Part A:
Simple Markovian Models

I

Markov Chains

1 Preliminaries

We consider a Markov chain $X_0, X_1, \ldots$ with discrete (i.e. finite or countable) state space $E = \{i, j, k, \ldots\}$ and specified by the transition matrix $\boldsymbol{P} = (p_{ij})_{i,j \in E}$. By this we mean that $\boldsymbol{P}$ is a given $E \times E$ matrix such that $\boldsymbol{p}_{i\cdot} = (p_{ij})_{j \in E}$ is a probability (vector) for each i, and that we study $\{X_n\}$ subject to exactly those governing probability laws $\mathbb{P} = \mathbb{P}_{\boldsymbol{\mu}}$ (*Markov probabilities*) for which

$$\mathbb{P}(X_0 = i_0,\, X_1 = i_1, \ldots, X_n = i_n) \;=\; \mu_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} \tag{1.1}$$

where $\mu_i = \mathbb{P}(X_0 = i)$. The particular value of the initial distribution $\boldsymbol{\mu}$ is unimportant in most cases and is therefore suppressed in the notation. An important exception is the case where X_0 is degenerate, say at i, and we write then $\mathbb{P}_i$ so that $\mathbb{P}_i(X_0 = i) = 1$.

Given $\boldsymbol{\mu}$, it is readily checked that (1.1) uniquely determines a probability distribution on $\mathscr{F}_n = \sigma(X_0, \ldots, X_n)$. Appealing to basic facts from the foundational theory of Markov processes (to be discussed in Section 8), this set of probabilities can be uniquely extended to a probability law $\mathbb{P}_{\boldsymbol{\mu}}$ governing the whole chain. Thus, since the transition matrix $\boldsymbol{P}$ is fixed here and in the following, the Markov probabilities are in one–to–one correspondence with the set of initial distributions.

If $\mathbb{P}$ is a Markov probability, then (with the usual a.s. interpretation of conditional probabilities and expectations)

$$p_{ij} \;=\; \mathbb{P}_i(X_1 = j) \;=\; \mathbb{P}\big(X_{n+1} = j \,\big|\, X_n = i\big), \tag{1.2}$$

$$\mathbb{P}(X_{n+1} = j \,|\, \mathscr{F}_n) \;=\; p_{X_n j} \;=\; \mathbb{P}_{X_n}(X_1 = j), \qquad (1.3)$$

$$\mathbb{E}\big[h(X_n, X_{n+1}, \ldots) \,\big|\, \mathscr{F}_n\big] \;=\; \mathbb{E}_{X_n} h(X_0, X_1, \ldots)\,. \qquad (1.4)$$

Conversely[1], either (1.3) or (1.4) is sufficient for $\mathbb{P}$ to be a Markov probability. The formal proof of these facts is an easy (though in part lengthy) exercise in conditioning arguments and will not be given here. However, equations (1.2), (1.3), (1.4) have important intuitive contents. Thus (1.4) means that at time n, the chain is restarted with the new initial value X_n. Equivalently, the post–n–chain $X_n, X_{n+1}, \ldots$ evolves as the Markov chain itself, started at X_n but otherwise independent of the past. Similarly, in simulation terminology (1.3) means that the chain can be stepwise constructed by at step n drawing X_{n+1} according to $\boldsymbol{p}_{X_n\cdot}$ (to get started, draw X_0 according to $\boldsymbol{\mu}$).

Recall from A10 (the Appendix) that a *stopping time* σ is a r.v. with values in $\mathbb{N} \cup \{\infty\}$ and satisfying $\{\sigma = n\} \in \mathscr{F}_n$ for all n, that $\mathscr{F}_\sigma$ denotes the σ–algebra which consists of all disjoint unions of the form $\cup_0^\infty A_n$ with $A_n \in \mathscr{F}_n$, $A_n \subseteq \{\sigma = n\}$ (here $n = \infty$ is included with the convention $\mathscr{F}_\infty = \sigma(X_0, X_1, \ldots)$), and that σ and X_σ are measurable w.r.t. to $\mathscr{F}_\sigma$. The important *strong Markov property* states that for the sake of predicting the future development of the chain a stopping time may be treated as a fixed deterministic point of time. For example, we have the following extension of (1.4):

Theorem 1.1 (STRONG MARKOV PROPERTY) *Let σ be a stopping time. Then a.s. on $\{\sigma < \infty\}$ it holds that*

$$\mathbb{E}\big[h(X_\sigma, X_{\sigma+1}, \ldots) \,\big|\, \mathscr{F}_\sigma\big] \;=\; \mathbb{E}_{X_\sigma} h(X_0, X_1, \ldots). \qquad (1.5)$$

Proof. We must show that for $A \in \mathscr{F}_\sigma$, $A \subseteq \{\sigma < \infty\}$ we have

$$\mathbb{E}\big[h(X_\sigma, X_{\sigma+1}, \ldots);\, A\big] \;=\; \mathbb{E}\big[\mathbb{E}_{X_\sigma} h(X_0, X_1, \ldots);\, A\big].$$

However, if $A \in \mathscr{F}_n$ and $\sigma = n$ on A, this is immediate from (1.4). Replace A by $A \cap \{\sigma = n\}$ and sum over n. □

The mth power (iterate) of the transition matrix is denoted by $\boldsymbol{P}^m = (p_{ij}^m)$. An easy calculation (e.g. let $n = nm$ in (1.1) and sum over the i_k with $k \notin \{0, m, \ldots, nm\}$) shows that $X_0, X_m, X_{2m}, \ldots$ is a Markov chain and that its transition matrix is simply $\boldsymbol{P}^m$.

Associated with each state is the hitting time

$$\tau(i) \;=\; \inf\{n \geq 1 :\, X_n = i\}$$

(with the usual convention $\tau(i) = \infty$ if no such n exists) and the number of visits $N_i = \sum_1^\infty I(X_n = i)$ to i. Clearly, $\{\tau(i) < \infty\} = \{N_i > 0\}$ and we

[1]The meaning of (1.4) is that this should hold for any $h : E \times E \times \cdots \to \mathbb{R}$ for which (1.4) makes sense, say h is bounded or nonnegative; similarly, (1.5) should hold for all n and j. In (1.3), $\mathbb{P}_{X_n}(X_1 = j)$ means $g(x) = \mathbb{P}_x(X_1 = j)$ evaulated at $x = X_n$.

call i *recurrent* if the recurrence time distribution $\mathbb{P}_i(\tau(i) = k)$ is proper, i.e. if $\mathbb{P}_i(\tau(i) < \infty) = 1$, and *transient* otherwise. The chain itself is recurrent (transient) if all states are so.

Proposition 1.2 *Let i be some fixed state. Then:*
(i) *The following assertions* (a), (b), (c) *are equivalent:* (a) *i is recurrent;* (b) *$N_i = \infty$ $\mathbb{P}_i$–a.s.;* (c) *$\mathbb{E}_i N_i = \sum_1^\infty p_{ii}^m = \infty$;*
(ii) *the following assertions* (a′), (b′), (c′) *are equivalent as well:* (a′) *i is transient;* (b′) *$N_i < \infty$ $\mathbb{P}_i$–a.s.;* (c′) *$\mathbb{E}_i N_i = \sum_1^\infty p_{ii}^m < \infty$.*

Proof. Define $\tau(i;1) = \tau(i)$,

$$\tau(i;k+1) \;=\; \inf\{n > \tau(i;k) :\, X_n = i\}\,, \quad \theta \;=\; \mathbb{P}_i(\tau(i;1) < \infty).$$

Then N_i is simply the number of k with $\tau(i;k) < \infty$, and by the strong Markov property and $X_{\tau(i;k)} = i$,

$$\begin{aligned}
\mathbb{P}_i(\tau(i;k+1) < \infty) \;&=\; \mathbb{E}_i\mathbb{P}\big(\tau(i;k+1) < \infty, \tau(i;k) < \infty \,\big|\, \mathscr{F}_{\tau(i;k)}\big) \\
&=\; \mathbb{E}_i\big[\mathbb{P}\big(\tau(i;k+1) < \infty \,\big|\, \mathscr{F}_{\tau(i;k)}\big);\, \tau(i;k) < \infty\big] \\
&=\; \mathbb{E}_i\big[\mathbb{P}_{X_{\tau(i;k)}}(\tau(i;1) < \infty);\, \tau(i;k) < \infty\big] \\
&=\; \theta\mathbb{P}_i(\tau(i,k) < \infty) \;=\; \cdots \;=\; \theta^{k+1}.
\end{aligned} \tag{1.6}$$

If (a) holds, then $\theta = 1$ so that it follows that all $\tau(i;k) < \infty$ $\mathbb{P}_i$–a.s., and (b) also holds. Clearly, (b)$\Rightarrow$(c) so that for part (i) it remains to prove (c)$\Rightarrow$(a) or equivalently (a′)$\Rightarrow$(c′). But if $\theta < 1$, then

$$\mathbb{E}_i N_i \;=\; \sum_{k=0}^{\infty} \mathbb{P}_i(N_i > k) \;=\; \sum_{k=1}^{\infty} \mathbb{P}_i(\tau(i;k) < \infty) \;=\; \sum_{k=1}^{\infty} \theta^k \;<\; \infty.$$

For part (ii), it follows by negation that (a′) $\Longleftrightarrow$ (c′) $\Longleftrightarrow$ (b″) $\mathbb{P}_i(N_i < \infty) > 0$. However, clearly (b′)$\Rightarrow$(b″) and from (1.6) it is seen that if (b″) holds, then $\theta < 1$. Thus (b″) $\Rightarrow$ (a′). $\square$

It should be noted that though Proposition 1.2 gives necessary and sufficient conditions for recurrence/transience, the criteria are almost always difficult to check: even for extremely simple transition matrices $\boldsymbol{P}$, it is usually impossible to find closed expressions for the p_{ii}^m. Some alternative general approaches are discussed in Section 5, but in many cases the recurrence/transience classification leads into arguments particular for the specific model.

Our emphasis in the following is on the recurrent case and we shall briefly discuss some aspects of the set–up. Two states i, j are said to communicate, written $i \leftrightarrow j$, if i can be reached from j (i.e. $p_{ji}^m > 0$ for some m) and vice versa. Clearly, the relation is transitive and symmetric. Now suppose i is recurrent and that j can be reached from i. Then also i can be reached from j. In fact even $\tau(i) < \infty$ $\mathbb{P}_j$–a.s. since otherwise $\mathbb{P}_i(\tau(i) = \infty) > 0$.

Furthermore, j is recurrent since

$$\sum_{m=1}^{\infty} p_{jj}^{m} \;\geq\; \sum_{m=1}^{\infty} p_{ji}^{m_1} p_{ii}^{m} p_{ij}^{m_2} \;=\; \infty$$

if m_1, m_2 are chosen with $p_{ji}^{m_1} > 0$, $p_{ij}^{m_2} > 0$. Obviously $i \leftrightarrow i$ by recurrence, and it follows that $\leftrightarrow$ is an equivalence relation on the recurrent states so that we may write

$$E \;=\; T \cup R_1 \cup R_2 \cdots, \tag{1.7}$$

where $R_1, R_2, \ldots$ are the equivalence classes (*recurrent classes*) and T the set of transient states. It is basic to note that the recurrent classes are *closed* (or *absorbing*), i.e.

$$\mathbb{P}_i(X_n \in R_k \text{ for all } n) = 1 \;\text{ when }\; i \in R_k$$

(this follows from the above characteriztion of R_k as the set of all states that can be reached from i). When started at $i \in R_k$ the chain therefore evolves within R_k only, and the state space may be reduced to R_k. If, on the other hand, $X_0 = i$ is transient, two types of paths may occur: either $X_n \in T$ for all n or at some stage the chain enters a recurrent class R_k and is *absorbed*, i.e. evolves from then on in R_k.

Most often one can restrict attention to *irreducible* chains, defined by the requirement that all states in E communicate. Such a chain is either transient or E consists of exactly one recurrent class. In fact, if a recurrent state, say i, exists at all, it follows from the above that any other state j is in the same recurrence class as i.

A recurrent state is called *positive recurrent* if the mean recurrence time $\mathbb{E}_i\tau(i)$ is finite. Otherwise i is *null recurrent*. The *period* $d = d(i)$ is the period of the recurrence–time distribution, i.e. the greatest integer d such that $\mathbb{P}_i(\tau(i) \in L_d) = 1$ where $L_d = \{d, 2d, 3d, \ldots\}$. If $d = 1$, i is *aperiodic*.

Proposition 1.3 *Let R be a recurrent class. Then the states in R* (i) *are either all positive recurrent or all null recurrent;* (ii) *have all the same period.*

Proof. (i) is deferred to Section 3. Let $i, j \in R$ and choose r, s with $p_{ij}^{r} > 0$, $p_{ji}^{s} > 0$. Then $p_{ii}^{r+s} > 0$, i.e. $r+s \in L_{d(i)}$, and whenever $p_{jj}^{n} > 0$, $p_{ii}^{r+s+n} > 0$ also, i.e. $r+s+n \in L_{d(i)}$ so that $n \in L_{d(i)}$ also. It follows that $\mathbb{P}_j(\tau(j) \in L_{d(i)}) = 1$, i.e. $d(j) \geq d(i)$. By symmetry, $d(i) \geq d(j)$. □

Proposition 1.4 *Let i be aperiodic and recurrent. Then:* (a) *there exists n_i such that $p_{ii}^{m} > 0$ for all $m \geq n_i$;* (b) *if j can be reached from i, then there exists n_j such that $p_{ij}^{m} > 0$ for all $m \geq n_j$.*

Proof. For (a), see A7.1(a). For (b), choose k_j with $p_{ij}^{k_j} > 0$ and let $n_j = n_i + k_j$. □

Problems

1.1 Explain that $\mathbb{P}_{\boldsymbol{\mu}} = \sum_{i\in E} \mu_i \mathbb{P}_i$.
1.2 Show that (1.2) implies (1.4).
1.3 Show that if $\theta = p_{ii} > 0$, then the exit time $\eta(i) = \inf\{n \ge 1 : X_n \neq i\}$ has a geometric distribution, $\mathbb{P}_i(\eta(i) = n) = (1-\theta)\theta^{n-1}$, $n = 1, 2, \ldots$.
1.4 In a number of population processes one encounters Markov chains with $E = \mathbb{N}$, X_n representing the population size at time n, state 0 absorbing and $\mathbb{P}_i(\tau(0) < \infty) > 0$ for all i. Explain why it is reasonable to denote $\{\tau(0) < \infty\}$ as the event of *extinction*. Show that any state $i \ge 1$ is transient and that $X_n \to \infty$ a.s. on the event $\{\tau(0) = \infty\}$ of nonextinction.

Notes In this book, we use the terminology that a Markov chain has discrete time and a Markov process has continuous time (the state space may be discrete as here or general as in Section 8). However, one should note that it is equally common to let "chain" refer to a discrete state space and "process" to a general one (time may be discrete or continuous).

One more convention: the bold typeface for say the initial distribution $\boldsymbol{\mu}$ indicates a representation as a (row) vector, but in many contexts it is more convenient to think of the measure interpretation, and we then write μ. Similarly, a function on the state space may be written either as a (column) vector $\boldsymbol{f} = (f_i)_{i\in E}$ or just as f (with value $f(i)$ at i) We will change freely between these notations; say we use whichever of $\nu(f), \boldsymbol{\nu f}$ which in a given context is convenient to represent $\sum \nu_i f(i)$. Accordingly, we can think of the transition matrix $\boldsymbol{P}$ as an operator acting on measures to the left and on functions to the right, and we sometimes write $\boldsymbol{\nu P}$ as νP and $\boldsymbol{Pf}$ as Pf. A particularly important function is the constant 1 which we write as $\mathbf{1}$ in vector notation.

Markov chains and processes with a discrete state space form in many ways a natural starting point of applied probability: when considering a specific phenomenon, the first attempt to formulate and solve a stochastic model is usually performed within the Markovian set–up, and also the mathematical question arising in connection with Markov chains are to a large extent the same as for more general models (in particular, this is so in queueing theory). The present text therefore starts with a treatment of the relevant features of discrete Markov chains and (in Chapter II) processes. The exposition is in principle self–contained, but the novice will miss examples, and thus the aim is more to provide a refresher and reference, covering also some topics that are not in all textbooks.

We will not list the many textbooks containing introductory chapters on Markov chains and processes. More advanced treatments of discrete Markov chains are in Brémaud (1999), Chung (1967), Freedman (1971), Kemeny *et al.* (1976) and Orey (1971), and of discrete Markov processes in Chung (1967) and Anderson (1991).

2 Aspects of Renewal Theory in Discrete Time

Let $f_1, f_2, \ldots$ be the point probabilities of a distribution on $\{1, 2, \ldots\}$. Then by a (discrete time) *renewal process* governed by $\{f_n\}$ we understand a point process (see A3 for the terminology) on $\mathbb{N}$ with epochs $S_0 = 0$, $S_n = Y_1 + \cdots + Y_n$, where the Y_i are i.i.d. with common distribution $\{f_n\}$. Instead of epochs, we usually speak of *renewals*. The associated *renewal sequence* $u_0, u_1, \ldots$ is defined by $u_k = \mathbb{P}(S_n = k \text{ for some } k \geq 0)$, i.e. the probability of a renewal at k.

A renewal occurs at $k > 0$ if either $Y_1 = k$ which happens w.p. $f_k = f_k u_0$, or if $Y_1 = \ell < k$ and $Y_2 + \cdots + Y_n = k - \ell$ for some n. The probability of this is $f_\ell u_{k-\ell}$, and so

$$u_k \;=\; f_k u_0 + f_{k-1} u_1 + \cdots + f_1 u_{k-1}, \quad k \geq 1, \tag{2.1}$$

i.e. in convolution equation $u = \delta_0 + u * f$ where $\delta_{0i} = I(i = 0)$. In conjunction with $u_0 = 1$, (2.1) clearly uniquely determines $\{u_n\}$.

