CHAPTER **2**

Doeblin's Theory for Markov Chains

In this chapter we begin in earnest our study of Markov processes. Like the random walks in Chapter 1, the processes with which we will be dealing here take only countably many values and have a discrete (as opposed to continuous) time parameter. In fact, in many ways, these processes are the simplest generalizations of random walks. To be precise, random walks proceed in such a way that the distribution of their increments are independent of everything which has happened before the increment takes place. The processes at which we will be looking now proceed in such a way that the distribution of their increments depends on where they are at the time of the increment but not on where they were in the past. A process with this sort of dependence property is said to have the *Markov property* and is called a *Markov chain*.[1]

The set $\mathbb{S}$ in which a process takes its values is called its *state space*, and, as we said, our processes will have state spaces which are either finite or countably infinite. Thus, at least for theoretical purposes, there is no reason for us not to think of $\mathbb{S}$ as the set $\{1, \dots, N\}$ or $\mathbb{Z}^+$, depending on whether $\mathbb{S}$ if finite or countably infinite. On the other hand, always taking $\mathbb{S}$ to be one of these has the disadvantage that it may mask important properties. For example, it would have been a great mistake to describe the nearest neighbor random walk on $\mathbb{Z}^2$ after mapping $\mathbb{Z}^2$ isomorphically onto $\mathbb{Z}^+$.

2.1 Some Generalities

Before getting started, there are a few general facts which we will need to know about Markov chains.

A *Markov chain* on a finite or countably infinite state space $\mathbb{S}$ is a family of $\mathbb{S}$-valued random variables $\{X_n : n \geq 0\}$ with the property that, for all $n \geq 0$ and $(i_0, \dots, i_n, j) \in \mathbb{S}^{n+2}$,

$$\mathbb{P}\big(X_{n+1} = j \,\big|\, X_0 = i_0, \dots, X_n = i_n\big) = (\mathbf{P})_{i_n j}, \tag{2.1.1}$$

where $\mathbf{P}$ is a matrix all of whose entries are non-negative and each of whose rows sums to 1. Equivalently (cf. §6.4.1)

$$\mathbb{P}\big(X_{n+1} = j \,\big|\, X_0, \dots, X_n\big) = (\mathbf{P})_{X_n j}. \tag{2.1.2}$$

[1] The term "chain" is commonly applied to processes with a time discrete parameter.

It should be clear that (2.1.2) is a mathematically precise expression of the idea that, when a Markov chain jumps, the distribution of where it lands depends only on where it was at the time when it jumped and not on where it was in the past.

2.1.1. Existence of Markov Chains: For obvious reasons, a matrix whose entries are non-negative and each of whose rows sum to 1 is called a *transition probability matrix*: it gives the probability that the Markov chain will move to the state j at time $n+1$ given that it is at state i at time n, independent of where it was prior to time n. Further, it is clear that only a transition probability matrix could appear on the right of (2.1.1). What may not be so immediate is that one can go in the opposite direction. Namely, let $\boldsymbol{\mu}$ be a *probability vector*[2] and $\mathbf{P}$ a transition probability matrix. Then there exists a Markov chain $\{X_n : n \geq 0\}$ with *initial distribution* $\boldsymbol{\mu}$ and transition probability matrix $\mathbf{P}$. That is, $\mathbb{P}(X_0 = i) = (\boldsymbol{\mu})_i$ and (2.1.1) holds.

To prove the preceding existence statement, one can proceed as follows. Begin by assuming, without loss in generality, that $\mathbb{S}$ is either $\{1, \dots, N\}$ or $\mathbb{Z}^+$. Next, given $i \in \mathbb{S}$, set $\beta(i,0) = 0$ and $\beta(i,j) = \sum_{k=1}^{j} (\mathbf{P})_{ik}$ for $j \geq 1$, and define $F : \mathbb{S} \times [0,1) \longrightarrow \mathbb{S}$ so that $F(i,u) = j$ if $\beta(i,j-1) \leq u < \beta(i,j)$. In addition, set $\alpha(0) = 0$ and $\alpha(i) = \sum_{k=1}^{i} (\boldsymbol{\mu})_k$ for $i \geq 1$, and define $f : [0,1) \longrightarrow \mathbb{S}$ so that $f(u) = i$ if $\alpha(i-1) \leq u < \alpha(i)$. Finally, let $\{U_n : n \geq 0\}$ be a sequence of mutually independent random variables (cf. Theorem 6.3.2) which are uniformly distributed on $[0,1)$, and set

$$X_n = \begin{cases} f(U_0) & \text{if } n = 0 \\ F(X_{n-1}, U_n) & \text{if } n \geq 1. \end{cases} \tag{2.1.3}$$

We will now show that the sequence $\{X_n : n \geq 0\}$ in (2.1.3) is a Markov chain with the required properties. For this purpose, suppose that $(i_0, \dots, i_n) \in \mathbb{S}^{n+1}$, and observe that

$$\begin{aligned} &\mathbb{P}\big(X_0 = i_0, \dots, X_n = i_n\big) \\ &= \mathbb{P}\Big(U_0 \in \big[\alpha(i_0 - 1), \alpha(i_0)\big) \\ &\qquad\qquad \& \; U_m \in \big[\beta(i_{m-1}, i_m - 1), \beta(i_{m-1}, i_m)\big) \text{ for } 1 \leq m \leq n\Big) \\ &= \boldsymbol{\mu}_{i_0} (\mathbf{P})_{i_0 i_1} \cdots (\mathbf{P})_{i_{n-1} i_n}. \end{aligned}$$

2.1.2. Transition Probabilities & Probability Vectors: Notice that the use of matrix notation here is clever. To wit, if $\boldsymbol{\mu}$ is the row vector with ith entry $(\boldsymbol{\mu})_i = \mathbb{P}(X_0 = i)$, then $\boldsymbol{\mu}$ is called the *initial distribution* of the chain and

$$(\boldsymbol{\mu}\mathbf{P}^n)_j = \mathbb{P}(X_n = j), \quad n \geq 0 \text{ and } j \in \mathbb{S}, \tag{2.1.4}$$

[2] A probability vector is a row vector whose coordinates are non-negative and sum to 1.

where we have adopted the convention that $\mathbf{P}^0$ is the identity matrix and $\mathbf{P}^n = \mathbf{P}\mathbf{P}^{n-1}$ $n \geq 1$.[3] To check (2.1.4), let $n \geq 1$ be given, and note that, by (2.1.1) and induction,

$$\mathbb{P}\big(X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = j\big) = (\boldsymbol{\mu})_{i_0}(\mathbf{P})_{i_0 i_1} \cdots (\mathbf{P})_{i_{n-1} j}.$$

Hence (2.1.4) results after one sums with respect to $(i_0, \dots, i_{n-1})$. Obviously, (2.1.4) is the statement that the row vector $\boldsymbol{\mu}\mathbf{P}^n$ *is the distribution of the Markov chain at time* n *if* $\boldsymbol{\mu}$ *is its initial distribution* (i.e., its distribution at time 0). Alternatively, $\mathbf{P}^n$ is the n-step transition probability matrix: $(\mathbf{P}^n)_{ij}$ is the conditional probability that $X_{m+n} = j$ given that $X_m = i$.

For future reference, we will introduce here an appropriate way in which to measure the length of row vectors when they are being used to represent measures. Namely, given a row vector $\boldsymbol{\rho}$, we set

$$\|\boldsymbol{\rho}\|_{\mathrm{v}} = \sum_{i \in \mathbb{S}} |(\boldsymbol{\rho})_i|, \tag{2.1.5}$$

where the subscript "v" is used in recognition that this is the notion of length which corresponds to the *variation norm* on the space of measures. The basic reason for our making this choice of norm is that

$$\|\boldsymbol{\rho}\mathbf{P}\|_{\mathrm{v}} \leq \|\boldsymbol{\rho}\|_{\mathrm{v}}, \tag{2.1.6}$$

since, by Theorem 6.1.15,

$$\|\boldsymbol{\rho}\mathbf{P}\|_{\mathrm{v}} = \sum_{j \in \mathbb{S}} \left| \sum_{i \in \mathbb{S}} (\boldsymbol{\rho})_i (\mathbf{P})_{ij} \right| \leq \sum_{i \in \mathbb{S}} \left(\sum_{j \in \mathbb{S}} |(\rho)_i| (\mathbf{P})_{ij} \right) = \|\boldsymbol{\rho}\|_{\mathrm{v}}.$$

Notice that this is a quite different way of measuring the length from the way Euclid would have: he would have used

$$\|\boldsymbol{\rho}\|_2 = \left(\sum_{i \in \mathbb{S}} (\boldsymbol{\rho})_i^2 \right)^{\frac{1}{2}}. \tag{2.1.7}$$

On the other hand, at least when $\mathbb{S}$ is finite, these two norms are comparable. Namely,

$$\|\boldsymbol{\rho}\|_2 \leq \|\boldsymbol{\rho}\|_{\mathrm{v}} \leq \sqrt{\#\mathbb{S}}\|\boldsymbol{\rho}\|_2, \quad \text{where } \#\mathbb{S} \text{ denotes the cardinality of } \mathbb{S}.$$

The first inequality is easily seen by squaring both sides, and the second is an application of Schwarz's inequality (cf. Exercise 1.3.1). Moreover, $\|\cdot\|_{\mathrm{v}}$ is a

[3] The reader should check for itself that $\mathbf{P}^n$ is again a transition probability matrix for all $n \in \mathbb{N}$: all entries are non-negative and each row sums to 1.

good *norm* (i.e., measure of length) in the sense that $\|\boldsymbol{\rho}\|_{\mathrm{v}} = 0$ if and only if $\boldsymbol{\rho} = \mathbf{0}$ and that it satisfies the *triangle inequality*: $\|\boldsymbol{\rho} + \boldsymbol{\rho}'\|_{\mathrm{v}} \le \|\boldsymbol{\rho}\|_{\mathrm{v}} + \|\boldsymbol{\rho}'\|_{\mathrm{v}}$. Finally, Cauchy's convergence criterion holds for $\|\cdot\|_{\mathrm{v}}$. That is, if $\{\boldsymbol{\rho}_n\}_1^\infty$ is a sequence in $\mathbb{R}^{\mathbb{S}}$, then there exists $\boldsymbol{\rho} \in \mathbb{R}^{\mathbb{S}}$ for which $\|\boldsymbol{\rho}_n - \boldsymbol{\rho}\|_{\mathrm{v}} \longrightarrow 0$ if and only $\{\boldsymbol{\rho}_n\}_1^\infty$ is *Cauchy convergent*

$$\lim_{m\to\infty} \sup_{n>m} \|\boldsymbol{\rho}_n - \boldsymbol{\rho}_m\|_{\mathrm{v}} = 0.$$

