Chapter 2
General Markov Chains

Aldous - Fill

September 10, 1999

The setting of this Chapter is a finite-state irreducible Markov chain (X_t), either in discrete time $(t = 0, 1, 2, \ldots)$ or in continuous time $(0 \leq t < \infty)$. Highlights of the elementary theory of general (i.e. not-necessarily-reversible) Markov chains are readily available in several dedicated textbooks and in chapters of numerous texts on introductory probability or stochastic processes (see the Notes), so we just give a rapid review in sections 1 and 1.2. Subsequent sections emphasize several specific topics which are useful for our purposes but not easy to find in any one textbook: using the fundamental matrix in mean hitting times and the central limit theorem, metrics on distributions and submultiplicativity, Matthews' method for cover times, and martingale methods.

1 Notation and reminders of fundamental results

We recommend the textbook of Norris [23] for a clear treatment of the basic theory and a wide selection of applications.

Write $I = \{i, j, k, \ldots\}$ for a finite state space. Write $\mathbf{P} = p_{i,j}$ for the transition matrix of a discrete-time Markov chain $(X_t : t = 0, 1, 2, \ldots)$. To avoid trivialities let's exclude the one-state chain (*two*-state chains are useful, because surprisingly often general inequalities are sharp for two-state chains). The t-step transition probabilities are $P(X_t = j | X_0 = i) = p_{ij}^{(t)}$, where $\mathbf{P}^{(t)} = \mathbf{P}\mathbf{P}\ldots\mathbf{P}$ is the t-fold matrix product. Write $P_i(\cdot)$ and $E_i(\cdot)$ for probabilities and expectations for the chain started at state i and time 0. More generally, write $P_\rho(\cdot)$ and $E_\rho(\cdot)$ for probabilities and expectations for the chain started at time 0 with distribution ρ. Write

$$T_i = \min\{t \geq 0 : X_t = i\}$$

for the *first hitting time* on state i, and write

$$T_i^+ = \min\{t \geq 1 : X_t = i\}.$$

Of course $T_i^+ = T_i$ unless $X_0 = i$, in which case we call T_i^+ the *first return time* to state i. More generally, a subset A of states has first hitting time

$$T_A = \min\{t \geq 0 : X_t \in A\}.$$

We shall frequently use without comment "obvious" facts like the following.

> Start a chain at state i, wait until it first hits j, then wait until the time (S, say) at which it next hits k. Then $E_i S = E_i T_j + E_j T_k$.

The elementary proof sums over the possible values t of T_j. The sophisticated proof appeals to the *strong Markov* property ([23] section 1.4) of the *stopping time* T_j, which implies

$$E_i(S | X_t, t \leq T_j) = T_j + E_j T_k.$$

Recall that the symbol $|$ is the probabilist's shorthand for "conditional on".

1.1 Stationary distribution and asymptotics

Now assume the chain is *irreducible*. A fundamental result ([23] Theorems 1.7.7 and 1.5.6) is that there exists a unique *stationary distribution* $\pi = (\pi_i : i \in I)$, i.e. a unique probability distribution satisfying the *balance equations*

$$\pi_j = \sum_i \pi_i p_{ij} \text{ for all } j. \tag{1}$$

One way to prove this existence (liked by probabilists because it extends easily to the countable state setting) is to turn Lemma 6 below into a definition. That is, fix arbitrary i_0, define $\tilde{\pi}(i_0) = 1$, and define

$$\tilde{\pi}(j) = E_{i_0}(\text{number of visits to } j \text{ before time } T_{i_0}^+),\ j \neq i_0.$$

It can then be checked that $\pi_i := \tilde{\pi}(i)/\sum_j \tilde{\pi}(j)$ is a stationary distribution. The point of stationarity is that, if the initial position X_0 of the chain is random with the stationary distribution π, then the position X_t at any subsequent non-random time t has the same distribution π, and the process $(X_t, t = 0, 1, 2, \ldots)$ is then called the *stationary chain*.

A highlight of elementary theory is that the stationary distribution plays the main role in asymptotic results, as follows.

Theorem 1 (The ergodic theorem: [23] Theorem 1.10.2) *Let $N_i(t)$ be the number of visits to state i during times $0, 1, \ldots, t-1$. Then for any initial distribution,*

$$t^{-1} N_i(t) \to \pi_i \text{ a.s., as } t \to \infty.$$

Theorem 2 (The convergence theorem: [23] Theorem 1.8.3) *For any initial distribution,*

$$P(X_t = j) \to \pi_j \text{ as } t \to \infty, \text{ for all } j$$

provided the chain is aperiodic.

Theorem 1 is the simplest illustration of the *ergodic principle* "time averages equal space averages". Many general identities for Markov chains can be regarded as aspects of the ergodic principle – in particular, in section 2.1 we use it to derive expressions for mean hitting times. Such identities are important and useful.

The most classical topic in mathematical probability is time-asymptotics for i.i.d. (independent, identically distributed) random sequences. A vast number of results are known, and (broadly speaking) have simple analogs for Markov chains. Thus the analog of the strong law of large numbers is Theorem 1, and the analog of the central limit theorem is Theorem 17 below. As mentioned in Chapter 1 section 2.1 (yyy 7/20/99 version) this book has a different focus, on results which say something about the behavior of the chain over some specific finite time, rather than what happens in the indefinite future.

1.2 Continuous-time chains

The theory of continuous-time Markov chains closely parallels that of the discrete-time chains discussed above. To the reader with background in algorithms or discrete mathematics, the introduction of continuous time may at first seem artificial and unnecessary, but it turns out that certain results are simpler in continuous time. See Norris [23] Chapters 2 and 3 for details on what follows.

A continuous-time chain is specified by *transition rates* ($q(i,j) = q_{ij}, j \neq i$) which are required to be non-negative but have no constraint on the sums. Given the transition rates, define

$$q_i := \sum_{j: j \neq i} q_{ij} \tag{2}$$

and extend (q_{ij}) to a matrix $\mathbf{Q}$ by putting $q_{ii} = -q_i$. The chain $(X_t : 0 \leq t < \infty)$ has two equivalent descriptions.

1. Infinitesimal description. Given that $X_t = i$, the chance that $X_{t+dt} = j$ is $q_{ij}dt$ for each $j \neq i$.

2. Jump-and-hold description. Define a transition matrix $\mathbf{J}$ by $J_{ii} = 0$ and

$$J_{ij} := q_{ij}/q_i, \ j \neq i. \tag{3}$$

Then the continuous-time chain may be constructed by the two-step procedure

(i) Run a discrete-time chain X^J with transition matrix $\mathbf{J}$.

(ii) Given the sequence of states $i_0, i_1, i_2, \ldots$ visited by X^J, the durations spent in states i_m are independent exponential random variables with rates q_{i_m}.

The discrete-time chain X^J is called the *jump chain* associated with X_t.

The results in the previous section go over to continuous-time chains with the following modifications.

(a) $P_i(X_t = j) = Q_{ij}^{(t)}$, where $\mathbf{Q}^{(t)} := \exp(\mathbf{Q}t)$.

(b) The definition of T_i^+ becomes

$$T_i^+ = \min\{t \geq T_{I\setminus i} : X_t = i\}.$$

(c) If the chain is irreducible then there exists a unique stationary distribution π characterized by

$$\sum_i \pi_i q_{ij} = 0 \text{ for all } j.$$

(d) In the ergodic theorem we interpret $N_i(t)$ as the total duration of time spent in state i during $[0, t]$:

$$N_i(t) := \int_0^t 1_{(X_s = i)} \, ds.$$

(e) In the convergence theorem the assumption of aperiodicity is unnecessary. [This fact is the one of the technical advantages of continuous time.]

(f) The evolution of $P(X_t = j)$ as a function of time is given by the *forwards equations*

$$\frac{d}{dt} P(X_t = j) = \sum_i P(X_t = i) q_{ij}. \tag{4}$$

Given a discrete-time chain X with some transition matrix $\mathbf{P}$, one can define the *continuized* chain $\widetilde{X}$ to have transition rates $q_{ij} = p_{ij}$, $j \neq i$. In other words, we replace the deterministic time-1 holds between jumps by holds with exponential(1) distribution. Many quantities are unchanged by the passage from the discrete time chain to the continuized chain. In particular the stationary distribution π and mean hitting times $E_i T_A$ are unchanged. Therefore results stated in continuous time can often be immediately applied in discrete time, and vice versa.

In different parts of the book we shall be working with discrete or continuous time as a current convention, mentioning where appropriate how results change in the alternate setting. Chapter 4 (yyy section to be written) will give a survey of the differences between these two settings.

2 Identities for mean hitting times and occupation times

2.1 Occupation measures and stopping times

The purpose of this section is to give a systematic "probabilistic" treatment of a collection of general identities by deriving them from a single result, Proposition 3. We work in discrete time, but give the corresponding continuous-time results in section 2.3. Intuitively, a stopping time is a random time which can be specified by some on-line algorithm, together (perhaps) with external randomization.

