limit, so, for each k, the bracketed term converges as $n \to \infty$ either to $\mu = 0$ or to $\mu = 1/m_j$.

Meanwhile, $\sum_{k=1}^{\infty} f_{ij}^{(k)} = r_{ij}$, so we can split the right-hand side of (23) to obtain first

$$\sum_{k=1}^{n-1} f_{ij}^{(k)} \left[\frac{1}{n} \sum_{m=k+1}^{n} P^{(m)}(j,j) - \mu \right] + \mu \sum_{k=1}^{n-1} f_{ij}^{(k)} + \frac{1}{n} \sum_{m=1}^{n} f_{ij}^{(m)}$$

and then

$$\sum_{k=1}^{n-1} f_{ij}^{(k)} \left[\frac{1}{n} \sum_{m=k+1}^{n} P^{(m)}(j,j) - \mu \right] + \mu r_{ij} - \mu \sum_{k=n+1}^{\infty} f_{ij}^{(k)} + \frac{1}{n} \sum_{m=1}^{n} f_{ij}^{(m)} \; . \tag{24}$$

This last expression works for any value of μ, but $\mu = 0$ and $\mu = 1/m_j$ are the ones we care about. Given any $\epsilon > 0$, because the $f_{ij}^{(m)}$ sum to r_{ij} you can pick N_1 large enough so that

$$\sum_{k=n+1}^{\infty} f_{ij}^{(k)} < \epsilon \text{ for all } n > N_1$$

and

$$\frac{1}{n}\sum_{m=1}^{n} f_{ij}^{(m)} < \epsilon \text{ for all } n > N_1 \ ,$$

which means the sum of the last two terms in (24) is less than $(\mu+1)\epsilon$ in absolute value when $n > N_1$. Now split the first sum in (24) as follows, assuming $n > N_1$:

$$\sum_{k=1}^{N_1} f_{ij}^{(k)} \left[\frac{1}{n}\sum_{m=k+1}^{k} P^{(m)}(j,j) - \mu\right] + \sum_{k=N_1+1}^{n} f_{ij}^{(k)} \left[\frac{1}{n}\sum_{m=k+1}^{n} P^{(m)}(j,j) - \mu\right] \ .$$

Let M be a finite upper bound on the absolute value of all the bracketed terms; M exists because the $P^{(m)}(j,j) \le 1$ for every m. By choice of N_1, the second sum is bounded above in absolute value by $M\epsilon$.

Finally, set $\mu = 0$ if j is transient and $\mu = 1/m_j$ if j is recurrent. In either case, by Theorem 2 you can then pick $N_2 > N_1$ large enough so that for $n > N_2$ the bracketed terms in the first sum are less than ϵ in absolute value. If $n > N_2$ the first sum is bounded above in absolute value by $\epsilon \sum_{k=1}^{N_1} f_{ij}^{(k)}$, which is in turn bounded above by ϵr_{ij}.

We have shown: given $\epsilon > 0$, we can find N_2 so that if $n > N_2$, (24) is within $\epsilon(\mu+1+M+r_{ij})$ of μr_{ij}. In other words, we can make all the terms in (24) save μr_{ij} as small as we want by choosing n large enough. This means that (24) converges to μr_{ij} as $n \to \infty$. But (24) is an alternative expression for the right-hand side of (23). If j is transient, then $\mu = 0$, and therefore

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(i,j) = 0 \ ,$$

which yields the first bullet point of the theorem because of (1). If j is recurrent, $\mu = 1/m_j$, and therefore

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(i,j) = \frac{r_{ij}}{m_j} \ ,$$

which together with (1) yields the second bullet point of the theorem. □

Proof of Lemma associated with Theorem 5

I proved this one in Section 4 for finite state spaces; now for a complete proof. First a re-statement of the result. Recall that I'm using q_j to denote $1/m_j$ for $j \in C$. Recall also that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{(m)}(i,j) = \lim_{n\to\infty} \frac{E_i(N_j(n))}{n} = \frac{1}{m_j} \ , \tag{25}$$

Lemma: Suppose a Markov chain has a positively recurrent class C. Define a distribution $\overline{\pi}$ for the Markov chain by setting

$$\overline{\pi}_j = \begin{cases} 1/m_j = q_j & \text{if } j \in C \\ 0 & \text{if} j \notin C \ , \end{cases}$$

Then $\overline{\pi}$ is a stationary distribution for the Markov chain. Furthermore, $\overline{\pi}$ is concentrated on the recurrence class C since $\overline{\pi}_j = 0$ for $j \notin C$.