As usual, the "only if" direction is an easy application of the triangle inequality:

$$\|\boldsymbol{\rho}_n - \boldsymbol{\rho}_m\|_{\mathrm{v}} \le \|\boldsymbol{\rho}_n - \boldsymbol{\rho}\|_{\mathrm{v}} + \|\boldsymbol{\rho} - \boldsymbol{\rho}_m\|_{\mathrm{v}}.$$

To go the other direction, suppose that $\{\boldsymbol{\rho}_n\}_1^\infty$ is Cauchy convergent, and observe that each coordinate of $\{\boldsymbol{\rho}_n\}_1^\infty$ must be Cauchy convergent as real numbers. Hence, by Cauchy's criterion for real numbers, there exists a $\boldsymbol{\rho}$ to which $\{\boldsymbol{\rho}_n\}_1^\infty$ converges in the sense that each coordinate of the $\boldsymbol{\rho}_n$'s tends to the corresponding coordinate of $\boldsymbol{\rho}$. Thus, by Fatou's Lemma, Theorem 6.1.10, as $m \to \infty$,

$$\|\boldsymbol{\rho} - \boldsymbol{\rho}_m\|_{\mathrm{v}} = \sum_{i\in\mathbb{S}} \bigl|(\boldsymbol{\rho})_i - (\boldsymbol{\rho}_m)_i\bigr| \le \varliminf_{n\to\infty} \sum_{i\in\mathbb{S}} \bigl|(\boldsymbol{\rho}_n)_i - (\boldsymbol{\rho}_m)_i\bigr| \longrightarrow 0.$$

2.1.3. Transition Probabilities and Functions: As we saw in §2.1.2, the representation of the transition probability as a matrix and the initial distributions as a row vector facilitates the representation of the distribution at later times. In order to understand how to get the analogous benefit when computing expectation values of functions, think of a function f on the state space $\mathbb{S}$ as the column vector $\mathbf{f}$ whose jth coordinate is the value of the function f at j. Clearly, if $\boldsymbol{\mu}$ is the row vector which represents the probability measure μ on $\{1,\dots,N\}$ and $\mathbf{f}$ is the column vector which represents a function f which is either non-negative or bounded, then $\boldsymbol{\mu}\mathbf{f} = \sum_{i\in\mathbb{S}} f(i)\mu(\{i\})$ is the expected value of f with respect to μ. Similarly, the column vector $\mathbf{P}^n\mathbf{f}$ represents that function whose value at i is the conditional expectation value of $f(X_n)$ given that $X_0 = i$. Indeed,

$$\begin{aligned}\mathbb{E}\bigl[f(X_n)\,\big|\,X_0 = i\bigr] &= \sum_{j\in\mathbb{S}} f(j)\mathbb{P}(X_n = j \,|\, X_0 = i)\\ &= \sum_{j\in\mathbb{S}} (\mathbf{P}^n)_{ij}(\mathbf{f})_j = (\mathbf{P}^n\mathbf{f})_i.\end{aligned}$$

More generally, if f is either a non-negative or bounded function on $\mathbb{S}$ and $\mathbf{f}$ is the column vector which it determines, then, for $0 \le m \le n$,

$$\begin{aligned}&\mathbb{E}\bigl[f(X_n)\,\big|\,X_0 = i_0,\dots,X_m = i_m\bigr] = (\mathbf{P}^{n-m}\mathbf{f})_{i_m},\\ &\text{or, equivalently, } \mathbb{E}\bigl[f(X_n)\,\big|\,X_0,\dots,X_m\bigr] = \bigl(\mathbf{P}^{n-m}\mathbf{f}\bigr)_{X_m}\end{aligned} \tag{2.1.8}$$

since

$$\begin{aligned}\mathbb{E}\big[f(X_n)\,\big|\,X_0=i_0,\dots,X_m=i_m\big]&\\ =\sum_{j\in\mathbb{S}} f(j)\mathbb{P}\big(X_n=j\,\big|\,X_0=i_0,\dots,X_m=i_m\big)&\\ =\sum_{j\in\mathbb{S}} f(j)(\mathbf{P}^{n-m})_{i_m j}=(\mathbf{P}^{n-m}\mathbf{f})_{i_m}.&\end{aligned}$$

In particular, if $\boldsymbol{\mu}$ is the initial distribution of $\{X_n : n\ge 0\}$, then

$$\mathbb{E}\big[f(X_n)\big]=\boldsymbol{\mu}\mathbf{P}^n\mathbf{f}, \tag{2.1.9}$$

since $\mathbb{E}[f(X_n)]=\sum_i(\boldsymbol{\mu})_i\mathbb{E}[f(X_n)|X_0=i]$.

Notice that, just as $\|\cdot\|_{\mathrm{v}}$ was the appropriate way to measure the length of row vectors when we were using them to represent measures, the appropriate way to measure the length of column vectors which represent functions is with the *uniform norm* $\|\cdot\|_{\mathrm{u}}$:

$$\|\mathbf{f}\|_{\mathrm{u}}=\sup_{j\in\mathbb{S}}|(\mathbf{f})_j|. \tag{2.1.10}$$

The reason why $\|\cdot\|_{\mathrm{u}}$ is the norm of choice here is that $|\boldsymbol{\mu}\mathbf{f}|\le\|\boldsymbol{\mu}\|_{\mathrm{v}}\|\mathbf{f}\|_{\mathrm{u}}$, since

$$|\boldsymbol{\mu}\mathbf{f}|\le\sum_{i\in\mathbb{S}}|(\boldsymbol{\mu})_i||(\mathbf{f})_i|\le\|\mathbf{f}\|_{\mathrm{u}}\sum_{i\in\mathbb{S}}|(\boldsymbol{\mu})_i|.$$

In particular, we have the complement to (2.1.6):

$$\|\mathbf{P}\mathbf{f}\|_{\mathrm{u}}\le\|\mathbf{f}\|_{\mathrm{u}}. \tag{2.1.11}$$

2.1.4. The Markov Property: By definition, if $\boldsymbol{\mu}$ is the initial distribution of $\{X_n : n\ge 0\}$, then

$$\mathbb{P}\big(X_0=i_0,\dots,X_n=i_n\big)=(\boldsymbol{\mu})_{i_0}(\mathbf{P})_{i_0i_1}\cdots(\mathbf{P})_{i_{n-1}i_n}. \tag{2.1.12}$$

Hence, if $m,n\ge 1$ and $F:\mathbb{S}^{n+1}\longrightarrow\mathbb{R}$ is either bounded or non-negative, then

$$\begin{aligned}&\mathbb{E}\big[F(X_m,\dots,X_{m+n}),\,X_0=i_0,\dots,X_m=i_m\big]\\ &=\sum_{j_1,\dots,j_n\in\mathbb{S}}F(i_m,j_1,\dots,j_n)\boldsymbol{\mu}_{i_0}(\mathbf{P})_{i_0i_1}\cdots(\mathbf{P})_{i_{m-1}i_m}(\mathbf{P})_{i_mj_1}\cdots(\mathbf{P})_{j_{1n-1}j_n}\\ &=\mathbb{E}\big[F(X_0,\dots,X_n)\,\big|\,X_0=i_m\big]\mathbb{P}\big(X_0=i_0,\dots,X_m=i_m\big).\end{aligned}$$

Equivalently, we have now proved the *Markov property* in the form

$$\begin{aligned}&\mathbb{E}\big[F(X_m,\dots,X_{m+n})\,\big|\,X_0=i_0,\dots,X_m=i_m\big]\\ &\quad=\mathbb{E}\big[F(X_0,\dots,X_n)\,\big|\,X_0=i_m\big].\end{aligned} \tag{2.1.13}$$

2.2 Doeblin's Theory

In this section we will introduce an elementary but basic technique, due to Doeblin, which will allow us to study the long time distribution of a Markov chain, particularly ones on a finite state space.

2.2.1. Doeblin's Basic Theorem: For many purposes, what one wants to know about a Markov chain is its distribution after a long time, and, at least when the state space is finite, it is reasonable to think that the distribution of the chain will stabilize. To be more precise, if one is dealing with a chain which can go in a single step from some state i to any state j with positive probability, then, because there are only a finite number of states, a pigeon hole argument shows that this state is going to visited again and again and that, after a while, the chain's initial distribution is going to get "forgotten." In other words, we are predicting for such a chain that $\boldsymbol{\mu}\mathbf{P}^n$ will, for sufficiently large n, be nearly independent of $\boldsymbol{\mu}$. In particular, this would mean that $\boldsymbol{\mu}\mathbf{P}^n = (\boldsymbol{\mu}\mathbf{P}^{n-m})\mathbf{P}^m$ is very nearly equal to $\boldsymbol{\mu}\mathbf{P}^m$ when m is large and therefore, by Cauchy's convergence criterion, that $\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\mu}\mathbf{P}^n$ exists. In addition, if this were the case, then we would have that $\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\mu}\mathbf{P}^{n+1} = \lim_{n\to\infty}(\boldsymbol{\mu}\mathbf{P}^n)\mathbf{P} = \boldsymbol{\pi}\mathbf{P}$. That is, $\boldsymbol{\pi}$ would have to be a left eigenvector for $\mathbf{P}$ with eigenvalue 1. A probability vector $\boldsymbol{\pi}$ is, for obvious reasons, called a *stationary distribution* for the transition probability matrix $\mathbf{P}$ if $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$.

Although we were thinking about finite state spaces in the preceding discussion, there are situations in which these musings apply even to infinite state spaces. Namely, if, no matter where the chain starts, it has a positive probability of visiting some fixed state, then, as the following theorem shows, it will stabilize.

