Graduate Texts in Mathematics

Daniel W. Stroock

An Introduction to Markov Processes

Springer

Graduate Texts in Mathematics 230

Graduate Texts in Mathematics

1 TAKEUTI/ZARING. Introduction to Axiomatic Set Theory. 2nd ed.
2 OXTOBY. Measure and Category. 2nd ed.
3 SCHAEFER. Topological Vector Spaces. 2nd ed.
4 HILTON/STAMMBACH. A Course in Homological Algebra. 2nd ed.
5 MAC LANE. Categories for the Working Mathematician. 2nd ed.
6 HUGHES/PIPER. Projective Planes.
7 J.-P. SERRE. A Course in Arithmetic.
8 TAKEUTI/ZARING. Axiomatic Set Theory.
9 HUMPHREYS. Introduction to Lie Algebras and Representation Theory.
10 COHEN. A Course in Simple Homotopy Theory.
11 CONWAY. Functions of One Complex Variable I. 2nd ed.
12 BEALS. Advanced Mathematical Analysis.
13 ANDERSON/FULLER. Rings and Categories of Modules. 2nd ed.
14 GOLUBITSKY/GUILLEMIN. Stable Mappings and Their Singularities.
15 BERBERIAN. Lectures in Functional Analysis and Operator Theory.
16 WINTER. The Structure of Fields.
17 ROSENBLATT. Random Processes. 2nd ed.
18 HALMOS. Measure Theory.
19 HALMOS. A Hilbert Space Problem Book. 2nd ed.
20 HUSEMOLLER. Fibre Bundles. 3rd ed.
21 HUMPHREYS. Linear Algebraic Groups.
22 BARNES/MACK. An Algebraic Introduction to Mathematical Logic.
23 GREUB. Linear Algebra. 4th ed.
24 HOLMES. Geometric Functional Analysis and Its Applications.
25 HEWITT/STROMBERG. Real and Abstract Analysis.
26 MANES. Algebraic Theories.
27 KELLEY. General Topology.
28 ZARISKI/SAMUEL. Commutative Algebra. Vol.I.
29 ZARISKI/SAMUEL. Commutative Algebra. Vol.II.
30 JACOBSON. Lectures in Abstract Algebra I. Basic Concepts.
31 JACOBSON. Lectures in Abstract Algebra II. Linear Algebra.
32 JACOBSON. Lectures in Abstract Algebra III. Theory of Fields and Galois Theory.
33 HIRSCH. Differential Topology.
34 SPITZER. Principles of Random Walk. 2nd ed.
35 ALEXANDER/WERMER. Several Complex Variables and Banach Algebras. 3rd ed.
36 KELLEY/NAMIOKA et al. Linear Topological Spaces.
37 MONK. Mathematical Logic.
38 GRAUERT/FRITZSCHE. Several Complex Variables.
39 ARVESON. An Invitation to C^*-Algebras.
40 KEMENY/SNELL/KNAPP. Denumerable Markov Chains. 2nd ed.
41 APOSTOL. Modular Functions and Dirichlet Series in Number Theory. 2nd ed.
42 J.-P. SERRE. Linear Representations of Finite Groups.
43 GILLMAN/JERISON. Rings of Continuous Functions.
44 KENDIG. Elementary Algebraic Geometry.
45 LOÈVE. Probability Theory I. 4th ed.
46 LOÈVE. Probability Theory II. 4th ed.
47 MOISE. Geometric Topology in Dimensions 2 and 3.
48 SACHS/WU. General Relativity for Mathematicians.
49 GRUENBERG/WEIR. Linear Geometry. 2nd ed.
50 EDWARDS. Fermat's Last Theorem.
51 KLINGENBERG. A Course in Differential Geometry.
52 HARTSHORNE. Algebraic Geometry.
53 MANIN. A Course in Mathematical Logic.
54 GRAVER/WATKINS. Combinatorics with Emphasis on the Theory of Graphs.
55 BROWN/PEARCY. Introduction to Operator Theory I: Elements of Functional Analysis.
56 MASSEY. Algebraic Topology: An Introduction.
57 CROWELL/FOX. Introduction to Knot Theory.
58 KOBLITZ. p-adic Numbers, p-adic Analysis, and Zeta-Functions. 2nd ed.
59 LANG. Cyclotomic Fields.
60 ARNOLD. Mathematical Methods in Classical Mechanics. 2nd ed.
61 WHITEHEAD. Elements of Homotopy Theory.
62 KARGAPOLOV/MERLZJAKOV. Fundamentals of the Theory of Groups.
63 BOLLOBAS. Graph Theory.

(continued after index)

Daniel W. Stroock

An Introduction to Markov Processes

Springer

Daniel W. Stroock
MIT
Department of Mathematics, Rm. 272
Massachusetts Ave 77
02139-4307 Cambridge, USA
dws@math.mit.edu

Mathematics Subject Classification (2000): 60-01, 60J10, 60J27

ISSN 0072-5285
ISBN 3-540-23499-3 Springer Berlin Heidelberg New York

Library of Congress Control Number: 20041113930

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by the translator
Cover design: *design & production* GmbH, Heidelberg
Printed on acid-free paper 41/3142 XT - 5 4 3 2 1 0

This book is dedicated to my longtime colleague:

Richard A. Holley

Contents

Preface

To some extent, it would be accurate to summarize the contents of this book as an intolerably protracted description of what happens when either one raises a transition probability matrix $\mathbf{P}$ (i.e., all entries $(\mathbf{P})_{ij}$ are non-negative and each row of $\mathbf{P}$ sums to 1) to higher and higher powers or one exponentiates $\mathbf{R}(\mathbf{P} - \mathbf{I})$, where $\mathbf{R}$ is a diagonal matrix with non-negative entries. Indeed, when it comes right down to it, that is all that is done in this book. However, I, and others of my ilk, would take offense at such a dismissive characterization of the theory of Markov chains and processes with values in a countable state space, and a primary goal of mine in writing this book was to convince its readers that our offense would be warranted.