Proposition 3 *Consider the chain started at state i. Let $0 < S < \infty$ be a stopping time such that $X_S = i$ and $E_i S < \infty$. Let j be an arbitrary state. Then*

$$E_i(\textit{number of visits to } j \textit{ before time } S) = \pi_j E_i S.$$

In the phrase "number of ... before time t", our convention is to include time 0 but exclude time t.

We shall give two different proofs. The first requires a widely-useful general theorem in stochastic processes.

Proof. Consider the renewal process whose inter-renewal time is distributed as S. The reward-renewal theorem (e.g. Ross [27] Thm. 3.6.1) says that the asymptotic proportion of time spent in state j equals

$$E_i(\text{number of visits to } j \text{ before time } S)/E_i S.$$

But this asymptotic average also equals π_j, by the ergodic theorem. □

We like that proof for philosophical reasons: a good way to think about general identities is that they show one quantity calculated in two different ways. Here is an alternative proof of a slightly more general assertion. We refer to Propositions 3 and 4 as *occupation measure identities*.

Proposition 4 *Let θ be a probability distribution on I. Let $0 < S < \infty$ be a stopping time such that $P_\theta(X_S \in \cdot) = \theta(\cdot)$ and $E_\theta S < \infty$. Let j be an arbitrary state. Then*

$$E_\theta(\textit{number of visits to } j \textit{ before time } S) = \pi_j E_\theta S.$$

Proof. Write $\rho_j = E_\theta$(number of visits to j before time S). We will show

$$\sum_j \rho_j p_{jk} = \rho_k \ \forall k. \tag{5}$$

Then by uniqueness of the stationary distribution, $\rho(\cdot) = c\pi(\cdot)$ for $c = \sum_k \rho_k = E_\theta S$.

Checking (5) is just a matter of careful notation.

$$\begin{aligned}
\rho_k &= \sum_{t=0}^{\infty} P_\theta(X_t = k, S > t) \\
&= \sum_{t=0}^{\infty} P_\theta(X_{t+1} = k, S > t) \text{ because } P_\theta(X_S = k) = P_\theta(X_0 = k) \\
&= \sum_{t=0}^{\infty} \sum_j P_\theta(X_t = j, S > t, X_{t+1} = k) \\
&= \sum_{t=0}^{\infty} \sum_j P_\theta(X_t = j, S > t)\ p_{jk} \text{ by the Markov property} \\
&= \sum_j \rho_j p_{jk}.
\end{aligned}$$

□

2.2 Mean hitting time and related formulas

The following series of formulas arise from particular choices of j and S in Proposition 3. For ease of later reference, we state them all together before starting the proofs. Some involve the quantity

$$Z_{ij} = \sum_{t=0}^{\infty} \left(p_{ij}^{(t)} - \pi_j \right) \tag{6}$$

In the periodic case the sum may oscillate, so we use the Cesaro limit or (equivalently, but more simply) the continuous-time limit (9). The matrix $\mathbf{Z}$ is called the *fundamental matrix* (see Notes for alternate standardizations). Note that from the definition

$$\sum_j Z_{ij} = 0 \text{ for all } i. \tag{7}$$

Lemma 5 $E_i T_i^+ = 1/\pi_i$.

Lemma 6

$$E_i(\textit{number of visits to } j \textit{ before time } T_i^+) = \pi_j/\pi_i.$$

Lemma 7 *For* $j \neq i$,

$$E_j(\textit{number of visits to } j \textit{ before time } T_i) = \pi_j(E_j T_i + E_i T_j).$$

Corollary 8 *For* $j \neq i$,

$$P_i(T_j < T_i^+) = \frac{1}{\pi_i(E_i T_j + E_j T_i)}.$$

Lemma 9 *For* $i \neq l$ *and arbitrary* j,

$$E_i(\textit{number of visits to } j \textit{ before time } T_l) = \pi_j(E_i T_l + E_l T_j - E_i T_j).$$

Corollary 10 *For* $i \neq l$ *and* $j \neq l$,

$$P_i(T_j < T_l) = \frac{E_i T_l + E_l T_j - E_i T_j}{E_j T_l + E_l T_j}.$$

Lemma 11 $\pi_i E_\pi T_i = Z_{ii}$.

Lemma 12 $\pi_j E_i T_j = Z_{jj} - Z_{ij}$.

Corollary 13 $\sum_j \pi_j E_i T_j = \sum_j Z_{jj}$ *for each* i.

Corollary 14 (The random target lemma) $\sum_j \pi_j E_i T_j$ *does not depend on* i.

Lemma 15

$$E_\pi(\textit{number of visits to } j \textit{ before time } T_i) = \frac{\pi_j}{\pi_i} Z_{ii} - Z_{ij}.$$

Lemmas 11 and 12, which will be used frequently throughout the book, will both be referred to as *the mean hitting time formula*. See the Remark following the proofs for a two-line heuristic derivation of Lemma 12. A consequence of the mean hitting time formula is that knowing the matrix $\mathbf{Z}$ is equivalent to knowing the matrix $(E_i T_j)$, since we can recover Z_{ij} as $\pi_j(E_\pi T_j - E_i T_j)$.

Proofs. The simplest choice of S in Proposition 3 is of course the first return time T_i^+. With this choice, the Proposition says

$$E_i(\text{number of visits to } j \text{ before time } T_i^+) = \pi_j E_i T_i^+.$$

Setting $j = i$ gives $1 = \pi_i E_i T_i^+$, which is Lemma 5, and then the case of general j gives Lemma 6.

Another choice of S is "the first return to i after the first visit to j". Then $E_i S = E_i T_j + E_j T_i$ and the Proposition becomes Lemma 7, because there are no visits to j before time T_j. For the chain started at i, the number of visits to i (including time 0) before hitting j has geometric distribution, and so

$$E_i(\text{number of visits to } i \text{ before time } T_j) = 1/P_i(T_j < T_i^+).$$

So Corollary 8 follows from Lemma 7 (with i and j interchanged).

Another choice of S is "the first return to i after the first visit to j after the first visit to l", where i, j, l are distinct. The Proposition says

$$\pi_j(E_i T_l + E_l T_j + E_j T_i) = E_i(\text{number of visits to } j \text{ before time } T_l)$$

$$+E_j(\text{number of visits to } j \text{ before time } T_i).$$

Lemma 7 gives an expression for the final expectation, and we deduce that (for distinct i, j, l)

$$E_i(\text{number of visits to } j \text{ before time } T_l) = \pi_j(E_i T_l + E_l T_j - E_i T_j).$$

This is the assertion of Lemma 9, and the identity remains true if $j = i$ (where it becomes Lemma 7) or if $j = l$ (where it reduces to $0 = 0$). We deduce Corollary 10 by writing

$$E_i(\text{number of visits to } j \text{ before time } T_l) =$$

$$P_i(T_j < T_l) E_j(\text{number of visits to } j \text{ before time } T_l)$$

and using Lemma 7 to evaluate the final expectation.

We now get slightly more ingenious. Fix a time $t_0 \geq 1$ and define S as the time taken by the following 2-stage procedure (for the chain started at i).

(i) wait time t_0

(ii) then wait (if necessary) until the chain next hits i.

Then the Proposition (with $j = i$) says

$$\sum_{t=0}^{t_0-1} p_{ii}^{(t)} = \pi_i(t_0 + E_\rho T_i) \tag{8}$$

where $\rho(\cdot) = P_i(X_{t_0} = \cdot)$. Rearranging,

$$\sum_{t=0}^{t_0-1} (p_{ii}^{(t)} - \pi_i) = \pi_i E_\rho T_i.$$

Letting $t_0 \to \infty$ we have $\rho \to \pi$ by the convergence theorem (strictly, we should give a separate argument for the periodic case, but it's simpler to translate the argument to continuous time where the periodicity issue doesn't arise) and we obtain Lemma 11.

For Lemma 12, where we may take $j \neq i$, we combine the previous ideas. Again fix t_0 and define S as the time taken by the following 3-stage procedure (for the chain started at i).

(i) wait until the chain hits k.

(ii) then wait a further time t_0.

(iii) then wait (if necessary) until the chain next hits i.

Applying Proposition 3 with this S and with $j = i$ gives

$$E_i(\text{number of visits to } i \text{ before time } T_k) + \sum_{t=0}^{t_0-1} p_{ki}^{(t)} = \pi_i(E_i T_k + t_0 + E_\rho T_i),$$

where $\rho(\cdot) = P_k(X_{t_0} = \cdot)$. Subtracting the equality of Lemma 7 and rearranging, we get

$$\sum_{t=0}^{t_0-1} (p_{ki}^{(t)} - \pi_i) = \pi_i(E_\rho T_i - E_k T_i).$$

Letting $t_0 \to \infty$, we have (as above) $\rho \to \pi$, giving

$$Z_{ki} = \pi_i(E_\pi T_i - E_k T_i).$$

Appealing to Lemma 11 we get Lemma 12. Corollary 13 follows from Lemma 12 by using (7).