2.2.1 Doeblin's Theorem. *Let $\mathbf{P}$ be a transition probability matrix with the property that, for some state $j_0 \in \mathbb{S}$ and $\epsilon > 0$, $(\mathbf{P})_{ij_0} \geq \epsilon$ for all $i \in \mathbb{S}$. Then $\mathbf{P}$ has a unique stationary probability vector $\boldsymbol{\pi}$, $(\boldsymbol{\pi})_{j_0} \geq \epsilon$, and, for all initial distributions $\boldsymbol{\mu}$,*

$$\|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \leq 2(1-\epsilon)^n, \quad n \geq 0.$$

Proof: The key to the proof lies the observations that if $\boldsymbol{\rho} \in \mathbb{R}^{\mathbb{S}}$ is a row vector with $\|\boldsymbol{\rho}\|_{\mathrm{v}} < \infty$, then

$$\begin{aligned} &\sum_{j\in\mathbb{S}} (\boldsymbol{\rho}\mathbf{P})_j = \sum_{i\in\mathbb{S}} (\boldsymbol{\rho})_i \quad \text{and} \\ &\sum_{i\in\mathbb{S}} (\boldsymbol{\rho})_i = 0 \implies \|\boldsymbol{\rho}\mathbf{P}^n\|_{\mathrm{v}} \leq (1-\epsilon)^n \|\boldsymbol{\rho}\|_{\mathrm{v}} \quad \text{for } n \geq 1. \end{aligned} \tag{2.2.2}$$

The first of these is trivial, because, by Theorem 6.1.15,

$$\sum_{j\in\mathbb{S}} (\boldsymbol{\rho}\mathbf{P})_j = \sum_{j\in\mathbb{S}} \left(\sum_{i\in\mathbb{S}} (\boldsymbol{\rho})_i (\mathbf{P})_{ij} \right) = \sum_{i\in\mathbb{S}} \left(\sum_{j\in\mathbb{S}} (\boldsymbol{\rho})_i (\mathbf{P})_{ij} \right) = \sum_{i\in\mathbb{S}} (\boldsymbol{\rho})_i.$$

As for the second, we note that, by an easy induction argument, it suffices to

check it when $n = 1$. Next, suppose that $\sum_i (\boldsymbol{\rho})_i = 0$, and observe that

$$\begin{aligned}\bigl|(\boldsymbol{\rho}\mathbf{P})_j\bigr| &= \left|\sum_{i\in\mathbb{S}} (\boldsymbol{\rho})_i (\mathbf{P})_{ij}\right| \\ &= \left|\sum_{i\in\mathbb{S}} (\boldsymbol{\rho})_i \bigl((\mathbf{P})_{ij} - \epsilon\delta_{j,j_0}\bigr)\right| \le \sum_{i\in\mathbb{S}} |(\boldsymbol{\rho})_i| \bigl((\mathbf{P})_{ij} - \epsilon\delta_{j,j_0}\bigr),\end{aligned}$$

and therefore that

$$\begin{aligned}\|\boldsymbol{\rho}\mathbf{P}\|_{\mathrm{v}} &\le \sum_{j\in\mathbb{S}} \left(\sum_{i\in\mathbb{S}} |(\boldsymbol{\rho})_i| \bigl((\mathbf{P})_{ij} - \epsilon\delta_{j,j_0}\bigr)\right) \\ &= \sum_{i\in\mathbb{S}} |(\boldsymbol{\rho})_i| \left(\sum_{j\in\mathbb{S}} \bigl((\mathbf{P})_{ij} - \epsilon\delta_{j,j_0}\bigr)\right) = (1-\epsilon)\|\boldsymbol{\rho}\|_{\mathrm{v}}.\end{aligned}$$

Now let $\boldsymbol{\mu}$ be a probability vector, and set $\boldsymbol{\mu}_n = \boldsymbol{\mu}\mathbf{P}^n$. Then, because $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n-m}\mathbf{P}^m$ and $\sum_i \bigl((\boldsymbol{\mu}_{n-m})_i - \boldsymbol{\mu}_i\bigr) = 1 - 1 = 0$,

$$\|\boldsymbol{\mu}_n - \boldsymbol{\mu}_m\|_{\mathrm{v}} \le (1-\epsilon)^m \|\boldsymbol{\mu}_{n-m} - \boldsymbol{\mu}\|_{\mathrm{v}} \le 2(1-\epsilon)^m$$

for $1 \le m < n$. Hence, $\{\boldsymbol{\mu}_n\}_1^\infty$ is Cauchy convergent; and therefore there exists a $\boldsymbol{\pi}$ for which $\|\boldsymbol{\mu}_n - \boldsymbol{\pi}\|_{\mathrm{v}} \longrightarrow 0$. Since each $\boldsymbol{\mu}_n$ is a probability vector, it is clear that $\boldsymbol{\pi}$ must also be a probability vector. In addition, $\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\mu}\mathbf{P}^{n+1} = \lim_{n\to\infty} (\boldsymbol{\mu}\mathbf{P}^n)\mathbf{P} = \boldsymbol{\pi}\mathbf{P}$, and so $\boldsymbol{\pi}$ is stationary. In particular,

$$(\boldsymbol{\pi})_{j_0} = \sum_{i\in\mathbb{S}} (\boldsymbol{\pi})_i (\mathbf{P})_{ij_0} \ge \epsilon \sum_{i\in\mathbb{S}} (\boldsymbol{\pi})_i = \epsilon.$$

Finally, if $\boldsymbol{\nu}$ is any probability vector, then

$$\|\boldsymbol{\nu}\mathbf{P}^m - \boldsymbol{\pi}\|_{\mathrm{v}} = \|(\boldsymbol{\nu} - \boldsymbol{\pi})\mathbf{P}^m\|_{\mathrm{v}} \le 2(1-\epsilon)^m,$$

which, of course, proves both the stated convergence result and the uniqueness of $\boldsymbol{\pi}$ as the only stationary probability vector for $\mathbf{P}$. □

It is instructive to understand what Doeblin's Theorem says in the language of *spectral theory*. Namely, as an operator on the space of bounded functions (a.k.a. column vectors with finite uniform norm), $\mathbf{P}$ has the function $\mathbf{1}$ as a right eigenfunction with eigenvalue 1: $\mathbf{P1} = \mathbf{1}$. Thus, at least if $\mathbb{S}$ is finite, general principles say that there should exist a row vector which is a left eigenvector of $\mathbf{P}$ with eigenvalue 1. Moreover, because 1 and the entries of $\mathbf{P}$ are real, this left eigenvector can be taken to have real components. Thus, from the spectral point of view, it is no surprise that there is a non-zero row vector $\boldsymbol{\mu} \in \mathbb{R}^{\mathbb{S}}$ with the property that $\boldsymbol{\mu}\mathbf{P} = \boldsymbol{\mu}$. On the other hand,

standard spectral theory would not predict that $\boldsymbol{\mu}$ can be chosen to have non-negative components, and this is the first place where Doeblin's Theorem gives information which is not readily available from spectral theory, even when $\mathbb{S}$ is finite. To interpret the estimate in Doeblin's Theorem, let $M_1(\mathbb{S};\mathbb{C})$ denote the space of row vectors $\boldsymbol{\nu} \in \mathbb{C}^{\mathbb{S}}$ with $\|\boldsymbol{\nu}\|_{\mathrm{v}} = 1$. Then,

$$\|\boldsymbol{\nu}\mathbf{P}\|_{\mathrm{v}} \leq 1 \quad \text{for all } \boldsymbol{\nu} \in M_1(\mathbb{S};\mathbb{C}),$$

and so

$$\sup\bigl\{|\alpha| : \alpha \in \mathbb{C} \,\&\, \exists\, \boldsymbol{\nu} \in M_1(\mathbb{S};\mathbb{C})\ \boldsymbol{\nu}\mathbf{P} = \alpha\boldsymbol{\nu}\bigr\} \leq 1.$$

Moreover, if $\boldsymbol{\nu}\mathbf{P} = \alpha\boldsymbol{\nu}$ for some $\alpha \neq 1$, then $\boldsymbol{\nu}\mathbf{1} = \boldsymbol{\nu}(\mathbf{P}\mathbf{1}) = (\boldsymbol{\nu}\mathbf{P})\mathbf{1} = \alpha\boldsymbol{\nu}\mathbf{1}$, and therefore $\boldsymbol{\nu}\mathbf{1} = 0$. Thus, the estimate in (2.2.2) says that all eigenvalues of $\mathbf{P}$ which are different from 1 have absolute value dominated by $1-\epsilon$. That is, the entire spectrum of $\mathbf{P}$ lies in the complex unit disk, 1 is a simple eigenvalue, and all the other eigenvalues lie in the disk of radius $1-\epsilon$. Finally, although general spectral theory fails to predict Doeblin's Theorem, it should be said that there is a spectral theory, the one initiated by Frobenius and developed further by Kakutani, which does cover Doeblin's results. The interested reader should consult Chapter VIII in [2].

2.2.2. A Couple of Extensions: An essentially trivial extension of Theorem 2.2.1 is provided by the observation that, for any $M \geq 1$ and $\epsilon > 0$,[4]

$$\sup_j \inf_i (\mathbf{P}^M)_{ij} \geq \epsilon \implies \|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \leq 2(1-\epsilon)^{[\frac{n}{M}]} \tag{2.2.3}$$

for all probability vectors $\boldsymbol{\mu}$ and a unique stationary probability vector $\boldsymbol{\pi}$. To see this, let $\boldsymbol{\pi}$ be the stationary probability vector for $\mathbf{P}^M$, the one guaranteed by Theorem 2.2.1, and note that, for any probability vector $\boldsymbol{\mu}$, any $m \in \mathbb{N}$, and any $0 \leq r < M$,

$$\|\boldsymbol{\mu}\mathbf{P}^{mM+r} - \boldsymbol{\pi}\|_{\mathrm{v}} = \|(\boldsymbol{\mu}\mathbf{P}^r - \boldsymbol{\pi})\mathbf{P}^{mM}\|_{\mathrm{v}} \leq 2(1-\epsilon)^m.$$

Thus (2.2.3) has been proved, and from (2.2.3) the argument needed to show that $\boldsymbol{\pi}$ is the one and only stationary measure for $\mathbf{P}$ is the same as the one given in the proof of Theorem 2.2.1.