The reason why I, and others of my persuasion, refuse to consider the theory here as no more than a subset of matrix theory is that to do so is to ignore the pervasive role that probability plays throughout. Namely, probability theory provides a model which both motivates and provides a context for what we are doing with these matrices. To wit, even the term "transition probability matrix" lends meaning to an otherwise rather peculiar set of hypotheses to make about a matrix. Namely, it suggests that we think of the matrix entry $(\mathbf{P})_{ij}$ as giving the probability that, in one step, a system in state i will make a transition to state j. Moreover, if we adopt this interpretation for $(\mathbf{P})_{ij}$, then we must interpret the entry $(\mathbf{P}^n)_{ij}$ of $\mathbf{P}^n$ as the probability of the same transition in n steps. Thus, as $n \to \infty$, $\mathbf{P}^n$ is encoding the long time behavior of a randomly evolving system for which $\mathbf{P}$ encodes the one-step behavior, and, as we will see, this interpretation will guide us to an understanding of $\lim_{n\to\infty}(\mathbf{P}^n)_{ij}$. In addition, and perhaps even more important, is the role that probability plays in bridging the chasm between mathematics and the rest of the world. Indeed, it is the probabilistic metaphor which allows one to formulate mathematical models of various phenomena observed in both the natural and social sciences. Without the language of probability, it is hard to imagine how one would go about connecting such phenomena to $\mathbf{P}^n$.

In spite of the propaganda at the end of the preceding paragraph, this book is written from a mathematician's perspective. Thus, for the most part, the probabilistic metaphor will be used to elucidate mathematical concepts rather than to provide mathematical explanations for non-mathematical phenomena. There are two reasons for my having chosen this perspective. First, and foremost, is my own background. Although I have occasionally tried to help people who are engaged in various sorts of applications, I have not accumulated a large store of examples which are easily translated into terms which are appropriate for a book at this level. In fact, my experience has taught me that people engaged in applications are more than competent to handle the routine problems which they encounter, and that they come to someone like me only as a last resort. As a consequence, the questions which they

ask me tend to be quite difficult and the answers to those few which I can solve usually involve material which is well beyond the scope of the present book. The second reason for my writing this book in the way that I have is that I think the material itself is of sufficient interest to stand on its own. In spite of what funding agencies would have us believe, mathematics *qua* mathematics is a worthy intellectual endeavor, and I think there is a place for a modern introduction to stochastic processes which is unabashed about making mathematics its top priority.

I came to this opinion after several semesters during which I taught the introduction to stochastic processes course offered by the M.I.T. department of mathematics. The clientele for that course has been an interesting mix of undergraduate and graduate students, less than half of whom concentrate in mathematics. Nonetheless, most of the students who stay with the course have considerable talent and appreciation for mathematics, even though they lack the formal mathematical training which is requisite for a modern course in stochastic processes, at least as such courses are now taught in mathematics departments to their own graduate students. As a result, I found no ready-made choice of text for the course. On the one hand, the most obvious choice is the classic text *A First Course in Stochastic Processes*, either the original one by S. Karlin or the updated version [4] by S. Karlin and H. Taylor. Their book gives a no nonsense introduction to stochastic processes, especially Markov processes, on a countable state space, and its consistently honest, if not always easily assimilated, presentation of proofs is complemented by a daunting number of examples and exercises. On the other hand, when I began, I feared that adopting Karlin and Taylor for my course would be a mistake of the same sort as adopting Feller's book for an undergraduate introduction to probability, and this fear prevailed the first two times I taught the course. However, after using, and finding wanting, two derivatives of Karlin's classic, I took the plunge and assigned Karlin and Taylor's book. The result was very much the one which I predicted: I was far more enthusiastic about the text than were my students.

In an attempt to make Karlin and Taylor's book more palatable for the students, I started supplementing their text with notes in which I tried to couch the proofs in terms which I hoped they would find more accessible, and my efforts were rewarded with a quite positive response from my students. In fact, as my notes became more and more extensive and began to diminish the importance of the book, I decided to convert them into what is now this book, although I realize that my decision to do so may have been stupid. For one thing, the market is already close to glutted with books which purport to cover this material. Moreover, some of these books are quite popular, although my experience with them leads me to believe that their popularity is not always correlated with the quality of the mathematics they contained. Having made that pejorative comment, I will not make public which are the books which led me to this conclusion. Instead, I will only mention the books on this topic, besides Karlin and Taylor's, which I very much liked. Namely,

J. Norris's book [5] is an excellent introduction to Markov processes which, at the same time, provides its readers with a good place to exercise their measure-theoretic skills. Of course, Norris's book is only appropriate for students who have measure-theoretic skills to exercise. On the other hand, for students who possess those skills, Norris's book is a place where they can see measure theory put to work in an attractive way. In addition, Norris has included many interesting examples and exercises which illustrate how the subject can be applied. The present book includes most of the mathematical material contained in [5], but the proofs here demand much less measure theory than his do. In fact, although I have systematically employed measure theoretic terminology (Lebesgue's Dominated Convergence Theorem, the Monotone Convergence Theorem, etc.), which is explained in Chapter 6, I have done so only to familiarize my readers with the jargon which they will encounter if they delve more deeply into the subject. In fact, because the state spaces in this book are countable, the applications which I have made of Lebesgue's theory are, with one notable exception, entirely trivial. The one exception, which is made in § 6.2, is that I have included a proof that there exist countably infinite families of mutually independent random variables. Be that as it may, the reader who is ready to accept that such families exist has no need to consult Chapter 6 except for terminology and the derivation of a few essentially obvious facts about series. For more advanced students, an excellent treatment of Markov chains on a general state space can be found in the book [6] by D. Revuz.