To prove Lemma 15, consider again the argument for (8), but now apply the Proposition with $j \neq i$. This gives

$$\sum_{t=0}^{t_0-1} p_{ij}^{(t)} + E_\rho(\text{number of visits to } j \text{ before time } T_i) = \pi_j(t_0 + E_\rho T_i)$$

where $\rho(\cdot) = P_i(X_{t_0} = \cdot)$. Rearranging,

$$\sum_{t=0}^{t_0-1} (p_{ij}^{(t)} - \pi_j) + E_\rho(\text{number of visits to } j \text{ before time } T_i) = \pi_j E_\rho T_i.$$

Letting $t_0 \to \infty$ gives

$$Z_{ij} + E_\pi(\text{number of visits to } j \text{ before time } T_i) = \pi_j E_\pi T_i.$$

Applying Lemma 11 gives Lemma 15.

Remark. We promised a two-line heuristic derivation of the mean hitting time formula, and here it is. Write

$$\sum_{t=0}^{\infty} \left(1_{(X_t=j)} - \pi_j\right) = \sum_{t=0}^{T_j-1} \left(1_{(X_t=j)} - \pi_j\right) + \sum_{t=T_j}^{\infty} \left(1_{(X_t=j)} - \pi_j\right).$$

Take $E_i(\cdot)$ of each term to get $Z_{ij} = -\pi_j E_i T_j + Z_{jj}$. Of course this argument doesn't make sense because the sums do not converge. Implicit in our (honest) proof is a justification of this argument by a limiting procedure.

Example 16 *Patterns in coin-tossing.*

This is a classical example for which $\mathbf{Z}$ is easy to calculate. Fix n. Toss a fair coin repeatedly, and let $X_0, X_1, X_2, \ldots$ be the successive overlapping n-tuples. For example (with $n = 4$)

tosses	H	T	H	H	T	T
$X_0 =$	H	T	H	H		
$X_1 =$		T	H	H	T	
$X_2 =$			H	H	T	T

So $\mathbf{X}$ is a Markov chain on the set $I = \{H, T\}^n$ of n-tuples $i = (i_1, \ldots, i_n)$, and the stationary distribution π is uniform on I. For $0 \leq d \leq n-1$ write

$I(i,j,d)$ for the indicator of the set "pattern j, shifted right by d places, agrees with pattern i where they overlap": formally, of the set

$$j_u = i_{u+d},\ 1 \le u \le n-d.$$

For example, with $n = 4$, $i = HHTH$ and $j = HTHH$,

d	0	1	2	3
$I(i,j,d)$	0	1	0	1

Then write

$$c(i,j) = \sum_{d=0}^{n-1} 2^{-d} I(i,j,d).$$

From the definition of $\mathbf{Z}$, and the fact that X_0 and X_t are independent for $t \ge n$,

$$Z_{ij} = c(i,j) - n2^{-n}.$$

So we can read off many facts about patterns in coin-tossing from the general results of this section. For instance, the mean hitting time formula (Lemma 11) says $E_\pi T_i = 2^n c(i,i) - n$. Note that "time 0" for the chain is the n'th toss, at which point the chain is in its stationary distribution. So the mean number of tosses until first seeing pattern i equals $2^n c(i,i)$. For $n = 5$ and $i = HHTHH$, the reader may check this mean number is 38. We leave the interested reader to explore further — in particular, find three patterns i, j, k such that

$$P(\text{pattern } i \text{ occurs before pattern } j) > 1/2$$

$$P(\text{pattern } j \text{ occurs before pattern } k) > 1/2$$

$$P(\text{pattern } k \text{ occurs before pattern } i) > 1/2.$$

Further results. One can of course obtain expressions in the spirit of Lemmas 5–15 for more complicated quantities. The reader may care to find expressions for

$$E_i \min(T_k, T_l)$$

$$E_i(\text{number of visits to } j \text{ before time } \min(T_k, T_l))$$

$$P_i(\text{hit } j \text{ before time } \min(T_k, T_l)).$$

Warning. Hitting times T_A on subsets will be studied later (e.g. Chapter 3 section 5.3) (yyy 9/2/94 version) in the reversible setting. It is important

to note that results often do not extend simply from singletons to subsets. For instance, one might guess that Lemma 11 could be extended to

$$E_\pi T_A = \frac{Z_{AA}}{\pi(A)}, \quad Z_{AA} := \sum_{t=0}^{\infty} (P_\pi(X_t \in A | X_0 \in A) - \pi(A)),$$

but it is easy to make examples where this is false.

2.3 Continuous-time versions

Here we record the continuous-time versions of the results of the previous section. Write

$$Z_{ij} = \int_0^\infty (P_i(X_t = j) - \pi_j) dt \tag{9}$$

This is consistent with (6) in that $\mathbf{Z}$ is the same for a discrete-time chain and its continuized chain. Recall from section 1.2 the redefinition (b) of T_i^+ in continuous time. In place of "number of visits to i" we use "total duration of time spent in i". With this substitution, Proposition 3 and the other results of the previous section extend to continuous time with only the following changes, which occur because the mean sojourn time in a state i is $1/q_i$ in continuous time, rather than 1 as in discrete time.

Lemma 5. $E_i T_i^+ = \frac{1}{q_i \pi_i}$.

Lemma 6.

$$E_i(\text{duration of time spent in } j \text{ before time } T_i^+) = \frac{\pi_j}{q_i \pi_i}.$$

Corollary 8. For $j \neq i$,

$$P_i(T_j < T_i^+) = \frac{1}{q_i \pi_i (E_i T_j + E_j T_i)}.$$

3 Variances of sums

In discrete time, consider the number $N_i(t)$ of visits to state i before time t. (Recall our convention is to count a visit at time 0 but not at time t.) For the *stationary* chain, we have (trivially)

$$E_\pi N_i(t) = t\pi_i.$$

It's not hard to calculate the variance:

$$\begin{aligned} \text{var}\ {}_\pi N_i(t) &= \sum_{r=0}^{t-1}\sum_{s=0}^{t-1}(P_\pi(X_r = i, X_s = i) - \pi_i^2) \\ &= \pi_i\left(\sum_{u=0}^{t-1} 2(t-u)(p_{ii}^{(u)} - \pi_i) - t(1-\pi_i)\right) \end{aligned}$$

setting $u = |s - r|$. This leads to the asymptotic result

$$\frac{\text{var}\ {}_\pi N_i(t)}{t} \to \pi_i(2Z_{ii} - 1 + \pi_i). \tag{10}$$

The fundamental matrix $\mathbf{Z}$ of (6) reappears in an apparently different context. Here is the more general result underlying (10). Take arbitrary functions $f : I \to R$ and $g : I \to R$ and center so that $E_\pi f(X_0) := \sum_i \pi_i f(i) = 0$ and $E_\pi g(X_0) = 0$. Write

$$S_t^f = \sum_{s=0}^{t-1} f(X_s)$$

and similarly for S_t^g. Then

$$E_\pi S_t^f S_t^g = \sum_i \sum_j f(i)g(j) \sum_{r=0}^{t-1}\sum_{s=0}^{t-1}(P_\pi(X_r = i, X_s = j) - \pi_i\pi_j).$$

The contribution to the latter double sum from terms $r \le s$ equals, putting $u = s - r$,

$$\pi_i \sum_{u=0}^{t-1}(t-u)(p_{ij}^{(u)} - \pi_j) \sim t\pi_i Z_{i,j}.$$

Collecting the other term and subtracting the twice-counted diagonal leads to the following result.

$$\frac{E_\pi S_t^f S_t^g}{t} \to f\Gamma g := \sum_i \sum_j f(i)\Gamma_{ij} g(j) \tag{11}$$

where Γ is the symmetric positive-definite matrix

$$\Gamma_{ij} := \pi_i Z_{ij} + \pi_j Z_{ji} + \pi_i\pi_j - \pi_i\delta_{ij}. \tag{12}$$

As often happens, the formulas simplify in continuous time. The asymptotic result (10) becomes

$$\frac{\text{var}\ {}_\pi N_i(t)}{t} \to 2\pi_i Z_{ii}$$

and the matrix Γ occurring in (11) becomes

$$\Gamma_{ij} := \pi_i Z_{ij} + \pi_j Z_{ji}.$$

Of course these asymptotic variances appear in the central limit theorem for Markov chains.

Theorem 17 *For centered f,*

$$t^{-1/2} S_t^f \quad \stackrel{d}{\to} \quad \text{Normal}(0, f\Gamma f) \text{ as } t \to \infty.$$

The standard proofs (e.g. [6] p. 378) don't yield any useful finite-time results, so we won't present a proof. We return to this subject in Chapter 4 section 4.1 (yyy 10/11/94 version) in the context of reversible chains. In that context, getting finite-time bounds on the approximation (10) for variances is not hard, but getting informative finite-time bounds on the Normal approximation remains quite hard.

Remark. Here's another way of seeing why asymptotic variances should relate (via $\mathbf{Z}$) to mean hitting times. Regard $N_i(t)$ as counts in a renewal process; in the central limit theorem for renewal counts ([6] Exercise 2.4.13) the variance involves the variance var $_i(T_i^+)$ of the inter-renewal time, and by (22) below this in turn relates to $E_\pi T_i$.