The next extension is a little less trivial. In order to appreciate the point which it is addressing, one should keep in mind the following example. Namely, consider the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{on } \{1,2\}.$$

Obviously, this two state chain goes in a single step from one state to the other. Thus, it certainly visits all its states. On the other hand, it does not satisfy

[4] Here and elsewhere, we use $[s]$ to denote the *integer part* $[s]$ *of* s of $s \in \mathbb{R}$. That is, $[s]$ is the largest integer dominated by s.

the hypothesis in (2.2.3): $(\mathbf{P}^n)_{ij} = 0$ if either $i = j$ and n is odd or if $i \neq j$ and n is even. Thus, it should not be surprising that the conclusion in (2.2.3) fails to hold for this $\mathbf{P}$. Indeed, it is easy to check that although $(\frac{1}{2}, \frac{1}{2})$ is the one and only stationary probability vector for $\mathbf{P}$, $\|(1,0)\mathbf{P}^n - (\frac{1}{2}, \frac{1}{2})\|_{\mathrm{v}} = 1$ for all $n \geq 0$. As we will see later (cf. §3.1.3), the problems encountered here stem from the fact that $(\mathbf{P}^n)_{11} > 0$ only if n is even.

In spite of the problems raised by the preceding example, one should expect that the chain corresponding to this $\mathbf{P}$ does equilibrate in some sense. To describe what we have in mind, set

$$\mathbf{A}_n = \frac{1}{n} \sum_{m=0}^{n-1} \mathbf{P}^m. \tag{2.2.4}$$

Although the matrix $\mathbf{A}_n$ is again a transition probability matrix, it is not describing transitions but instead it is giving the average amount of time that the chain will visit states. To be precise, because

$$(\mathbf{A}_n)_{ij} = \frac{1}{n} \sum_{m=0}^{n-1} \mathbb{P}(X_m = j \,|\, X_0 = i) = \mathbb{E}\left[\frac{1}{n} \sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) \,\middle|\, X_0 = i \right],$$

$(\mathbf{A}_n)_{ij}$ is the expected value of the average time spent at state j during the time interval $[0, n-1]$ given that i was the state from which the chain started. Experience teaches us that data becomes much more forgiving when it is averaged, and the present situation is no exception. Indeed, continuing with the example given above, observe that, for any probability vector $\boldsymbol{\mu}$,

$$\left\| \boldsymbol{\mu}\mathbf{A}_n - \left(\tfrac{1}{2}, \tfrac{1}{2}\right) \right\|_{\mathrm{v}} \leq \frac{1}{n} \quad \text{for } n \geq 1.$$

What follows is a statement which shows that this sort of conclusion is quite general.

2.2.5 Theorem. *Suppose that $\mathbf{P}$ is a transition probability matrix on $\mathbb{S}$. If for some $M \in \mathbb{Z}^+$, $j_0 \in \mathbb{S}$, and $\epsilon > 0$, $(\mathbf{A}_M)_{ij_0} \geq \epsilon$ for all $i \in \mathbb{S}$, then there is precisely one stationary probability vector $\boldsymbol{\pi}$ for $\mathbf{P}$, $(\boldsymbol{\pi})_{j_0} \geq \epsilon$, and*

$$\left\| \boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi} \right\|_{\mathrm{v}} \leq \frac{M-1}{n\epsilon}$$

for any probability vector $\boldsymbol{\mu}$.

To get started, let $\boldsymbol{\pi}$ be the unique stationary probability which Theorem (2.2.3) guarantees for $\mathbf{A}_M$. Then, because any $\boldsymbol{\mu}$ which is stationary for $\mathbf{P}$ is certainly stationary for $\mathbf{A}_M$, it is clear that $\boldsymbol{\pi}$ is the only candidate for $\mathbf{P}$-stationarity. Moreover, to see that $\boldsymbol{\pi}$ is $\mathbf{P}$-stationary, observe that, because $\mathbf{P}$ commutes with $\mathbf{A}_M$, $(\boldsymbol{\pi}\mathbf{P})\mathbf{A}_M = (\boldsymbol{\pi}\mathbf{A}_M)\mathbf{P} = \boldsymbol{\pi}\mathbf{P}$. Hence, $\boldsymbol{\pi}\mathbf{P}$ is stationary for $\mathbf{A}_M$ and therefore, by uniqueness, must be equal to $\boldsymbol{\pi}$. That is, $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$.

In order to prove the asserted convergence result, we will need an elementary property of averaging procedures. Namely, for any probability vector $\boldsymbol{\mu}$,

$$\|\boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_m - \boldsymbol{\mu}\mathbf{A}_n\|_{\mathrm{v}} \le \frac{m-1}{n} \quad \text{for all } m, n \ge 1. \tag{2.2.6}$$

To check this, first note that, by the triangle inequality,

$$\begin{aligned}\|\boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_m - \boldsymbol{\mu}\mathbf{A}_n\|_{\mathrm{v}} &= \frac{1}{m}\left\|\sum_{k=0}^{m-1}\bigl(\boldsymbol{\mu}\mathbf{A}_n\mathbf{P}^k - \boldsymbol{\mu}\mathbf{A}_n\bigr)\right\|_{\mathrm{v}} \\ &\le \frac{1}{m}\sum_{k=0}^{m-1}\|\boldsymbol{\mu}\mathbf{A}_n\mathbf{P}^k - \boldsymbol{\mu}\mathbf{A}_n\|_{\mathrm{v}}.\end{aligned}$$

Second, for each $k \ge 0$,

$$\boldsymbol{\mu}\mathbf{A}_n\mathbf{P}^k - \boldsymbol{\mu}\mathbf{A}_n = \frac{1}{n}\sum_{\ell=0}^{n-1}\bigl(\boldsymbol{\mu}\mathbf{P}^{\ell+k} - \boldsymbol{\mu}\mathbf{P}^{\ell}\bigr) = \frac{1}{n}\left(\sum_{\ell=k}^{n+k-1}\boldsymbol{\mu}\mathbf{P}^{\ell} - \sum_{\ell=0}^{n-1}\boldsymbol{\mu}\mathbf{P}^{\ell}\right),$$

and so $\|\boldsymbol{\mu}\mathbf{P}^k\mathbf{A}_n - \boldsymbol{\mu}\mathbf{A}_n\|_{\mathrm{v}} \le \frac{2k}{n}$. Hence, after combining this with the first observation, we are lead to

$$\|\boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_m - \boldsymbol{\mu}\mathbf{A}_n\|_{\mathrm{v}} \le \frac{2}{mn}\sum_{k=0}^{m-1} k = \frac{m-1}{n},$$

which is what we wanted.

To complete the proof of Theorem 2.2.5 from here, assume that $(\mathbf{A}_M)_{ij_0} \ge \epsilon$ for all i, and, as above, let $\boldsymbol{\pi}$ be the unique stationary probability vector for $\mathbf{P}$. Then, $\boldsymbol{\pi}$ is also the unique stationary probability vector for $\mathbf{A}_M$, and so, by the estimate in the second line of (2.2.2) applied to $\mathbf{A}_M$, $\|\boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_M - \boldsymbol{\pi}\|_{\mathrm{v}} = \|(\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi})\mathbf{A}_M\|_{\mathrm{v}} \le (1-\epsilon)\|\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi}\|_{\mathrm{v}}$, which, in conjunction with (2.2.6), leads to

$$\begin{aligned}\|\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi}\|_{\mathrm{v}} &\le \|\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_M\|_{\mathrm{v}} + \|\boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_M - \boldsymbol{\pi}\|_{\mathrm{v}} \\ &\le \frac{M-1}{n} + (1-\epsilon)\|\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi}\|_{\mathrm{v}}.\end{aligned}$$

Finally, after elementary rearrangement, this gives the required result.

2.3 Elements of Ergodic Theory

In the preceding section we saw that, under suitable conditions, either $\boldsymbol{\mu}\mathbf{P}^n$ or $\boldsymbol{\mu}\mathbf{A}_n$ converge and that the limit is the unique stationary probability vector $\boldsymbol{\pi}$ for $\mathbf{P}$. In the present section, we will provide a more probabilistically oriented interpretation of these results. In particular, we will give a probabilistic

interpretation of $\boldsymbol{\pi}$. This will be done again, by entirely different methods, in Chapter 3.

Before going further, it will be useful to have summarized our earlier results in the form (cf. (2.2.3) and remember that $|\boldsymbol{\mu}\mathbf{f}| \leq \|\boldsymbol{\mu}\|_{\mathrm{v}}\|f\|_{\mathrm{u}}$)[5]

$$\sup_j \inf_i (\mathbf{P}^M)_{ij} \geq \epsilon \implies \|\mathbf{P}\mathbf{f} - \boldsymbol{\pi}\mathbf{f}\|_{\mathrm{u}} \leq 2(1-\epsilon)^{[\frac{n}{M}]}\|\mathbf{f}\|_{\mathrm{u}} \tag{2.3.1}$$

and (cf. Theorem 2.2.5)

$$\sup_j \inf_i (\mathbf{A}_M)_{ij} \geq \epsilon \implies \|\mathbf{A}_n\mathbf{f} - \boldsymbol{\pi}\mathbf{f}\|_{\mathrm{u}} \leq \frac{M-1}{n\epsilon}\|\mathbf{f}\|_{\mathrm{u}} \tag{2.3.2}$$

when $\mathbf{f}$ is a bounded column vector.

2.3.1. The Mean Ergodic Theorem: Let $\{\mathbf{X}_n : n \geq 0\}$ be a Markov chain with transition probability $\mathbf{P}$. Obviously,

$$\overline{T}_j^{(n)} \equiv \frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) \tag{2.3.3}$$

is that average amount of time that the chain spends at j before time n. Thus, if $\boldsymbol{\mu}$ is the initial distribution of the chain (i.e., $(\boldsymbol{\mu})_i = \mathbb{P}(X_0 = i)$), then $(\boldsymbol{\mu}\mathbf{A}_n)_j = \mathbb{E}[\bar{T}_j^{(n)}]$, and so, when it applies, Theorem 2.2.5 implies that $\mathbb{E}[\bar{T}_j^{(n)}] \longrightarrow (\boldsymbol{\pi})_j$ as $n \to \infty$. Here we will be proving that the random variables $T_j^{(n)}$ themselves, not just their expected values, tend to $(\boldsymbol{\pi})_j$ as $n \to \infty$. Such results come under the heading of *ergodic theory*. Ergodic theory is the mathematics of the principle, first enunciated by the physicist J.W. Gibbs in connection with the kinetic theory of gases, which asserts that the time-average over a particular trajectory of a random dynamical system will approximate the equilibrium state of that system. Unfortunately, in spite of results, like those given here, confirming this principle, even now, nearly 150 years after Gibbs, there are essentially no physically realistic situations in which Gibbs's principle has been mathematically confirmed.