The organization of this book should be more or less self-evident from the table of contents. In Chapter 1, I give a bare hands treatment of the basic facts, with particular emphasis on recurrence and transience, about nearest neighbor random walks on the square, d-dimensional lattice $\mathbb{Z}^d$. Chapter 2 introduces the study of ergodic properties, and this becomes the central theme which ties together Chapters 2 through 5. In Chapter 2, the systems under consideration are Markov chains (i.e., the time parameter is discrete), and the driving force behind the development there is an idea which was introduced by Doeblin. Restricted as the applicability of Doeblin's idea may be, it has the enormous advantage over the material in Chapters 3 and 4 that it provides an estimate on the rate at which the chain is converging to its equilibrium distribution. After giving a reasonably thorough account of Doeblin's theory, in Chapter 3 I study the ergodic properties of Markov chains which do not necessarily satisfy Doeblin's condition. The main result here is the one summarized in equation (3.2.15). Even though it is completely elementary, the derivation of (3.2.15), is, without doubt, the most demanding piece of analysis in the entire book. So far as I know, every proof of (3.2.15) requires work at some stage. In supposedly "simpler" proofs, the work is hidden elsewhere (either measure theory, as in [5] and [6], or in operator theory, as in [2]). The treatment given here, which is a re-working of the one in [4] based on Feller's renewal theorem, demands nothing more of the reader than a thorough understanding of arguments involving limits superior, limits inferior, and their

role in proving that limits exist. In Chapter 4, Markov chains are replaced by continuous-time Markov processes (still on a countable state space). I do this first in the case when the rates are bounded and therefore problems of possible explosion do not arise. Afterwards, I allow for unbounded rates and develop criteria, besides boundedness, which guarantee non-explosion. The remainder of the chapter is devoted to transferring the results obtained for Markov chains in Chapter 3 to the continuous-time setting. Aside from Chapter 6, which is more like an appendix than an integral part of the book, the book ends with Chapter 5. The goal in Chapter 5 is to obtain quantitative results, reminiscent of, if not as strong as, those in Chapter 2, when Doeblin's theory either fails entirely or yields rather poor estimates. The new ingredient in Chapter 5 is the assumption that the chain or process is reversible (i.e., the transition probability is self-adjoint in the L^2-space of its stationary distribution), and the engine which makes everything go is the associated Dirichlet form. In the final section, the power of the Dirichlet form methodology is tested in an analysis of the Metropolis (a.k.a. as simulated annealing) algorithm. Finally, as I said before, Chapter 6 is an appendix in which the ideas and terminology of Lebesgue's theory of measure and integration are reviewed. The one substantive part of Chapter 6 is the construction, alluded to earlier, in § 6.2.1.

Finally, I have reached the traditional place reserved for thanking those individuals who, either directly or indirectly, contributed to this book. The principal direct contributors are the many students who suffered with various and spontaneously changing versions of this book. I am particularly grateful to Adela Popescu whose careful reading brought to light many minor and a few major errors which have been removed and, perhaps, replaced by new ones. Thanking, or even identifying, the indirect contributors is trickier. Indeed, they include all the individuals, both dead and alive, from whom I received my education, and I am not about to bore you with even a partial list of who they were or are. Nonetheless, there is one person who, over a period of more than ten years, patiently taught me to appreciate the sort of material treated here. Namely, Richard A. Holley, to whom I have dedicated this book, is a *true probabilist.* To wit, for Dick, intuitive understanding usually precedes his mathematically rigorous comprehension of a probabilistic phenomenon. This statement should lead no one to to doubt Dick's powers as a rigorous mathematician. On the contrary, his intuitive grasp of probability theory not only enhances his own formidable mathematical powers, it has saved me and others from blindly pursuing flawed lines of reasoning. As all who have worked with him know, reconsider what you are saying if ever, during some diatribe into which you have launched, Dick quietly says "I don't follow that."

In addition to his mathematical prowess, every one of Dick's many students will attest to his wonderful generosity. I was not his student, but I was his colleague, and I can assure you that his generosity is not limited to his students.

Daniel W. Stroock, August 2004

CHAPTER 1

Random Walks

A Good Place to Begin

The purpose of this chapter is to discuss some examples of Markov processes which can be understood even before the term "Markov process" is. Indeed, anyone who has been introduced to probability theory will recognize that these processes all derive from consideration of elementary "coin tossing."