4 Two metrics on distributions

A major theme of this book is quantifying the convergence theorem (Theorem 2) to give estimates of how close the distribution of a chain is to the stationary distribution at *finite* times. Such quantifications require some explicit choice of "distance" between distributions, and two of the simplest choices are explained in this section. We illustrate with a trivial

Example 18 *Rain or shine?*

Suppose the true probability of rain tomorrow is 80% whereas we think the probability is 70%. How far off are we? In other words, what is the "distance" between π and θ, where

$$\pi(\text{rain}) = 0.8, \ \pi(\text{shine}) = 0.2$$

$$\theta(\text{rain}) = 0.7, \ \theta(\text{shine}) = 0.3.$$

Different notions of distance will give different numerical answers. Our first notion abstracts the idea that the "additive error" in this example is $0.8 - 0.7 = 0.1$.

4.1 Variation distance

Perhaps the simplest notion of distance between probability distributions is *variation distance*, defined as

$$||\theta_1 - \theta_2|| = \max_{A \subseteq I} |\theta_1(A) - \theta_2(A)|.$$

So variation distance is just the maximum *additive* error one can make, in using the "wrong" distribution to evaluate the probability of an event. In example 18, variation distance is 0.1. Several equivalent definitions are provided by

Lemma 19 *For probability distributions θ_1, θ_2 on a finite state space I,*

$$\begin{aligned}
\frac{1}{2}\sum_i |\theta_1(i) - \theta_2(i)| &= \sum_i (\theta_1(i) - \theta_2(i))^+ \\
&= \sum_i (\theta_1(i) - \theta_2(i))^- \\
&= 1 - \sum_i \min(\theta_1(i), \theta_2(i)) \\
&= \max_{A \subseteq I} |\theta_1(A) - \theta_2(A)| \\
&= \min P(V_1 \neq V_2)
\end{aligned}$$

the minimum taken over random pairs (V_1, V_2) such that V_m has distribution θ_m $(m = 1, 2)$. So each of these quantities equals the variation distance $||\theta_1 - \theta_2||$.

Proof. The first three equalities are clear. For the fourth, set $B = \{i : \theta_1(i) > \theta_2(i)\}$. Then

$$\begin{aligned}
\theta_1(A) - \theta_2(A) &= \sum_{i \in A} (\theta_1(i) - \theta_2(i)) \\
&\leq \sum_{i \in A \cap B} (\theta_1(i) - \theta_2(i)) \\
&\leq \sum_{i \in B} (\theta_1(i) - \theta_2(i)) \\
&= \sum_i (\theta_1(i) - \theta_2(i))^+
\end{aligned}$$

with equality when $A = B$. This, and the symmetric form, establish the fourth equality. In the final equality, the "$\leq$" follows from

$$|\theta_1(A) - \theta_2(A)| = |P(V_1 \in A) - P(V_2 \in A)| \leq P(V_2 \neq V_1).$$

And equality is attained by the following joint distribution. Let $\theta(i) = \min(\theta_1(i), \theta_2(i))$ and let

$$P(V_1 = i, V_2 = i) = \theta(i)$$

$$P(V_1 = i, V_2 = j) = \frac{(\theta_1(i) - \theta(i))(\theta_2(j) - \theta(j))}{1 - \sum_k \theta(k)}, \ i \neq j.$$

(If the denominator is zero, then $\theta_1 = \theta_2$ and the result is trivial.) $\square$

In the context of Markov chains we may use

$$d_i(t) := ||P_i(X_t = \cdot) - \pi(\cdot)|| \tag{13}$$

as a measure of deviation from stationarity at time t, for the chain started at state i. Also define

$$d(t) := \max_i d_i(t) \tag{14}$$

as the worst-case deviation from stationarity. Finally, it is technically convenient to introduce also

$$\bar{d}(t) = \max_{ij} ||P_i(X_t = \cdot) - P_j(X_t = \cdot)||. \tag{15}$$

In Chapter 4 we discuss, for reversible chains, relations between these "variation distance" notions and other measures of closeness-to-stationarity, and discuss parameters τ measuring "time until $d(t)$ becomes small" and their relation to other parameters of the chain. For now, let's just introduce a fundamental technical fact, the submultiplicativity property.

Lemma 20
(a) $\bar{d}(s+t) \leq \bar{d}(s)\bar{d}(t), \ s,t \geq 0$ **[the submultiplicativity property]**.
(b) $d(s+t) \leq 2d(s)d(t), \ s,t \geq 0$.
(c) $d(t) \leq \bar{d}(t) \leq 2d(t), \ t \geq 0$.
(d) $d(t)$ *and* $\bar{d}(t)$ *decrease as* t *increases.*

Proof. We use the characterization of variation distance as

$$||\theta_1 - \theta_2|| = \min P(V_1 \neq V_2), \tag{16}$$

the minimum taken over random pairs (V_1, V_2) such that V_m has distribution θ_m $(m = 1, 2)$.

Fix states i_1, i_2 and times s, t, and let $\mathbf{Y}^1, \mathbf{Y}^2$ denote the chains started at i_1, i_2 respectively. By (16) we can construct a joint distribution for (Y_s^1, Y_s^2) such that

$$\begin{aligned} P(Y_s^1 \neq Y_s^2) &= ||P_{i_1}(X_s = \cdot) - P_{i_2}(X_s = \cdot)|| \\ &\leq \bar{d}(s). \end{aligned}$$

Now for each pair (j_1, j_2), we can use (16) to construct a joint distribution for (Y_{s+t}^1, Y_{s+t}^2) given $(Y_s^1 = j_1, Y_s^2 = j_2)$ with the property that

$$P(Y_{s+t}^1 \neq Y_{s+t}^2 | Y_s^1 = j_1, Y_s^2 = j_2) = ||P_{j_1}(X_t = \cdot) - P_{j_2}(X_t = \cdot)||.$$

The right side is 0 if $j_1 = j_2$, and otherwise is at most $\bar{d}(t)$. So unconditionally

$$P(Y_{s+t}^1 \neq Y_{s+t}^2) \leq \bar{d}(s)\bar{d}(t)$$

and (16) establishes part (a) of the lemma. For part (b), the same argument (with $\mathbf{Y}^2$ now being the stationary chain) shows

$$d(s+t) \leq d(s)\bar{d}(t) \tag{17}$$

so that (b) will follow from the upper bound $\bar{d}(t) \leq 2d(t)$ in (c). But this upper bound is clear from the triangle inequality for variation distance. And the lower bound in (c) follows from the fact that $\mu \to ||\theta - \mu||$ is a convex function, so that averaging over j with respect to π in (15) can only decrease distance. Finally, the "decreasing" property for $\bar{d}(t)$ follows from (a), and for $d(t)$ follows from (17). □

The assertions of this section hold in either discrete or continuous time. But note that the numerical value of $d(t)$ changes when we switch from a discrete-time chain to the continuized chain. In particular, for a discrete-time chain with period q we have $d(t) \to (q-1)/q$ as $t \to \infty$ (which incidently implies, taking $q = 2$, that the factor 2 in Lemma 20(b) cannot be reduced) whereas for the continuized chain $d(t) \to 0$.

One often sees slightly disguised corollaries of the submultiplicativity property in the literature. The following is a typical one.

Corollary 21 *Suppose there exists a probability measure μ, a real $\delta > 0$ and a time t such that*

$$p_{ij}^{(t)} \geq \delta \mu_j \ \forall i, j.$$

Then

$$d(s) \leq (1-\delta)^{\lfloor s/t \rfloor}, \ s \geq 0.$$

Proof. The hypothesis implies $\bar{d}(t) \leq 1 - \delta$, by the third equality in Lemma 19, and then the conclusion follows by submultiplicativity.

4.2 L^2 distance

Another notion of distance, which is less intuitively natural but often more mathematically tractable, is L^2 distance. This is defined with respect to some fixed reference probability distribution π on I, which for our purposes will be the stationary distribution of some irreducible chain under consideration (and so $\pi_i > 0 \ \forall i$). The L^2 norm of a *function* $f : I \to R$ is

$$||f||_2 = \sqrt{\sum_i \pi_i f^2(i)}. \tag{18}$$

We define the L^2 norm of a *signed measure* ν on I by

$$||\nu||_2 = \sqrt{\sum_i \nu_i^2/\pi_i}. \tag{19}$$

This may look confusing, because a signed measure ν and a function f are in a sense "the same thing", being determined by values $(f(i); i \in I)$ or $(\nu_i; i \in I)$ which can be chosen arbitrarily. But the measure ν can also be determined by its density function $f(i) = \nu_i/\pi_i$, and so (18) and (19) say that the L^2 norm of a signed measure is defined to be the L^2 norm of its density function.