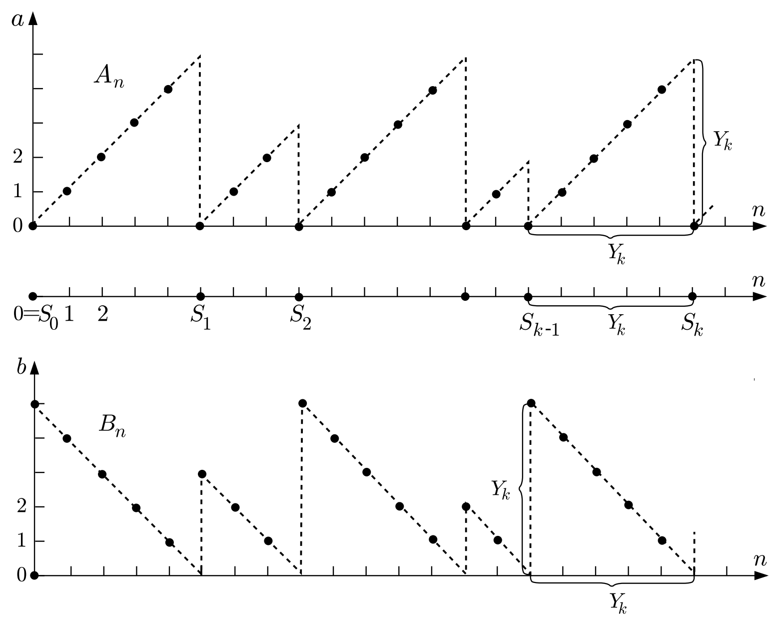

Figure 2.1

These concepts are intimately related to Markov chains. Consider some fixed recurrent state i, let $Y_1 = \tau(i)$ and more generally let Y_k be the inter–occurence time between the $(k-1)$th and kth visit to i. Then $Y_1, Y_2, \ldots$ are i.i.d. w.r.t. $\mathbb{P}_i$ according to the strong Markov property, the common distribution $\{f_n\}$ is the recurrence time distribution of i and the renewals are the visits to i so that $u_n = p_{ii}^n$. Conversely, *any* renewal processs can be constructed in this way from a Markov chain which we shall denote by $\{A_n\}$. Indeed, define $A_n = n - \sup\{S_k : S_k \leq n\}$ as the *backward recurrence time* at n, i.e. the time passed since the last renewal; see Fig. 2.1. Then the paths of $\{A_n\}$ are at 0 exactly at the renewals, i.e. the renewals are the recurrence times of 0, and the Markov property follows by noting that $\{A_n\}$ moves from i to either $i+1$ or 0, the probability of $i+1$ being $\mathbb{P}(Y_k > i+1 \,|\, Y_k > i)$

independently of $A_0, \ldots, A_{n-1}$. The state space E is $\mathbb{N}$ if $\{f_k\}$ has infinite support and $\{0, 1, \ldots, K-1\}$ with $K = \inf\{k : f_1 + \cdots + f_k = 1\}$ otherwise. A closely related important Markov chain is the *forward recurrence time chain* $\{B_n\}$, i.e. B_n is the waiting time until the next renewal after n; see again Fig. 2.1. The Markov property is even more immediate since the paths decrease deterministically from i to $i-1$ if $i > 1$, whereas the value of B_{n+1}, following $B_n = 1$, is chosen according to $\{f_k\}$ independently of the past. The state space is $\{1, 2, \ldots\}$ in the infinite support case and $\{1, \ldots, K\}$ otherwise, and a renewal occurs at n if and only if $B_{n-1} = 1$.

Lemma 2.1 *$\{u_n\}$ and $\{f_n\}$ have the same period d.*

Proof. Since $u_n \geq f_n$, it is clear that the period d_f of $\{f_n\}$ is at least that d_u of $\{u_n\}$. Conversely, it is only possible that $\mathbb{P}(S_k = n) > 0$ and hence $u_n > 0$ if n is a multiple of d_f. Hence $d_u \geq d_f$. □

If $d = 1$ in Lemma 2.1, we will call the renewal sequence (process) *aperiodic.*

Renewal processes with the Y_k having a possible continuous distribution will play a major role in later parts of the book. We shall here exploit the connection between (discrete) renewal processes and Markov chains in the limit theory. Within the framework of renewal processes, the main result is as follows (to be translated to Markov chains in Section 4):

Theorem 2.2 *Let $\{u_n\}$ be an aperiodic renewal sequence governed by $\{f_n\}$ and define $\mu = \sum_1^\infty n f_n = \mathbb{E}Y_1$. Then $u_n \to 1/\mu$ as $n \to \infty$* (here $1/\infty = 0$)*.*

Proof. Define $r_n = f_{n+1} + f_{n+2} + \cdots = \mathbb{P}(Y_1 > n)$ and let L be the index of the last renewal in $\{0, \ldots, n\}$. Then $L = \ell$ if there is a renewal at ℓ and the next Y is $> n - \ell$, i.e. the probability is $u_\ell r_{n-\ell}$ so that

$$1 \;=\; \mathbb{P}(L \leq n) \;=\; r_0 u_n + r_1 u_{n-1} + \cdots + r_n u_0. \tag{2.2}$$

Now let $\lambda = \limsup u_n$ and choose $n(k)$ such that $u_{n(k)} \to \lambda$. Let i satisfy $f_i > 0$. Choosing N such that $r_N < \epsilon$, we obtain from (2.1) and $u_n \leq 1$ that for k sufficiently large

$$\lambda - \epsilon \;\leq\; u_{n(k)} \;\leq\; r_N + \sum_{j=1}^{N} f_j u_{n(k)-j} \tag{2.3}$$

$$\leq\; \epsilon + (1 - f_i)(\lambda + \epsilon) + f_i u_{n(k)-i}. \tag{2.4}$$

Letting first $k \to \infty$ and next $\epsilon \downarrow 0$ yields $\liminf u_{n(k)-i} \geq \lambda$ which is only possible if $u_{n(k)-i} \to \lambda$. Repeating the argument we see that this also holds for any i of the form $i = x_1 a_1 + \cdots + x_t a_t$ where $x_k \in \mathbb{N}$, $f_{a_k} > 0$. But since $\{f_n\}$ is aperiodic, it follows by A7.1(a) (see also Proposition 1.4) that any sufficiently large i, say $i \geq a$, can be represented in this form. Thus letting

$n = n(k) - a$ in (2.2) we obtain for any N

$$1 \geq \sum_{j=0}^{N} r_j u_{n(k)-a-j} \to \lambda \sum_{j=0}^{N} r_j. \tag{2.5}$$

Since $r_0 + r_1 + \cdots = \mu$, this proves $1 \geq \lambda\mu$.

It remains to show that $\nu = \liminf u_n \geq \mu^{-1}$. This is clear if $\mu = \infty$ and can be proved similarly as above if $\mu < \infty$. In fact, if $\{m(k)\}$ is chosen such that $u_{m(k)} \to \nu$, we obtain, instead of (2.3),

$$\begin{aligned} \nu + \epsilon \ \geq \ u_{m(k)} \ &\geq \ \sum_{j=1}^{N} f_j u_{m(k)-j} \ \geq \ (\nu - \epsilon) \sum_{j \leq N, j \neq i} f_j + f_i u_{m(k)-i} \\ &= \ (1 - f_i)(\nu - \epsilon) - r_N(\nu - \epsilon) + f_i u_{m(k)-i}. \end{aligned}$$

As above, this implies $\limsup u_{m(k)-i} \leq \nu$ and $u_{m(k)-i} \to \nu$. Hence for fixed N

$$1 \leq \sum_{j=0}^{N} r_j u_{m(k)-a-j} + \sum_{j=N+1}^{\infty} r_j \to \nu \sum_{j=0}^{N} r_j + \sum_{j=N+1}^{\infty} r_j,$$

which tends to $\nu\mu + 0$ as $N \to \infty$. □

Corollary 2.3 *Let $\{u_n\}$, $\{f_n\}$ have period $d > 1$. Then:* (i) *$\{u_{nd}\}_{n=1}^{\infty}$ is an aperiodic renewal sequence governed by $\{f_{nd}\}_{n=1}^{\infty}$;* (ii) *$u_m = 0$ whenever m is not of the form $m = nd$;* (iii) *$u_{nd} \to d/\mathbb{E}Y = d/\mu$ as $n \to \infty$.*

Proof. Here (i) and (ii) are obvious, and from Theorem 2.2 we get

$$u_{nd} \to \left(f_d + 2f_{2d} + 3f_{3d} + \cdots\right)^{-1} = d/\mathbb{E}Y.$$

□

Sometimes one also encounters defective governing distributions $\{f_n\}$, i.e. $f_\infty = 1 - f_1 - f_2 - \cdots > 0$. The corresponding renewal sequence is still uniquely determined by $u_0 = 1$ and (2.1), and can be interpreted in terms of a *terminating* or *transient* renewal process. This is defined simply by attaching the Y_k mass f_∞ at ∞. If $f_\infty > 0$, then $\sigma = \inf\{n \geq 1 : Y_n = \infty\}$ is finite a.s., and $S_n < \infty$ for $n = 0, \ldots, \sigma - 1$, $= \infty$ for $n \geq \sigma$. In particular, the number σ of renewals is finite a.s., and hence the probability u_n of a renewal at n tends to zero as $n \to \infty$. More precisely:

Proposition 2.4 *If $f_\infty > 0$, then the expected number of renewals is given by $\mathbb{E}\sigma = \sum_0^\infty u_n = 1/f_\infty$.*

Proof. Since u_n is the probability of a renewal at n, the expected number of renewals is indeed $\sum_0^\infty u_n$. But it is also

$$\begin{aligned} \mathbb{E}\sigma \ &= \ \sum_{n=1}^{\infty} \mathbb{P}(\sigma \geq n) \ = \ \sum_{n=1}^{\infty} \mathbb{P}(Y_k < \infty, k = 1, \ldots, n-1) \\ &= \ \sum_{n=1}^{\infty} (1 - f_\infty)^{n-1} \ = \ 1/f_\infty. \end{aligned}$$

□

Problems

2.1 Define the generating function of $\{f_n\}$ by $\widehat{f}[s] = \sum_0^\infty s^n f_n$ $(f_0 = 0)$. Show that $\widehat{u}[s] = \sum_0^\infty s^n u_n = (1 - \widehat{f}[s])^{-1}$.
2.2 Consider the geometric case $f_n = (1-\theta)\theta^{n-1}$. Show that u_n is constant for $n > 0$, $u_n = 1 - \theta$.
2.3 Show that $\{u_n v_n\}$ is a renewal sequence if $\{u_n\}$, $\{v_n\}$ are so.
2.4 Let $\{u_n\}$ be a renewal sequence with $\sum_1^\infty f_n \neq 1$. Assume that $\sum_1^\infty \rho^n f_n = 1$ for some ρ. Show that $\{\rho^n u_n\}$ is a renewal sequence and that $u_n \sim c\rho^{-n}$ for some $c \geq 0$ (provided $\{f_n\}$ is aperiodic).

Notes The proof of Theorem 2.2 is a classical argument due to Erdös *et al.* (1949) (many texts today use coupling instead and we return to this in VII.2). Additional material on renewal sequences and related topics can be found in Kingman (1972).

3 Stationarity

Let $\boldsymbol{\nu} = (\nu_i)_{i\in E}$ be any nonnegative measure on E (it is not assumed that $\boldsymbol{\nu}$ is a distribution, $|\boldsymbol{\nu}| = \sum \nu_i = 1$, neither that $\boldsymbol{\nu}$ is finite, $|\boldsymbol{\nu}| < \infty$, but just that all $\nu_i < \infty$). We can then define a new measure $\boldsymbol{\nu P}$ by usual matrix multiplication (viewing $\boldsymbol{\nu}$ as a row vector), so that $\boldsymbol{\nu P}$ attaches mass $\sum_{i\in E} \nu_i p_{ij}$ to j. We call $\boldsymbol{\nu} \neq \mathbf{0}$ *stationary* if all $\nu_i < \infty$ and $\boldsymbol{\nu P} = \boldsymbol{\nu}$, i.e. if in algebraic terms $\boldsymbol{\nu}$ is a left eigenvector of the transition matrix $\boldsymbol{P}$ corresponding to the eigenvalue 1.

Of particular importance is the case where $\boldsymbol{\nu}$ is a distribution. Irrespective of whether $\boldsymbol{\nu}$ is stationary or not, we then have

$$\mathbb{P}_{\boldsymbol{\nu}}(X_1 = j) \;=\; \sum_{i\in E} \mathbb{P}_{\boldsymbol{\nu}}(X_0 = i)p_{ij} \;=\; \sum_{i\in E} \nu_i p_{ij} \;=\; (\boldsymbol{\nu P})_j.$$

Thus $\boldsymbol{\nu P}$ can be interpreted as the $\mathbb{P}_{\boldsymbol{\nu}}$–distribution of X_1, and in a similar manner the $\mathbb{P}_{\boldsymbol{\nu}}$–distribution of X_m is $\boldsymbol{\nu P}^m$. In particular, if $\boldsymbol{\nu}$ is stationary, then $\boldsymbol{\nu P}^m = \boldsymbol{\nu}$ for all m so that the distribution of X_m is independent of m. More generally:

Theorem 3.1 *Suppose that $\boldsymbol{\nu}$ is a stationary distribution. Then:*
(i) *The chain is strictly stationary w.r.t. $\mathbb{P}_{\boldsymbol{\nu}}$, i.e. the $\mathbb{P}_{\boldsymbol{\nu}}$–distribution of $(X_n, X_{n+1}, \ldots)$ does not depend on n;*
(ii) *there exists a strictly stationary version $\{X_n\}_{n\in\mathbb{Z}}$ of the chain with doubly infinite time, such that $\mathbb{P}_{\boldsymbol{\nu}}(X_n = i) = \nu_i$ for all $n \in \mathbb{Z}$.*

Proof. (i) Clearly $(X_n, X_{n+1}, \ldots)$ is a Markov chain with transition matrix $\boldsymbol{P}$ w.r.t. $\mathbb{P}_{\boldsymbol{\nu}}$. Then the distribution of the whole sequence is uniquely given by the initial distribution which is $\boldsymbol{\nu P}^n = \boldsymbol{\nu}$, hence independent of n.

(ii) This is a standard construction based upon Kolmogorov's consistency theorem and valid for general stationary sequences: let $\mathbb{P}^{n(1),\ldots,n(k)}$ be the

$\mathbb{P}_{\boldsymbol{\nu}}$–distribution of $\big(X_0, X_{n(2)-n(1)}, \ldots, X_{n(k)-n(1)}\big)$, $n(1) < n(2) < \cdots < n(k)$, and note that (by stationarity) $\big\{\mathbb{P}^{n(1),\ldots,n(k)}\big\}$ is a consistent family (see Breiman, 1968, p. 105, for more detail). □

Question of existence and uniqueness of stationary distributions is one of the main topics of Markov chain theory. We start by an explicit construction (a generalization of which will also turn out to be basic for non–Markovian processes; cf. VI.1 and VII.6):

Theorem 3.2 *Let i be a fixed recurrent state. Then a stationary measure $\boldsymbol{\nu}$ can be defined by letting ν_j be the expected number of visits to j in between two consecutive visits to i,*

$$\nu_j \;=\; \mathbb{E}_i \sum_{n=0}^{\tau(i)-1} I(X_n = j) \;=\; \sum_{n=0}^{\infty} \mathbb{P}_i(X_n = j, \tau(i) > n). \tag{3.1}$$

The proof is based upon the following lemma:

Lemma 3.3 *Let $\boldsymbol{\lambda}$ be an arbitrary initial distribution and σ a stopping time, and define new measures $\boldsymbol{\lambda}(\sigma)$, $\boldsymbol{\mu}(0)$, $\boldsymbol{\mu}(1)$ by $\lambda_j(\sigma) = \mathbb{P}_{\boldsymbol{\lambda}}(X_\sigma = j)$,*

$$\mu_j(0) \;=\; \mathbb{E}_{\boldsymbol{\lambda}} \sum_{n=0}^{\sigma-1} I(X_n = j), \quad \mu_j(1) \;=\; \mathbb{E}_{\boldsymbol{\lambda}} \sum_{n=1}^{\sigma} I(X_n = j).$$

Then $\boldsymbol{\lambda} + \boldsymbol{\mu}(1) = \boldsymbol{\mu}(0) + \boldsymbol{\lambda}(\sigma)$, $\boldsymbol{\mu}(1) = \boldsymbol{\mu}(0)\boldsymbol{P}$.