1.1 Nearest Neighbor Random Walks on $\mathbb{Z}$

Let p be a fixed number from the open interval $(0,1)$, and suppose that[1] $\{B_n : n \in \mathbb{Z}^+\}$ is a sequence of $\{-1,1\}$-valued, identically distributed *Bernoulli random variables*[2] which are 1 with probability p. That is, for any $n \in \mathbb{Z}^+$ and any $E \equiv (\epsilon_1, \dots, \epsilon_n) \in \{-1,1\}^n$,

$$\mathbb{P}\big(B_1 = \epsilon_1, \dots, B_n = \epsilon_n\big) = p^{N(E)} q^{n-N(E)} \text{ where } q \equiv 1-p \text{ and}$$
$$N(E) \equiv \#\big\{m : \epsilon_m = 1\big\} = \frac{n + S_n(E)}{2} \quad \text{when } S_n(E) \equiv \sum_1^n \epsilon_m. \tag{1.1.1}$$

Next, set

$$X_0 = 0 \quad \text{and} \quad X_n = \sum_{m=1}^n B_m \quad \text{for } n \in \mathbb{Z}^+. \tag{1.1.2}$$

The existence of the family $\{B_n : n \in \mathbb{Z}^+\}$ is the content of § 6.2.1.

The above family of random variables $\{X_n : n \in \mathbb{N}\}$ is often called a *nearest neighbor random walk* on $\mathbb{Z}$. Nearest neighbor random walks are examples of Markov processes, but the description which we have just given is the one which would be given in elementary probability theory, as opposed to a course, like this one, devoted to stochastic processes. Namely, in the study of stochastic processes the description should emphasize the dynamic aspects

[1] $\mathbb{Z}$ is used to denote the set of all integers, of which $\mathbb{N}$ and $\mathbb{Z}^+$ are, respectively, the non-negative and positive members.

[2] For historical reasons, mutually independent random variables which take only two values are often said to be Bernoulli random variables.

of the family. Thus, a stochastic process oriented description might replace (1.1.2) by

$$\mathbb{P}(X_0 = 0) = 1 \text{ and}$$
$$\mathbb{P}\big(X_n - X_{n-1} = \epsilon \,\big|\, X_0, \dots, X_{n-1}\big) = \begin{cases} p & \text{if } \epsilon = 1 \\ q & \text{if } \epsilon = -1, \end{cases} \tag{1.1.3}$$

where $\mathbb{P}\big(X_n - X_{n-1} = \epsilon \,\big|\, X_0, \dots, X_{n-1}\big)$ denotes the *conditional probability* (cf. § 6.4.1) that $X_n - X_{n-1} = \epsilon$ given $\sigma(\{X_0, \dots, X_{n-1}\})$. Notice that (1.1.3) is indeed more dynamic a description than the one in (1.1.2). Specifically, it says that the process starts from 0 at time $n = 0$ and proceeds so that, at each time $n \in \mathbb{Z}^+$, it moves one step forward with probability p or one step backward with probability q, independent of where it has been before time n.

1.1.1. Distribution at Time n: In this subsection, we will present two approaches to computing $\mathbb{P}(X_n = m)$. The first computation is based on the description given in (1.1.2). Namely, from (1.1.2) it is clear that $\mathbb{P}\big(|X_n| \le n\big) = 1$. In addition, it is clear that

$$n \text{ odd} \implies \mathbb{P}(X_n \text{ is odd}) = 1 \quad \text{and} \quad n \text{ even} \implies \mathbb{P}(X_n \text{ is even}) = 1.$$

Finally, given $m \in \{-n, \dots, n\}$ with the same parity as n and a string $E = (\epsilon_1, \dots, \epsilon_n) \in \{-1, 1\}^n$ with (cf. (1.1.1)) $S_n(E) = m$, $N(E) = \frac{n+m}{2}$ and so

$$\mathbb{P}\big(B_1 = \epsilon_1, \dots, B_n = \epsilon_n\big) = p^{\frac{n+m}{2}} q^{\frac{n-m}{2}}.$$

Hence, because, when $\binom{\ell}{k} \equiv \frac{\ell!}{k!(\ell-k)!}$ is the *binomial coefficient* "ℓ choose k," there are $\binom{n}{\frac{m+n}{2}}$ such strings E, we see that

$$\mathbb{P}\big(X_n = m\big) = \binom{n}{\frac{m+n}{2}} p^{\frac{n+m}{2}} q^{\frac{n-m}{2}} \tag{1.1.4}$$
$$\text{if } m \in \mathbb{Z},\ |m| \le n, \text{ and } m \text{ has the same parity as } n$$

and is 0 otherwise.

Our second computation of the same probability will be based on the more dynamic description given in (1.1.3). To do this, we introduce the notation $(P^n)_m \equiv \mathbb{P}(X_n = m)$. Obviously, $(P^0)_m = \delta_{0,m}$, where $\delta_{k,\ell}$ is the *Kronecker symbol* which is 1 when $k = \ell$ and 0 otherwise. Further, from (1.1.3), we see that $\mathbb{P}\big(X_n = m\big)$ equals

$$\begin{aligned} &\mathbb{P}\big(X_{n-1} = m - 1 \ \&\ X_n = m\big) + \mathbb{P}\big(X_{n-1} = m + 1 \ \&\ X_n = m\big) \\ &\quad = p\mathbb{P}\big(X_{n-1} = m - 1\big) + q\mathbb{P}\big(X_{n-1} = m + 1\big). \end{aligned}$$

That is,

$$(P^0)_m = \delta_{0,m} \quad \text{and} \quad (P^n)_m = p(P^{n-1})_{m-1} + q(P^{n-1})_{m+1}. \tag{1.1.5}$$

Obviously, (1.1.5) provides a complete, albeit implicit, prescription for computing the numbers $(P^n)_m$, and one can easily check that the numbers given by (1.1.4) satisfy this prescription. Alternatively, one can use (1.1.5) plus induction on n to see that $(P^n)_m = 0$ unless $m = 2\ell - n$ for some $0 \le \ell \le n$ and that $(C^n)_\ell = (C^n)_{\ell-1} + (C^n)_{\ell-1}$ when $(C^n)_\ell \equiv p^{-\ell}q^{n-\ell}(P^n)_{2\ell-n}$. In other words, the coefficients $\{(C^n)_\ell : n \in \mathbb{N} \ \& \ 0 \le \ell \le n\}$ are given by Pascal's triangle and are therefore the binomial coefficients.