Here goes. Since $q_j > 0$ for every $j \in C$, all the elements in $\overline{\pi}$ are nonnegative. What we need to show is that they sum to 1 and satisfy the stationarity condition (5) in Definition 4. First observe that since C is a closed set of states, $P(i,j) = 0$ when $i \in C$ and $j \notin C$, so $\sum_{j \in C} P(i,j) = 1$ for every $i \in C$, and, when $j \notin C$,

$$\sum_{i \in \mathcal{S}} \overline{\pi}_i P(i,j) = \sum_{i \in C} q_i P(i,j) = 0 = \overline{\pi}_j \ .$$

So to prove that $\overline{\pi}$ satisfies (5), since $\overline{\pi}_j = q_j$ for $j \in C$, we need to show only that

$$q_j = \sum_{i \in C} q_i P(i,j) \tag{26}$$

for every $j \in C$.

The first step is to deduce that $\sum_{j \in \mathcal{S}} q_j \leq 1$. Let D be any finite subset of C. For every $n > 0$, we have

$$\sum_{j \in D} \frac{E_j(N_j)}{n} \leq 1 \ .$$

Taking the limit as $n \to \infty$ — which we can interchange with the sum because it's a finite sum — yields by (25)

$$\sum_{j \in D} q_j \leq 1 \ .$$

Since this last inequality holds for every finite subset D of C, we can conclude that

$$\sum_{j \in C} q_j = \sum_{j \in \mathcal{S}} q_j \leq 1 \ .$$

The next step is to show that

$$q_j \geq \sum_{i \in C} q_i P(i,j) \ \text{ for all } \ j \in C \ . \tag{27}$$

To that end, again let D be any finite subset of C. Observe that for every $m > 0$ and k, $j \in C$,

$$P^{(m+1)}(j,j) \geq \sum_{i \in D} P^{(m)}(j,i) P(i,j) \ .$$

Accordingly, for every $n > 0$,

$$\frac{1}{n} \sum_{m=1}^{n} P^{(m+1)}(j,j) \geq \sum_{i \in D} \left(\frac{1}{n} \sum_{m=1}^{n} P^{(m)}(j,i) \right) P(i,j) \ .$$

The left-hand side is the same as

$$\frac{n+1}{n}\left(\frac{1}{n+1}\sum_{m=1}^{n+1}P^{(m)}(j,j)\right)-\frac{1}{n}P(j,j)\ .$$

Taking limits as $n\to\infty$ using (25) yields

$$q_j\geq\sum_{i\in D}q_iP(i,j)\ \text{ for all }\ j\in C\ .$$

This holds for every finite subset D of C, and equation (27) follows.

Summing (27) over $j\in C$ yields

$$\sum_{j\in C}q_j\geq\sum_{j\in C}\sum_{i\in C}q_iP(i,j)=\sum_{i\in C}q_i\sum_{j\in C}P(i,j)=\sum_{i\in C}q_i\ .$$

Interchanging the order of the double sum is okay because it converges and all its terms are nonnegative, so it converges absolutely. Since the first and last items in the chain are equal, all the items in the chain must be equal. In particular, the inequality must hold with equality. Because of (27), which establishes a bounding relationship between corresponding terms in the two equal sums, the corresponding terms must actually be equal, i.e.

$$q_j=\sum_{i\in C}q_iP(i,j)\ \text{ for all }\ j\in C\ ,$$

which is just (26), the stationarity condition we needed to prove.