Proof. The first statement follows by computing $\mathbb{E}_{\boldsymbol{\lambda}} \sum_0^\sigma I(X_n = j)$ by splitting first into the contribution from $n = 0$ and the sum from 1 to σ, and next into the sum from 0 to $\sigma - 1$ and the contribution from $n = \sigma$. The second follows from

$$\begin{aligned}
\mu_j(1) \;&=\; \sum_{n=1}^{\infty} \mathbb{P}_{\boldsymbol{\lambda}}(X_n = j, \sigma \ge n) \;=\; \sum_{n=1}^{\infty} \mathbb{E}_{\boldsymbol{\lambda}}\big[\mathbb{P}\big(X_n = j, \sigma \ge n \,\big|\, \mathscr{F}_{n-1}\big)\big] \\
&=\; \sum_{n=1}^{\infty} \mathbb{E}_{\boldsymbol{\lambda}}\big[\mathbb{P}\big(X_n = j \,\big|\, \mathscr{F}_{n-1}\big);\, \sigma \ge n\big] \;=\; \sum_{n=1}^{\infty} \mathbb{E}_{\boldsymbol{\lambda}}\big[p_{X_{n-1}j};\, \sigma \ge n\big] \\
&=\; \sum_{k \in E} p_{kj} \sum_{n=0}^{\infty} \mathbb{P}_{\boldsymbol{\lambda}}(\sigma > n, X_n = k) \;=\; \sum_{k \in E} p_{kj} \mu_k(0) \;=\; (\boldsymbol{\mu}(0)\boldsymbol{P})_j.
\end{aligned}$$

Here in the third step we used the $\mathscr{F}_{n-1}$–measurability of $I(\sigma \ge n)$. □

Proof of Theorem 3.2. If in Lemma 3.3 we take $\boldsymbol{\lambda}$ as the one–point distribution at i and $\sigma = \tau(i)$, we have $\boldsymbol{\mu}(0) = \boldsymbol{\nu}$ and $\boldsymbol{\lambda}(\sigma) = \boldsymbol{\lambda}$. The conclusion of the lemma can be written $\boldsymbol{\lambda} + \boldsymbol{\nu}\boldsymbol{P} = \boldsymbol{\nu} + \boldsymbol{\lambda}$. Hence $\boldsymbol{\nu}\boldsymbol{P} = \boldsymbol{\nu}$, and we need only to check that $\nu_j < \infty$ for any j. Clearly, $\nu_i = 1$ and $\nu_j = 0$ if j is not in the same recurrent class as i. Otherwise observe first that $p_{ji}^m > 0$ for some m so that $\nu_j < \infty$ follows from

$$\nu_i \;=\; \sum_{k\in E} \nu_k p_{ki}^m \;\geq \nu_j p_{ji}^m. \tag{3.2}$$

□

Theorem 3.4 *If the chain is irreducible and recurrent, then a stationary measure $\boldsymbol{\nu}$ exists, satisfies $\nu_j > 0$ for all j and is unique up to a multiplicative constant.*

Here existence is immediate from Theorem 3.2 (we denote in the following the measure in (3.1) by $\boldsymbol{\nu}^{(i)}$). Also, $\nu_i > 0$ for any $i \in E$ and any stationary measure $\boldsymbol{\nu}$ is clear from (3.2) since we may choose j with $\nu_j > 0$. The key step for uniqueness is the following:

Lemma 3.5 *Let i be some fixed state and let $\boldsymbol{\nu}$ be superstationary* (i.e. $\boldsymbol{\nu P} \leq \boldsymbol{\nu}$) *with $\nu_i \geq 1$. Then $\nu_j \geq \nu_j^{(i)}$ for all $j \in E$.*

Proof. With $\widetilde{\boldsymbol{P}}$ the matrix obtained from $\boldsymbol{P}$ by replacing the ith column by zeros, it is easily seen by induction that $\widetilde{p}_{kj^n}$ is the *taboo probability* $\mathbb{P}_k(X_n = j, \tau(i) > n)$. In particular, if we let $k = i$ and sum over n, we get $\boldsymbol{\nu}^{(i)} = \boldsymbol{\delta}^{(i)} \sum_0^\infty \widetilde{\boldsymbol{P}}^n$ where $\boldsymbol{\delta}^{(i)}$ is the distribution degenerate at i. We next claim that $\nu_j \geq \delta_j^{(i)} + (\boldsymbol{\nu}\widetilde{\boldsymbol{P}})_j$. Indeed, for $j = i$ this follows from $\nu_i \geq 1 = \delta_i^{(i)}$, and for $j \neq i$ we have $(\boldsymbol{\nu}\widetilde{\boldsymbol{P}})_j = (\boldsymbol{\nu P})_j \leq \nu_j$. Hence

$$\begin{aligned} \boldsymbol{\nu} \;&\geq\; \boldsymbol{\delta}^{(i)} + \boldsymbol{\nu}\widetilde{\boldsymbol{P}} \;\geq\; \boldsymbol{\delta}^{(i)}(\boldsymbol{I} + \widetilde{\boldsymbol{P}}) + \boldsymbol{\nu}\widetilde{\boldsymbol{P}}^2 \;\geq\; \cdots \\ &\geq\; \boldsymbol{\delta}^{(i)} \sum_{n=0}^N \widetilde{\boldsymbol{P}}^n + \boldsymbol{\nu}\widetilde{\boldsymbol{P}}^{N+1} \;\geq\; \boldsymbol{\delta}^{(i)} \sum_{n=0}^N \widetilde{\boldsymbol{P}}^n, \end{aligned}$$

and letting $N \to \infty$, $\boldsymbol{\nu} \geq \boldsymbol{\nu}^{(i)}$ follows. □

Proof of Theorem 3.4. If $\boldsymbol{\nu}$ is stationary, then $\nu_i > 0$ as observed above. Thus we may assume $\nu_i = 1$ and the proof will be complete if we can show $\boldsymbol{\nu} = \boldsymbol{\nu}^{(i)}$. But according to the lemma, we have $\boldsymbol{\nu} \geq \boldsymbol{\nu}^{(i)}$. Hence $\boldsymbol{\mu} = \boldsymbol{\nu} - \boldsymbol{\nu}^{(i)}$ is nonnegative and $\boldsymbol{\mu P} = \boldsymbol{\mu}$. As noted above $\mu_i = 0$ then implies $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\nu} = \boldsymbol{\nu}^{(i)}$. □

Clearly, the total mass of the stationary measure $\boldsymbol{\nu}^{(i)}$ given by (3.1) is

$$|\boldsymbol{\nu}^{(i)}| \;=\; \sum_{j\in E} \nu_j^{(i)} \;=\; \mathbb{E}_i \sum_{n=0}^{\tau(i)-1} 1 \;=\; \mathbb{E}_i \tau(i). \tag{3.3}$$

Now if the chain is irreducible and recurrent, it follows by uniqueness that the $|\boldsymbol{\nu}^{(i)}| = \mathbb{E}_i\tau(i)$ are either all finite or all infinite, i.e. that the states are all positive recurrent or all null recurrent, proving the remaining part of Proposition 1.3. In the first case, $\boldsymbol{\nu}$ hence can be normalized to a stationary distribution $\boldsymbol{\pi} = \boldsymbol{\nu}/|\boldsymbol{\nu}|$ which is unique. In particular, for each j we have $\pi_j = \nu_j^{(j)}/|\boldsymbol{\nu}^{(j)}| = 1/\mathbb{E}_j\tau(j)$ which yields an expression for $\boldsymbol{\pi}$ independent of the reference state i. In summary:

Corollary 3.6 *If the chain is irreducible and positive recurrent, there exists a unique stationary distribution* $\boldsymbol{\pi}$ *given by*

$$\pi_j \;=\; \frac{1}{\mathbb{E}_i\tau(i)}\mathbb{E}_i \sum_{n=0}^{\tau(i)-1} I(X_n = j) \;=\; \frac{1}{\mathbb{E}_j\tau(j)} \tag{3.4}$$

Corollary 3.7 *Any irreducible Markov chain with a finite state space is positive recurrent.*

Proof. With $S_i = \sum_0^\infty I(X_n = i)$, we have $\sum_{i\in E} S_i = \infty$ so that by finiteness $S_i = \infty$ for at least one i. But then i is recurrent, and therefore by irreducibility the chain is recurrent. Since obviously the stationary measure cannot have infinite mass if E is finite, we have positive recurrence. □

Example 3.8 Consider the backward and forward recurrence time chains $\{A_n\}$, $\{B_n\}$ of a renewal process governed by $\{f_n\}$. It is clear from the discussion in Section 2 that both chains are irreducible on the appropriate state spaces. It is also clear that 0 is recurrent for $\{A_n\}$ and 1 for $\{B_n\}$ with $\{f_n\}$ as recurrence time distribution in both cases. In particular, positive recurrence is equivalent to $\mu = \sum nf_n < \infty$. For $\{A_n\}$, the stationary measure (3.1) with $i = 0$ becomes $\nu_n = r_n = f_{n+1} + f_{n+2} + \cdots$, $n = 0, 1, \ldots$. Indeed, n is visited once in between two consecutive visits to 0 if the recurrence time is $\geq n+1$. This occurs w.p. r_n and otherwise n is not visited. In particular, if $\mu < \infty$, then the stationary distribution is $\pi_n = r_n/\mu$. In an entirely similar manner it is seen that the stationary measure for $\{B_n\}$ is $\nu_n = r_{n-1}$, $n = 1, 2, \ldots$, and if $\mu < \infty$ then $\pi_n = r_{n-1}/\mu$ defines the stationary distribution. □

The above assumption of irreducibility and recurrence (i.e. one recurrent class) can easily be weakened by invoking the decomposition (1.7) of the state space. For example, if $\boldsymbol{\nu}^{(r)}$ is a stationary measure on the rth recurrent class R_r, it is easy to see that $\boldsymbol{\nu} = \sum_r \boldsymbol{\nu}^{(r)}$ is stationary for the whole chain. Conversely, the restriction of a stationary $\boldsymbol{\nu}$ to R_r is stationary (for the chain restricted to R_r). Also, some *transient* chains have a stationary measure. The theory is more difficult than for the recurrent case and will not be discussed here. We remark only that a stationary *distribution* always attaches mass zero to the transient states because $\mathbb{P}(X_n = i) \to 0$ when i is transient. It is then easy to see that the most general form of a stationary distribution is a convex combination of the unique stationary distributions on the positively recurrent classes.

An alternative proof of the uniqueness of the stationary measure will be given in VII.3. It relies on restricting the Markov chain to a subset F of the state space, a procedure that also has other applications and which we now take the opportunity to discuss briefly. Let $\tau(F;k)$ be the time of the kth visit of $\{X_n\}$ to F, and define $\tau(F) = \tau(F;1)$, $X_k^F = X_{\tau(F;k)}$. In the recurrent case, $\tau(F;k) < \infty$ for all k, and by the strong Markov property $\{X_k^F\}$ is a Markov chain. The transition matrix has elements $p_{k\ell}^F = \mathbb{P}_k(X_{\tau(F)} = \ell)$, $k,\ell \in F$, but these cannot in general be found explicitly in terms of the p_{ij} (but see Problem 3.8). Nevertheless, we have the following result:

Proposition 3.9 *If $\{X_n\}$ is irreducible and recurrent with stationary measure $\boldsymbol{\nu}$, then $\{X_k^F\}$ is also irreducible and recurrent, and the stationary measure $\boldsymbol{\nu}^F = (\nu_\ell^F)_{\ell\in F}$ can be obtained by restricting $\boldsymbol{\nu}$ to F, i.e.* (up to a multiplicative constant) *$\nu_\ell^F = \nu_\ell$, $\ell \in F$. In particular, if $\{X_k^F\}$ is positive recurrent, then the stationary distribution is given by $\pi_\ell^F = \nu_\ell / \sum_{k\in F} \nu_k$.*

Proof. The first assertion is obvious. If we choose the initial state i in (3.1) in F, then both $\{X_n\}$ and $\{X_k^F\}$ visit $\ell \in F$ the same number of times in between visits to i. Hence, also constructing $\boldsymbol{\nu}^F$ according to (3.1) with the same i yields $\nu_\ell^F = \nu_\ell$. □

The formula which conversely expresses $\boldsymbol{\nu}$ in terms of $\boldsymbol{\nu}^F$ (and $\boldsymbol{P}$) is given in VII.5.

Occasionally, the following criterion is useful:

Lemma 3.10 *Let $\{X_n\}$ be irreducible and F a finite subset of the state space. Then the chain is positive recurrent if $\mathbb{E}_i\tau(F) < \infty$ for all $i \in F$.*

Proof. Define $\sigma(i) = \inf\{k \geq 1 : X_k^F = i\}$, $\tau(F;0) = 0$, $Y_k = \tau(F;k) - \tau(F;k-1)$. Then with $m = \max_{j\in F} \mathbb{E}_j\tau(F)$ we have for $i \in F$ that

$$\begin{aligned} \mathbb{E}_i\tau(i) &= \mathbb{E}_i \sum_{k=1}^{\sigma(i)} Y_k \;=\; \sum_{k=1}^{\infty} \mathbb{E}_i\big[\mathbb{E}\big[Y_k \,\big|\, \mathscr{F}_{\tau(F;k-1)}\big];\, k \leq \sigma(i)\big] \\ &\leq m \sum_{k=1}^{\infty} \mathbb{P}_i\big(k \leq \sigma(i)\big) \;=\; m\mathbb{E}_i\sigma(i). \end{aligned}$$

Since E is finite, $\{X_n^F\}$ is positive recurrent. Thus $\mathbb{E}_i\sigma(i) < \infty$, implying $\mathbb{E}_i\tau(i) < \infty$ and positive recurrence of $\{X_n\}$. □

Problems

3.1 Compute a stationary measure if $\boldsymbol{P}$ is *doubly stochastic*, i.e. both the rows and columns sum to 1.

3.2 Show that a Bernoulli random walk ($E = \mathbb{Z}$, $p_{n(n+1)} = \theta$, $p_{n(n-1)} = 1-\theta$) is doubly stochastic and, if in addition $\theta \neq 1/2$, transient. Show that both $\nu_n = 1$ and $\mu_n = \theta^n/(1-\theta)^n$ are stationary.

3.3 (Continuation of Problem 2.1). Show that the generating function $\widehat{\nu}[s]$ of the stationary measure of the backward recurrence–time chain of a renewal process is given by $\widehat{\nu}[s] = (\widehat{f}[s]-1)/(s-1)$.

3.4 A set A of states i called an *atom* if $\boldsymbol{p}_{i\cdot}$ is the same for all $i \in A$. Show that $\tau(A)$ is finite $\mathbb{P}_i$–a.s. either for all $i \in A$ or for no $i \in A$, and that in the first case a stationary measure can be defined by

$$\nu_j \;=\; \mathbb{E}_i \sum_{n=1}^{\tau(A)} I(X_n = j) \quad \text{with } i \in A \text{ arbitrary.}$$

3.5 Consider the recurrence times A_n, B_n of a renewal process. Show that $\{(A_n, B_n)\}$ is Markov with the set of states of the form $(i, 1)$ being an atom, and that the stationary measure is given by $\nu_{ij} = f_{i+j}$.

3.6 Show that $\{(X_n, X_{n+1})\}$ is a Markov chain, and compute the stationary measure in terms of that of $\{X_n\}$.

3.7 Let $\{X_n\}$ have stationary distribution $\boldsymbol{\pi}$ and let $\tau = \inf\{n \geq 1 : X_n = X_0\}$ be the time of return to the initial state. Evaluate $\mathbb{E}_{\boldsymbol{\pi}}\tau$.

3.8 In block notation corresponding to $E = F + F^c$, write the transition matrix as

$$\boldsymbol{P} \;=\; \begin{pmatrix} \boldsymbol{P}_{FF} & \boldsymbol{P}_{FF^c} \\ \boldsymbol{P}_{F^cF} & \boldsymbol{P}_{F^cF^c} \end{pmatrix}.$$

Show that $\left\{X_n^F\right\}$ has transition matrix

$$\boldsymbol{P}^F \;=\; \boldsymbol{P}_{FF} + \boldsymbol{P}_{FF^c}(\boldsymbol{I} - \boldsymbol{P}_{F^cF^c})^{-1}\boldsymbol{P}_{F^cF}.$$

Notes A concept somewhat related to a stationary distribution is that of a *quasi–stationary distribution*. For the precise definition, assume that a special state, say $0 \in E$, is absorbing, and write $E_0 = E\backslash\{0\}$. Then $\boldsymbol{\lambda} = (\lambda_i)_{i \in E_0}$ is called quasi–stationary if $\mathbb{P}_{\boldsymbol{\lambda}}(X_1 = j \,|\, \tau(0) > 1) = \lambda_j$. Closely related are *Yaglom limits*, defined as limits λ_j of $\mathbb{P}_i(X_n = j \,|\, \tau(0) > n)$. A main result in the area states that a (proper) Yaglom limit is necessarily quasi–stationary. However, it is more difficult to assess when a quasi–stationary distribution or a Yaglom limit is unique (the finite case is, however, easy).

Under weak irreducibility conditions, it is trivial to check that when a quasi–stationary distribution $\boldsymbol{\lambda}$ exists, then $\mathbb{P}_{\boldsymbol{\lambda}}(\tau(0) > n) = \theta^n$ where $\theta = \mathbb{P}_{\boldsymbol{\lambda}}(\tau(0) > 1)$ $= \mathbb{P}_{\boldsymbol{\lambda}}(X_1 \neq 0)$. This implies in particular $\mathbb{E}_i R^{\tau(0)} < \infty$ for $R < 1/\theta$. A recent result of Ferrari *et al.* (1995) goes the other way and states that under mild additional conditions, $\mathbb{E}_i R^{\tau(0)} < \infty$ for some $R > 1$ is necessary and sufficient for the existence of a quasi–stationary distribution. Further recent references in the area include Seneta (1994) and Glynn and Thorisson (2001).

4 Limit Theory

The aim is to obtain the limiting behaviour of the p_{ij}^n. We start by noting that this is nontrivial only in the positive recurrent case:

Proposition 4.1 *If state j is either transient or null recurrent, then $p_{ij}^n \to 0$ for any $i \in E$.*

Proof. In the transient case, $I(X_n = j) = 0$ eventually so that the $\mathbb{P}_i$–expectation p_{ij}^n must tend to zero. In the null recurrent case, write

$$p_{ij}^n \;=\; \sum_{k=1}^{n} \mathbb{P}_i(\tau(j) = k) u_{n-k} \quad \text{where} \quad u_n = p_{jj}^n. \tag{4.1}$$

Now $\{u_n\}$ is a renewal sequence governed by a distribution by infinite mean and therefore by Corollary 2.3 $u_n \to 0$. Letting $n \to \infty$ in (4.1) and appealing to dominated convergence yields $p_{ij}^n \to 0$. □

Theorem 4.2 (ERGODIC THEOREM FOR MARKOV CHAINS) *Suppose that the chain is irreducible, positive recurrent and aperiodic with stationary distribution* $\boldsymbol{\pi}$. *Then* $p_{ij}^n \to \pi_j$ *for all* j. *That is,* $\boldsymbol{P}^n \to \mathbf{1}\boldsymbol{\pi}$.