1.1.2. Passage Times via the Reflection Principle: More challenging than the computation in §1.1.1 is finding the distribution of the first passage time to a point $a \in \mathbb{Z}$. That is, given $a \in \mathbb{Z} \setminus \{0\}$, set[3]

$$\zeta_a = \inf\{n \ge 1 : X_n = a\} \ (\equiv \infty \text{ when } X_n \ne a \text{ for any } n \ge 1). \tag{1.1.6}$$

Then ζ_a is the *first passage time* to a, and our goal here is to find its distribution. Equivalently, we want an expression for $\mathbb{P}(\zeta_a = n)$, and clearly, by the considerations in §1.1.1, we need only worry about n's which satisfy $n \ge |a|$ and have the same parity as a.

Again we will present two approaches to this problem, here based on (1.1.2) and in §1.1.5 on (1.1.3). To carry out the one based on (1.1.2), assume that $a \in \mathbb{Z}^+$, suppose that $n \in \mathbb{Z}^+$ has the same parity as a, and observe first that

$$\mathbb{P}(\zeta_a = n) = \mathbb{P}\big(X_n = a \ \& \ \zeta_a > n-1\big) = p\mathbb{P}\big(\zeta_a > n-1 \ \& \ X_{n-1} = a-1\big).$$

Hence, it suffices for us to compute $\mathbb{P}\big(\zeta_a > n-1 \ \& \ X_{n-1} = a-1\big)$. For this purpose, note that for any $E \in \{-1,1\}^{n-1}$ with $S_{n-1}(E) = a-1$, the event $\{(B_1, \dots, B_{n-1}) = E\}$ has probability $p^{\frac{n+a}{2}-1}q^{\frac{n-a}{2}}$. Thus,

$$(*) \qquad \mathbb{P}(\zeta_a = n) = \mathcal{N}(n,a)p^{\frac{n+a}{2}}q^{\frac{n-a}{2}}$$

where $\mathcal{N}(n,a)$ is the number of $E \in \{-1,1\}^{n-1}$ with the properties that $S_\ell(E) \le a-1$ for $0 \le \ell \le n-1$ and $S_{n-1}(E) = a-1$. That is, everything comes down to the computation of $\mathcal{N}(n,a)$. Alternatively, since $\mathcal{N}(n,a) = \binom{n-1}{\frac{n+a}{2}-1} - \mathcal{N}'(n,a)$, where $\mathcal{N}'(n,a)$ is the number of $E \in \{-1,1\}^{n-1}$ such that $S_{n-1}(E) = a-1$ and $S_\ell(E) \ge a$ for some $\ell \le n-1$, we need only compute $\mathcal{N}'(n,a)$. For this purpose we will use a beautiful argument known as the *reflection principle*. Namely, consider the set $P(n,a)$ of paths $(S_0, \dots, S_{n-1}) \in \mathbb{Z}^n$ with the properties that $S_0 = 0$, $S_\ell - S_{m-1} \in \{-1,1\}$ for $1 \le m \le n-1$, and $S_m \ge a$ for some $1 \le m \le n-1$. Clearly, $\mathcal{N}'(n,a)$ is the numbers of paths in the set $L(n,a)$ consisting of those $(S_0, \dots, S_{n-1}) \in P(n,a)$ for which $S_{n-1} = a-1$, and, as an application of the reflection principle, we will show that the set $L(n,a)$ has the same number of elements as the set $U(n,a)$ whose elements are those paths $(S_0, \dots, S_{n-1}) \in P(n,a)$ for which $S_{n-1} = a+1$. Since $(S_0, \dots, S_{n-1}) \in U(n,a)$ if and only if $S_0 = 0$, $S_m - S_{m-1} \in \{-1,1\}$

[3] As the following indicates, we take the infemum over the empty set to be $+\infty$.

for all $1 \le m \le n-1$, and $S_{n-1} = a+1$, we already know how to count them: there are $\binom{n-1}{\frac{n+a}{2}}$ of them. Hence, all that remains is to provide the advertised application of the reflection principle. To this end, for a given $\mathbf{S} = (S_0, \dots, S_{n-1}) \in P(n,a)$, let $\ell(\mathbf{S})$ be the smallest $0 \le k \le n-1$ for which $S_k \ge a$, and define the *reflection* $\mathfrak{R}(\mathbf{S}) = (\hat{S}_0, \dots, \hat{S}_{n-1})$ of $\mathbf{S}$ so that $\hat{S}_m = S_m$ if $0 \le m \le \ell(\mathbf{S})$ and $\hat{S}_k = 2a - S_k$ if $\ell(\mathbf{S}) < m \le n-1$. Clearly, $\mathfrak{R}$ maps $L(n,a)$ into $U(n,a)$ and $U(n,a)$ into $L(n,a)$. In addition, $\mathfrak{R}$ is idempotent: its composition with itself is the identity map. Hence, as a map from $L(n,a)$ to $U(n,a)$, $\mathfrak{R}$ it must be both one-to-one and onto, and so $L(n,a)$ and $U(n,a)$ have the same numbers of elements.