It remains to show that the q_j sum to 1. It follows easily from (26) that

$$q_j=\sum_{i\in C}q_iP^{(m)}(i,j)$$

for every $j\in C$ and $m>0$. Averaging over m yields

$$q_j=\sum_{i\in C}q_i\left(\frac{1}{n}\sum_{m=1}^{n}P^{(m)}(i,j)\right)$$

for every $n>0$ and $j\in C$. Since the q_j have a finite sum, for any $\delta>0$ we can find a subset C_δ of C such that $\sum_{i\notin C_\delta}q_i<\delta$. Since the terms in parentheses lie between 0 and 1 for every i and j,

$$\sum_{i\in C_\delta}q_i\left(\frac{1}{n}\sum_{m=1}^{n}P^{(m)}(i,j)\right)\leq q_j\leq\sum_{i\in C_\delta}q_i\left(\frac{1}{n}\sum_{m=1}^{n}P^{(m)}(i,j)\right)+\delta\ .$$

Taking the limit as $n\to\infty$ using (25) yields

$$\left(\sum_{i\in C_\delta}q_i\right)q_j\leq q_j\leq\left(\sum_{i\in C_\delta}q_i\right)q_j+\delta\ .$$

Using the fact that $C_\delta\subset C$ we can massage the right-hand inequality into

$$q_j\leq\left(\sum_{i\in C}q_i\right)q_j+\delta\ \text{ for all }\ \delta>0\ .$$

Dividing by the positive number q_j reveals that

$$\sum_{i\in C}q_i\geq 1-\delta/q_j\ \text{ for all }\ \delta>0\ ,$$

Because j is arbitrary, $\sum_{i\in C} q_i \geq 1$. We've seen already that $\sum_{i\in C} q_i \leq 1$, so we're done. □

Proof of Theorem 8

First a re-statement of the theorem.

Theorem 8: Suppose a Markov chain with state space $\mathcal{S}$ is irreducible and has only positively recurrent states. Let π^* be the unique stationary distribution guaranteed by Theorem 5. Let $f : \mathcal{S} \to \mathbb{R}$ be any function for which

$$E_{\pi^*}(f) = \sum_{j\in\mathcal{S}} f(j)\pi^*_j$$

is finite. Then, for any initial distribution $\pi(0)$,

$$S_n(f) = \frac{1}{n}\sum_{m=1}^{n} f(X_m)$$

converges with probability 1 as $n \to \infty$ to $E_{\pi^*}(f)$, where $X(m)$ is the state of the Markov chain at time m.

Now for the proof. Let i be any state and assume we start the chain in state i at time 0, which is the same as saying that we endow the chain with initial distribution concentrated entirely on state i. Thus, as in the proof of Theorem 6, E_i and total expectation E are the same. For every $q > 0$ let $T_i^{(q)}$ be the qth time the chain returns to state i after time 0; set $T_i^{(0)} = 0$. Let $f : \mathcal{S} \to \mathbb{R}$ be any function. For each $l \geq 0$, define

$$W_l = \sum_{m=T_i^{(l-1)}+1}^{T_i^{(l)}} f(X_m) \; .$$

By the standard reasoning, $\{W_l\}$ is an independent identically distributed sequence of random variables. All the W_l have the same distribution as

$$W_1 = \sum_{m=1}^{T_i^{(1)}} f(X_m) \; .$$

During the time interval $0 < m \leq T_i^{(1)}$, the Markov chain hits state i exactly once and hits each other states j exactly $N_j(T_i^{(1)})$ times. Accordingly,

$$W_1 = f(i) + \sum_{j\neq i} f(j)N_j\left(T_i^{(1)}\right) \; . \tag{28}$$

Let's focus for a moment on the case $f = \chi_{\{j\}}$ for some state $j \neq i$. This makes

$$W_1 = N_j(T_i^{(1)}) \; .$$

Note that $N_j\left(T_i^{(1)}\right) < T_i^{(1)}$. Furthermore, $E\left(T_i^{(1)}\right) < \infty$ since i is positively recurrent. Hence

$$E(W_1) = E\left(N_j\left(T_i^{(1)}\right)\right) < \infty\ .$$

So SLLN applies to the W_k-sequence in this special case, yielding

$$\lim_{q\to\infty} \frac{1}{q}\sum_{l=1}^{q} W_l = E(W_1)$$

with probability 1. Re-write the left-hand side as follows:

$$\lim_{q\to\infty}\left(\frac{T_i^{(q)}}{q}\right)\left(\frac{1}{T_i^{(q)}}N_j\left(T_i^{(q)}\right)\right)\ .$$