2.3.4 Mean Ergodic Theorem. *Under the hypotheses in Theorem 2.2.5,*

$$\sup_{j\in\mathbb{S}} \mathbb{E}\Big[\big(\overline{T}_j^{(n)} - (\boldsymbol{\pi})_j\big)^2\Big] \leq \frac{2(M-1)}{n\epsilon} \quad \textit{for all } n \geq 1.$$

(See (2.3.10) below for a more refined, less quantitative version.) More generally, for any bounded function f on $\mathbb{S}$ and all $n \geq 1$:

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} f(X_m) - \boldsymbol{\pi}\mathbf{f}\right)^2\right] \leq \frac{2(M-1)\|\mathbf{f}\|_{\mathrm{u}}^2}{n\epsilon},$$

where $\mathbf{f}$ denotes the column vector determined by f.

[5] Here, and elsewhere, we abuse notation by using a constant to stand for the associated constant function.

Proof: Let $\bar{\mathbf{f}}$ be the column vector determined by the function $\bar{f} = f - \boldsymbol{\pi}\mathbf{f}$. Obviously,

$$\frac{1}{n}\sum_{m=0}^{n-1} f(X_m) - \boldsymbol{\pi}\mathbf{f} = \frac{1}{n}\sum_{m=0}^{n-1} \bar{f}(X_m),$$

and so

$$\begin{aligned}\left(\frac{1}{n}\sum_{m=0}^{n-1} f(X_m) - \boldsymbol{\pi}\mathbf{f}\right)^2 &= \frac{1}{n^2}\left(\sum_{m=0}^{n-1}\bar{f}(X_m)\right)^2 = \frac{1}{n^2}\sum_{k,\ell=0}^{n-1}\bar{f}(X_k)\bar{f}(X_\ell)\\ &= \frac{2}{n^2}\sum_{0\le k\le \ell<n}\bar{f}(X_k)\bar{f}(X_\ell) - \frac{1}{n^2}\sum_{k=0}^{n-1}\bar{f}(X_k)^2\\ &\le \frac{2}{n^2}\sum_{0\le k\le \ell<n}\bar{f}(X_k)\bar{f}(X_\ell).\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} f(X_m) - \boldsymbol{\pi}\mathbf{f}\right)^2\right] &\le \frac{2}{n^2}\sum_{k=0}^{n-1}\mathbb{E}\left[\bar{f}(X_k)\sum_{\ell=0}^{n-k-1}\bar{f}(X_{k+\ell})\right]\\ &= \frac{2}{n^2}\sum_{k=0}^{n-1}\mathbb{E}\left[\bar{f}(X_k)\sum_{\ell=0}^{n-k-1}(\mathbf{P}^\ell\bar{\mathbf{f}})_{X_k}\right]\\ &= \frac{2}{n^2}\sum_{k=0}^{n-1}(n-k)\mathbb{E}\Big[\bar{f}(X_k)(\mathbf{A}_{n-k}\bar{\mathbf{f}})_{X_k}\Big]\end{aligned}$$

But, by (2.3.2), $\|\mathbf{A}_{n-k}\bar{\mathbf{f}}\|_{\mathrm{u}} \le \frac{M-1}{(n-k)\epsilon}\|\bar{\mathbf{f}}\|_{\mathrm{u}}$, and so, since $\|\bar{\mathbf{f}}\|_{\mathrm{u}} \le \|\mathbf{f}\|_{\mathrm{u}}$,

$$(n-k)\mathbb{E}\Big[\bar{f}(X_k)(\mathbf{A}_{n-k}\bar{\mathbf{f}})_{X_k}\Big] \le \frac{(M-1)\|\mathbf{f}\|_{\mathrm{u}}^2}{\epsilon}.$$

After plugging this into the preceding, we get the second result. To get the first, simply take $f = \mathbf{1}_{\{j\}}$ and observe that, in this case, $\|\mathbf{f}\|_{\mathrm{u}} \le 1$. □

2.3.2. Return Times: As the contents of §§1.1 and 1.2 already indicate, return times ought to play an important role in the analysis of the long time behavior of Markov chains. In particular, if $\rho_j^{(0)} \equiv 0$ and, for $m \ge 1$, the *time of mth return* to j is defined so that $\rho_j^{(m)} = \infty$ if $\rho_j^{(m-1)} = \infty$ or $X_n \ne j$ for every $n > \rho^{(m-1)}$ and $\rho_j^{(m)} = \inf\{n > \rho_j^{(m-1)} : X_n = j\}$ otherwise, then we say that j is *recurrent* or *transient* depending on whether $\mathbb{P}(\rho_j^{(1)} < \infty | X_0 = j) = 1$ or not; and we can hope that when j is recurrent, then the history of the chain breaks into epochs which are punctuated by the successive returns to j. In this subsection we will provide evidence which bolsters that hope.

Notice that $\rho_j \equiv \rho_j^{(1)} \geq 1$ and, for $n \geq 1$,

$$\begin{aligned} &\mathbf{1}_{(n,\infty]}(\rho_j) = F_{n,j}(X_0, \dots, X_n) \\ &\text{where } F_{n,j}(i_0, \dots, i_n) = \begin{cases} 1 & \text{if } i_m \neq j \text{ for } 1 \leq m \leq n \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{2.3.5}$$

In particular, this shows that the event $\{\rho_j > n\}$ is a measurable function of $(X_0, \dots, X_n)$. More generally, because

$$\mathbf{1}_{(n,\infty]}(\rho_j^{(m+1))}) = \mathbf{1}_{[n,\infty]}(\rho_j^{(m)}) + \sum_{\ell=1}^{n-1} \mathbf{1}_{\{\ell\}}(\rho_j^{(m)}) F_{n-\ell,j}(X_\ell, \dots, X_n),$$

an easy inductive argument shows that, for each $m \in \mathbb{N}$ and $n \in \mathbb{N}$, $\{\rho_j^{(m)} > n\}$ is a measurable function of $(X_0, \dots, X_n)$.

2.3.6 Theorem. *For all $m \in \mathbb{Z}^+$ and $(i,j) \in \mathbb{S}^2$,*

$$\mathbb{P}\big(\rho_j^{(m)} < \infty \,\big|\, X_0 = i\big) = \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = i\big)\mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = j\big)^{m-1}.$$

In particular, if j is recurrent, then $\mathbb{P}(\rho_j^{(m)} < \infty | X_0 = j) = 1$ for all $m \in \mathbb{N}$. In fact, if j is recurrent, then, conditional on $X_0 = j$, $\{\rho_j^{(m)} - \rho_j^{(m-1)} : m \geq 1\}$ is a sequence of mutually independent random variables each of which has the same distribution as ρ_j.

Proof: To prove the first statement, we apply (2.1.13) and the Monotone Convergence Theorem, Theorem 6.1.9, to justify

$$\begin{aligned} \mathbb{P}\big(\rho_j^{(m)} < \infty \,\big|\, X_0 = i\big) &= \sum_{n=1}^{\infty} \mathbb{P}\big(\rho_j^{(m-1)} = n \;\&\; \rho_j^{(m)} < \infty \,\big|\, X_0 = i\big) \\ &= \sum_{n=1}^{\infty} \lim_{N\to\infty} \mathbb{E}\Big[1 - F_{N,j}(X_n, \dots, X_{n+N}),\, \rho_j^{(m-1)} = n \,\Big|\, X_0 = i\Big] \\ &= \sum_{n=1}^{\infty} \lim_{N\to\infty} \mathbb{E}\big[1 - F_{N,j}(X_0, \dots, X_N) \,\big|\, X_0 = j\big]\mathbb{P}\big(\rho_j^{(m-1)} = n \,\big|\, X_0 = i\big) \\ &= \sum_{n=1}^{\infty} \lim_{N\to\infty} \mathbb{P}\big(\rho_j \leq N \,\big|\, X_0 = j\big)\mathbb{P}\big(\rho_j^{(m-1)} = n \,\big|\, X_0 = i\big) \\ &= \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = j\big)\mathbb{P}\big(\rho_j^{(m-1)} < \infty \,\big|\, X_0 = i\big). \end{aligned}$$

Turning to the second statement, note that it suffices for us prove that

$$\begin{aligned} &\mathbb{P}\big(\rho_j^{(m+1)} > n + n_m \,\big|\, X_0 = j, \rho_j^{(1)} = n_1, \dots, \rho_j^{(m)} = n_m\big) \\ &\quad = \mathbb{P}\big(\rho_j > n \,\big|\, X_0 = j\big). \end{aligned}$$

But, again by (2.1.13), the expression on the left is equal to

$$\begin{aligned}&\mathbb{E}\big[F_{n,j}(X_{n_m},\dots,X_{n_m+n})\,\big|\,X_0=j,\rho_j^{(1)}=n_1,\dots,\rho_j^{(m)}=n_m\big]\\&\quad=\mathbb{E}\big[F_{n,j}(X_0,\dots,X_n)\,\big|\,X_0=j\big]=\mathbb{P}\big(\rho_j>n\,\big|\,X_0=j\big).\quad\square\end{aligned}$$

Reasoning as we did in §1.2.2, we can derive from the first part of Theorem 2.3.6:

$$\begin{aligned}\mathbb{E}\big[T_j\,\big|\,X_0=i\big]&=\delta_{i,j}+\frac{\mathbb{P}(\rho_j<\infty|X_0=i)}{\mathbb{P}(\rho_j=\infty|X_0=j)}\\\mathbb{E}\big[T_j\,\big|\,X_0=j\big]=\infty\iff\mathbb{P}\big(T_j=\infty\,\big|\,X_0=j\big)&=1\\\mathbb{E}\big[T_j\,\big|\,X_0=j\big]<\infty\iff\mathbb{P}\big(T_j<\infty\,\big|\,X_0=j\big)&=1,\end{aligned}\tag{2.3.7}$$

where $T_j=\sum_{m=0}^{\infty}\mathbf{1}_{\{j\}}(X_m)$ is the total time the chain spends in the state j. Indeed, because

$$\mathbb{P}\big(T_j>m\,\big|\,X_0=i\big)=\begin{cases}\mathbb{P}\big(\rho_j^{(m)}<\infty\,\big|\,X_0=j\big)&\text{if } i=j\\\mathbb{P}\big(\rho_j^{(m+1)}<\infty\,\big|\,X_0=i\big)&\text{if } i\neq j,\end{cases}$$

all three parts of (2.3.7) follow immediately from the first part of Theorem 2.3.6.