So $||\theta - \mu||_2$ is the "L^2" measure of distance between probability distributions θ, μ. In particular, the distance between θ and the reference distribution π is

$$||\theta - \pi||_2 = \sqrt{\sum_i \frac{(\theta_i - \pi_i)^2}{\pi_i}} = \sqrt{\sum_i \frac{\theta_i^2}{\pi_i} \quad - 1}.$$

In Example 18 we find $||\theta - \pi||_2 = 1/4$.

Writing $\theta(t)$ for the distribution at time t of a chain with stationary distribution π, it is true (cf. Lemma 20(d) for variation distance) that $||\theta(t) - \pi||_2$ is decreasing with t. Since there is a more instructive proof in the reversible case (Chapter 3 Lemma 23) (yyy 9/2/94 version) we won't prove the general case (see Notes).

Analogous to the L^2 norms are the L^1 norms

$$||f||_1 = \sum_i \pi_i |f(i)|$$

$$||\nu||_1 = \sum_i |\nu_i|.$$

The Cauchy-Schwarz inequality gives $||\cdot||_1 \le ||\cdot||_2$. Note that in the definition of $||\nu||_1$ the reference measure π has "cancelled out". Lemma 19 shows that for probability measures θ_1, θ_2 the L^1 distance is the same as variation distance, up to a factor of 2:

$$||\theta_1 - \theta_2|| = \tfrac{1}{2}||\theta_1 - \theta_2||_1.$$

As a trivial example in the Markov chain setting, consider

Example 22 *Take $I = \{0, 1, \ldots, n-1\}$, fix a parameter $0 < a < 1$ and define a transition matrix*

$$p_{ij} = a 1_{(j = i+1 \bmod n)} + \frac{1-a}{n}.$$

In this example the t-step transition probabilities are

$$p_{ij}^{(t)} = a^t 1_{(j = i+t \bmod n)} + \frac{1-a^t}{n}$$

and the stationary distribution π is uniform. We calculate (for arbitrary $j \ne i$)

$$d(t) = ||P_i(X_t \in \cdot) - \pi|| = (1 - n^{-1})a^t$$
$$\bar{d}(t) = ||P_i(X_t \in \cdot) - P_j(X_t \in \cdot)|| = a^t$$
$$||P_i(X_t \in \cdot) - \pi||_2 = (n-1)^{1/2} a^t.$$

4.3 Exponential tails of hitting times

The submultiplicative property of $\bar{d}(t)$ is one instance of a general principle:

> because our state space is *finite*, many quantities which converge to zero as $t \to \infty$ must converge exponentially fast, by iterating over worst-case initial states.

Here's another instance, tails of hitting time distributions.

Consider the first hitting time T_A on a subset A. Define $t_A^* := \max_i E_i T_A$. For any initial distribution μ, any time $s > 0$ and any integer $m \ge 1$,

$$\begin{aligned} P_\mu(T_A > ms | T_A > (m-1)s) &= P_\theta(T_A > s) \text{ for some dist. } \theta \\ &\le \max_i P_i(T_A > s) \\ &\le t_A^*/s. \end{aligned}$$

So by induction on m

$$P_\mu(T_A > js) \le (t_A^*/s)^j$$

implying

$$P_\mu(T_A > t) \le (t_A^*/s)^{\lfloor t/s \rfloor}, \quad t > 0.$$

In continuous time, a good (asymptotically optimal) choice of s is $s = et_A^*$, giving the exponential tail bound

$$\sup_\mu P_\mu(T_A > t) \le \exp\left(-\left\lfloor \frac{t}{et_A^*} \right\rfloor\right), \quad 0 < t < \infty. \tag{20}$$

A messier bound holds in discrete time, where we have to choose s to be an integer.

5 Distributional identities

It is much harder to get useful information about distributions (rather than mere expectations). Here are a few general results.

5.1 Stationarity consequences

A few useful facts about stationary Markov chains are, to experts, just specializations of facts about arbitrary (i.e. not-necessarily-Markov) stationary processes. Here we give a bare-hands proof of one such fact, the relation between the distribution of return time to a subset A and the distribution of first hitting time to A from a stationary start. We start in discrete time.

Lemma 23 *For $t = 1, 2, \ldots$,*

$$P_\pi(T_A = t - 1) = P_\pi(T_A^+ = t) = \pi(A) P_{\pi_A}(T_A^+ \ge t)$$

where $\pi_A(i) := \pi_i/\pi(A)$, $i \in A$.

Proof. The first equality is obvious. Now let (X_t) be the chain started with its stationary distribution π. Then

$$\begin{aligned}
P_\pi(T_A^+ = t) &= P(X_1 \notin A, \ldots, X_{t-1} \notin A, X_t \in A) \\
&= P(X_1 \notin A, \ldots, X_{t-1} \notin A) - P(X_1 \notin A, \ldots, X_t \notin A) \\
&= P(X_1 \notin A, \ldots, X_{t-1} \notin A) - P(X_0 \notin A, \ldots, X_{t-1} \notin A) \\
&= P(X_0 \in A, X_1 \notin A, \ldots, X_{t-1} \notin A) \\
&= \pi(A) P_{\pi_A}(T_A^+ \ge t),
\end{aligned}$$

establishing the Lemma.

We'll give two consequences of Lemma 23. Summing over t gives

Corollary 24 (Kac's formula) $\pi(A)E_{\pi_A}T_A^+ = 1$

which extends the familiar fact $E_iT_i^+ = 1/\pi_i$. Multiplying the identity of Lemma 23 by t and summing gives

$$\begin{aligned} E_\pi T_A + 1 &= \sum_{t\geq 1} tP_{\pi_A}(T_A = t-1) \\ &= \pi(A)\sum_{t\geq 1} tP_{\pi_A}(T_A^+ \geq t) \\ &= \pi(A)\sum_{m\geq 1} \frac{1}{2}m(m+1)P_{\pi_A}(T_A^+ = m) \\ &= \frac{\pi(A)}{2}\left(E_{\pi_A}T_A^+ + E_{\pi_A}(T_A^+)^2\right). \end{aligned}$$

Appealing to Kac's formula and rearranging,

$$E_{\pi_A}(T_A^+)^2 = \frac{2E_\pi T_A + 1}{\pi(A)}, \tag{21}$$

$$\mathrm{var}_{\pi_A}(T_A^+) = \frac{2E_\pi T_A + 1}{\pi(A)} - \frac{1}{\pi^2(A)}. \tag{22}$$

More generally, there is a relation between $E_{\pi_A}(T_A^+)^p$ and $E_\pi(T_A^+)^{p-1}$.

In continuous time, the analog of Lemma 23 is

$$P_\pi(T_A \in (t, t+dt)) = Q(A, A^c)P_{\rho_A}(T_A > t)dt,\ t > 0 \tag{23}$$

where

$$Q(A, A^c) := \sum_{i\in A}\sum_{j\in A^c} q_{ij}, \quad \rho_A(j) := \sum_{i\in A} q_{ij}/Q(A, A^c), j \in A^c.$$

Integrating over $t > 0$ gives the analog of Kac's formula

$$Q(A, A^c)E_{\rho_A}T_A = \pi(A^c). \tag{24}$$

5.2 A generating function identity

Transform methods are useful in analyzing special examples, though that is not the main focus of this book. We record below just the simplest "transform fact". We work in discrete time and use generating functions – the corresponding result in continuous time can be stated using Laplace transforms.

Lemma 25 *Define*

$$G_{ij}(z) = \sum_{t \geq 0} P_i(X_t = j) z^t, \;\; F_{ij}(z) = \sum_{t \geq 0} P_i(T_j = t) z^t.$$

Then $F_{ij} = G_{ij}/G_{jj}$.

Analysis proof. Conditioning on T_j gives

$$p_{ij}^{(t)} = \sum_{l=0}^{t} P_i(T_j = l) p_{jj}^{(t-l)}$$

and so

$$\sum_{t \geq 0} p_{ij}^{(t)} z^t = \sum_{l \geq 0} \sum_{t-l \geq 0} P_i(T_j = l) z^l p_{jj}^{(t-l)} z^{t-l}$$

Thus $G_{ij}(z) = F_{ij}(z) G_{jj}(z)$, and the lemma follows. □

Probability proof. Let ζ have geometric(z) law $P(\zeta > t) = z^t$, independent of the chain. Then

$$\begin{aligned} G_{ij}(z) &= E_i(\text{number of visits to } j \text{ before time } \zeta) \\ &= P_i(T_j < \zeta) \; E_j(\text{number of visits to } j \text{ before time } \zeta) \\ &= F_{ij}(z) G_{jj}(z). \end{aligned}$$

□

Note that, differentiating term by term,

$$E_i T_j = \tfrac{d}{dz} F_{ij}(z) \Big|_{z=1} .$$

This and Lemma 25 can be used to give an alternative derivation of the mean hitting time formula, Lemma 12.