The first term in parentheses approaches m_i with probability 1 as $q\to\infty$ by (10). The second term in parentheses approaches $1/m_j$ with probability 1 as $q\to\infty$ by Theorem 6. The applicability of Theorem 6 in this case hinges on the fact that i and j are in the same recurrence class, so $r_{ij} = 1$. Conclude that in the special case $f = \chi_{\{j\}}$ for some state $j \neq i$,

$$E(W_1) = E\left(N_j\left(T_i^{(1)}\right)\right) = \frac{m_i}{m_j}$$

For more general f, expression (28) reveals that

$$\begin{aligned} E(W_1) &= f(i) + \sum_{j\neq i} f(j) E\left(N_j\left(T_i^{(1)}\right)\right) \\ &= f(i) + m_i \sum_{i\neq j} \frac{f(j)}{m_j} \\ &= m_i \sum_{j\in\mathcal{S}} \frac{f(j)}{m_j}\ . \end{aligned}$$

In order to apply SLLN to the W_k-sequence, we need for $E(W_1)$ to be finite. The condition $E(W_1) < \infty$ is the same as

$$\sum_{j\in\mathcal{S}} \frac{f(j)}{m_j} = E_{\pi^*}(f) < \infty\ . \tag{29}$$

Given that f satisfies (29), we can conclude from SLLN that

$$\lim_{q\to\infty} \frac{1}{q}\sum_{l=1}^{q} W_l = E(W_1) = m_i E_{\pi^*}(f) \tag{30}$$

with probability 1. We can relate the left-hand side of the last equation to the time average of f as follows. Assume for the moment that f is nonnegative-valued. Given n, suppose $q(n)$ is such that $T_i^{(q(n))} < n \leq T_i^{(q(n)+1)}$. As $n\to\infty$, $n/q(n) \to m_i$ with probability 1 by (10) (compare the argument leading up to (11)). Furthermore, since f is nonnegative-valued,

$$\frac{1}{n}\sum_{m=1}^{T_i^{(q(n))}} f(X_m) \leq \frac{1}{n}\sum_{m=1}^{n} f(X_m) \leq \frac{1}{n}\sum_{m=1}^{T_i^{(q(n)+1)}} f(X_m)\ .$$

Re-writing the outer bounds in terms of the W_l, you get

$$\left(\frac{q(n)}{n}\right)\left(\frac{1}{q(n)}\sum_{l=1}^{q(n)} W_l\right) \le \frac{1}{n}\sum_{m=1}^{n} f(X_m) \le \left(\frac{q(n)+1}{n}\right)\left(\frac{1}{q(n)+1}\sum_{l=1}^{q(n)+1} W_l\right) .$$

As $n \to \infty$, the first terms in parentheses in each of the outer bounds approach $1/m_i$ and the second terms approach $E(W_1)$, both with probability 1. Hence the central time average, being sandwiched between the outer bounds, approaches $E(W_1)/m_i$ as $n \to \infty$. Applying (30) yields

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} f(X_m) = E_{\pi^*}(f)$$

with probability 1.

Now, the foregoing argument depended on starting the Markov chain in some specific state i, which is the same as prescribing for the chain special initial distribution. For more general initial distributions $\pi(0)$, the argument demonstrates that

$$\text{Prob}_i\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} f(X_m) = E_{\pi^*}(f)\right\} = 1 \text{ for all } i \in \mathcal{S} .$$

for every $i \in \mathcal{S}$. Since for any $\pi(0)$ it is the case that

$$\text{Prob}\{\text{any event}\} = \sum_{i\in\mathcal{S}} \pi_i(0)\text{Prob}_i\{\text{that event}\} ,$$

it follows that, for arbitrary $\pi(0)$,

$$\text{Prob}\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} f(X_m) = E_{\pi^*}(f)\right\} = 1 .$$

During the course of the proof, we assumed that f was nonnegative. To extend the argument to a general f satisfying (29), all you have to do is write $f = f_+ - f_-$, where f_+ and f_- are nonnegative functions both of which satisfy (29), and then run the argument separately for f_+ and f_-. □

Proof of Theorem 9

Here's a re-statement of Theorem 9.

Theorem 9: Suppose a Markov chain with state space $\mathcal{S}$ is irreducible and has only aperiodic positively recurrent states. Let π^* be the unique stationary distribution guaranteed by Theorem 6. Then

$$\lim_{n\to\infty} P^{(n)}(i,j) = \frac{1}{m_j} = \pi_j^* \text{ for all } i \text{ and } j \text{ in } \mathcal{S} .$$

It follows that for any initial distribution $\pi(0)$, $\pi(n)$ converges to π^* as $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \pi_j(n) = \pi_j^*$$

for every $j \in \mathcal{S}$.