Proof. We use again (4.1). By Theorem 2.2, $u_n \to \mu^{-1}$ where μ is the mean recurrence time $\mathbb{E}_j\tau(j) = \pi_j^{-1}$. Appeal to dominated convergence once more to get $p_{ij}^n \to \pi_j$. □

The conclusion is that the limiting distribution of X_n is $\boldsymbol{\pi}$, irrespective of the initial state. Replacing $\mathbb{P}_i$ by $\mathbb{P}_{\boldsymbol{\nu}}$ shows that the same conclusion more generally holds for *any* initial distribution $\boldsymbol{\nu}$.

The case $d > 1$ can be quite easily reduced to the case $d = 1$. To this end, we need the concept of *cyclic classes*, i.e. a partitioning of E into disjoint sets $E_0, \ldots, E_{d-1}$ with the property that the only possible transitions are of the form $E_r \to E_{r+1}$ (here we identify E_d with E_0, E_{d+1} with E_1 and so on).

Proposition 4.3 *Consider an irreducible chain with period* $d > 1$, *let* i *be some arbitrary but fixed state and define*

$$E_r \;=\; \left\{ j \in E : P_{ij}^{nd+r} > 0 \text{ for some } n \geq 0 \right\}, \quad r = 0, \ldots, d-1.$$

Then $E_0, \ldots, E_{d-1}$ *partition* E *into nonempty disjoint sets, and if* $j \in E_r$, *then* $\mathbb{P}_j(X_1 \in E_{r+1}) = 1$ *and more generally* $\mathbb{P}_j(X_m \in E_{r+m}) = 1$. *Furthermore, these properties determine the* E_r *uniquely up to a cyclic rotation.*

Proof. It is obvious that $E_r \neq \emptyset$ (take $n = 0$). By irreducibility, each j is in some E_r so that $\cup_0^{d-1} E_r = E$. Suppose that p_{ij}^{nd+r} and p_{ij}^{md+s} are both > 0, and choose t with $p_{ji}^t > 0$. Then $nd+r+t$ and $md+s+t$ must both be multiples of d, so that $r - s = 0$ (mod d), showing that the E_r are disjoint. Clearly, $j \in E_r$ and $p_{jk}^m > 0$ implies $k \in E_{r+m}$. Summing over all such k yields $\mathbb{P}_j(X_m \in E_{r+m}) = 1$. Uniqueness is easy and is omitted. □

It follows that if $d > 1$, then the chain $X_0, X_d, X_{2d}, \ldots$ has $E_0, \ldots, E_{d-1}$ as disjoint closed sets. In the irreducible positive recurrent case it is furthermore clear that $\{X_{nd}\}$ is aperiodic positive recurrent on each E_r, i.e. admits a unique stationary distribution $\boldsymbol{\pi}^{(r)}$ concentrated on E_r. Now if $\boldsymbol{\pi}$ is stationary for $\{X_n\}$, its restriction to E_r is also stationary for $\{X_{nd}\}$, and thus by uniqueness $\boldsymbol{\pi}$ is a convex combination $\sum_0^{d-1} \alpha_r \boldsymbol{\pi}^{(r)}$ of the $\boldsymbol{\pi}^{(r)}$. Since

$$\alpha_{r+1} \;=\; \mathbb{P}_{\boldsymbol{\pi}}(X_1 \in E_r) \;=\; \mathbb{P}_{\boldsymbol{\pi}}(X_0 \in E_r) \;=\; \alpha_r,$$

we must even have $\alpha_r = d^{-1}$. Also, the limiting behaviour of p^n_{jk} can easily be seen from $p^{nd}_{j\ell} \to \pi^{(r)}_\ell$ if $j, \ell \in E_r$. Indeed, if $j \in E_r$ then $p^{nd+s}_{jk} = 0$ for all n if $k \notin E_{r+s}$, whereas if $k \in E_{r+s}$ then by dominated convergence

$$p^{nd+s}_{jk} \;=\; \sum_{\ell \in E_{r+s}} p^s_{j\ell} p^{nd}_{\ell k} \;\to\; \sum_{\ell \in E_{r+s}} p^s_{j\ell} \pi^{(r+s)}_k \;=\; \pi^{(r+s)}_k \;=\; d\pi_k. \tag{4.2}$$

In view of this discussion one can assume aperiodicity in most cases. An irreducible aperiodic positive recurrent chain is simply called *ergodic*.

A further noteworthy property of the stationary distribution is as the limit of time averages (aperiodicity is not required),

$$\frac{1}{n}\sum_{k=0}^{n} f(X_k) \;\to\; \pi(f) \;=\; \boldsymbol{\pi f} \;=\; \mathbb{E}_{\boldsymbol{\pi}} f(X_k) \;=\; \sum_{i \in E} f(i)\pi_i, \tag{4.3}$$

which holds if f is say bounded or nonnegative. The (easy) proof is carried out in a more general setting in VI.3; a corresponding CLT is in Section 7.

It is reasonable to ask what is the rate of convergence of p^n_{ij} to π_i. In particular, there has been considerable interest in *geometrical ergodicity*, defined by the requirement $p^n_{ij} - \pi_j = \mathrm{O}(\delta^n)$ for some $\delta < 1$ independent of i, j. One has:

Proposition 4.4 (a) *An ergodic Markov chain is geometrically ergodic provided* $\mathbb{E}_i z^{\tau(i)} < \infty$ *for some* $i \in E$ *and some* $z > 1$; (b) *any irreducible aperiodic finite Markov chain is geometrically ergodic.*

Proof. Part (a) is a contained in the more general VII.2.11 proved later. For (b), we can choose m_{ki} such that $p^m_{ki} > \pi_i/2$ for all $m \geq m_{ki}$. By finiteness, this implies the existence of $\epsilon > 0$ and $M < \infty$ such that $p^m_{ki} > \epsilon$ for all $m \geq M$ and all k. Hence $\mathbb{P}_i(\tau(i) > (n+1)M \,|\, \tau(i) \geq nM) \leq 1 - \epsilon$, and hence by the geometrical trials lemma A6.1 $\mathbb{E}_i z^{\tau(i)} < \infty$ when $z > 1$ is chosen with $z^M(1-\epsilon) < 1$. Now just appeal to (a). □

Also in the null recurrent case it is sometimes possible in various ways to obtain limit statements in terms of the stationary measure which are more refined than just $p^n_{ij} \to 0$. For example:

Proposition 4.5 *If the chain is irreducible recurrent with stationary measure* $\boldsymbol{\nu}$, *then for all* $i, j, k, \ell \in E$

$$\frac{\sum_{n=0}^m p^n_{ij}}{\sum_{n=0}^m p^n_{\ell k}} \;\to\; \frac{\nu_j}{\nu_k}, \quad m \to \infty. \tag{4.4}$$

For the proof, we need two lemmas (the proof of the first is a straightforward verification and omitted; generalizations are in Problem 5.1 and Section 6).

Lemma 4.6 *The matrix* $\widetilde{\boldsymbol{P}}$ *with elements* $\widetilde{p}_{ij} = \nu_j p_{ji}/\nu_i$ *is a transition matrix. Furthermore, the* ij*th element* $\widetilde{p}_{ij}^m$ *of* $\widetilde{\boldsymbol{P}}^m$ *is given by* $\widetilde{p}_{ij}^m = \nu_j p_{ji}^m/\nu_i$.

Lemma 4.7 *Define* $N_i^m = \sum_{n=0}^m I(X_n = i)$ *as the number of visits to* i *before time* m. *Then in the irreducible recurrent case,* $\lim_{m\to\infty} \mathbb{E}_j N_i^m / \mathbb{E}_k N_i^m = 1$ *for any* $j, k \in E$.

Proof. It may be assumed that $k = i$. By recurrence, $N_i^m \uparrow \infty$ and hence $\mathbb{E}_i N_i^m \uparrow \infty$. Since $N_i^{m-n} = N_i^m + \mathrm{O}(1)$, dominated convergence yields

$$\frac{\mathbb{E}_j N_i^m}{\mathbb{E}_i N_i^m} \;=\; \sum_{n=0}^m \mathbb{P}_j(\tau(i) = n) \frac{\mathbb{E}_i N_i^{m-n}}{\mathbb{E}_i N_i^m} \;\to\; \sum_{n=0}^\infty \mathbb{P}_j(\tau(i) = n) \;=\; 1. \qquad \square$$

Proof of Proposition 4.5. Consider a Markov chain $\{\widetilde{X}_n\}$ with transition matrix $\widetilde{\boldsymbol{P}}$ given by Lemma 4.6. The expression for $\widetilde{p}_{ij}^n$ shows that $\{X_n\}$ and $\{\widetilde{X}_n\}$ are irreducible at the same time and (sum over n and use Proposition 1.2) recurrent at the same time. Hence $\{\widetilde{X}_n\}$ satisfies the assumptions of Lemma 4.7, and we obtain

$$\begin{aligned} 1 \;&=\; \lim_{m\to\infty} \frac{\widetilde{\mathbb{E}}_j N_i^m}{\widetilde{\mathbb{E}}_i N_i^m} \cdot \frac{\mathbb{E}_i N_k^m}{\mathbb{E}_\ell N_k^m} \;=\; \lim_{m\to\infty} \frac{\sum_{n=0}^m \widetilde{p}_{ji}^n}{\sum_{n=0}^m \widetilde{p}_{ki}^n} \cdot \frac{\sum_{n=0}^m p_{ik}^n}{\sum_{n=0}^m p_{\ell k}^n} \\ &=\; \frac{\nu_k}{\nu_j} \lim_{m\to\infty} \frac{\sum_{n=0}^m p_{ij}^n}{\sum_{n=0}^m p_{ik}^n} \cdot \frac{\sum_{n=0}^m p_{ik}^n}{\sum_{n=0}^m p_{\ell k}^n} \;=\; \frac{\nu_k}{\nu_j} \lim_{m\to\infty} \frac{\sum_{n=0}^m p_{ij}^n}{\sum_{n=0}^m p_{\ell k}^n}. \end{aligned}$$

$\square$

Notes The terminology "ergodic" as used above is standard, but one should beware not to confuse it with the meaning it has in general stationary process theory (e.g. Breiman, 1968, Ch. 6), namely that the invariant σ–field is trivial. In the Markov chain setting, this does not require aperiodicity, whereas the tail σ–field of a positive recurrent Markov chain being trivial is equivalent to aperiodicity; see e.g. Freedman (1971).

For further results on geometric convergence rates, see VII.2.10. Studying convergence rates via asymptotics of $p_{ij}^n - \pi_j$ as $n \to \infty$ is not the only possible point of view. For example, in a number of models one has observed that $\|\boldsymbol{\nu}\boldsymbol{P}^n - \boldsymbol{\pi}\|$ (t.v. distance) changes from $\|\boldsymbol{\nu} - \boldsymbol{\pi}\|$ to 0 rather abrubtly at a certain time point N, and this N may be a more appropriate measure of the convergence rate than sharp estimates of the deviation of p_{ij}^n from π_j when n is so large that the difference is negligible anyway. Surveys of such broader aspects are in Rosenthal (1995) and Saloff–Coste (1996).

One might expect from Proposition 4.5 and the ergodic theorem for Markov chains that if the chain is also aperiodic, then the *strong ratio property* $p_{ij}^n / p_{\ell k}^n \to \nu_j/\nu_k$ holds. This is, however, not true for all null recurrent chains and presents in fact difficult and not completely solved problems; see Orey (1971). There has also been much discussion of the strong ratio property in relation to quasi–stationarity; see Kesten (1995).

5 Harmonic Functions, Martingales and Test Functions

There is a concept dual to that of a stationary measure, namely that of a *harmonic function* h defined as a right eigenvector $\boldsymbol{h}$ of $\boldsymbol{P}$ corresponding to the eigenvalue 1.[2] The requirement $\boldsymbol{Ph} = \boldsymbol{h}$ means

$$h(i) \;=\; \sum_{j\in E} p_{ij}h(j) \;=\; \mathbb{E}_i h(X_1) \;=\; \mathbb{E}\big[h(X_{n+1})\,\big|\,X_n = i\big],$$

i.e. that $\{h(X_n)\}$ is a martingale. Similarly, one defines h to be *subharmonic* if $\boldsymbol{Ph} \geq \boldsymbol{h}$, i.e. $\{h(X_n)\}$ is a submartingale, and *superharmonic* or *excessive* if $\boldsymbol{Ph} \leq \boldsymbol{h}$, i.e. $\{h(X_n)\}$ is a supermartingale.

Proposition 5.1 *If the chain is irreducible and recurrent, then any non-negative superharmonic function h is necessarily constant. Similarly, any bounded subharmonic function h is constant.*

Proof. We must show that $h(i) = h(j)$ for $i \neq j$. Now from the convergence of any non–negative supermartingale we have that $Z = \lim h(X_n)$ exists $\mathbb{P}_i$–a.s. Since $\mathbb{P}_i(X_n = i \text{ i.o.}) = 1$, it follows that $Z = h(i)$ $\mathbb{P}_i$–a.s. Similarly, $\mathbb{P}_i(X_n = j \text{ i.o.}) = 1$ implies that $Z = h(j)$ $\mathbb{P}_i$–a.s. and hence $h(i) = h(j)$. The subharmonic case is similar, using the a.s. convergence of any bounded submartingale. □

When concerned with the recurrent case as in most of this book, the implication is that (super– or sub–) harmonic functions do not play a major role. In the rest of this section we will see, however, that a number of useful recurrence/transience criteria and other properties can be stated in terms of functions h (commonly referred to as *test functions* or *Lyapounov functions*), having properties which are rather similar and allowing for arguments along the lines of the proof of Proposition 5.1.

The problems we study are trivial if E is finite, and in the infinite case we write $h(j) \to \infty$ if the set $\{j : h(j) \leq a\}$ is finite for any $a < \infty$.

Proposition 5.2 *Suppose the chain is irreducible and let i be some fixed state. Then the chain is transient if and only if there is a bounded nonzero function $h:\, E\backslash\{i\} \to \mathbb{R}$ satisfying*

$$h(j) \;=\; \sum_{k\neq i} p_{jk}h(k), \quad j \neq i. \tag{5.1}$$

Proof. Obviously $h(j) = \mathbb{P}_j(\tau(i) = \infty)$ is bounded and satisfies (5.1). If the chain is transient, then furthermore $h \neq 0$. Suppose, conversely, that there is an h as stated and define $\widetilde{h}(j) = h(j)$, $j \neq i$, $\widetilde{h}(i) = 0$, $\alpha = P\widetilde{h}(i)$. By changing the sign if necessary, we may assume $\alpha \geq 0$ so that $P\widetilde{h}(i) \geq \widetilde{h}(i)$.

[2] See Notes to Section 1 for notation, identication of h with $\boldsymbol{h}$, of $\boldsymbol{Ph}$ with Ph, etc.

Since $P\widetilde{h}(j) = \widetilde{h}(j)$ for $j \neq i$, $\widetilde{h}$ is thus subharmonic. Hence if the chain is recurrent, we have by Proposition 5.1 that $h(j) = \widetilde{h}(j) = \widetilde{h}(i) = 0$ for all $j \neq i$, contradicting $h \neq 0$. Hence the chain is transient. □

Proposition 5.3 *Suppose the chain is irreducible and let E_0 be a finite subset of the state space E. Then:*
(i) *the chain is recurrent if there exists a function $h : E \to \mathbb{R}$ such that $h(x) \to \infty$ and*

$$\sum_{k\in E} p_{jk}h(k) \;\le\; h(j), \quad j \notin E_0. \tag{5.2}$$

(ii) *the chain is positive recurrent if for some $h : E \to \mathbb{R}$ and some $\epsilon > 0$ we have $\inf_x h(x) > -\infty$ and*

$$\sum_{k\in E} p_{jk}h(k) \;<\; \infty, \quad j \in E_0, \tag{5.3}$$

$$\sum_{k\in E} p_{jk}h(k) \;\le\; h(j) - \epsilon, \quad j \notin E_0. \tag{5.4}$$

An often encountered compact way to write (5.3)–(5.4) is

$$Ph(j) \;\le\; h(j) - \epsilon + bI(j \in E_0).$$

The intuitive content of (5.2) is that the "center" of the state space in the h–scale corresponds to small values, and that the drift points to the center; similarly, (5.4) can be interpreted as a uniformly positive drift towards the center.