5.3 Distributions and continuization

The distribution at time t of the continuization $\hat{X}$ of a discrete-time chain X is most simply viewed as a Poisson mixture of the distributions (X_s). That is, $\hat{X}_t \stackrel{d}{=} X_{N_t}$ where N_t has Poisson(t) distribution independent of X. At greater length,

$$P_i(\hat{X}_t = j) = \sum_{s=0}^{\infty} \frac{e^{-t} t^s}{s!} P_i(X_s = j).$$

This holds because we can construct $\hat{X}$ from X by replacing the deterministic "time 1" holds by random, exponential(1), holds (ξ_j) between jumps, and then the number N_t of jumps before time t has Poisson(t) distribution. Now write $S_n = \sum_{j=1}^{n} \xi_j$ for the time of the n'th jump. Then the hitting time $\hat{T}_A$ for the continuized chain is related to the hitting time T_A of the discrete-time chain by $\hat{T}_A = S_{T_A}$. Though these two hitting time distributions are different, their expectations are the same, and their variances are related in a simple way. To see this, the conditional distribution of $\hat{T}_A$ given T_A is the distribution of the sum of T_A independent ξ's, so (using the notion of conditional expectation given a random variable)

$$E(\hat{T}_A|T_A) = T_A, \text{ var } (\hat{T}_A|T_A) = T_A.$$

Thus (for any initial distribution)

$$E\hat{T}_A = EE(\hat{T}_A|T_A) = ET_A.$$

And the *conditional variance formula* ([6] p. 198)

$$\text{var } Z = E \text{ var } (Z|Y) + \text{var } E(Z|Y)$$

tells us that

$$\begin{aligned} \text{var } \hat{T}_A &= E\text{var } (\hat{T}_A|T_A) + \text{var } E(\hat{T}_A|T_A) \\ &= ET_A + \text{var } T_A. \end{aligned} \tag{25}$$

6 Matthews' method for cover times

Theorem 26 below is the only non-classical result in this Chapter. We make extensive use of this *Matthews' method* in Chapter 6 to analyze cover times for random walks on graphs.

Consider the *cover time* $C := \max_j T_j$ of the chain, i.e. the time required to visit every state. How can we bound E_iC in terms of the mean hitting times E_iT_j? To appreciate the cleverness of Theorem 26 let us first consider a more routine argument. Write $t^* := \max_{i,j} E_iT_j$. Since E_iC is unaffected by continuization, we may work in continuous time. By (20)

$$P_i(T_j > ket^*) \le e^{-k}, \ k = 1, 2, 3, \ldots.$$

By Boole's inequality, for an n-state chain

$$P_i(C > ket^*) \le ne^{-k}, \ k = 1, 2, 3, \ldots.$$

One can rewrite this successively as

$$\begin{aligned} P_i\left(\frac{C}{et^*} > x\right) &\leq ne \cdot e^{-x}, \quad 0 \leq x < \infty \\ P_i\left(\frac{C}{et^*} - \log(en) > x\right) &\leq e^{-x}, \quad 0 \leq x < \infty. \end{aligned}$$

In words, this says that the distribution of $\frac{C}{et^*} - \log(en)$ is stochastically smaller that the exponential(1) distribution, implying $E_i\left(\frac{C}{et^*} - \log(en)\right) \leq 1$ and hence

$$\max_i E_i C \leq (2 + \log n) et^*.$$

This argument does lead to a bound, but one suspects the factors 2 and e are artifacts of the proof; also, it seems hard to obtain a lower bound this way. The following result both "cleans up" the upper bound and gives a lower bound.

Theorem 26 (Matthews [20]) *For any n-state Markov chain,*

$$\max_v E_v C \leq h_{n-1} \max_{i,j} E_i T_j$$

$$\min_v E_v C \geq h_{n-1} \min_{i \neq j} E_i T_j$$

where $h_{n-1} := \sum_{m=1}^{n-1} \frac{1}{m}$.

Proof. We'll prove the lower bound — the upper bound proof is identical. Let $J_1, J_2, \ldots, J_n$ be a uniform random ordering of the states, independent of the chain. Define $C_m := \max_{i \leq m} T_{J_i}$ to be the time until all of $\{J_1, J_2, \ldots, J_m\}$ have been visited, in some order. The key identity is

$$E(C_m - C_{m-1} | J_1, \ldots, J_m; X_t, t \leq C_{m-1}) = t(L_{m-1}, J_m) 1_{(L_m = J_m)} \qquad (26)$$

where $t(i,j) := E_i T_j$ and

$$L_m \text{ is the state amongst } \{J_1, J_2, \ldots, J_m\} \text{ hit last.}$$

To understand what this says, suppose we are told which are the states $\{J_1, J_2, \ldots, J_m\}$ and told the path of the chain up through time C_{m-1}. Then we know whether or not $L_m = J_m$: if not, then $C_m = C_{m-1}$, and if so, then the conditional distribution of $C_m - C_{m-1}$ is the distribution of the time to hit J_m from the state at time C_{m-1}, which we are told is state L_{m-1}.

Writing $t_* := \min_{i \neq j} t(i,j)$, the right side of (26) is $\geq t_* 1_{(L_m = J_m)}$, and so taking expectations

$$E(C_m - C_{m-1}) \geq t_* P(L_m = J_m).$$

But obviously $P(L_m = J_m) = 1/m$ by symmetry. So

$$E_v C = E_v C_1 + \sum_{m=2}^{n} E_v(C_m - C_{m-1}) \geq E_v C_1 + t_* \sum_{m=2}^{n} \frac{1}{m}.$$

Allowing for the possibility $J_1 = v$ we see $E_v C_1 \geq (1 - \frac{1}{n}) t_*$, and the lower bound follows.

7 New chains from old

Consider a chain (X_t) on state-space I, and fix $A \subseteq I$. There are many different constructions of new chains whose state space is (exactly or roughly) just A, and it's important not to confuse them. Three elementary constructions are described here. Anticipating the definition of *reversible* from Chapter 3, it is easy to check that if the original chain is reversible then each new chain is reversible.

7.1 The chain watched only on A

This is the chain (Y_n) defined by

$$S_0 = T_A = \min\{t \geq 0 : X_t \in A\}$$

$$S_n = \min\{t > S_{n-1} : X_t \in A\}$$

$$Y_n = X_{S_n}.$$

The chain (Y_n) has state space A and transition matrix

$$\bar{P}_A(i,j) = P_i(X_{T_A} = j), \ i,j \in A.$$

From the ergodic theorem (Theorem 1) it is clear that the stationary distribution π_A of (Y_t) is just π conditioned on A, that is

$$\pi_A(i) = \pi_i / \pi(A), \ i \in A. \tag{27}$$

7.2 The chain restricted to A

This is the chain with state space A and transition matrix $\hat{P}_A$ defined by

$$\begin{aligned} \hat{P}_A(i,j) &= P(i,j),\ i,j \in A, i \neq j \\ \hat{P}_A(i,i) &= 1 - \sum_{j \in A, j \neq i} P(i,j),\ i \in A. \end{aligned}$$

In general there is little connection between this chain and the original chain (X_t), and in general it is not true that the stationary distribution is given by (27). However, when the original chain is reversible, it is easy to check that the restricted chain does have the stationary distribution (27).

7.3 The collapsed chain

This chain has state space $I^* = A \cup \{a\}$ where a is a new state. We interpret the new chain as "the original chain with states A^c collapsed to a single state a". *Warning. In later applications we switch the roles of A and A^c, i.e. we collapse A to a single state a and use the collapsed chain on states $I^* = A^c \cup \{a\}$.* The collapsed chain has transition matrix

$$\begin{aligned} p^*_{ij} &= p_{ij},\ i,j \in A \\ p^*_{ia} &= \sum_{k \in A^c} p_{ik},\ i \in A \\ p^*_{ai} &= \frac{1}{\pi(A^c)} \sum_{k \in A^c} \pi_k p_{ki},\ i \in A \\ p^*_{aa} &= \frac{1}{\pi(A^c)} \sum_{k \in A^c} \sum_{l \in A^c} \pi_k p_{kl}. \end{aligned}$$

The collapsed chain has stationary distribution π^* given by

$$\pi^*_i = \pi_i, i \in A; \quad \pi^*_a = \pi(A^c).$$

Obviously the $\mathbf{P}$-chain started at i and run until T_{A^c} is the same as the $\mathbf{P}^*$-chain started at i and run until T_a. This leads to the general *collapsing principle*

> To prove a result which involves the behavior of the chain only up to time T_{A^c}, we may assume A^c is a singleton.

For we may apply the singleton result to the $\mathbf{P}^*$-chain run until time T_a, and the same result will hold for the $\mathbf{P}$-chain run until time T_{A^c}.

It is important to realize that typically (even for reversible chains) all three constructions give different processes. Loosely, the chain restricted to A "rebounds off the boundary of A^c where the boundary is hit", the collapsed chain "exits A^c at a random place independent of the hitting place", and the chain watched only on A "rebounds at a random place *dependent* on the hitting place".

8 Miscellaneous methods

8.1 Martingale methods

Modern probabilists regard the martingale optional stopping theorem as one of the most important results in their subject. As propaganda for martingales we give below four quick applications of that theorem, and a few more will appear later. All of these results could be proved in alternative, elementary ways. For the reader unfamiliar with martingales, Durrett [6] Chapter 4 contains much more than you need to know: Karlin and Taylor [14] Chapter 6 is a gentler introduction.