It remains to prove the the convergence of the $P^{(n)}(i,j)$. Let X_n be the state of the Markov chain at time $n \geq 0$, and start the chain off in state i at time $n = 0$. At the same time as you're running the X-chain, run another Markov chain with state space $\mathcal{S}$ and the same transition probabilities as the X-chain, except start this auxiliary chain off with initial distribution π^*. Let Y_n be the state at time n of the auxiliary chain. The X- and Y-chains run independently of each other but synchronously in time.

You can regard the two synchronously running Markov chains as constituting a compound Markov chain with state $Z_n = (X_n, Y_n)$ at every time $n \geq 0$. The state space of the Z-chain is $\mathcal{S} \times \mathcal{S}$ and its one-step transition probabilities are

$$P((h,k),(j,q)) = P(h,j)P(k,q) \; .$$

The initial distribution of the Z-chain is

$$\pi_{(h,k)} = \delta_{ih}\pi^*_k = \begin{cases} \pi^*_j & \text{when } h = i \\ 0 & \text{when } h \neq i \end{cases}$$

for $(h,k) \in \mathcal{S} \times \mathcal{S}$. Because $h \to j$ for any states h and j of the X-chain and $k \to q$ for any states k and q of the Y-chain, and all these states are aperiodic, $P^{(m)}(h,j) > 0$ and $P^{(m)}(k,q) > 0$ for sufficiently large m by Fact 4. So for the Z-chain

$$P^{(m)}((h,k),(j,q)) = P^{(m)}(h,j)P^{(m)}(k,q) > 0$$

for sufficiently large $m > 0$. It follows that $(h,k) \to (j,q)$ for any states (h,k) and (j,q) of the Z-chain, and the Z-chain is therefore irreducible. Furthermore, every state of the Z-chain is positively recurrent because the distribution $\overline{\pi}$ defined by

$$\overline{\pi}_{(h,k)} = \pi^*_h \pi^*_k$$

is a stationary distribution for the Z-chain and is positive for every (h,k) in $\mathcal{S} \times \mathcal{S}$ (compare Theorem 5 and its Corollary). Finally, the Z-chain is aperiodic because

$$P^{(m)}((h,k),(h,k)) = P^{(m)}(h,h)P^{(m)}(k,k)$$

for every $m > 0$. For any h and k both terms on the right-hand side are positive for sufficiently large m by Fact 4, so the left-hand side is positive for sufficiently large m, which implies that every state (h,k) of the Z-chain is aperiodic.

Since the Z-chain is irreducible and all its states are recurrent, the Z-chain will visit every state in $\mathcal{S} \times \mathcal{S}$ infinitely often with probability 1. In particular, with probability 1, the Z-chain will visit every state of the form (k,k). What this means in terms of the X- and Y-chains is that, with probability 1, for every $k \in \mathcal{S}$ there will exist times $n > 0$ when $X_n = Y_n = k$. Fix some $k_o \in \mathcal{S}$ and let T_c be the first $n > 0$ for which $X_n = Y_n = k_o$. Note that T_c is a random quantity, and $T_c < \infty$ with probability 1, so

$$\sum_{m=1}^{\infty} \text{Prob}\{T_c = m\} = 1 \; . \tag{31}$$

Next, for any $n > 0$,

$$\begin{aligned} P^{(n)}(i,j) &= \text{Prob}\{X_n = j\} \\ &= \sum_{k\in\mathcal{S}} \text{Prob}\{Z_n = (j,k)\} \\ &= \sum_{m=1}^{\infty}\sum_{k\in\mathcal{S}} \text{Prob}\{Z_n = (j,k)|T_c = m\}\text{Prob}\{T_c = m\} \ . \end{aligned}$$

The equality on the first line holds because the X-chain starts in state i at time 0 with probability 1. The other two lines follow from basic facts about probability. Similarly, for every $n > 0$,

$$\begin{aligned} \pi^*_j &= \text{Prob}\{Y_n = j\} \\ &= \sum_{k\in\mathcal{S}} \text{Prob}\{Z_n = (k,j)\} \\ &= \sum_{m=1}^{\infty}\sum_{k\in\mathcal{S}} \text{Prob}\{Z_n = (k,j)|T_c = m\}\text{Prob}\{T_c = m\} \ . \end{aligned}$$

The equality in the first line holds because the Y-chain starts — and therefore stays — with stationary distribution π^*.