Of course, from (2.3.7) we know that j is recurrent if and only if $\mathbb{E}[T_j|X_0=j]=\infty$. In particular, under the conditions in Theorem 2.2.5, we know that $(\mathbf{A}_n)_{j_0j_0}\longrightarrow(\boldsymbol{\pi})_{j_0}>0$, and so

$$\mathbb{E}\big[T_{j_0}\,\big|\,X_0=j_0\big]=\sum_{m=0}^{\infty}(\mathbf{P}^m)_{j_0j_0}=\lim_{n\to\infty}n(\mathbf{A}_n)_{j_0j_0}=\infty.$$

That is, the conditions in Theorem 2.2.5 imply that j_0 is recurrent, and as we are about to demonstrate, we can say much more.

To facilitate the statement of the next result, we will say that j is *accessible* from i and will write $i\to j$ if $(\mathbf{P}^n)_{ij}>0$ for some $n\geq 0$. Equivalently, $i\to j$ if and only if $i=j$ or $\mathbb{P}(\rho_j<\infty|X_0=i)>0$.

2.3.8 Theorem. *Assume that* $\inf_i(\mathbf{A}_M)_{ij_0}\geq\epsilon$ *for some* $M\geq 1$, j_0, *and* $\epsilon>0$. *Then* j *is recurrent if and only if* $j_0\to j$. *Moreover, if* $j_0\to j$, *then* $\mathbb{E}[\rho_j^p|X_0=j]<\infty$ *for all* $p\in(0,\infty)$.

Proof: First suppose that $j_0\not\to j$. Equivalently, $\mathbb{P}(\rho_j=\infty|X_0=j_0)=1$. At the same time, because $(\mathbf{A}_M)_{jj_0}\geq\epsilon$, there exists an $1\leq m<M$ such that $(\mathbf{P}^m)_{jj_0}>0$, and so

$$\begin{aligned}\mathbb{P}\big(\rho_j^{(m)}=\infty\,\big|\,X_0=j\big)&\geq\mathbb{P}\big(\rho_j^{(m)}=\infty\;\&\;X_m=j_0\,\big|\,X_0=j\big)\\&=\lim_{N\to\infty}\mathbb{E}\big[F_{N,j}(X_m,\dots,X_{m+N}),\,X_m=j_0\big|\,X_0=j\big]\\&=\lim_{N\to\infty}\mathbb{E}\big[F_{N,j}(X_0,\dots,X_N)\,\big|\,X_0=j_0\big]\mathbb{P}\big(X_m=j_0\,\big|\,X_0=j\big)\\&=\mathbb{P}\big(\rho_j=\infty\,\big|\,X_0=j_0\big)(\mathbf{P}^m)_{jj_0}>0.\end{aligned}$$

Hence, by Theorem 2.3.6, j cannot be recurrent.

We next show that

$$(*) \qquad j_0 \to j \implies \inf_i (\mathbf{A}_{M'})_{ij} > 0 \quad \text{for some } M' \geq 1.$$

To this end, choose $m \in \mathbb{N}$ so that $(\mathbf{P}^m)_{j_0 j} > 0$. Then, for all $i \in \mathbb{S}$,

$$\begin{aligned}(\mathbf{A}_{m+M})_{ij} &= \frac{1}{m+M} \sum_{\ell=0}^{M+m-1} (\mathbf{P}^\ell)_{ij} \geq \frac{1}{m+M} \sum_{\ell=0}^{M-1} (\mathbf{P}^\ell)_{ij_0} (\mathbf{P}^m)_{j_0 j} \\ &= \frac{M}{m+M} (\mathbf{A}_M)_{ij_0} (\mathbf{P}^m)_{j_0 j} \geq \frac{M\epsilon}{m+M} (\mathbf{P}^m)_{j_0 j} > 0.\end{aligned}$$

In view of (*) and what we have already shown, it suffices to show that $\mathbb{E}[\rho_j^p \,|\, X_0 = j] < \infty$ if $\inf_i (\mathbf{A}_M)_{ij} \geq \epsilon$ for some $\epsilon > 0$ and $M \in \mathbb{Z}^+$. For this purpose, set $u(n,i) = \mathbb{P}(\rho_j > nM | X_0 = i)$ for $n \in \mathbb{Z}^+$ and $i \in \mathbb{S}$. Then, by (2.1.13),

$$\begin{aligned}u(n+1,i) &= \sum_{k \in \mathbb{S}} \mathbb{P}\big(\rho_j > (n+1)M \;\&\; X_{nM} = k \,\big|\, X_0 = i\big) \\ &= \sum_{k \in \mathbb{S}} \mathbb{E}\Big[F_{M,j}(X_{nM}, \dots, X_{(n+1)M}),\, \rho_j > nM \;\&\; X_{nM} = k \,\Big|\, X_0 = i\Big] \\ &= \sum_{k \in \mathbb{S}} \mathbb{P}\big(\rho_j > M \,\big|\, X_0 = k\big) \mathbb{P}\big(\rho_j > nM \;\&\; X_{nM} = k \,\big|\, X_0 = i\big) \\ &= \sum_{k \in \mathbb{S}} u(1,k) \mathbb{P}\big(\rho_j > nM \;\&\; X_{nM} = k \,\big|\, X_0 = i\big).\end{aligned}$$

Hence, $u(n+1,i) \leq U u(n,i)$ where $U \equiv \max_{k \in \mathbb{S}} u(1,k)$. Finally, since $u(1,k) = 1 - \mathbb{P}(\rho_j \leq M | X_0 = k)$ and

$$\mathbb{P}\big(\rho_j \leq M \,\big|\, X_0 = k\big) \geq \max_{0 \leq m < M} (\mathbf{P}^m)_{kj} \geq (\mathbf{A}_M)_{kj} \geq \epsilon,$$

$U \leq 1 - \epsilon$. In particular, this means that $u(n+1,j) \leq (1-\epsilon) u(n,j)$, and therefore that $\mathbb{P}(\rho_j > nM | X_0 = j) \leq (1-\epsilon)^n$, from which

$$\begin{aligned}\mathbb{E}\big[\rho_j^p \,\big|\, X_0 = j\big] &= \sum_{n=1}^{\infty} n^p \mathbb{P}(\rho_j = n | X_0 = j) \\ &\leq \sum_{m=1}^{\infty} (mM)^p \sum_{n=(m-1)M+1}^{mM} \mathbb{P}(\rho_j = n \,|\, X_0 = j) \\ &\leq M^p \sum_{m=1}^{\infty} m^p \mathbb{P}(\rho_j > (m-1)M | X_0 = j) \\ &\leq M^p \sum_{m=1}^{\infty} m^p (1-\epsilon)^{m-1} < \infty\end{aligned}$$

follows immediately. □

2.3.3. Identification of $\boldsymbol{\pi}$: Under the conditions in Theorem 2.2.5, we know that there is precisely one $\mathbf{P}$-stationary probability vector $\boldsymbol{\pi}$. In this section, we will give a probabilistic interpretation of $(\boldsymbol{\pi})_j$. Namely, we will show that

$$\begin{aligned}&\sup_{M\geq 1}\sup_{j\in\mathbb{S}}\inf_{i\in\mathbb{S}}(\mathbf{A}_M)_{ij}>0\\ &\qquad\Longrightarrow\ (\boldsymbol{\pi})_j=\frac{1}{\mathbb{E}[\rho_j|X_0=j]}\ (\equiv 0 \text{ if } j \text{ is transient}).\end{aligned}\tag{2.3.9}$$

The idea for the proof of (2.3.9) is that, on the one hand, (cf. (2.3.3))

$$\mathbb{E}\big[\overline{T}_j^{(n)}\,\big|\,X_0=j\big]=(\mathbf{A}_n)_{jj}\longrightarrow(\boldsymbol{\pi})_j,$$

while, on the other hand,

$$X_0=j\implies\overline{T}_j^{(\rho_j^{(m)})}=\frac{1}{\rho_j^{(m)}}\sum_{\ell=0}^{\rho_j^{(m)}-1}\mathbf{1}_{\{j\}}(X_\ell)=\frac{m}{\rho_j^{(m)}}.$$

Thus, since $\rho_j^{(m)}$ is the sum of m mutually independent copies of ρ_j, the preceding combined with the Weak Law of Large Numbers should lead

$$(\boldsymbol{\pi})_j=\lim_{m\to\infty}\mathbb{E}\big[\overline{T}_j^{(\rho_j^{(m)})}\,\big|\,X_0=j\big]=\frac{1}{\mathbb{E}[\rho_j|X_0=j]}.$$

To carry out the program suggested above, we will actually prove a stronger result. Namely, we will show that, for each $j\in\mathbb{S}$,[6]

$$\mathbb{P}\left(\lim_{n\to\infty}\overline{T}_j^{(n)}=\frac{1}{\mathbb{E}[\rho_j\,|\,X_0=j]}\,\middle|\,X_0=j\right)=1.\tag{2.3.10}$$

In particular, because $0\leq\overline{T}^{(n)}\leq 1$, Lebesgue's Dominated Convergence Theorem, Theorem 6.1.11, says that

$$(\boldsymbol{\pi})_j=\lim_{n\to\infty}(\mathbf{A}_n)_{jj}=\lim_{n\to\infty}\mathbb{E}\big[\overline{T}_j^{(n)}\,\big|\,X_0=j\big]=\frac{1}{\mathbb{E}[\rho_j\,|\,X_0=j]}$$

follows from (2.3.10). Thus, we need only prove (2.3.10). To this end, choose j_0, M, and $\epsilon>0$ so that $(\mathbf{A}_M)_{ij_0}\geq\epsilon$ for all i. If $j_0\not\to j$, then, by Theorem 2.3.8, j is transient, and so, by (2.3.7), $\mathbb{P}(T_j<\infty\,|\,X_0=j)=1$. Hence,

[6] Statements like the one which follows are called *individual ergodic theorems* because they, as distinguished from the first part of Theorem 2.3.4, are about convergence with probability 1 as opposed to convergence in mean. See Exercise 3.3.9 below for more information.

conditional on $X_0 = j$, $\overline{T}_j^{(n)} \leq \frac{1}{n} T_j \longrightarrow 0$ with probability 1. At the same time, because j is transient, $\mathbb{P}(\rho_j = \infty \,|\, X_0 = j) > 0$, and so $\mathbb{E}[\rho_j \,|\, X_0 = j] = \infty$. Hence, we have proved (2.3.10) in the case when $j_0 \not\to j$.