Proof. By adding a constant if necessary, we may assume $h \ge 0$. Write $T = \tau(E_0)$ and define $Y_n = h(X_n)I(T > n)$.
(i) Note first that (5.2) may be rewritten as $\mathbb{E}[h(X_{n+1}) \,|\, X_n = j] \le h(j)$ for $j \notin E_0$. Let $X_0 = i \notin E_0$. Then on $\{T > n\}$, $X_n \notin E_0$ (this fails for $n = 0$ if $X_0 \in E_0$) and hence

$$\begin{aligned}\mathbb{E}_i[Y_{n+1} \,|\, \mathscr{F}_n] \;&\le\; \mathbb{E}_i\big[h(X_{n+1});\, T > n \,\big|\, \mathscr{F}_n\big] \\ &=\; I(T > n)\mathbb{E}_i[h(X_{n+1}) \,|\, \mathscr{F}_n] \;\le\; I(T > n)h(X_n) \;=\; Y_n.\end{aligned} \tag{5.5}$$

If $T \le n$, then $Y_n = Y_{n+1} = 0$, and thus $\mathbb{E}_i[Y_{n+1} \,|\, \mathscr{F}_n] \le Y_n$, i.e. $\{Y_n\}$ is a nonnegative supermartingale and hence converges a.s., $Y_n \stackrel{\text{a.s.}}{\to} Y_\infty$. Suppose the chain is transient. Then $h(X_n) \le a$ only finitely often, i.e. $h(X_n) \to \infty$, and since $Y_\infty < \infty$, we must have $\mathbb{P}_i(T = \infty) = 0$. But $\mathbb{P}_i(T < \infty) = 1$ for all $i \notin E_0$ implies that some $j \in E_0$ is recurrent, a contradiction.
(ii) Again let $X_0 = i \notin E_0$. Then as in (5.5), we get on $\{T > n\}$ that

$$\mathbb{E}_i[Y_{n+1} \,|\, \mathscr{F}_n] \;\le\; I(T > n)\mathbb{E}_i[h(X_{n+1}) \,|\, \mathscr{F}_n] \;\le\; Y_n - \epsilon I(T > n).$$

Again, the same is obvious on $\{T \le n\}$ and hence

$$0 \le\; \mathbb{E}_iY_{n+1} \;\le\; \mathbb{E}_iY_n - \epsilon\mathbb{P}_i(T > n) \;\le\; \cdots \;\le \mathbb{E}_iY_0 - \epsilon\sum_{k=0}^{n}\mathbb{P}_i(T > k).$$

Letting $n \to \infty$ and using $Y_0 = h(i)$ yields $\mathbb{E}_i T \le \epsilon^{-1} h(i)$. Thus for $j \in E_0$,

$$\mathbb{E}_j T \;=\; \sum_{i \in E_0} p_{ji} + \sum_{i \notin E_0} p_{ji} \mathbb{E}_i (T+1) \;\le\; 1 + \epsilon^{-1} \sum_{i \notin E_0} p_{ji} h(i)$$

which is finite by (5.3). That the chain is positive recurrent now follows by Lemma 3.10. □

Proposition 5.4 *Suppose the chain is irreducible and let E_0 be a finite subset of E and h a function such that*

$$\sum_{k \in E} p_{jk} h(k) \;\ge\; h(j), \quad j \notin E_0, \tag{5.6}$$

and that $h(i) > h(j)$ for some $i \notin E_0$ and all $j \in E_0$. Then: (i) *if h is bounded, then the chain is transient;* (ii) *if h is bounded below and*

$$\sum_{k \in E} p_{jk} |h(k) - h(j)| \;\le A, \quad j \in E, \tag{5.7}$$

for some $A < \infty$, then the chain is null recurrent or transient .

Proof. Define T as above but let now $Y_n = h(X_{n \wedge T})$. It is then readily verified that $\{Y_n\}$ is a submartingale when $X_0 = i \notin E_0$. In (i), boundedness then implies $Y_n \stackrel{\text{a.s.}}{\to} Y_\infty$ where $\mathbb{E}_i Y_\infty \ge \mathbb{E}_i Y_0 = h(i)$. But $Y_\infty < h(i)$ on $\{T < \infty\}$ so that $\mathbb{P}_i(T < \infty) < 1$, showing transience.

For (ii), we can choose $j \in E_0$ such that $\alpha = \mathbb{P}_j(\tau(i) < T) > 0$. Then $\mathbb{E}_j \tau(j) \ge \mathbb{E}_j T \ge \alpha \mathbb{E}_i T$ so that is suffices to show $\mathbb{E}_i T = \infty$. Suppose $\mathbb{E}_i T < \infty$. Then in particular, $T < \infty$ $\mathbb{P}_i$–a.s. and by (5.7),

$$\mathbb{E}_i \sum_{n=1}^{T} |Y_n - Y_{n-1}| \;=\; \mathbb{E}_i \sum_{n=1}^{\infty} I(T \ge n) \mathbb{E}\big[|Y_n - Y_{n-1}| \,\big|\, \mathscr{F}_{n-1}\big] \;\le\; A \mathbb{E}_i T \;<\; \infty.$$

Thus we can interchange summation and expectation to get

$$\begin{aligned} \mathbb{E}_i Y_T \;&=\; \mathbb{E}_i Y_0 + \mathbb{E}_i \sum_{n=1}^{T} (Y_n - Y_{n-1}) \;=\; h(i) + \sum_{n=1}^{\infty} \mathbb{E}_i [Y_n - Y_{n-1};\, T \ge n] \\ &=\; h(i) + \sum_{n=1}^{\infty} \mathbb{E}_i \big(I(T \ge n) \mathbb{E}\big[Y_n - Y_{n-1} \,\big|\, \mathscr{F}_{n-1}\big]\big) \;\ge\; h(i), \end{aligned}$$

using the submartingale property in the last step. This is a contradiction since $Y_T < h(i)$. □

Proposition 5.5 *Suppose the chain is irreducible and recurrent, and let E_0 be a finite subset of the state space E. Then the chain is geometrically ergodic if for some $h \ge 0$ with $h(i) > A > 0$, $i \in E_0$, and some $r > 1$*

$$\sum_{k \in E} p_{jk} h(k) \quad < \quad \infty, \quad j \in E_0, \tag{5.8}$$

$$\sum_{k \in E} p_{jk} h(k) \quad \leq \quad h(j)/r, \quad j \notin E_0. \tag{5.9}$$

Proof. Let $X_0 = i \notin E_0$, $Y_n = r^n h(X_{n \wedge T})$. Then it follows easily from (5.9) that $\{Y_n\}$ is a nonnegative supermartingale. By recurrence, the limit is $Y_\infty = r^T h(X_T) \geq Ar^T$. On the other hand, $\mathbb{E}_i Y_\infty \leq \mathbb{E}_i Y_0 = h(i)$. For $j \in E_0$, (5.8) then yields

$$\mathbb{E}_j r^T \;\leq\; r + r \sum_{i \notin E_0} p_{ji} \mathbb{E}_i r^T \;\leq\; 1 + A^{-1} \sum_{i \notin E_0} p_{ji} h(i) \;<\; \infty.$$

It remains to show that $\mathbb{E}_j r^T \leq$ for all $j \in E_0$ implies geometric ergodicity. By Proposition 4.4, this will follow if we can show $\mathbb{E}_i r^T < \infty$ for all $i \in E_0$. This in turn follows by a variant of the proof of Lemma 3.10, left as Problem 5.3. □

Proposition 5.6 *Suppose the chain is irreducible and positive recurrent with stationary distribution π, and let f, g, h be nonnegative functions on E such that*

$$\sum_{j \in E} p_{ij} h(j) \;\leq\; h(i) - f(i) + g(i), \quad i \in E. \tag{5.10}$$

If $\pi(g) < \infty$, $\pi(h) < \infty$, then also $\pi(f) < \infty$.

Proof. We can rewrite (5.10) as $f \leq h - Ph + g$. Thus $P^k f \leq P^k h - P^{k+1} h + P^k g$ and for any i,

$$\sum_{k=1}^{n} P^k f(i) \leq \; Ph(i) - P^{n+1} h(i) + \sum_{k=1}^{n} P^k g(i) \leq \; Ph(i) + \sum_{k=1}^{n} P^k g(i).$$

Applying π to the left and noting that $\pi(Ph)/n = \pi(h)/n \to 0$ yields $\pi(f) \leq \pi(g) < \infty$. □

Example 5.7 Consider a queue where service takes place at a discrete sequence of instants $n = 0, 1, 2, \ldots$, let X_n be the queue length at time n, B_n the number of customers arriving between n and $n+1$ and A_n the maximal number of customers that can be served at the $(n+1)$th service epoch. Thus with $Y_n = B_n - A_n$

$$X_{n+1} \;=\; (X_n + Y_n)^+, \tag{5.11}$$

a recurrence relation (the Lindley recursion) also typical for many other queueing situations and discussed in length in III.6. For example, this could describe the queue at the stop of a bus with regular schedule, with A_n the number of free seats in the nth bus.

Assume further that the random vectors (A_n, B_n) are i.i.d.; then $\{X_n\}$ is a Markov chain on $\mathbb{N}$. Let $\mu = \mathbb{E} Y_n$. With $h(i) = i$, (5.11) then yields $\mathbb{E}_i W_1 = \mathbb{E}(i + Y_1)^+ \geq i + \mu$. Thus, if $\mu \geq 0$, Proposition 5.4(ii) shows immediately that $\{X_n\}$ cannot be positive recurrent. Suppose on the other hand that $\mu < 0$ and let $\mu_i = \mathbb{E}[Y_n; Y_n > -i]$. Then $\mu_i \to \mu$, $i \to \infty$, and hence for i so large, say $i > i_0$, that $\mu_i \le \mu/2$,

$$\mathbb{E}(i + Y_1)^+ \ = \ \mathbb{E}[i + Y_1; Y_1 > -i] \ \le \ i + \mu_i \ \le \ i + \mu/2.$$

Thus Proposition 5.3(ii) with $E_0 = \{0, \ldots, i_0\}$, $h(i) = i$, $\epsilon = -\mu/2$ yields positive recurrence.

For geometrical ergodicity, assume $\mu < 0$ and that $\mathbb{E}z^{B_1} < \infty$ for some $z > 1$. By replacing z by a smaller z if necessary, we may assume $r_1 = \mathbb{E}z^{Y_1} < 1$. We have $\mathbb{E}z^{i+Y_1} = z^i r_1$, and as above, one then gets $\mathbb{E}_i z^{W_1} < z^i r$ for $i \ge i_0$ and some $r \in (r_1, 1)$. Thus Proposition 5.5 with $h(i) = z^i$ yields geometric ergodicity.

Finally, assume $\mu < 0$, $\mu_2 = \mathbb{E}Y_n^2 < \infty$. With $h(i) = i^2$, we then have $\mathbb{E}h(i+Y_1) = h(i)+\mu_2+2i\mu$. As above, this implies $Ph(i) \le h(i)-f(i)+g(i)$ for $i \ge i_0$ where $g(i) = \mu_2/2$, $f(i) = -i\mu$. Since $\pi(g) < \infty$, Proposition 5.6 yields $\pi(f) < \infty$. I.e., the stationary distribution has finite mean when $\mu_2 < \infty$ [see further X.2]. □

Problems

5.1 (DOOB'S h–TRANSFORM) Suppose the chain is irreducible and $h \ge 0$ harmonic with $h \ne 0$. Show that $h(i) > 0$ for all i and that the matrix $\widetilde{\boldsymbol{P}}$ with elements $\widetilde{p}_{ij} = h(j)p_{ij}/h(i)$ is a transition matrix.

5.2 Consider a population process satisfying the assumptions of Problem 1.4 and with all states $i, j \ge 1$ communicating. Show that the extinction probability $q_i = \mathbb{P}_i(\tau(0) < \infty)$ is either 1 for all $i \ge 1$ or 0 for all $i \ge 1$. Let E_0 be finite and suppose (5.2) holds. Show that $q_i = 1$ if $h(j) \to \infty$ and that $q_i < 1$ if h is bounded with $h(i) < h(j)$ for some $i \notin E_0$ and all $j \in E_0$. [*Hint*: Consider $\{\widetilde{X}_n\}$ evolving as $\{X_n\}$ except that $\widetilde{p}_{01} = 1$ rather than $\widetilde{p}_{01} = 0$, and use h as test function for $\{\widetilde{X}_n\}$.] Show in particular that if $\mathbb{E}[X_{n+1} \,|\, X_n] \le X_n$ (i.e. the expected number of children per individual does not exceed 1), then extinction occurs a.s.

5.3 Carry out the last part of the proof of Proposition 5.5.

Notes Results of type Proposition 5.3, 5.4(i) have a long history and are often referred to as *Foster's criteria*. A main reference for test function techniques is Meyn and Tweedie (1993), who also treat the case of an uncountable state space (essentially, all results carry over at the cost of more tedious proofs and formulations). It is known that many of the sufficient conditions given above are also necessary in the sense that a test function with the stated properties must exist. However, finding the appropriate one is far from easy in more complicated models; Brémaud (1999) surveys a number of examples dealing with nonstandard queueing models. See also Fayolle *et al.* (1995).

6 Nonnegative Matrices

Finite square matrices with nonnegative elements occur in a variety of contexts in applied probability. The so–called *Perron–Frobenius theory* of such matrices describes in quite some detail their spectral properties (and therefore also the asymptotic properties of their powers), and is therefore a powerful and indispensable tool for many applications. We shall here develop this theory by exploiting the intimate connection to Markov chains with a finite number of states.

We start by recalling some facts from linear algebra. Let $\boldsymbol{A}$ be any $p \times p$ matrix and define for $\lambda \in \mathbb{C}$ $E_\lambda = \{\boldsymbol{x} \in \mathbb{C}^p : \boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}\}$. Thus $\mathrm{sp}(\boldsymbol{A}) = \{\lambda : E_\lambda \neq \emptyset\}$ is the set of eigenvalues of $\boldsymbol{A}$ or the *spectrum* of $\boldsymbol{A}$, and $\mathrm{spr}(\boldsymbol{A}) = \sup\bigl\{|\lambda| : \lambda \in \mathrm{sp}(\boldsymbol{A})\bigr\}$ is the *spectral radius* of $\boldsymbol{A}$. If $\lambda \in \mathrm{sp}(\boldsymbol{A})$, then λ is a root in the characteristic polynomial $\det(\boldsymbol{A} - \lambda \boldsymbol{I})$, and if the multiplicity is 1, we call λ *simple*. Then also the geometric multiplicity $\dim\bigl(E_\lambda \cup \{0\}\bigr)$ is 1, i.e. the eigenvector is unique up to a constant. If $\lambda \in \mathrm{sp}(\boldsymbol{A})$, then λ is also eigenvalue for the transposed matrix $\boldsymbol{A}^\mathsf{T}$. The existence of an eigenvector for $\boldsymbol{A}^\mathsf{T}$ then means that $\boldsymbol{\nu}\boldsymbol{A} = \lambda \boldsymbol{\nu}$ for some row vector $\boldsymbol{\nu} \neq \boldsymbol{0}$, called a *left eigenvector* for $\boldsymbol{A}$ ($\boldsymbol{x} \in E_\lambda$ is a *right eigenvector*). The following lemma is standard (all statements are easy to verify if one writes $\boldsymbol{A}$ on the Jordan canonical form):

Lemma 6.1 (i) $\mathrm{sp}(\boldsymbol{A}^m) = \bigl\{\lambda^m : \lambda \in \mathrm{sp}(\boldsymbol{A})\bigr\}$*; (ii) the* $\boldsymbol{A}^m$*–multiplicity of* $\lambda \in \mathrm{sp}(\boldsymbol{A})$ *is the sum of the* $\boldsymbol{A}$*–multiplicities of the* $\lambda_i \in \mathrm{sp}(\boldsymbol{A})$ *with* $\lambda_i^m = \lambda$*; (iii) if* $\lambda \in \mathrm{sp}(\boldsymbol{A})$ *is not simple, then either* $\dim\bigl(E_\lambda \cup \{0\}\bigr) > 1$ *or for any* $\boldsymbol{h} \in E_\lambda$ *we can find* $\boldsymbol{k}$ *with* $\boldsymbol{A}\boldsymbol{k} = \boldsymbol{h} + \lambda \boldsymbol{k}$*; (iv)* $\boldsymbol{A}^n = \mathrm{O}\bigl(n^k[\mathrm{spr}(\boldsymbol{A})]^n\bigr)$ *for some* $k = 0, 1, 2, \ldots$*.*

We start by examining the spectral properties of ergodic transition matrices:

Proposition 6.2 *Let* $\boldsymbol{P} = (p_{ij})_{i,j=1,\ldots,p}$ *be an ergodic* $p \times p$ *transition matrix with stationary distribution* $\boldsymbol{\pi}$*. Then* $\mathrm{spr}(\boldsymbol{P}) = 1$ *and* 1 *is a simple eigenvalue of* $\boldsymbol{P}$ *with* $\boldsymbol{\pi}$ *and* $\boldsymbol{1} = (1 \cdots 1)^\mathsf{T}$ *as corresponding left and right eigenvectors. Furthermore for* $\lambda \in \mathrm{sp}(\boldsymbol{P})$*,* $\lambda \neq 1$*, we have* $|\lambda| < 1$ *and with* $\lambda_1 = \max\bigl\{|\lambda| : \lambda \in \mathrm{sp}(\boldsymbol{P}), \lambda \neq 1\bigr\}$ *it holds for some* k *that the powers* $\boldsymbol{P}^n = (p_{ij}^n)$ *satisfy*

$$p_{ij}^n \;=\; \pi_j + \mathrm{O}(n^k \lambda_1^n), \quad n \to \infty. \tag{6.1}$$

Proof. It is clear that $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$, $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{1}$ and hence $1 \in \mathrm{sp}(\boldsymbol{P})$. Also $\boldsymbol{h} \in E_1$ means that $\boldsymbol{h}$ is harmonic and thus $\boldsymbol{h} = c\boldsymbol{1}$ (cf. Proposition 5.1; the extension to the complex case is easy). Thus if 1 is not simple, Lemma 6.1(iii) shows that we can find $\boldsymbol{k}$ with $\boldsymbol{P}\boldsymbol{k} = \boldsymbol{1} + \boldsymbol{k}$. But then $\boldsymbol{P}^n\boldsymbol{k} = n\boldsymbol{1} + \boldsymbol{k}$ which in Markov chain terms means that $\mathbb{E}_i k_{X_n} = n + k_i$, contradicting that $\boldsymbol{k}$ is bounded in the finite case. Similarly, the ergodic theorem means that $\boldsymbol{P}^n \to \mathbf{1}\boldsymbol{\pi}$ and hence if $\lambda \in \mathrm{sp}(\boldsymbol{P})$, $\boldsymbol{k} \in E_\lambda$, we have $\lambda^n \boldsymbol{k} = \boldsymbol{P}^n \boldsymbol{k} \to \mathbf{1}\boldsymbol{\pi}\boldsymbol{k}$. But $\lambda^n \boldsymbol{k}$ can only converge if $|\lambda| < 1$ or $\lambda = 1$.