We have now shown that $\mathcal{N}'(n,a) = \binom{n-1}{\frac{n+a}{2}}$ and therefore that

$$\mathcal{N}(n,a) = \binom{n-1}{\frac{n+a}{2}-1} - \binom{n-1}{\frac{n+a}{2}}.$$

Finally, after plugging this into (*), we arrive at

$$\mathbb{P}(\zeta_a = n) = \left[\binom{n-1}{\frac{n+a}{2}-1} - \binom{n-1}{\frac{n+a}{2}}\right] p^{\frac{n+a}{2}} q^{\frac{n-a}{2}},$$

which simplifies to the remarkably simple expression

$$\mathbb{P}(\zeta_a = n) = \frac{a}{n}\binom{n}{\frac{n+a}{2}} p^{\frac{n+a}{2}} q^{\frac{n-a}{2}} = \frac{a}{n}\mathbb{P}(X_n = a).$$

The computation when $a < 0$ can be carried out either by repeating the argument just given or, after reversing the roles of p and q, applying the preceding result to $-a$. However one arrives at it, the general result is that

$$a \neq 0 \implies \mathbb{P}(\zeta_a = n) = \frac{|a|}{n}\binom{n}{\frac{n+a}{2}} p^{\frac{n+a}{2}} q^{\frac{n-a}{2}} = \frac{|a|}{n}\mathbb{P}(X_n = a) \tag{1.1.7}$$

for $n \ge |a|$ with the same parity as a and is 0 otherwise.

1.1.3. Some Related Computations: Although the formula in (1.1.7) is elegant, it is not particularly transparent. In particular, it is not at all evident how one can use it to determine whether $\mathbb{P}(\zeta_a < \infty) = 1$. To carry out this computation, let $a > 0$ be given, and write of $\zeta_a = f_a(B_1, \dots, B_n, \dots)$, where f_a is the function which maps $\{-1,1\}^{\mathbb{Z}^+}$ into $\mathbb{Z}^+ \cup \{\infty\}$ so that, for each $n \in \mathbb{N}$,

$$f_a(\epsilon_1, \dots, \epsilon_n, \dots) > n \iff \sum_{\ell=1}^{m} \epsilon_\ell < a \quad \text{for } 1 \le m \le n.$$

Because the event $\{\zeta_a = m\}$ depends only on $(B_1, \dots, B_m)$ and

$$\begin{aligned} \zeta_a = m &\implies \zeta_{a+1} = m + \zeta_1 \circ \Sigma^m \\ \text{where } \zeta_1 \circ \Sigma^m &\equiv f_1(B_{m+1}, \dots, B_{m+n}, \dots), \end{aligned} \tag{1.1.8}$$

$\{\zeta_a = m \ \& \ \zeta_{a+1} < \infty\} = \{\zeta_a = m\} \cap \{\zeta_1 \circ \Sigma^m < \infty\}$, and $\{\zeta_a = m\}$ is independent of $\{\zeta_1 \circ \Sigma^m < \infty\}$. In particular, this leads to

$$\begin{aligned}\mathbb{P}(\zeta_{a+1} < \infty) &= \sum_{m=1}^{\infty} \mathbb{P}(\zeta_a = m \ \& \ \zeta_{a+1} < \infty) \\ &= \sum_{m=1}^{\infty} \mathbb{P}(\zeta_a = m)\mathbb{P}(\zeta_1 \circ \Sigma^m < \infty) \\ &= \mathbb{P}(\zeta_1 < \infty) \sum_{m=1}^{\infty} \mathbb{P}(\zeta_a = m) = \mathbb{P}(\zeta_1 < \infty)\mathbb{P}(\zeta_a < \infty),\end{aligned}$$

since $(B_{m+1}, \dots, B_{m+n}, \dots)$ and $(B_1, \dots, B_n, \dots)$ have the same distribution and therefore so do $\zeta_1 \circ \Sigma^m$ and ζ_1. The same reasoning applies equally well when $a < 0$, only now with -1 playing the role of 1. In other words, we have proved that

$$\mathbb{P}(\zeta_a < \infty) = \mathbb{P}\big(\zeta_{\mathrm{sgn}(a)} < \infty\big)^{|a|} \quad \text{for } a \in \mathbb{Z} \setminus \{0\}, \tag{1.1.9}$$

where $\mathrm{sgn}(a)$, the *signum* of a, is 1 or -1 according to whether $a > 0$ or $a < 0$. In particular, this shows that $\mathbb{P}(\zeta_1 < \infty) = 1 \implies \mathbb{P}(\zeta_a < \infty) = 1$ and $\mathbb{P}(\zeta_{-1} < \infty) = 1 \implies \mathbb{P}(\zeta_{-a} < \infty) = 1$ for all $a \in \mathbb{Z}^+$.