Lemma 27 *Given a non-empty subset $A \subset I$ and a function $f(i)$ defined for $i \in A$, there exists a unique extension of f to all I satisfying*

$$f(i) = \sum_j p_{ij} f(j), \ i \notin A.$$

Proof. If f satisfies the equations above then for any initial distribution the process $M_t := f(X_{\min(t,T_A)})$ is a martingale. So by the optional stopping theorem

$$f(i) = E_i f(X_{T_A}) \text{ for all } i. \tag{28}$$

This establishes uniqueness. Conversely, if we define f by (28) then the desired equations hold by conditioning on the first step.

Corollary 28 *If h is* harmonic, *i.e. if*

$$h(i) = \sum_j p_{ij} h(j) \text{ for all } i$$

then h is constant.

Proof. Clearly a constant function is harmonic. So the Corollary follows from the uniqueness assertion of Lemma 27, taking A to be some singleton.

Lemma 29 (The random target lemma) *The sum $\sum_j E_i T_j \pi_j$ does not depend on i.*

Proof. This repeats Corollary 14 with a different argument. The first-step recurrence for $g_j(i) := E_i T_j$ is

$$g_j(i) = 1_{(i \neq j)} + 1_{(i \neq j)} \sum_k p_{ik} g_j(k).$$

By Corollary 28 it is enough to show that $h(i) := \sum_j \pi_j g_j(i)$ is a harmonic function. We calculate

$$\begin{aligned} h(i) &= 1 - \pi_i + \sum_{j,k} \pi_j p_{ik} g_j(k) 1_{(i \neq j)} \\ &= 1 - \pi_i + \sum_k p_{ik}(h(k) - \pi_i g_i(k)) \text{ by definition of } h(k) \\ &= \sum_k p_{ik} h(k) \quad + 1 - \pi_i \left(1 + \sum_k p_{ik} g_i(k)\right). \end{aligned}$$

But $1/\pi_i = E_i T_i^+ = 1 + \sum_k p_{ik} g_i(k)$, so h is indeed harmonic.

Lemma 30 *For any stopping time S and any states i, j, k,*

$$E_i(\textit{number of transitions } j \to k \textit{ starting before time } S)$$
$$= p_{jk} E_i(\textit{number of visits to } j \textit{ before time } S).$$

Proof. Recall that "before" means strictly before. The assertion of the lemma is intuitively obvious, because each time the chain visits j it has chance p_{jk} to make a transition $j \to k$, and one can formalize this as in the proof of Proposition 4. A more sophisticated proof is to observe that $M(t)$ is a martingale, where

$$M(t) := N_{jk}(t) - p_{jk} N_j(t).$$

$$N_j(t) := \text{ number of visits to } j \text{ before time } t$$
$$N_{jk}(t) := \text{ number of transitions } j \to k \text{ starting before time } t \ .$$

And the assertion of the lemma is just the optional stopping theorem applied to the martingale M and the stopping time S.

Lemma 31 *Let A be a non-empty subset of I and let $h : I \to R$ satisfy*
(i) $h(i) \geq 0,\ i \in A$
(ii) $h(i) \geq 1 + \sum_j p_{ij} h(j),\ i \in A^c$.
Then $E_i T_A \leq h(i),\ i \in I$.

Proof. For arbitrary h, define g by

$$h(i) = 1 + \sum_j p_{ij} h(j) + g(i)$$

and then define

$$M_t = t + h(X_t) + \sum_{s=0}^{t-1} g(X_s).$$

Then $M_{\min(t,T_A)}$ is a martingale, so the optional sampling theorem says

$$E_i M_{T_A} = E_i M_0 = h(i).$$

But the hypotheses on h imply $M_{T_A} \geq T_A$.

8.2 A comparison argument

A theme running throughout the book is the idea of getting inequalities for a "hard" chain by making a comparison with some "easier" chain for which we can do exact calculations. Here is a simple example.

Lemma 32 *Let X be a discrete-time chain on states $\{0, 1, 2, \ldots, n\}$ such that $p_{ij} = 0$ whenever $j > i$. Write $m(i) = i - E_i X_1$, and suppose $0 < m(1) \leq m(2) \leq \ldots \leq m(n)$. Then $E_n T_0 \leq \sum_{j=1}^{n} \frac{1}{m(j)}$.*

Proof. The proof implicitly compares the given chain to the continuous-time chain with $q_{i,i-1} = m(i)$. Write $h(i) = \sum_{j=1}^{i} 1/m(j)$, and extend h by linear interpolation to real $0 \leq x \leq n$. Then h is concave and for $i \geq 1$

$$\begin{aligned} E_i h(X_1) &\leq h(E_i X_1) \text{ by concavity} \\ &= h(i - m(i)) \\ &\leq h(i) - m(i) h'(i) \text{ by concavity} \\ &= h(i) - 1 \end{aligned}$$

where h' is the left derivative. The result now follows from Lemma 31.

8.3 Wald equations

As mentioned previously, the results above don't really require martingales. Next we record a genuine martingale result, not directly involving Markov chains but ultimately useful in their analysis. Part (c) is *Wald's equation* and part (b) is *Wald's equation for martingales.* The result is a standard consequence of the optional sampling theorem: see [6] (3.1.6) for (c) and [6] Theorem 4.7.5 for (a).

Lemma 33 *(a) Let* $0 = Y_0 \le Y_1 \le Y_2 \ldots$ *be such that*

$$E(Y_{i+1} - Y_i | Y_j, j \le i) \le c, \ i \ge 0$$

for a constant c*. Then for any stopping time* T*,*

$$EY_T \le cET.$$

(b) If in the hypothesis we replace "$\le c$*" by "*$= c$*", then* $EY_T = cET$*.*
(c) In particular, if $Y_n = \sum_{i=1}^{n} \xi_i$ *for i.i.d. nonnegative* (ξ_i) *then* $EY_T = (E\xi_i)(ET)$*.*

9 Notes on Chapter 2.

Textbooks on Markov chains.

> It is easy to write books on ...or finite Markov chains, or on any of the other well-understood topics for which no further expositions are needed. *G.-C. Rota*
>
> `Your search for the Subject: MARKOV PROCESSES retrieved 273 records.` *U.C. Berkeley Library book catalog, September 1999.*

Almost every introductory textbook on stochastic processes has a chapter or two about Markov chains: among the best are Karlin-Taylor [14, 15], Grimmett-Stirzaker [8] and, slightly more advanced, Asmussen [3]. In addition to Norris [23] there are several other undergraduate-level textbooks entirely or mostly devoted to Markov chains: Adke-Manjanuth [1], Hunter [9], Iosifescu [10], Isaacson-Madsen [11], Kemeny-Snell [17], Romanovsky [26]. At the graduate level, Durrett [6] has a concise chapter on the modern approach to the basic limit theory. Several more advanced texts which overlap our material were mentioned in Chapter 1 section 2.3 (yyy 7/20/99

version); other texts are Freedman [7], Anderson [2], and the treatize of Syski [30] on hitting times. Most textbooks leave an exaggerated impression of the difference between discrete- and continuous-time chains.

Section 1.2. Continuized is an ugly neologism, but no-one has collected my \$5 prize for suggesting a better name!

Section 2. Elementary matrix treatments of results like those in section 2.2 for finite state space can be found in [9, 17]. On more general spaces, this is part of *recurrent potential* theory: see [5, 18] for the countable-state setting and Revuz [25] for continuous space. Our treatment, somewhat novel at the textbook level, Pitman [24] studied occupation measure identities more general than thos in section 2.1 and their applications to hitting time formulas, and we follow his approach in sectionMHTF. We are being slightly dishonest in treating Lemmas 5 and 6 this way, because these facts figure in the "right" proof of the ergodic theorems we use. We made a special effort *not* to abbreviate "number of visits to j before time S" as $N_j(S)$, which forces the reader to decode formulas.

Kemeny and Snell [17] call $\mathbf{Z}$+Π the fundamental matrix, and use $(E_i T_j^+)$ rather than $(E_i T_j)$ as the matrix of mean hitting times. Our set-up seems a little smoother – cf. Meyer [13] who calls $\mathbf{Z}$ the *group inverse* of $\mathbf{I} - \mathbf{P}$.

The name "random target lemma" for Corollary 14 was coined by Lovász and Winkler [19]; the result itself is classical ([17] Theorem 4.4.10).