It only remains to prove (6.1). Write $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$ with $\boldsymbol{P}_1 = \mathbf{1}\boldsymbol{\pi}$, $\boldsymbol{P}_2 = \boldsymbol{P} - \mathbf{1}\boldsymbol{\pi}$. It is then readily checked that $\boldsymbol{P}_1\boldsymbol{P}_2 = \boldsymbol{P}_2\boldsymbol{P}_1 = \mathbf{0}$ and hence $\boldsymbol{P}^n = \boldsymbol{P}_1^n + \boldsymbol{P}_2^n$. It is also easily seen that $\boldsymbol{P}_1^n = \boldsymbol{P}_1 = \mathbf{1}\boldsymbol{\pi}$. Hence by Lemma 6.1(iv) it suffices to show that if $\lambda \in \mathrm{sp}(\boldsymbol{P}_2)\setminus\{0\}$, then $|\lambda| \le \lambda_1$. But from $\boldsymbol{P}_2\boldsymbol{k} = \lambda\boldsymbol{k}$ we get $\boldsymbol{P}\boldsymbol{k} = (\boldsymbol{P}_1 + \boldsymbol{P}_2)\lambda^{-1}\boldsymbol{P}_2\boldsymbol{k} = \lambda\boldsymbol{k}$, i.e. $\lambda \in \mathrm{sp}(\boldsymbol{P})$. If $\lambda = 1$, we would have $\boldsymbol{k} = c\mathbf{1}$ and hence $\boldsymbol{P}_2\boldsymbol{k} = 0$ which is impossible. Hence $|\lambda| \le \lambda_1$. □

If $\lambda_1 < \delta < 1$, (6.1) may be rewritten as $p_{ij}^n = \pi_j + \mathrm{O}(\delta^n)$, and we have obtained a second proof of Proposition 4.4(b), stating that any irreducible finite Markov chain is geometrically ergodic.

A matrix $\boldsymbol{Q}$ is *nonnegative* if $q_{ij} \ge 0$ for all i,j, and *substochastic* if also $\boldsymbol{Q}\mathbf{1} \le \mathbf{1}$, i.e. the rows sums are at most 1. The following result is often used and holds under weaker conditions than irreducibility:

Proposition 6.3 *Let $\boldsymbol{Q}$ be substochastic, such that to each i there is a k and $j_1,\ldots,j_m$ with $\sum_\ell q_{k\ell} < 1$ and $q_{ij_1}q_{j_1j_2}\ldots q_{j_mk} > 0$. Then* $\mathrm{spr}(\boldsymbol{Q}) < 1$.

Proof. Let λ be an eigenvalue of absolute value $\mathrm{spr}(\boldsymbol{Q})$ and let $\boldsymbol{h} \in E_\lambda$. Consider a Markov chain $\{X_n\}$ on $\{0,1,\ldots,p\}$ such that 0 is absorbing, and the probability of a transition $i \to j$ is q_{ij} for $i,j \ge 1$ and $1 - \sum_\ell q_{i\ell}$ for $j = 0$. The assumptions on $\boldsymbol{Q}$ and a geometrical trials argument (cf. A6.1) then easily yield that $X_n = 0$ eventually and that taking $h_0 = 0$ makes $\lambda^{-n}h_{X_n}$ a martingale. If $|\lambda| \ge 1$, boundedness would imply L_1–convergence (necessarily to h_0) so that taking $X_0 = i$ yields $h_i = h_0 = 0$ which contradicts $\boldsymbol{h} \ne \mathbf{0}$. Hence $|\lambda| < 1$ and $\mathrm{spr}(\boldsymbol{Q}) < 1$. □

We shall now derive a close analogue of Proposition 6.2 for nonnegative matrices $\boldsymbol{A}$. We shall adopt the definitions of irreducibility and the period d from transition matrices to nonnegative matrices by noting that they depend only on the pattern of entries i,j with $a_{ij} > 0$. Thus $\boldsymbol{A}$ is *irreducible* if for any i,j we can find m such that $a_{ij}^m > 0$, and we have:

Lemma 6.4 *If $\boldsymbol{A}$ is an irreducible nonnegative matrix, then the greatest common divisor d of the m with $a_{ii}^m > 0$ does not depend on i. If $d = 1$, then it holds for all sufficiently large m that $a_{ij}^m > 0$ for all i,j.*

Proof. Choose a transition matrix $\boldsymbol{P}$ with $p_{ij} > 0$ for exactly the same i,j as for which $a_{ij} > 0$. Then $a_{ij}^m > 0$ precisely when $p_{ij}^m > 0$ and results from Section 1 complete the proof. □

The d in Lemma 6.4 is called the *period* of $\boldsymbol{A}$, and $\boldsymbol{A}$ is *aperiodic* if $d = 1$.

Theorem 6.5 (PERRON–FROBENIUS) *Let $\boldsymbol{A}$ be an irreducible non–negative $p \times p$ matrix. Then:*
(a) *the spectral radius λ_0 of $\boldsymbol{A}$ is strictly positive and a simple eigenvalue of $\boldsymbol{A}$ with the corresponding left and right eigenvectors $\boldsymbol{\nu}$, $\boldsymbol{h}$ satisfying $\nu_i > 0$,*

$h_i > 0$ for all i;
(b) *if $\boldsymbol{A}$ is also aperiodic, then $\lambda_1 = \max\bigl\{|\lambda| : \lambda \in \mathrm{sp}(\boldsymbol{A})\backslash\{\lambda_0\}\bigr\} < \lambda_0$. Furthermore, if we normalize $\boldsymbol{\nu}, \boldsymbol{h}$ by $\boldsymbol{\nu h} = \sum_1^p \nu_i h_i = 1$, then for some k*

$$\boldsymbol{A}^n \;=\; \lambda^n \boldsymbol{h\nu} + \mathrm{O}(n^k \lambda_1^n), \quad n \to \infty; \tag{6.2}$$

(c) *if $\boldsymbol{A}$ has period $d > 1$, then $|\lambda| \le \lambda_0$ for any $\lambda \in \mathrm{sp}(\boldsymbol{A})$. Furthermore, $\lambda \in \mathrm{sp}(\boldsymbol{A})$, $|\lambda| = \lambda_0$ holds exactly when λ is of the form $\lambda_0 \theta^k$, $k = 0, 1, \ldots, d-1$, with $\theta^k = \mathrm{e}^{2\pi k/d}$ the roots of unity.*

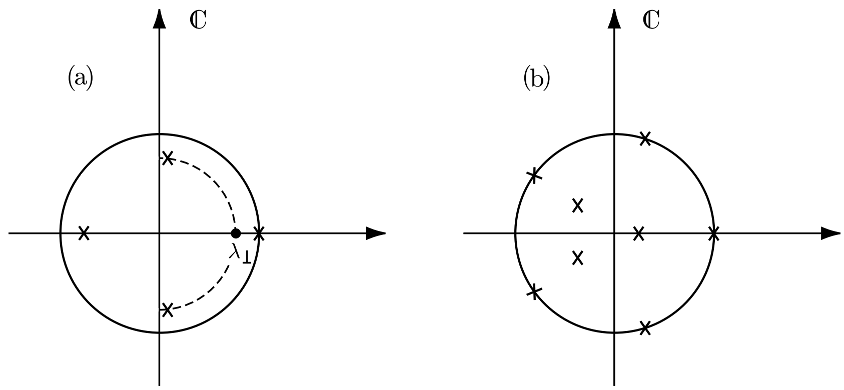

Figure 6.1

Figure 6.1 depicts $\mathrm{sp}(A)$ for the aperiodic case in (a) and for the periodic case $d = 5$ in (b). The eigenvalues fall in pairs of complex conjugates since $\boldsymbol{A}$ is real. We shall refer to λ_0 as the *Perron–Frobenius root* of $\boldsymbol{A}$.

The proof of the Perron–Frobenius theorem will be reduced to the Markov case in Proposition 6.2. We need some lemmas.

Lemma 6.6 *If $\boldsymbol{A}$ has all $a_{ij} > 0$, then there exists $\lambda \in \mathrm{sp}(\boldsymbol{A})$, $x \in E_\lambda$ with $\lambda > 0$, $x_i > 0$, $i = 1, \ldots, p$.*

Proof. The basic observation is that all $a_{ij} > 0$ implies

$$x_i \ge 0,\ \sum_{i=1}^p x_i > 0 \quad \Rightarrow \quad \text{all components of } \boldsymbol{Ax} \text{ are } > 0. \tag{6.3}$$

Define

$$\begin{aligned} K &= \Bigl\{\boldsymbol{x} \in \mathbb{R}^p : 0 \le x_i \le 1, \sum_{i=1}^p x_i = 1\Bigr\}, \\ S &= \bigl\{\mu \ge 0 : \boldsymbol{Ax} \ge \mu \boldsymbol{x} \text{ for some } \boldsymbol{x} \in K\bigr\}, \end{aligned}$$

$\lambda = \sup\{\mu : \mu \in S\}$. Since $\boldsymbol{A}K$ is compact, $\lambda < \infty$. For a given $\boldsymbol{x} \in K$, (6.3) implies $\boldsymbol{Ax} \ge \epsilon \boldsymbol{x}$ for small enough ϵ, and hence $\lambda > 0$. Now choose $\lambda_n \in S$, $\boldsymbol{x}_n \in K$ with $\lambda_n \uparrow \lambda$, $\boldsymbol{Ax}_n \ge \lambda_n \boldsymbol{x}_n$. Passing to a subsequence if necessary, we may assume that $\boldsymbol{x} = \lim \boldsymbol{x}_n$ exists. Then $\boldsymbol{Ax} \ge \lambda \boldsymbol{x}$ and we shall complete the proof by showing that indeed $\boldsymbol{Ax} = \lambda \boldsymbol{x}$ ($x_i > 0$ is then ensured by (6.3)). Otherwise let $\boldsymbol{y} = c\boldsymbol{A}\boldsymbol{x}$ with $c > 0$ chosen so that $\boldsymbol{y} \in K$. Then $\boldsymbol{A}\boldsymbol{y} - \lambda\boldsymbol{y} = c\boldsymbol{A}(\boldsymbol{A}\boldsymbol{x} - \lambda\boldsymbol{x})$ has all components > 0 by (6.3). Hence $\boldsymbol{A}\boldsymbol{y} \geq (\lambda + \epsilon)\boldsymbol{y}$ for some $\epsilon > 0$, a contradiction. □

Lemma 6.7 *Suppose that $\boldsymbol{A}\boldsymbol{k} = \lambda\boldsymbol{k}$ with $\lambda > 0$ and all $k_i > 0$. Then the matrix $\boldsymbol{P}$ with elements $a_{ij}k_j/\lambda k_i$ is a transition matrix, $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{1}$, and the formulas*

$$\lambda^{\boldsymbol{A}} = \lambda\lambda^{\boldsymbol{P}}, \quad h_i^{\boldsymbol{A}} = k_i h_i^{\boldsymbol{P}}, \quad \pi_i^{\boldsymbol{A}} = \pi_i^{\boldsymbol{P}}/k_i$$

establish a one–to–one correspondence between $\lambda^{\boldsymbol{A}} \in \mathrm{sp}(\boldsymbol{A})$ and $\lambda^{\boldsymbol{P}} \in \mathrm{sp}(\boldsymbol{P})$ and the corresponding right and left eigenvectors ($\boldsymbol{\pi}^{\boldsymbol{A}}\boldsymbol{A} = \lambda^{\boldsymbol{A}}\boldsymbol{\pi}^{\boldsymbol{A}}$ etc.). Furthermore, $\lambda^{\boldsymbol{A}}$ is simple for $\boldsymbol{A}$ if and only if $\lambda^{\boldsymbol{P}}$ is simple for $\boldsymbol{P}$.

Proof. Everything is a straightforward verification except for the last statement which follows from

$$\det(\boldsymbol{P} - \mu\boldsymbol{I}) \;=\; \det(\lambda^{-1}\boldsymbol{A} - \mu\boldsymbol{I}) \;=\; \lambda^{-p}\det(\boldsymbol{A} - \mu\lambda\boldsymbol{I}).$$

Indeed, multiplying the ith row by k_i and the jth column by k_j^{-1} leaves the determinant unchanged and transform $\boldsymbol{P}$ into $\lambda^{-1}\boldsymbol{A}$, $\boldsymbol{I}$ into $\boldsymbol{I}$. □

Proof of Theorem 6.5 *in the aperiodic case.* Choose first m with all $a_{ij}^m > 0$, cf. Lemma 6.4, and next λ, k with $\boldsymbol{A}^m\boldsymbol{k} = \lambda\boldsymbol{k}$, $\lambda > 0$, all $k_i > 0$, cf. Lemma 6.6. Then by Lemma 6.7 1 is simple for $\boldsymbol{P}^m = (a_{ij}^m k_j/k_i)$ and hence λ simple for $\boldsymbol{A}^m$. If $\lambda_0 \in \mathrm{sp}(\boldsymbol{A})$ satisfies $\lambda_0^m = \lambda$, then by Lemma 6.1(ii) λ_0 is simple for $\boldsymbol{A}$. Choose $\boldsymbol{h} \in E_{\lambda_0}$. Then $\boldsymbol{A}^m\boldsymbol{h} = \lambda_0^m\boldsymbol{h} = \lambda\boldsymbol{h}$, and since λ is simple for $\boldsymbol{A}^m$, it follows that we may take $\boldsymbol{h} = \boldsymbol{k}$. Then by nonnegativity, $\boldsymbol{A}\boldsymbol{h} = \lambda_0\boldsymbol{h}$ implies $\lambda_0 > 0$ and $\boldsymbol{P} = (a_{ij}k_j/\lambda_0 k_i)$ is a transition matrix. Applying Proposition 6.2 and Lemma 6.7 everything then comes out in a straightforward manner. For (6.2), note that if $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$, $\boldsymbol{\pi}\boldsymbol{1} = 1$ and we let $\nu_i = \pi_i/h_i$, then $\boldsymbol{\nu}\boldsymbol{A} = \lambda_0\boldsymbol{\nu}$, $\boldsymbol{\nu}\boldsymbol{h} = 1$ and

$$a_{ij}^n \;=\; \lambda_0^n \frac{p_{ij}^n h_i}{h_j} \;=\; \lambda_0^n \frac{h_i}{h_j}\left\{\pi_j + \mathrm{O}\Big(n^k\Big(\frac{\lambda_1}{\lambda_0}\Big)^n\Big)\right\} \;=\; \lambda_0^n h_i \nu_j + \mathrm{O}(n^k\lambda_1^n).$$

□

Proof of Theorem 6.5 *in the periodic case.* We can reorder the coordinates by a cyclic class argument so that $\boldsymbol{A}$ has the form

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{A}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & & \boldsymbol{0} \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{A}_{d-1} \\ \boldsymbol{A}_d & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} \end{pmatrix}.$$

Letting $\boldsymbol{B}_k = \boldsymbol{A}_k\boldsymbol{A}_{k+1}\cdots\boldsymbol{A}_d\boldsymbol{A}_1\cdots\boldsymbol{A}_{k-1}$, it follows that $\boldsymbol{A}^d$ is block–diagonal with diagonal elements $\boldsymbol{B}_k$ which are irreducible aperiodic. Let μ_k be the Perron–Frobenius root of $\boldsymbol{B}_k$ and $\boldsymbol{B}_k\boldsymbol{h}^{(k)} = \mu_k\boldsymbol{h}^{(k)}$ with $h_i^{(k)} > 0$.

Now

$$\boldsymbol{B}_k \boldsymbol{A}_k \boldsymbol{h}^{(k+1)} \;=\; \boldsymbol{A}_k \boldsymbol{B}_{k+1} \boldsymbol{h}^{(k+1)} \;=\; \mu_{k+1} \boldsymbol{A}_k \boldsymbol{h}^{(k+1)}$$

(identifying $d+1$ with 1). Since $\boldsymbol{A}_k \boldsymbol{h}^{(k+1)} \neq \boldsymbol{0}$, it follows that $\mu_{k+1} \in \mathrm{sp}(\boldsymbol{B}_k)$ and hence $\mu_{k+1} \le \mu_k$. Hence all μ_k are equal, say $\mu_k = \mu$, and we may take $\boldsymbol{h}^{(k)} = \boldsymbol{A}_k \boldsymbol{h}^{(k+1)} = \boldsymbol{A}_k \ldots \boldsymbol{A}_{d-1} \boldsymbol{h}^{(d)}$. Now

$$\det(\boldsymbol{A}^d - \eta \boldsymbol{I}) \;=\; \prod_{k=1}^{d} \det(\boldsymbol{B}_k - \eta \boldsymbol{I}).$$

This shows that if $\lambda \in \mathrm{sp}(\boldsymbol{A})$, then $\eta = \lambda^d$ is in $\mathrm{sp}(\boldsymbol{B}_k)$ for some k. Hence $|\lambda| = |\eta|^{1/d} \le |\mu|^{1/d} = \lambda_0$ (say) and $|\lambda| = \lambda_0$ can only occur if $\lambda^d = \mu$, i.e. λ is of the form $\lambda_0 \theta^k$ for some k. Also the $\boldsymbol{A}^d$–multiplicity of μ is exactly d. By Lemma 6.1(ii) the proof is now complete if we can show that each $\lambda_0 \theta^k$ is an eigenvalue and that $\boldsymbol{z}^{(0)} \in E_{\lambda_0}$ may be taken with all $z_i^{(0)} > 0$. But an easy calculation shows that

$$\boldsymbol{z}^{(k)} \;=\; \Bigl((\lambda_0 \theta^k)^0 \boldsymbol{h}^{(1)\mathsf{T}} \;\cdots\; (\lambda_0 \theta^k)^{d-1} \boldsymbol{h}^{(d)\mathsf{T}} \Bigr)^{\mathsf{T}}$$

satisfies $\boldsymbol{A}\boldsymbol{z}^{(k)} = \lambda_0 \theta^k \boldsymbol{z}^{(k)}$. $\square$

Problems

6.1 Is it true that if $\boldsymbol{P}$ is an *infinite* ergodic transition matrix, then all $p_{ij}^n > 0$ for some n?

6.2 Suppose that $\boldsymbol{A}$ is an irreducible aperiodic nonnegative matrix such that $\boldsymbol{A}^m$ is a transition matrix for some $m = 1, 2, \ldots$ Show that then $\boldsymbol{A}$ is itself a transition matrix. Show also that the result fails in the periodic case.