In view of the preceding, we need only look at $\mathbb{P}(\zeta_1 < \infty)$. Moreover, by the Monotone Convergence Theorem, Theorem 6.1.9,

$$\mathbb{P}(\zeta_1 < \infty) = \lim_{s \nearrow 1} \mathbb{E}\big[s^{\zeta_1}\big] = \lim_{s \nearrow 1} \sum_{n=1}^{\infty} s^{2n-1}\mathbb{P}(\zeta_1 = 2n-1).$$

Applying (1.1.7) with $a = 1$, we know that

$$\mathbb{P}(\zeta_1 = 2n-1) = \frac{1}{2n-1}\binom{2n-1}{n} p^n q^{n-1}.$$

Next, note that

$$\begin{aligned}\frac{1}{2n-1}\binom{2n-1}{n} &= \frac{(2(n-1))!}{n!(n-1)!} = \frac{2^{n-1}}{n!}\prod_{m=1}^{n-1}(2m-1) \\ &= \frac{4^{n-1}}{n!}\prod_{m=1}^{n-1}\left(m - \tfrac{1}{2}\right) = (-1)^{n-1}\frac{4^n}{2}\binom{\frac{1}{2}}{n},\end{aligned}$$

where[4], for any $\alpha \in \mathbb{R}$,

$$\binom{\alpha}{n} \equiv \begin{cases} 1 & \text{if } n = 0 \\ \frac{1}{n!}\prod_{m=0}^{n-1}(\alpha - m) & \text{if } n \in \mathbb{Z}^+ \end{cases}$$

[4] In the preceding, we have adopted the convention that $\prod_{j=k}^{\ell} a_j = 1$ if $\ell < k$.

is the *generalized binomial coefficient* which gives the coefficient of x^n in the Taylor's expansion of $(1+x)^\alpha$ around $x = 0$. Hence,

$$\sum_{n=1}^{\infty} s^{2n-1}\mathbb{P}(\zeta_1 = 2n-1) = -\frac{1}{2qs}\sum_{n=1}^{\infty}\binom{\frac{1}{2}}{n}(-4pqs^2)^n = \frac{1-\sqrt{1-4pqs^2}}{2qs},$$

and so

$$\mathbb{E}\big[s^{\zeta_1}\big] = \frac{1-\sqrt{1-4pqs^2}}{2qs} \quad \text{for } |s| < 1. \tag{1.1.10}$$

Of course, by symmetry, one can reverse the roles of p and q to obtain

$$\mathbb{E}\big[s^{\zeta_{-1}}\big] = \frac{1-\sqrt{1-4pqs^2}}{2ps} \quad \text{for } |s| < 1. \tag{1.1.11}$$

By letting $s \nearrow 1$ in (1.1.10) and noting that $1 - 4pq = (p+q)^2 - 4pq = (p-q)^2$, we see that[5]

$$\lim_{s\nearrow 1}\mathbb{E}\big[s^{\zeta_1}\big] = \frac{1-|p-q|}{2q} = \frac{p\wedge q}{q},$$

and so

$$\mathbb{P}(\zeta_1 < \infty) = \begin{cases} 1 & \text{if } p \geq q \\ \frac{p}{q} & \text{if } p < q. \end{cases}$$

Of course, $\mathbb{P}(\zeta_{-1} < \infty)$ is given by the same formula, only with the roles of p and q reversed. Thus,

$$\mathbb{P}(\zeta_a < \infty) = \begin{cases} 1 & \text{if } a \in \mathbb{Z}^+ \ \&\ p \geq q \text{ or } -a \in \mathbb{Z}^+ \ \&\ p \leq q \\ \left(\frac{p}{q}\right)^a & \text{if } a \in \mathbb{Z}^+ \ \&\ p < q \text{ or } -a \in \mathbb{Z}^+ \ \&\ p > q. \end{cases} \tag{1.1.12}$$

1.1.4. Time of First Return: Having gone to so much trouble to arrive at (1.1.12), it is only reasonable to draw from it a famous conclusion about the *recurrence* properties of nearest neighbor random walks on $\mathbb{Z}$. Namely, let

$$\rho_0 \equiv \inf\{n \geq 1 : X_n = 0\} \ (\equiv \infty \text{ if } X_n \neq 0 \text{ for all } n \geq 1)$$

be the *time of first return* to 0. Then, by precisely the same sort of reasoning which allowed us to arrive at (1.1.9), we see that $\mathbb{P}(X_1 = 1 \ \&\ \rho_0 < \infty) = p\mathbb{P}(\zeta_{-1} < \infty)$ and $\mathbb{P}(X_1 = -1 \ \&\ \rho_0 < \infty) = q\mathbb{P}(\zeta_1 < \infty)$, and so, by (1.1.12),

$$\mathbb{P}(\rho_0 < \infty) = 2(p \wedge q). \tag{1.1.13}$$

[5] We use $a \wedge b$ to denote the minimum $\min\{a, b\}$ of $a, b \in \mathbb{R}$.

In other words, *the random walk* $\{X_n : n \geq 0\}$ *will return to* 0 *with probability* 1 *if and only if it is symmetric* in the sense that $p = \frac{1}{2}$.