Now for a crucial observation. For any m and n with $n > m$ and any $k \in \mathcal{S}$,

$$\begin{aligned} &\text{Prob}\{Z_n = (j,k)|T_c = m\} \\ &\qquad = \text{Prob}\{Z_n = (j,k)|Z_m = (k_o,k_o) \text{ and } Z_l \neq (k_o,k_o) \text{ if } 0 < l < m\} \\ &\qquad = \text{Prob}\{Z_n = (j,k)|Z_m = (k_o,k_o)\} \\ &\qquad = P^{(n-m)}(k_o,j)P^{(n-m)}(k_o,k) \ . \end{aligned}$$

The first line holds because of the meaning of $T_c = m$. The second line holds because the Z-process is a Markov chain, so the history of the chain before time m is superfluous for determining the probability that $Z_n = (j,k)$ once we know that $Z_m = (k_o,k_o)$. Similarly, for any m and n with $n > m$ and any $k \in \mathcal{S}$,

$$\begin{aligned} &\text{Prob}\{Z_n = (k,j)|T_c = m\} \\ &\qquad = \text{Prob}\{Z_n = (k,j)|Z_m = (k_o,k_o) \text{ and } Z_l \neq (k_o,k_o) \text{ if } 0 < l < m\} \\ &\qquad = \text{Prob}\{Z_n = (k,j)|Z_m = (k_o,k_o)\} \\ &\qquad = P^{(n-m)}(k_o,k)P^{(n-m)}(k_o,j) \ . \end{aligned}$$

Thus for every $m > 0$ we have

$$\text{Prob}\{Z_n = (j,k)|T_c = m\} = \text{Prob}\{Z_n = (k,j)|T_c = m\} \ \text{ for all } \ n > m \ .$$

Plug this result into the formulas for $P^{(n)}(i,j)$ and π^*_j and take the difference between them to obtain for every $n > 1$ the following expression for $P^{(n)}(i,j) - \pi^*_j$:

$$\sum_{m=n+1}^{\infty}\sum_{k\in\mathcal{S}} \left(\text{Prob}\{Z_n = (j,k)|T_c = m\} - \text{Prob}\{Z_n = (k,j)|T_c = m\}\right)\text{Prob}\{T_c = m\} \ .$$

This yields in turn

$$P^{(n)}(i,j) - \pi_j^*$$
$$= \sum_{m=n+1}^{\infty} (\text{Prob}\{X_n = j | T_c = m\} - \text{Prob}\{Y_n = j | T_c = m\}) \, \text{Prob}\{T_c = m\} \, .$$

Thus

$$\left| P^{(n)}(i,j) - \pi_j^* \right|$$
$$\leq \sum_{m=n+1}^{\infty} |\text{Prob}\{X_n = j | T_c = m\} - \text{Prob}\{Y_n = j | T_c = m\}| \, \text{Prob}\{T_c = m\}$$
$$\leq \text{Prob}\{T_c > n\} \text{ for all } n > 1 \, ,$$

where the last line holds because the difference inside the absolute value sign is at most 1. Now, $\text{Prob}\{T_c > n\} \to 0$ as $n \to \infty$ because of (31), so

$$\lim_{n\to\infty} P^{(n)}(i,j) = \pi_j^* = \frac{1}{m_j} \text{ for all } i \text{ and } j \text{ in } \mathcal{S} \, ,$$

which is the theorem's principal assertion. I proved the second part of the theorem in Section 6.

Proof of Theorem 12

First, a restatement of Theorem 12, which extends Theorem 11.