6.3 Let $\boldsymbol{A}$ be irreducible and nonnegative, and assume that $\boldsymbol{A}\boldsymbol{x} \le \lambda \boldsymbol{x}$ with $\boldsymbol{x} \ge \boldsymbol{0}$, $\boldsymbol{x} \neq \boldsymbol{0}$ and $\lambda > 0$. Show that $\mathrm{spr}(\boldsymbol{A}) \le \lambda$ provided either (i) $\boldsymbol{A}$ is irreducible, or (ii) all $x_i > 0$. Show also in case (i) that $\mathrm{spr}(\boldsymbol{A}) < \lambda$ if in addition $\boldsymbol{A}\boldsymbol{x} \neq \lambda \boldsymbol{x}$.

Notes Standard references for nonnegative matrices are Berman and Plemmons (1994) and Seneta (1994). Of extensions of the Perron–Frobenius theorem, we mention in particular operator versions such as the Krein–Rutman theorem, e.g. Schaefer (1970), and the more probabilistic inspired discussion of Nummelin (1984).

7 The Fundamental Matrix, Poisson's Equation and the CLT

We assume throughout this section that $\{X_n\}$ is irreducible positive recurrent with stationary distribution $\boldsymbol{\pi}$.

Let f be a real–valued function on E, sometimes written as a column vector $\boldsymbol{f}$ (see Notes to Section 1 for this and other notational issues like $\pi(f)$ versus $\boldsymbol{\pi f}$, Pf versus $\boldsymbol{Pf}$, etc.). The equation

$$\boldsymbol{g} \;=\; \boldsymbol{f} + \boldsymbol{Pg}, \tag{7.1}$$

with $\boldsymbol{g}$ the unknown, is referred to as *Poisson's equation*.

Proposition 7.1 *Assume that f is π–integrable. Then:* (i) *a necessary condition for the existence of a $\boldsymbol{\pi}$–integrable solution to Poisson's equation is $\pi(f) = 0$;* (ii) *a π–integrable solution is unique up to a multiple of $\mathbf{1}$;* (iii) *if $\pi(f) = 0$, then for any k $g(i) = \mathbb{E}_i \sum_0^{\tau(k)-1} f(X_n)$ is a finite solution satisfying $g(k) = 0$.*

Proof. Multiplying (7.1) by $\boldsymbol{\pi}$ immediately gives (i). If $\boldsymbol{g}_1, \boldsymbol{g}_2$ are solutions, then $\boldsymbol{d} = \boldsymbol{g}_1 - \boldsymbol{g}_2$ satifies $\boldsymbol{d} = \boldsymbol{Pd}$, i.e. $\boldsymbol{d}$ is harmonic and must therefore be constant by Proposition 5.1, showing (ii). In (iii), we have from Corollary 3.6 that

$$\pi(|f|)\mathbb{E}_k\tau(k) \;=\; \mathbb{E}_k \sum_{n=0}^{\tau(k)-1} |f|(X_n) \;\geq\; \mathbb{P}_k\big(\tau(j) < \tau(k)\big)\, \mathbb{E}_j \sum_{n=0}^{\tau(k)-1} |f|(X_n).$$

This shows first that $g(j)$ is finite and next, upon replacing $|f|$ by f in the left identity, that $g(k) = 0$. Conditioning upon X_1 and using the definition of g then gives

$$g(i) \;=\; f(i) + \sum_{j\neq k} p_{ij} g(j) \;=\; f(i) + \sum_{j\in E} p_{ij} g(j) \;=\; f(i) + Pg(i),$$

which is the same as (7.1). □

Theorem 7.2 *Let $\boldsymbol{f}$ be a $\boldsymbol{\pi}$–integrable function on E and define $\widetilde{f}(i) = f(i) - \pi(f)$. Assume that g is a solution of $g = \widetilde{f} + Pg$ and that $\pi(g^2) < \infty$. Then*

$$\frac{1}{\sqrt{n}}\big(f(X_0) + \cdots + f(X_{n-1}) - n\pi(f)\big) \;\overset{\mathscr{D}}{\to}\; N(0, \sigma^2(f)) \tag{7.2}$$

where $\sigma^2(f) = \pi(g^2) - \pi\big((Pg)^2\big)$.

Proof. We may assume w.l.o.g. that $\pi(f) = 0$ so that $f = \widetilde{f}$. Let $\Delta_k = g(X_k) - Pg(X_{k-1})$. Then $g = f + Pg$ implies

$$\sum_{k=0}^{n} f(X_k) \;=\; g(X_0) - Pg(X_n) + \sum_{k=1}^{n} \Delta_k. \tag{7.3}$$

Since $Pg(X_{k-1}) = \mathbb{E}(g(X_k)\,|\,\mathscr{F}_{k-1})$, the sequence $\{\Delta_k\}$ is a martingale difference sequence, and we have

$$\mathbb{V}ar(\Delta_k\,|\,\mathscr{F}_{k-1}) \;=\; \mathbb{V}ar(g(X_k)\,|\,\mathscr{F}_{k-1}) \;=\; \omega^2(X_{k-1})$$

where $\omega^2(i) = g^2(i) - (Pg)^2(i)$. Here $\pi(\omega^2) = \sigma^2(f)$ is finite by assumption so that $\sum_1^n \mathbb{V}ar(\Delta_k \mid \mathscr{F}_{k-1})/n \to \sigma^2(f)$ by the LLN (4.3). Therefore an appropriate martingale CLT (e.g. Hall and Heyde, 1980, p. 58, or Shiryaev, 1996, p. 541) shows that $\sum_1^n \Delta_k/n^{1/2}$ has a limiting $N(0, \sigma^2(f))$ distribution. In view of (7.3), this is equivalent to the assertion of the theorem.

□

Now assume that E is finite and define the *fundamental matrix* $\boldsymbol{Z}$ by

$$\boldsymbol{Z} \;=\; (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{1\pi})^{-1} \;=\; \sum_{n=0}^{\infty} (\boldsymbol{P} - \boldsymbol{1\pi})^n \;=\; \boldsymbol{I} + \sum_{n=1}^{\infty} (\boldsymbol{P}^n - \boldsymbol{1\pi}). \quad (7.4)$$

Note that by Proposition 6.2 we have $|\lambda| < 1$ for any eigenvalue of $\boldsymbol{P} - \boldsymbol{1\pi}$ when $\boldsymbol{P}$ is aperiodic, so that the first series converges and equals the inverse; the last expression for $\boldsymbol{Z}$ follows by verifying by induction that $(\boldsymbol{P} - \boldsymbol{1\pi})^n = \boldsymbol{P}^n - \boldsymbol{1\pi}$ (we omit the easy proof that (7.4) also holds in the periodic case). Some easily verified identities are

$$\boldsymbol{\pi Z} = \boldsymbol{\pi}, \quad \boldsymbol{Z1} = \boldsymbol{1}, \quad \boldsymbol{PZ} = \boldsymbol{ZP} = \boldsymbol{Z} - \boldsymbol{I} + \boldsymbol{1\pi}. \quad (7.5)$$

Proposition 7.3 *Assume that E is finite. Then if $\boldsymbol{\pi f} = \boldsymbol{0}$, the unique solution $\boldsymbol{g}$ of Poisson's equation satisying $\boldsymbol{\pi g} = 0$ is $\boldsymbol{g} = \boldsymbol{Zf}$.*

Proof. From (7.5), we first get $\boldsymbol{\pi g} = \boldsymbol{\pi f} = 0$ and next

$$\boldsymbol{Pg} \;=\; (\boldsymbol{Z} - \boldsymbol{I} + \boldsymbol{1\pi})\boldsymbol{f} \;=\; \boldsymbol{g} - \boldsymbol{f} + \boldsymbol{0}. \qquad \square$$

Proposition 7.4 $z_{ij} \;=\; \begin{cases} \pi_j \mathbb{E}_{\boldsymbol{\pi}} \tau(j) & i = j \\ \pi_j \mathbb{E}_{\boldsymbol{\pi}} \tau(j) - \pi_j \mathbb{E}_i \tau(j) & i \neq j \end{cases}.$

Note in particular that whereas the calculation of $\mathbb{E}_i \tau(i) = 1/\pi_i$ is easy, so is not the case for $\mathbb{E}_i \tau(j)$, and the answer $(z_{jj} - z_{ij})/\pi_j$ is provided by Proposition 7.4.

Proof. Define $\boldsymbol{f} = \boldsymbol{1}_j - \pi_j \boldsymbol{1}$. Then $\pi(f) = 0$, and so by Proposition 7.1(iii) the solution g of Poisson's equation with $g(j) = 0$ is given by

$$g(i) \;=\; \mathbb{E}_i \sum_{n=0}^{\tau(j)-1} I(X_n = j) - \pi_j \mathbb{E}_i \tau(j) \;=\; \delta_{ij} - \pi_j \mathbb{E}_i \tau(j).$$

Thus the solution g^* satifying $\pi(g^*) = 0$ is

$$g^*(i) \;=\; g(i) - \pi(g) \;=\; \delta_{ij} - \pi_j \mathbb{E}_i \tau(j) - \pi_j + \pi_j \mathbb{E}_{\boldsymbol{\pi}} \tau(j).$$

On the other hand, by Proposition 7.3 we have

$$g^*(i) \;=\; \boldsymbol{1}_i' \boldsymbol{Zf} \;=\; z_{ij} - \pi_j.$$

Equating these two expressions yields the result (if $i = j$, note that $\pi_j \mathbb{E}_i \tau(j) = \delta_{ij} = 1$). □

Corollary 7.5 *In the finite case, $\sigma^2(\boldsymbol{f}) = \boldsymbol{\pi}(2\boldsymbol{f} \bullet \boldsymbol{Zf} - \boldsymbol{f} \bullet \boldsymbol{f}) - \overline{f}^2$ where $\overline{f} = \boldsymbol{\pi f}$ and $\bullet$ denotes multiplication element by element, $(\boldsymbol{a} \bullet \boldsymbol{b})_i = a_i b_i$.*

Proof. We have $\pi(\boldsymbol{f}) = \overline{f}$, $\boldsymbol{g} = \boldsymbol{Z}(\boldsymbol{f} - \overline{f}\boldsymbol{1})$ in Theorem 7.2 and thus

$$\begin{aligned}
\sigma^2(\boldsymbol{f}) &= \boldsymbol{\pi}(\boldsymbol{g} \bullet \boldsymbol{g} - \boldsymbol{P}\boldsymbol{g} \bullet \boldsymbol{P}\boldsymbol{g}) \\
&= \boldsymbol{\pi}\big((\boldsymbol{Z}\boldsymbol{f} - \overline{f}\boldsymbol{1}) \bullet (\boldsymbol{Z}\boldsymbol{f} - \overline{f}\boldsymbol{1}) - (\boldsymbol{Z}\boldsymbol{f} - \boldsymbol{f}) \bullet (\boldsymbol{Z}\boldsymbol{f} - \boldsymbol{f})\big) \\
&= \boldsymbol{\pi}\big(\overline{f}^2\boldsymbol{1} - 2\overline{f}\boldsymbol{Z}\boldsymbol{f} - \boldsymbol{f} \bullet \boldsymbol{f} + 2\boldsymbol{f} \bullet \boldsymbol{Z}\boldsymbol{f}\big) \\
&= \overline{f}^2 - 2\overline{f}^2 + \boldsymbol{\pi}(-\boldsymbol{f} \bullet \boldsymbol{f} + 2\boldsymbol{f} \bullet \boldsymbol{Z}\boldsymbol{f}),
\end{aligned}$$

where we used (7.5) repeatedly in the second step. □

8 Foundations of the General Theory of Markov Processes

We shall consider two generalizations, first that of a general (not necessarily countable) state space E, and next that of a continuous time parameter $t \in [0, \infty)$.

In the general state space case, one needs to assume that E is equipped with a measurable structure, i.e. a σ–algebra $\mathscr{E}$ to which all subsets of E considered in the following are assumed to belong. Instead of the transition matrix we have a *transition* (or *Markov*) *kernel*, i.e. a function $P(x, A)$ of $x \in E$ and $A \in \mathscr{E}$ such that $P(x, \cdot)$ is a probability on $(E, \mathscr{E})$ for each x and $P(\cdot, A)$ is $\mathscr{E}$–measurable for each A.

Markov chains with transition kernel P and the corresponding Markov probabilities $\mathbb{P}_\mu$ (with μ a distribution on $(E, \mathscr{E})$) are defined by the requirements $\mathbb{P}_\mu(X_0 \in A) = \mu(A)$,

$$\mathbb{P}_\mu(X_{n+1} \in A \,|\, \mathscr{F}_n) \;=\; P(X_n, A) \tag{8.1}$$

where $\mathscr{F}_n = \sigma(X_0, \ldots, X_n)$. With the usual a.s. interpretation of conditional probabilities, it follows from (8.1) that

$$\mathbb{P}_\mu(X_{n+1} \in A \,|\, X_n = x) \;=\; P(x, A) \tag{8.2}$$

Also, say by induction, one easily gets

$$\begin{aligned}
&\mathbb{P}_\mu(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) \\
&\quad = \int_{A_0} \mu(\mathrm{d}x_0) \int_{A_1} P(x_0, \mathrm{d}x_1) \cdots \int_{A_{n-1}} P(x_{n-2}, \mathrm{d}x_{n-1}) P(x_{n-1}, A_n).
\end{aligned} \tag{8.3}$$

This formula also immediately suggests how to define the Markov probabilities and the Markov chain: take $X_0, X_1, \ldots$ as the projections $E^{\mathbb{N}} \to E$ and let

$$\mathscr{E}_n = \sigma(X_0, \ldots, X_n), \quad \mathscr{E}_\infty = \sigma(X_0, X_1, \ldots) = \mathscr{E}^{\mathbb{N}}.$$

Then by standard arguments from measure theory it can be seen that the r.h.s. of (8.3) in a unique way corresponds to a probability $\mathbb{P}_\mu^n$ on

$(E^{\mathbb{N}}, \mathscr{E}_n)$. The $\mathbb{P}_\mu^n$ have the consistency property $\mathbb{P}_\mu^n(A) = \mathbb{P}_\mu^m(A)$, $m \le n$, $A \in \mathscr{E}_m$, and hence define a finitely additive probability on the algebra $\cup_0^\infty \mathscr{E}_n$. The desired $\mathbb{P}_\mu$ is the (necessarily unique) extension to $\mathscr{E}_\infty = \sigma\left(\cup_0^\infty \mathscr{E}_n\right)$. The existence, i.e. the σ–additivity on $\cup_0^\infty \mathscr{E}_n$, may be seen either from Kolmogorov's consistency theorem which requires some topological assumptions like E being Polish and $\mathscr{E}$ the Borel σ–algebra, or by a measure–theoretic result of Ionescu Tulcea (see Neveu, 1965).

The continuous–time case is substantially more involved. What will be needed in later chapters is, however, only a few basic facts and we shall therefore just outline a theory which needs several amendments when pursuing Markov process theory in its full generality.

One does not get very far without topology, so we assume right from the start that E is Polish with $\mathscr{E}$ the Borel σ–algebra. That a process $\{X_t\}_{t\ge 0}$ with state space E is Markov means intuitively just the same as in discrete time: given the history $\mathscr{F}_t = \sigma(X_s;\, s \le t)$, the process evolves from then on as restarted at time 0 in state X_t and depending on $\mathscr{F}_t$ through X_t only. Formally, this may be expressed by the existence of a family of probability measures $\mathbb{P}_\mu$ with the property $\mathbb{P}_\mu(X_0 \in A) = \mu(A)$,

$$\mathbb{E}_\mu\big[h(X_{s+t};\, t \ge 0) \,\big|\, \mathscr{F}_s\big] \;=\; \mathbb{E}_{X_s} h(X_t;\, t \ge 0) \tag{8.4}$$

where $\mathbb{P}_x, \mathbb{E}_x$ refer to $X_0 = x$ and (8.4) should hold for a class of functions h of the process sufficiently rich to determine the distribution of $\{X_t\}_{t\ge 0}$. For example, it would suffice to consider the class $\mathscr{H}$ of all h of the form

$$h(x_t;\, t \ge 0) \;=\; \prod_{i=0}^{n} I(x_{t_i} \in A_i). \tag{8.5}$$

If $\{X_t\}_{t\ge 0}$ has paths say in $D = D([0,\infty), E)$, then (8.4) for all $h \in \mathscr{H}$ will be equivalent to (8.4) to hold for all bounded measurable $h: D \to \mathbb{R}$. In fact, an easy induction argument shows that it is even sufficient to let $n = 0$ in (8.5), and the Markov property in this equivalent formulation then becomes

$$\mathbb{P}(X_{s+t} \in A \,|\, \mathscr{F}_s) \;=\; P^t(X_s, A) \quad \text{where} \quad P^t(x, A) = \mathbb{P}_x(X_t \in A). \tag{8.6}$$

Given a Markov process, it is clear that $P^t(x, A)$ as defined by (8.6) is a transition kernel. Using the Markov property we get

$$P^{t+s}(x, A) \;=\; \mathbb{E}_x \mathbb{P}(X_{s+t} \in A \,|\, \mathscr{F}_s) \;=\; \mathbb{E} P^t(X_s, A) \;=\; \int P^t(y, A) P^s(x, \mathrm{d}y),$$

which in operator notation is written $P^{t+s} = P^t P^s$ and referred to as the *Chapman–Kolmogorov equations* (or the *semi–group property*). Conversely, given a family $\{P^t\}_{t\ge 0}$ satisfying the Chapman–Kolmogorov equations, it is possible to construct a corresponding Markov process. To this end, we proceed as in discrete time: let $X_t : E^{[0,\infty)} \to E$ be the projection and define for $0 = t_0 < t_1 < \cdots < t_n$ a probability on the sub–σ–algebra

$\sigma(X_{t_i};\, i = 0, \ldots, n)$ by

$$\begin{aligned}
&\mathbb{P}_\mu(X_{t_0} \in A_0, X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) \\
&\quad = \int_{A_0} \mu(\mathrm{d}x_0) \int_{A_1} P^{t_1 - t_0}(x_0, \mathrm{d}x_1) \\
&\qquad \cdots \int_{A_{n-1}} P^{t_{n-1} - t_{n-2}}(x_{n-2}, \mathrm{d}x_{n-1}) P^{t_n - t_{n-1}}(x_{n-1}, A_n). \qquad (8.7)
\end{aligned}$$

That this defines a semigroup is readily apparent from the Chapman–Kolmogorov equations. Since E is Polish, there thus exists a unique extension to $\mathscr{E}^{[0,\infty)}$, and the Markov property (8.4) with $h \in \mathscr{H}$ is inherent in the definition (8.7).