Next assume that $j_0 \to j$. Then, again by Theorem 2.3.8, $\mathbb{E}[\rho_j^4 | X_0 = j] < \infty$ and, conditional on $X_0 = j$, $\{\rho_j^{(m)} - \rho_j^{(m-1)} : m \geq 1\}$ is a sequence of mutually independent random variables with the same distribution as ρ_j. In particular, by the Strong Law of Large Numbers (cf. Exercise 1.3.4)

$$\mathbb{P}\left(\lim_{m \to \infty} \frac{\rho_j^{(m)}}{m} = r_j \,\middle|\, X_0 = j \right) = 1 \quad \text{where } r_j \equiv \mathbb{E}[\rho_j | X_0 = j].$$

On the other hand, for any $m \geq 1$,

$$\left| \overline{T}_j^{(n)} - r_j^{-1} \right| \leq \left| \overline{T}_j^{(n)} - \overline{T}_j^{(\rho_j^{(m)})} \right| + \left| \overline{T}_j^{(\rho_j^{(m)})} - r_j^{-1} \right|,$$

and

$$\begin{aligned} \left| \overline{T}_j^{(n)} - \overline{T}_j^{(\rho_j^{(m)})} \right| &\leq \frac{|T_j^{(n)} - T_j^{(\rho_j^{(m)})}|}{n} + \left| 1 - \frac{\rho_j^{(m)}}{n} \right| \overline{T}_j^{(\rho_j^{(m)})} \\ &\leq 2 \left| 1 - \frac{\rho_j^{(m)}}{n} \right| \leq 2 \left| 1 - \frac{m r_j}{n} \right| + \frac{2m}{n} \left| \frac{\rho_j^{(m)}}{m} - r_j \right| \end{aligned}$$

while, since $\rho_j^{(m)} \geq m$,

$$\left| \overline{T}_j^{(\rho_j^{(m)})} - r_j^{-1} \right| \leq \frac{1}{r_j} \left| \frac{\rho_j^{(m)}}{m} - r_j \right|.$$

Hence,

$$\left| \overline{T}_j^{(n)} - r_j^{-1} \right| \leq 2 \left| 1 - \frac{m r_j}{n} \right| + \left(\frac{2m}{n} + \frac{1}{r_j} \right) \left| \frac{\rho_j^{(m)}}{m} - r_j \right|.$$

Finally, by taking $m_n = \left[\frac{n}{r_j} \right]$ we get

$$\left| \overline{T}_j^{(n)} - r_j^{-1} \right| \leq \frac{2}{n} + \frac{3}{r_j} \left| \frac{\rho_j^{(m_n)}}{m_n} - r_j \right| \longrightarrow 0 \quad \text{as } n \to \infty.$$

Notice that (2.3.10) is precisely the sort of statement for which Gibbs was looking. That is, it says that, with probability 1, when one observes an individual path, the average time which it spends in each state tends, as one observes for a longer and longer time, to the probability which the equilibrium (i.e., stationary) distribution assigns to that state.

2.4 Exercises

Exercise 2.4.1. In this exercise we will give a probabilistic interpretation of the *adjoint* of a transition probability matrix with respect to a stationary distribution. That is, suppose that the transition probability matrix $\mathbf{P}$ admits a stationary distribution $\boldsymbol{\mu}$, assume $(\boldsymbol{\mu})_i > 0$ for each $i \in \mathbb{S}$, and determine the matrix $\mathbf{P}^\top$ by $(\mathbf{P}^\top)_{ij} = \frac{(\boldsymbol{\mu})_j}{(\boldsymbol{\mu})_i}(\mathbf{P})_{ji}$.

(**a**) Show that $\mathbf{P}^\top$ is a transition probability matrix for which $\boldsymbol{\mu}$ is again a stationary distribution.

(**b**) Use $\mathbb{P}$ and $\mathbb{P}^\top$ to denote probabilities computed for the chains determined, respectively, by $\mathbf{P}$ and $\mathbf{P}^\top$ with initial distribution $\boldsymbol{\mu}$, and show that these chains are the *reverse* of one another in the sense that, for each $n \geq 0$ the distribution of $(X_0, \ldots, X_n)$ under $\mathbb{P}^\top$ is the same as the distribution of $(X_n, \ldots, X_0)$ under $\mathbb{P}$. That is,

$$\mathbb{P}^\top\big(X_0 = i_0, \ldots, X_n = i_n\big) = \mathbb{P}\big(X_n = i_0, \ldots, X_0 = i_n\big)$$

for all $n \geq 0$ and $(i_0, \ldots, i_n) \in \mathbb{S}^{n+1}$.

Exercise 2.4.2. The Doeblin theory applies particularly well to chains on a finite state. For example, suppose that $\mathbf{P}$ a transition probability matrix on an N element state space $\mathbb{S}$, and show that there exists an $\epsilon > 0$ such that $(\mathbf{A}_N)_{ij_0} \geq \epsilon$ for all $i \in \mathbb{S}$ if and only if $i \to j_0$ for all $i \in \mathbb{S}$. In particular, if such a j_0 exists, conclude that, for all probability vectors $\boldsymbol{\mu}$,

$$\|\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi}\|_{\mathrm{v}} \leq \frac{2(N-1)}{n\epsilon}, \quad n \geq 1,$$

where $\boldsymbol{\pi}$ is the unique stationary probability vector for $\mathbf{P}$.

Exercise 2.4.3. Here is a version of Doeblin's Theorem which sometimes gives a slightly better estimate. Namely, assume that $(\mathbf{P})_{ij} \geq \epsilon_j$ for all (i, j), and set $\epsilon = \sum_j \epsilon_j$. If $\epsilon > 0$, show that the conclusion of Theorem 2.2.1 holds and that $(\boldsymbol{\pi})_i \geq \epsilon_i$ for each $i \in \mathbb{S}$.

Exercise 2.4.4. Assume that $\mathbf{P}$ is a transition probability matrix on the finite state space $\mathbb{S}$, and show that $j \in \mathbb{S}$ is recurrent if and only if $\mathbb{E}[\rho_j | X_0 = j] < \infty$. Of course, the "if" part is trivial and has nothing to do with the finiteness of the state space.

Exercise 2.4.5. Again assume that $\mathbf{P}$ is a transition probability matrix on the finite state space $\mathbb{S}$. In addition, assume that $\mathbf{P}$ is *doubly stochastic* in the sense that each of its columns as well as each of its rows sums to 1. Under the condition that every state is accessible from every other state, show that $\mathbb{E}[\rho_j | X_0 = j] = \#\mathbb{S}$ for each $j \in \mathbb{S}$.

EXERCISE 2.4.6. In order to test how good Doeblin's Theorem is, consider the case when $\mathbb{S} = \{1, 2\}$ and

$$\mathbf{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \quad \text{for some } (\alpha, \beta) \in (0, 1).$$

Show that $\boldsymbol{\pi} = (\alpha + \beta)^{-1}(\beta, \alpha)$ and that

$$\max\bigl\{\|\boldsymbol{\nu}\mathbf{P} - \boldsymbol{\pi}\|_{\mathrm{v}} : \boldsymbol{\nu} \text{ is a probability vector}\bigr\} = \frac{2(\alpha \vee \beta)|\alpha + \beta - 1|}{\alpha + \beta}.$$

EXERCISE 2.4.7. One of the earliest examples of Markov processes are the *branching processes* introduced, around the end of the nineteenth century, by Galton and Watson to model demographics. In this model, $\mathbb{S} = \mathbb{N}$, the state $i \in \mathbb{N}$ representing the number of members in the population, and the process evolves so that, at each stage, every individual, independently of all other members of the population, dies and is replaced by a random number of offspring. Thus, 0 is an absorbing state, and, given that there are $i \geq 1$ individuals alive at time n, the number of individuals alive at time $n + 1$ will be distributed like $-i$ plus the sum of i mutually independent, $\mathbb{N}$-valued, identically distributed random variables. To be more precise, if $\boldsymbol{\mu} = (\mu_0, \dots, \mu_k, \dots)$ is the probability vector giving the number of offspring each individual produces, define the m-fold convolution power $\boldsymbol{\mu}^{\star m}$ so that $(\boldsymbol{\mu}^{\star 0})_j = \delta_{0,j}$ and, for $m \geq 1$,

$$(\boldsymbol{\mu}^{\star m})_j = \sum_{i=0}^{j} (\boldsymbol{\mu}^{\star(m-1)})_{j-i}\mu_i.$$

Then the transition probability matrix $\mathbf{P}$ is given by $(\mathbf{P})_{ij} = (\boldsymbol{\mu}^{\star i})_j$.

The first interesting question which one should ask about this model is what it predicts will be the probability of eventual *extinction*. That is, what is $\lim_{n\to\infty} \mathbb{P}(X_n = 0)$? A naïve guess is that eventual extinction should occur or should not occur depending on whether the expected number $\gamma \equiv \sum_{k=0}^{\infty} k\mu_k$ of progeny is strictly less or strictly greater than 1, with the case when the expected number is precisely 1 being more ambiguous. In order to verify this guess and remove trivial special cases, we make the assumptions that $(\boldsymbol{\mu})_0 > 0$, $(\boldsymbol{\mu})_0 + (\boldsymbol{\mu})_1 < 1$, and $\gamma \equiv \sum_{k=0}^{\infty} k(\boldsymbol{\mu})_k < \infty$.