Theorem 12: Suppose every state of a Markov chain is either transient or positively recurrent and aperiodic. Let $\mathcal{S}_T$ be the set of transient states and $\mathcal{S}_R$ the set of recurrent states. Then for any $i \in \mathcal{S}_T$

$$\lim_{n\to\infty} P^{(n)}(i,j) = \begin{cases} 0 & \text{when } j \in \mathcal{S}_T \\ \frac{r_{ij}}{m_j} & \text{when } j \in \mathcal{S}_R \, . \end{cases}$$

Furthermore, for any initial distribution $\pi(0)$, $\pi(n)$ converges as $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \pi_j(n) = \sum_{C\in\Pi} \lambda_C \overline{\pi}_j^C \text{ for all } j \in \mathcal{S} \, ,$$

where Π is the set of all recurrence classes and

$$\lambda_C = \sum_{i\in C} \pi_i(0) + \sum_{i\in\mathcal{S}_T} \pi_i(0) r_i^C \text{ for all } C \in \Pi \, .$$

In particular, $\lim_{n\to\infty} \pi_j(n) = 0$ when $j \in \mathcal{S}_T$.

It remains to prove the assertion about the convergence of $P^{(n)}(i,j)$. Suppose first that j is transient. By Theorem 1 and equation (3), $\sum_{n=1}^{\infty} P^{(n)}(i,j)$ converges, so it must be the case that

$$\lim_{n\to\infty} P^{(n)}(i,j) = 0 \text{ when } j \in \mathcal{S}_T \, .$$

When j is recurrent, the proof of the limiting formula builds on Theorem 11 similarly to the way the proof of Theorem 3 builds on Theorem 2. Start with

$$P^{(n)}(i,j) = \sum_{k=1}^{n-1} f_{ij}^{(k)} P^{(n-k)}(j,j) + f_{ij}^{(n)} \text{ for all } n > 0 \text{ and } i \in \mathcal{S} , \tag{32}$$

which is essentially a repeat of equation (13). Because

$$\sum_{k=1}^{\infty} f_{ij}^{(k)} = r_{ij} ,$$

we can massage (32) to obtain for every $n > 0$

$$P^{(n)}(i,j) - \frac{r_{ij}}{m_j} = \sum_{k=1}^{n-1} f_{ij}^{(k)} \left(P^{(n-k)}(j,j) - \frac{1}{m_j} \right) + \frac{1}{m_j} \sum_{k=n}^{\infty} f_{ij}^{(k)} + f_{ij}^{(n)} \tag{33}$$

Fix $\epsilon > 0$ and pick N_1 so large that $\sum_{k=n}^{\infty} f_{ij}^{(k)} < \epsilon$ when $n > N_1$. The sum of the last two terms on the right-hand side of (33) then satisfies

$$\frac{1}{m_j} \sum_{k=n}^{\infty} f_{ij}^{(k)} + f_{ij}^{(n)} < \left(\frac{1}{m_j} + 1 \right) \epsilon \text{ for all } n > N_1 .$$

The first term on the right-hand side of (33) parses as

$$\sum_{k=1}^{N_1} f_{ij}^{(k)} \left(P^{(n-k)}(j,j) - \frac{1}{m_j} \right) + \sum_{k=N_1+1}^{n-1} f_{ij}^{(k)} \left(P^{(n-k)}(j,j) - \frac{1}{m_j} \right) ;$$

the second piece is less than $M\epsilon$ in absolute value when $n > N_1$ if M is a common upper bound on all the terms in parentheses. Now pick N_2 so that

$$\left| P^{(n-k)}(j,j) - \frac{1}{m_j} \right| < \epsilon \text{ when } n > N_1 + N_2 ,$$

which you can do because of Theorem 11. This choice of N_2 means

$$\left| \sum_{k=1}^{N_1} f_{ij}^{(k)} \left(P^{(n-k)}(j,j) - \frac{1}{m_j} \right) \right| < \epsilon \sum_{k=1}^{N_1} f_{ij}^{(k)} \leq \epsilon r_{ij} \text{ when } n > N_1 + N_2 .$$

Referring back to (33), we see that given $\epsilon > 0$ we can find an N-value so large that

$$\left| P^{(n)}(i,j) - \frac{r_{ij}}{m_j} \right| < \left(M + \frac{1}{m_j} + 1 + r_{ij} \right) \epsilon \text{ when } n > N ,$$

from which it follows that

$$\lim_{n \to \infty} P^{(n)}(i,j) = \frac{r_{ij}}{m_j} \text{ when } j \in \mathcal{S}_R .$$

Just for a reality check, note that this conclusion agrees with Theorem 11 when the Markov chain has no transient states, because in that case $r_{ij} = 1$ if and only if i and j lie in the same recurrence class and $r_{ij} = 0$ if and only if i and j lie in different recurrence classes.