There are, however, severe difficulties associated with this approach. First, the intuitive description of a particular model is seldom in terms of the P^t. Next, the construction makes $\mathscr{E}^{[0,\infty)}$ the collection of measurable sets, i.e. when $A \notin \mathscr{E}^{[0,\infty)}$ one cannot make sense of $\mathbb{P}_\mu(A)$. But $\mathscr{E}^{[0,\infty)}$ is not very rich since one can easily see that $A \in \mathscr{E}^{[0,\infty)}$ implies that A depends on the X_t for t in a countable collection $T_A \subset [0, \infty)$ of time points. Thus for example sets like

$$\big\{\omega :\ X_t(\omega) \text{ is a continuous function of } t\big\}$$

is not in $\mathscr{E}^{[0,\infty)}$, and (when say $E = \mathbb{R}$) similarly $\max_{0 \le t \le T} X_t$ and $\inf\{t :\ X_t = 0\}$ are not measurable. Hence it is necessary to construct versions of the process with sample paths say in D. This requires further properties of the P^t, typically continuity requirements. We shall not go into this since the explicit examples that we shall encounter will almost a priori satisfy such path regularity properties. For example, queues are constructed by simple transformations of sequences of service times and interarrival times, and not starting from semi–groups, consistent families and so on.

Now let σ be a stopping time w.r.t. $\{\mathscr{F}_t\}_{t \ge 0}$ amd let $\mathscr{F}_\sigma$ be the stopping time σ–algebra, cf. A10. We say that $\{X_t\}_{t \ge 0}$ has the *strong Markov property* w.r.t. σ if a.s. on $\{\sigma < \infty\}$

$$\mathbb{P}\big(X_{\sigma+t} \in A \,\big|\, \mathscr{F}_\sigma\big) \;=\; P^t(X_\sigma, A); \qquad (8.8)$$

again, this implies a functional form

$$\mathbb{E}_\mu\big[h(X_{\sigma+t};\, t \ge 0) \,\big|\, \mathscr{F}_\sigma\big] \;=\; \mathbb{E}_{X_\sigma} h(X_t;\, t \ge 0).$$

The process is *strong Markov* if it has the strong Markov property w.r.t. any stopping time σ.

Proposition 8.1 *A Markov process $\{X_t\}_{t \ge 0}$ has the strong Markov property w.r.t. any stopping time σ which assumes only a countable number of values, $\sigma \in \{\infty, s_1, s_2, \ldots\}$.*

Proof. We must show that for $A \in \mathscr{E}$, $F \in \mathscr{F}_\sigma$

$$\mathbb{P}_\mu\big(X_{\sigma+t} \in A;\, F,\, \sigma < \infty\big) \;=\; \mathbb{E}_\mu\big[P^t(X_\sigma, A);\, F,\, \sigma < \infty\big].$$

However, if $F \subseteq \{\sigma = s_k\}$ this is immediate from the Markov property (8.4). In the general case, decompose $F \cap \{\sigma < \infty\}$ as the disjoint union of the sets $F \cap \{\sigma = s_k\}$ and sum over k. □

As an immediate consequence, we have:

Corollary 8.2 *Any discrete time Markov chain* (with discrete or general state space) *has the strong Markov property.*

Also in continuous time, Proposition 8.1 is greatly helpful in establishing the strong Markov property. A typical example is the following:

Corollary 8.3 *Assume that $\{X_t\}_{t\geq 0}$ has right–continuous paths and that for any bounded continuous $f: E \to \mathbb{R}$ and any s it holds that $\mathbb{E}_x f(X_s)$ is a continuous function of x or, more generally, that the paths of $\mathbb{E}_{X_t} f(X_s)$ are right–continuous functions of t. Then the strong Markov property holds.*

Proof. Let σ be a given stopping time and define $\sigma(k) = n2^{-k}$ on $\big\{(n-1)2^{-k} < \sigma \leq n2^{-k}\big\}$. Then the $\sigma(k)$ are stopping times and $\sigma(k) \downarrow \sigma$ as $k \to \infty$. By Proposition 8.1 we have furthermore

$$\mathbb{E}_\mu\big[f(X_{\sigma(k)+s}) \,\big|\, \mathscr{F}_{\sigma(k)}\big] \;=\; \mathbb{E}_{X_{\sigma(k)}} f(X_s). \tag{8.9}$$

If $F \in \mathscr{F}_\sigma$, then $F \in \mathscr{F}_{\sigma(k)}$, and hence (8.9) implies

$$\mathbb{E}_\mu[f(X_{\sigma(k)+s});\, F] \;=\; \mathbb{E}_\mu[\mathbb{E}_{X_{\sigma(k)}} f(X_s);\, F].$$

A check of the assumptions show that the integrands converge pointwise. Thus by dominated convergence,

$$\mathbb{E}_\mu[f(X_{\sigma+s});\, F] \;=\; \mathbb{E}_\mu[\mathbb{E}_{X_\sigma} f(X_s);\, F].$$

The truth of this for all bounded continuous f and all $F \in \mathscr{F}_\sigma$ implies (8.8). □

We next consider the *hitting time* $\tau(A)$ of a Borel subset A, $\tau(A) = \inf\{t > 0 :\, X_t \in A\}$. That $\tau(A)$ is a stopping time is a triviality in discrete time since then obviously

$$\{\tau(A) \leq n\} \;=\; \bigcup_{k=1}^{n} \{X_k \in A\}\,.$$

However, in continuous time some (perhaps unexpected) difficulties arise even for elementary sets like closed and open ones, and this is in fact one of the reasons that one needs to amend and extend the theory that has been discussed so far and which may still appear reasonably simple and intuitive. We discuss these points briefly below, but first state and prove a more elementary result that is sufficient to deal with virtually all the processes to be met and all the questions to be asked in this book.

Proposition 8.4 *Suppose the paths of $\{X_t\}$ are piecewise continuous with right limits. Then:*
(a) *the jump times $0 < \iota(1) < \iota(2) < \cdots$ are stopping times w.r.t. $\{\mathscr{F}_t\}$;*
(b) *if A is closed, then $\tau(A)$ is a stopping time w.r.t. $\{\mathscr{F}_t\}$.*

Proof. Let $\mathbb{Q}(t)$ be the set of numbers of the form qt with q rational and $0 \le q \le 1$, and let d be some metric on E. Then the sets $G_1 = \{\iota(1) \le t\}$ and

$$G_2 \;=\; \bigcup_{m=1}^{\infty} \bigcap_{n=1}^{\infty} \bigcup_{u,s \in \mathbb{Q}(t)} \big\{|u-s| \le 1/n,\, d(X_u, X_s) > 1/m\big\}$$

coincide. In fact, on G_1 we have for some m a jump of size at least m^{-1}, and this easily gives $G_1 \subseteq G_2$. Conversely, the uniform continuity of $\{X_s\}_{s \le t}$ on G_1^c easily shows $G_1^c \subseteq G_2^c$. Since $d(X_u, X_s)$ is $\mathscr{F}_t$–measurable for $u, s \le t$, we have $G_1 = G_2 \in \mathscr{F}_t$, and thus $\iota(1)$ is a stopping time. For $\iota(2)$, just add the requirement $u, s \ge \iota(1)$ in the definition of G_2, and so on.

To prove (b), define $m(S) = \inf\{d(X_u, A);\ u \in S\}$, $S \subseteq [0,\infty)$. If A is closed, we have $X_{\tau(A)} \in A$ by right–continuity, and hence in the special case of continuous paths

$$\{\tau(A) \le t\} \;=\; \big\{m\big([0,t]\big) = 0\big\} \;=\; \big\{m\big(\mathbb{Q}(t)\big) = 0\big\} \in \mathscr{F}_t. \tag{8.10}$$

But if $I_{k,n} \;=\; \big\{u :\, \iota(k) - 1/n \le u < \iota(k) \le t\big\}$, then

$$\{u \in I_{k,n}\} \;=\; \big\{u < \iota(k) \le t \wedge (u + 1/n)\big\} \;\in\; \mathscr{F}_t.$$

Thus as in (8.10)

$$\begin{aligned} \{\tau(A) \le t\} \;&=\; \lim_{n\to\infty} \Big\{\tau(A) \in [0,t] \Big\backslash \bigcup_{k=1}^{\infty} I_{k,n}\Big\} \\ &=\; \lim_{n\to\infty} \Big\{m\Big(\mathbb{Q}(t) \Big\backslash \bigcup_{k=1}^{\infty} I_{k,n}\Big) = 0\Big\} \;\in\; \mathscr{F}_t. \end{aligned}$$

□

We conclude with a brief discussion of some more difficult topics which, however, are not essential for the rest of the book. Define $\mathscr{F}_{t+} = \cap_{s>t}\mathscr{F}_s$ and let $\mathscr{G}^{(\mu)}$ denote the $\mathbb{P}_\mu$–completion of $\mathscr{G}$ (some arbitrary σ–field), i.e. the smallest σ–field containing $\mathscr{G}$ and all $\mathbb{P}_\mu$–null sets. Then:

Proposition 8.5 *Suppose that $\{X_t\}$ has right–continuous paths. Then:*
(a) *If A is open, then $\tau(A)$ is a stopping time w.r.t. $\{\mathscr{F}_{t+}\}$;*
(b) *For any Borel set A, $\tau(A)$ is a stopping time w.r.t. $\{\mathscr{F}_{t+}^{(\mu)}\}$.*

Proof of (a). If A is open and $X_u \in A$, then $X_{u+v} \in A$ for all small positive v. Hence the event $\{\tau(A) \le t\}$ may be written as

$$\bigcap_{n=1}^{\infty} \bigcup_{s \le t+1/n} \{X_s \in A\} \;=\; \bigcap_{n=1}^{\infty} \bigcup_{s \in \mathbb{Q}(t+1/n)} \{X_s \in A\},$$

and here the event on the r.h.s. is clearly in $\mathscr{F}_{t+}$.

The proof of (b) is far beyond the present scope (and need!), and we refer, e.g., to Dellacherie and Meyer (1975–93). □

One now defines a *history* of the process as an increasing family $\{\mathscr{G}_t\}_{t\geq 0}$ of σ–fields (a filtration) with $\mathscr{F}_t \subseteq \mathscr{G}_t$ (or equivalenty X_t $\mathscr{G}_t$–measurable) for all t, and say that $\{X_t\}$ is Markov with transition semigroup $\{P^t\}$ w.r.t. $\{\mathscr{G}_t\}$ and some fixed governing probability measure if

$$\mathbb{P}(X_{t+s} \in A \,|\, \mathscr{G}_s) \;=\; P^t(X_s, A). \tag{8.11}$$

Apart from $\{\mathscr{F}_t\}$, some main candidates for the history are $\{\mathscr{F}_{t+}\}$ and $\{\mathscr{F}^{(\mu)}_{t+}\}$. It follows immediately from the chain rule for conditional expectations that if $\{X_t\}$ is Markov w.r.t. some history, then $\{X_t\}$ is Markov w.r.t. $\{\mathscr{F}_t\}$ as well. Conversely:

Proposition 8.6 *Let $\{X_t\}$ be Markov w.r.t. $\{\mathscr{F}_t\}$ and satisfy the regularity conditions of Corollary 8.3. Then:*
(a) *for each μ and each bounded measurable h, we have $\mathbb{P}_\mu$–a.s. that*

$$\begin{aligned}\mathbb{E}_\mu\big[h(X_{s+t}; t \geq 0) \,\big|\, \mathscr{F}_s\big] \;&=\; \mathbb{E}_\mu\big[h(X_{s+t}; t \geq 0) \,\big|\, \mathscr{F}_{s+}\big]\\ &=\; \mathbb{E}_\mu\big[h(X_{s+t}; t \geq 0) \,\big|\, \mathscr{F}^{(\mu)}_{s+}\big];\end{aligned}$$

(b) (BLUMENTHAL'S 0–1 LAW) *if $A \in \mathscr{F}_{0+}$, then for a fixed $x \in E$ either $\mathbb{P}_x(A) = 0$ or $\mathbb{P}_x(A) = 1$.*
(c) *$\{X_t\}$ is Markov w.r.t. $\{\mathscr{F}_{t+}\}$ and $\{\mathscr{F}^{(\mu)}_{t+}\}$ as well.*

Proof. (a) The second identity is just a general property of the completion operator. For the first, arguments similar to those used many times above show that it suffices to take h of the form $h(X_t)$ with $t > 0$ and h continuous and bounded. Since then $h(X_{s+t+1/n}) \stackrel{\text{a.s.}}{\to} h(X_{s+t})$, it follows from a continuity result for conditional expectations (Chung, 1974, p. 340) that indeed

$$\begin{aligned}\mathbb{E}_\mu\big[h(X_{s+t}) \,\big|\, \mathscr{F}_{s+}\big] \;&=\; \lim_{n\to\infty} \mathbb{E}_\mu\big[h(X_{s+t+1/n}) \,\big|\, \mathscr{F}_{s+1/n}\big]\\ =\; \lim_{n\to\infty} \mathbb{E}_{X_{s+1/n}} h(X_t) \;&=\; \mathbb{E}_{X_s} h(X_t) \;=\; \mathbb{E}_\mu\big[h(X_{s+t}) \,\big|\, \mathscr{F}_s\big],\end{aligned}$$

and the proof of (a) is complete. For (b), let $t = 0$ and $h = I(A)$ in (a) to obtain $\mathbb{P}_x(A|\mathscr{F}_{0+}) = \mathbb{P}_x(A|\mathscr{F}_0)$ a.s. Here the l.h.s. is just $I(A)$ and since $\mathscr{F}_0$ is $\mathbb{P}_x$–trivial, the r.h.s. is constant. Hence $I(A)$ is constant a.s. which is only possible if the probability is either 0 or 1. Finally (c) is an immediate consequence of (a). □

We stop the discussion of the foundations of the general theory of Markov processes at this point. As for the topics discussed in Sections 2–4, classification of states and limit theory will be discussed in Chapter II for a discrete state space and continuous time process. The case of a general E is much more complicated even in discrete time. For example, it is not clear what recurrence should mean since even in simple–minded continuous state space models, $\mathbb{P}_x(\tau(x) < \infty)$ will most often be 0. Some results (more or less the best known ones) are given in VII.3 and can, somewhat surprisingly, be derived as simple consequences of the ergodic theorm for discrete Markov chains. In continuous time, the existing theory is hardly equally satisfying, but a number of special cases will be encountered. For example, the main problem within the whole area of renewal theory (Chapter V) will be seen to be equivalent to the ergodicity question for the continuous–time and –state version of the recurrence time chains in Section 2.

Notes General Markov chains in discrete time are discussed, e.g., in Neveu (1965), Meyn and Tweedie (1993) and Revuz (1984). For up–to–date and readable accounts of the continuous–time case, see Rogers and Williams (1994) or Revuz and Yor (1999).

A topic not treated above but used at a few places in the book is the *generator* $\mathscr{A}$ of a continuous-time Markov process, a certain operator on a subspace $\mathscr{D}_{\mathscr{A}}$ of functions on E. There are many variants of the definition around, but the intutition behind them all is that one should have

$$\mathbb{E}_x f(X_h) \;=\; f(x) + \mathscr{A}f(x)h + \mathrm{o}(h), \quad f \in \mathscr{D}_{\mathscr{A}}. \tag{8.12}$$

The domain $\mathscr{D}_{\mathscr{A}}$ is specified by additional requirements in (8.12), one classical variant (see e.g. Karlin and Taylor, 1981) being that f should be bounded and the convergence in (8.12) uniform. Note that the identification of $\mathscr{D}_{\mathscr{A}}$ in this set–up is tedious even in such a basic case as standard Brownian motion where $\mathscr{A}$ is a restriction of the differential operator $f \to f''/2$. Note also that $\mathscr{D}_{\mathscr{A}}$ actually may contain crucial information on the process. For example, for reflecting Brownian motion with reflection at 0 or absorbtion at 0, $\mathscr{A}f = f''/2$ in both cases, but $f \in \mathscr{D}_{\mathscr{A}}$ requires $f'(0) = 0$ in the reflected case and $f(0) = 0$ in the absorbing case.

Typically, $f(X_t) - \int_0^t \mathscr{A}f(X_s)\,\mathrm{d}s$ is a martingale (the *Dynkin martingale*) for $f \in \mathscr{D}_{\mathscr{A}}$, and a modern variant of the definition is that $f \in \mathscr{D}_{\mathscr{A}}$, $g = \mathscr{A}f$ means that $f(X_t) - \int_0^t g(X_s)\,\mathrm{d}s$ is a local martingale.

The most basic case is a Markov jump process as in Chapter II, where in the finite case it holds for any of the possible definitions that $\mathscr{D}_{\mathscr{A}}$ is the set of all functions on E and $\mathscr{A}$ is the operator $\boldsymbol{f} \to \boldsymbol{\Lambda f}$ where $\boldsymbol{\Lambda}$ is the intensity matrix.

http://www.springer.com/978-0-387-00211-8

Applied Probability and Queues
Asmussen, S.
2003, XII, 438 p., Hardcover
ISBN: 978-0-387-00211-8