(**a**) Set $f(s) = \sum_{k=0}^{\infty} s^k\mu_k$ for $s \in [0, 1]$, and define $f^{\circ n}(s)$ inductively so that $f^{\circ 0}(s) = s$ and $f^{\circ n} = f \circ f^{\circ(n-1)}$ for $n \geq 1$. Show that $\gamma = f'(1)$ and that

$$f^{\circ n}(s)^i = \mathbb{E}\bigl[s^{X_n} \,\big|\, X_0 = i\bigr] = \sum_{j=0}^{\infty} s^j(\mathbf{P}^n)_{ij} \quad \text{for } s \in [0, 1] \text{ and } i \geq 0.$$

Hint: Begin by showing that $f(s)^i = \sum_{j=0}^{\infty} s^j(\boldsymbol{\mu}^{\star i})_j$.

(**b**) Observe that $s \in [0,1] \longmapsto f(s) - s$ is a continuous function which is positive at $s = 0$, zero at $s = 1$, smooth and strictly convex (i.e., $f'' > 0$) on $(0,1)$. Conclude that either $\gamma \leq 1$ and $f(s) > s$ for all $s \in [0,1)$ or $\gamma > 1$ and there is exactly one $\alpha \in (0,1)$ at which $f(\alpha) = \alpha$.

(**c**) Referring to the preceding, show that

$$\gamma \leq 1 \implies \lim_{n\to\infty} \mathbb{E}\big[s^{X_n} \,\big|\, X_0 = i\big] = 1 \quad \text{for all } s \in (0,1]$$

and that

$$\gamma > 1 \implies \lim_{n\to\infty} \mathbb{E}\big[s^{X_n} \,\big|\, X_0 = i\big] = \alpha^i \quad \text{for all } s \in (0,1)$$

.

(**d**) Based on (**c**), conclude that $\gamma \leq 1 \implies \mathbb{P}(X_n = 0|X_0 = i) \longrightarrow 1$ and that $\gamma > 1 \implies \lim_{n\to\infty} \mathbb{P}(X_n = 0|X_0 = i) = \alpha^i$ and

$$\lim_{n\to\infty} \mathbb{P}(1 \leq X_n \leq L|X_0 = i) = 0 \text{ for all } L \geq 1.$$

The last conclusion has the ominous implication that, when the expected number of progeny is larger than 1, then the population either becomes extinct or, what may be worse, grows indefinitely.

Exercise 2.4.8. Continue with the setting and notion in Exercise 2.4.7. We want to show in this exercise that there are significant differences between the cases when $\gamma < 1$ and $\gamma = 1$.

(**a**) Show that $\mathbb{E}\big[X_n \,\big|\, X_0 = i\big] = i\gamma^n$. Hence, when $\gamma < 1$, the expected size of the population goes to 0 at an exponential rate. On the other hand, when $\gamma = 1$, the expected size remains constant, this in spite of the fact that $\mathbb{P}(X_n = 0|X_0 = i) \longrightarrow 1$. Thus, when $\gamma = 1$, we have a typical situation of the sort which demonstrates why Lebesgue had to make the hypotheses he did in his dominated convergence theorem, Theorem 6.1.11. In the present case, the explanation is simple: as $n \to \infty$, with large probability $X_n = 0$ but, nonetheless, with positive probability X_n is enormous.

(**b**) Let ρ_0 be the time of first return to 0. Show that

$$\mathbb{P}(\rho_0 \leq n|X_0 = i) = \mathbb{P}(X_n = 0|X_0 = i) = \big(f^{\circ\,(n-1)}(\mu_0)\big)^i,$$

and use this to get the estimate

$$\mathbb{P}(\rho_0 > n|X_0 = i) \leq i\gamma^{n-1}(1 - \mu_0).$$

In particular, this shows that $\mathbb{E}[\rho_0|X_0 = i] < \infty$ when $\gamma < 1$.

(**c**) Now assume that $\gamma = 1$. Under the additional condition that $\beta \equiv f''(1) = \sum_{k\geq 2} k(k-1)\mu_k < \infty$, start from $\mathbb{P}(\rho_0 \leq n|X_0 = 1) = f^{\circ\,(n-1)}(\mu_0)$, and show that $\mathbb{E}[\rho_0|X_0 = i] = \infty$ for all $i \geq 1$.

Hint: Begin by showing that

$$1 - f^{\circ\, n}(\mu_0) \geq \left(\prod_{\ell=m}^{n-1}\Big(1 - \beta\big(1 - f^{\circ\,\ell}(\mu_0)\big)\Big)\right)\big(1 - f^{\circ\, m}(\mu_0)\big)$$

for $n > m$. Next, use this to show that

$$\infty > \mathbb{E}[\rho_0 | X_0 = 1] = 1 + \sum_0^\infty \big(1 - f^{\circ\, n}(\mu_0)\big)$$

would lead to a contradiction.

(**d**) Here we want to show that the conclusion in (**c**) will, in general, be false without the finiteness condition on the second derivative. To see this, let $\theta \in (0,1)$ be given, and check that $f(s) \equiv s + \frac{(1-s)^{1+\theta}}{1+\theta} = \sum_{k=0}^\infty s^k \mu_k$, where $\boldsymbol{\mu} = \big(\mu_0, \ldots, \mu_k, \ldots\big)$ is a probability vector for which $\mu_k > 0$ unless $k = 1$. Now use this choice of $\boldsymbol{\mu}$ to see that, when the second derivative condition in (**c**) fails, $\mathbb{E}[\rho_0 | X_0 = 1]$ can be finite even though $\gamma = 1$.

Hint: Set $a_n = 1 - f^{\circ\, n}(\mu_0)$, note that $a_n - a_{n+1} = \mu_0 a_n^{1+\theta}$, and use this first to see that $\frac{a_{n+1}}{a_n} \longrightarrow 1$ and then that there exist $0 < c_2 < c_2 < \infty$ such that $c_1 \leq a_{n+1}^{-\theta} - a_n^{-\theta} \leq c_2$ for all $n \geq 1$. Conclude that $\mathbb{P}(\rho_0 > n | X_0 = 1)$ tends to 0 like $n^{-\frac{1}{\theta}}$.

Exercise 2.4.9. The idea underlying this exercise was introduced by J.L. Doob and is called[7] *Doob's h-transformation*. Let $\mathbf{P}$ is a transition probability matrix on the state space $\mathbb{S}$. Next, let $\emptyset \neq \Gamma \subsetneq \mathbb{S}$ be given, set $\rho_\Gamma = \inf\{n \geq 1 : X_n \in \Gamma\}$, and assume that

$$h(i) \equiv \mathbb{P}\big(\rho_\Gamma = \infty \,\big|\, X_0 = i\big) > 0 \quad \text{for all } i \in \hat{\mathbb{S}} \equiv \mathbb{S} \setminus \Gamma.$$

(**a**) Show that $h(i) = \sum_{j \in \hat{\mathbb{S}}} (\mathbf{P})_{ij} h(j)$ for all $i \in \hat{\mathbb{S}}$, and conclude that the matrix $\hat{\mathbf{P}}$ given by $(\hat{\mathbf{P}})_{ij} = \frac{1}{h(i)} (\mathbf{P})_{ij} h(j)$ for $(i,j) \in (\hat{\mathbb{S}})^2$ is a transition probability matrix on $\hat{\mathbb{S}}$.

(**b**) For all $n \in \mathbb{N}$ and $(j_0, \ldots, j_n) \in (\hat{\mathbb{S}})^{n+1}$, show that, for each $i \in \hat{\mathbb{S}}$,

$$\begin{aligned}\hat{\mathbb{P}}\big(X_0 = j_0, \ldots, X_n = j_n \,\big|\, X_0 = i\big) \\ = \mathbb{P}\big(X_0 = j_0, \ldots, X_n = j_n \,\big|\, \rho_\Gamma = \infty \;\&\; X_0 = i\big),\end{aligned}$$

where $\hat{\mathbb{P}}$ is used here to denote probabilities computed for the Markov chain on $\hat{\mathbb{S}}$ whose transition probability matrix is $\hat{\mathbf{P}}$. That is, the Markov chain determined by $\hat{\mathbf{P}}$ is the Markov chain determined by $\mathbf{P}$ conditioned to never hit Γ.

[7] The "h" comes from the connection with harmonic functions.

Exercise 2.4.10. Here is another example of an h-transform. Namely, assume that $j_0 \in \mathbb{S}$ is transient but that $i \to j_0$ for all $i \in \mathbb{S}$.[8] Set $h(i) = \mathbb{P}(\rho_{j_0} < \infty | X_0 = i)$ for $i \neq j_0$ and $h(j_0) = 1$.

(**a**) After checking that $h(i) > 0$ for all $i \in \mathbb{S}$, define $\hat{\mathbf{P}}$ so that

$$(\hat{\mathbf{P}})_{ij} = \begin{cases} (\mathbf{P})_{j_0 j} & \text{if } i = j_0 \\ h(i)^{-1}(\mathbf{P})_{ij} h(j) & \text{if } i \neq j_0. \end{cases}$$

Show that $\hat{\mathbf{P}}$ is again a transition probability matrix.

(**b**) Using $\hat{\mathbb{P}}$ to denote probabilities computed relative to the chain determined by $\hat{\mathbf{P}}$, show that

$$\hat{\mathbb{P}}\big(\rho_{j_0} > n \,\big|\, X_0 = i\big) = \frac{1}{h(i)} \mathbb{P}\big(n < \rho_{j_0} < \infty \,\big|\, X_0 = i\big)$$

for all $n \in \mathbb{N}$ and $i \neq j_0$.

(**c**) Starting from the result in (**b**), show that j_0 is recurrent for the chain determined by $\hat{\mathbf{P}}$.

[8] By Exercise 2.4.2, this is possible only if $\mathbb{S}$ in infinite.

http://www.springer.com/978-3-540-23499-9

An Introduction to Markov Processes
Stroock, D.W.
2005, XIV, 178 p., Hardcover
ISBN: 978-3-540-23499-9