Open Problem 34 *Portmanteau theorem for occupation times.*

Can the results of section 2.2 be formulated as a single theorem? To explain the goal by analogy, consider the use [12] of *Feynman diagrams* to calculate quantities such as $E(A^3BC^2)$ for dependent mean-zero Gaussian (A, B, C). One rewrites the expectation as $E\prod_{i=1}^6 \xi_i$ for $\xi_1 = \xi_2 = \xi_3 = A; \xi_4 = B, \xi_5 = \xi_6 = C$, and then applies the formula

$$E\prod_{i=1}^{6} \xi_i = \sum_M \nu(M)$$

where the sum is over *matchings* $M = \{\{u_1, v_1\}, \{u_2, v_2\}, \{u_3, v_3\}\}$ of $\{1, 2, 3, 4, 5, 6\}$ and where

$$\nu(M) = \prod_{j=1}^{3} E(\xi_{u_j}\xi_{v_j}).$$

By analogy, we seek a general rule which associates an expression like

$$E_i(\text{number of visits to } j \text{ before time } \min(T_k, T_l))$$

with a combinatorial structure involving $\{i,j,k,l\}$; then associates with the combinatorial structure some function of variables $\{p_v, z_{vw},\ v,w \in \{i,j,k,l\}\}$; then shows that the value of the expression applied to a finite Markov chain equals the function of $\{\pi_v, Z_{vw},\ v,w \in \{i,j,k,l\}\}$.

Section 4.1. Corollary 21 and variants are the basis for the theory of positive-recurrent chains on continuous spaces: see [6] section 5.6 and Meyn and Tweedie [22].

Section 4.2. The fact that $||\theta(t) - \pi||_2$ is decreasing is a special case $(H(u) = u^2)$ of the following result (e.g. [16] Theorem 1.6).

Lemma 35 *Let $H : [0,\infty) \to [0,\infty)$ be concave [convex]. Let $\theta(t)$ be the distribution of an irreducible chain with stationary distribution π. Then $\sum_i \pi_i H(\theta_i(t)/\pi_i)$ is increasing [decreasing].*

Section 6. Matthews [20, 21] introduced his method (Theorem 26) to study some highly symmetric walks (cf. Chapter 7) and to study some continuous-space Brownian motion covering problems.

Section 7. A more sophisticated notion is "the chain conditioned never to hit A", which can be formalized using Perron-Frobenius theory.

Section 8.1. Applying the optional stopping theorem involves checking side conditions (involving integrability of the martingale or the stopping time), but these are trivially satisfied in our applications.

Numerical methods. In many applications of non-reversible chains, e.g. to queueing-type processes, one must resort to numerical computations of the stationary distribution: see Stewart [29]. We don't discuss such issues because in the reversible case we have conceptually simple expressions for the stationary distribution,

Matrix methods. There is a curious dichotomy between textbooks on Markov chains which use matrix theory almost everywhere and textbooks which use matrix theory almost nowhere. Our style is close to the latter; matrix formalism obscures more than it reveals. For our purposes, the one piece of matrix theory which is really essential is the spectral decomposition of reversible transition matrices in Chapter 3. Secondarily useful is the theory surrounding the Perron-Frobenius theorem, quoted for reversible chains in Chapter 3 section 6.5. (yyy 9/2/94 version)

References

[1] S.R. Adke and S.M. Manjunath. *An Introduction to Finite Markov Processes.* Wiley, 1984.

[2] W.J. Anderson. *Continuous-Time Markov Chains.* Springer–Verlag, 1991.

[3] S. Asmussen. *Applied Probability and Queues.* Wiley, 1987.

[4] J.R. Baxter and R.V. Chacon. Stopping times for recurrent Markov processes. *Illinois J. Math.*, 20:467–475, 1976.

[5] K.L. Chung. *Markov Chains with Stationary Transition Probabilities.* Springer–Verlag, second edition, 1967.

[6] R. Durrett. *Probability: Theory and Examples.* Wadsworth, Pacific Grove CA, 1991.

[7] D. Freedman. *Markov Chains.* Springer–Verlag, 1983. Reprint of 1971 Holden-Day edition.

[8] G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, 1982.

[9] J.J. Hunter. *Mathematical Techniques of Applied Probability.* Academic Press, 1983.

[10] M. Iosifescu. *Finite Markov Processes and Their Applications.* Wiley, 1980.

[11] D. Isaacson and R. Madsen. *Markov Chains: Theory and Applications.* Wiley, 1976.

[12] S. Janson. *Gaussian Hilbert Spaces.* Number 129 in Cambridge Tracts in Mathematics. Cambridge University Press, 1997.

[13] C.D. Meyer Jr. The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Review*, 17:443–464, 1975.

[14] S. Karlin and H.M. Taylor. *A First Course in Stochastic Processes.* Academic Press, 1975.

[15] S. Karlin and H.M. Taylor. *A Second Course in Stochastic Processes.* Academic Press, 1981.

[16] F.P. Kelly. *Reversibility and Stochastic Networks.* Wiley, 1979.

[17] J.G. Kemeny and J.L. Snell. *Finite Markov Chains.* Van Nostrand, 1960.

[18] J.G. Kemeny, J.L Snell, and A.W. Knapp. *Denumerable Markov Chains*. Springer–Verlag, 2nd edition, 1976.

[19] L. Lovász and P. Winkler. Efficient stopping rules for Markov chains. In *Proc. 27th ACM Symp. Theory of Computing*, pages 76–82, 1995.

[20] P.C. Matthews. Covering problems for Brownian motion on spheres. *Ann. Probab.*, 16:189–199, 1988.

[21] P.C. Matthews. Covering problems for Markov chains. *Ann. Probab.*, 16:1215–1228, 1988.

[22] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer–Verlag, 1993.

[23] J.R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[24] J.W. Pitman. Occupation measures for Markov chains. *Adv. in Appl. Probab.*, 9:69–86, 1977.

[25] D. Revuz. *Markov Chains*. North-Holland, second edition, 1984.

[26] V. I. Romanovsky. *Discrete Markov Chains*. Wolthers-Noordhoff, 1970. English translation of Russian original.

[27] S. Ross. *Stochastic Processes*. Wiley, 1983.

[28] H. Rost. The stopping distributions of a Markov process. *Inventiones Math.*, 14:1–16, 1971.

[29] W.J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1995.

[30] R. Syski. *Passage Times for Markov Chains*. IOS Press, Amsterdam, 1992.

yyy move both subsections to Chapter 8 "A Second Look ...".

10 Move to other chapters

10.1 Attaining distributions at stopping times

We quote a result, Theorem 36, which may look superficially like the identities in section 2.1 but which in fact is deeper, in that it cannot be proved by mere matrix manipulations or by Proposition 3. The result goes back to Baxter and Chacon [4] (and is implicit in Rost [28]) in the more general continuous-space setting: a proof tailored to the finite state space case has recently been given by Lovász and Winkler [19].

Given distributions σ, μ, consider a stopping time T such that

$$P_\sigma(X_T \in \cdot) = \mu(\cdot). \tag{29}$$

Clearly, for any state j we have $E_\sigma T_j \le E_\sigma T + E_\mu T_j$, which rearranges to $E_\sigma T \ge E_\sigma T_j - E_\mu T_j$. So if we define

$$\bar{t}(\sigma, \mu) = \inf\{E_\sigma T : \ T \text{ a stopping time satisfying (29)}\}$$

then we have shown that $\bar{t}(\sigma, \mu) \ge \max_j(E_\sigma T_j - E_\mu T_j)$. Surprisingly, this inequality turns out to be an equality.

Theorem 36 $\bar{t}(\sigma, \mu) = \max_j(E_\sigma T_j - E_\mu T_j)$.

10.2 Differentiating stationary distributions

From the definition (6) of the fundamental matrix Z we can write, in matrix notation,

$$(\mathbf{I} - \mathbf{P})\mathbf{Z} = \mathbf{Z}(\mathbf{I} - \mathbf{P}) = \mathbf{I} - \Pi \tag{30}$$

where Π is the matrix with (i, j)-entry π_j. The matrix $\mathbf{I} - \mathbf{P}$ is not invertible but (30) expresses $\mathbf{Z}$ as a "generalized inverse" of $\mathbf{I} - \mathbf{P}$, and one can use matrix methods to verify general identities in the spirit of section 2.1. See e.g. [9, 17]. Here is a setting where such matrix methods work well.

Lemma 37 *Suppose* $\mathbf{P}$ *(and hence* π *and* $\mathbf{Z}$*) depend on a real parameter* α*, and suppose* $\mathbf{R} = \frac{d}{d\alpha}\mathbf{P}$ *exists. Then, at* α *such that* $\mathbf{P}$ *is irreducible,*

$$\frac{d}{d\alpha}\pi = \pi\mathbf{R}\mathbf{Z}.$$

Proof. Write $\eta = \frac{d}{d\alpha}\pi$. Differentiating the balance equations $\pi = \pi\mathbf{P}$ gives $\eta = \eta\mathbf{P} + \pi\mathbf{R}$, in other words $\eta(\mathbf{I} - \mathbf{P}) = \pi\mathbf{R}$. Right-multiply by $\mathbf{Z}$ to get

$$\pi\mathbf{R}\mathbf{Z} = \eta(\mathbf{I} - \mathbf{P})\mathbf{Z} = \eta(\mathbf{I} - \Pi) = \eta - \eta\Pi.$$

But $\eta\Pi = 0$ because $\sum_i \eta_i = \frac{d}{d\alpha}(\sum_i \pi_i) = 0$.