By sharpening the preceding a little, one sees that $\mathbb{P}(X_1 = 1 \ \& \ \rho_0 = 2n) = p\mathbb{P}(\zeta_{-1} = 2n - 1)$ and $\mathbb{P}(X_1 = -1 \ \& \ \rho_0 = 2n) = q\mathbb{P}(\zeta_1 = 2n - 1)$, and so, by (1.1.10) and (1.1.11),

$$\mathbb{E}\big[s^{\rho_0}\big] = 1 - \sqrt{1 - 4pqs^2} \quad \text{for } |s| < 1. \tag{1.1.14}$$

Hence,

$$\mathbb{E}\big[\rho_0 s^{\rho_0}\big] = s\frac{d}{ds}\mathbb{E}\big[s^{\rho_0}\big] = \frac{4pqs^2}{\sqrt{1 - 4pqs^2}} \quad \text{for } |s| < 1,$$

and therefore, since[6] $\mathbb{E}[\rho_0 s^{\rho_0}] \nearrow \mathbb{E}[\rho_0,\ \rho_0 < \infty]$ as $s \nearrow 1$,

$$\mathbb{E}\big[\rho_0,\ \rho_0 < \infty\big] = \frac{4pq}{|p - q|},$$

which, in conjunction with (1.1.13), means that[7]

$$\mathbb{E}\big[\rho_0 \,\big|\, \rho_0 < \infty\big] = \frac{2p \vee q}{|p - q|} = 1 + \frac{1}{|p - q|}. \tag{1.1.15}$$

The conclusions drawn in the preceding provide significant insight into the behavior of nearest neighbor random walks on $\mathbb{Z}$. In the first place, they say that when the random walk is symmetric, it returns to 0 with probability 1 but the expected amount of time it takes to do so is infinite. Secondly, when the random walk is not symmetric, it will, with positive probability, fail to return. On the other hand, in the non-symmetric case, the behavior of the trajectories is interesting. Namely, (1.1.13) in combination with (1.1.15) say that either they fail to return at all or they return relatively quickly.

1.1.5. Passage Times via Functional Equations: We close this discussion of passage times for nearest neighbor random walks with a less computational derivation of (1.1.10). For this purpose, set $u_a(s) = \mathbb{E}[s^{\zeta_a}]$ for $a \in \mathbb{Z}\backslash\{0\}$ and $s \in (-1, 1)$. Given $a \in \mathbb{Z}^+$, we use the ideas in §1.1.3, especially (1.1.8), to arrive at

$$\begin{aligned} u_{a+1}(s) &= \sum_{m=1}^{\infty} s^m \mathbb{E}\big[s^{\zeta_1 \circ \Sigma^m},\ \zeta_a = m\big] = \sum_{m=1}^{\infty} s^m \mathbb{P}(\zeta_a = m)\mathbb{E}\big[s^{\zeta_1 \circ \Sigma^m}\big] \\ &= \sum_{m=1}^{\infty} s^m \mathbb{P}(\zeta_a = m)u_1(s) = u_a(s)u_1(s). \end{aligned}$$

[6] When X is a random variable and A is an event, we will often use $\mathbb{E}[X,\ A]$ to denote $\mathbb{E}[X\mathbf{1}_A]$.

[7] $a \vee b$ is used to denote the maximum $\max\{a, b\}$ of $a,\ b \in \mathbb{R}$.

Similarly, if $-a \in \mathbb{Z}^+$, then $u_{a-1}(s) = u_a(s)u_{-1}(s)$. Hence

$$u_a(s) = u_{\text{sgn}(a)}(s)^{|a|} \quad \text{for } a \in \mathbb{Z} \setminus \{0\} \text{ and } |s| < 1. \tag{1.1.16}$$

Continuing with the same line of reasoning and using (1.1.16) with $a = 1$, we also have

$$\begin{aligned} u_1(s) &= \mathbb{E}\left[s^{\zeta_1}, X_1 = 1\right] + \mathbb{E}\left[s^{\zeta_1}, X_1 = -1\right] \\ &= ps + qs\mathbb{E}\left[s^{\zeta_2 \circ \Sigma^1}, X_1 = -1\right] = ps + qsu_2(s) = ps + qsu_1(s)^2. \end{aligned}$$

Hence, by the quadratic formula,

$$u_1(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2qs}.$$

Because $\mathbb{P}(\zeta_1 \text{ is odd}) = 1$, $u_1(-s) = -u_1(s)$. At the same time,

$$s \in (0,1) \implies \frac{1 + \sqrt{1 - 4pqs^2}}{2qs} > \frac{1 + \sqrt{1 - 4pq}}{2q} = \frac{p \vee q}{q} \geq 1.$$

Hence, since $s \in (0,1) \implies u_1(s) < 1$, we can eliminate the "+" solution and thereby arrive at a second derivation of (1.1.10). In fact, after combining this with (1.1.16), we have shown that

$$\mathbb{E}\left[s^{\zeta_a}\right] = \begin{cases} \left(\frac{1-\sqrt{1-4pqs^2}}{2qs}\right)^a & \text{if } a \in \mathbb{Z}^+ \\ \left(\frac{1-\sqrt{1-4pqs^2}}{2ps}\right)^{-a} & \text{if } -a \in \mathbb{Z}^+ \end{cases} \quad \text{for } |s| < 1. \tag{1.1.17}$$

1.2 Recurrence Properties of Random Walks

In §1.1.4, we studied the time ρ_0 of first return of a nearest neighbor random walk to 0. As we will see in Chapters 2 and 3, times of first return are critical (cf. §2.3.2) for an understanding of the long time behavior of random walks and related processes. Indeed, when the random walk returns to 0, it starts all over again. Thus, if it returns with probability 1, then the entire history of the walk will consist of epochs, each epoch being a sojourn which begins and ends at 0. Because it marks the time at which one epoch ends and a second, identically distributed, one begins, a time of first return is often called a *recurrence time*, and the walk is said to be *recurrent* if $\mathbb{P}(\rho_0 < \infty) = 1$. Walks which are not recurrent are said to be *transient*.

In this section, we will discuss the recurrence properties of nearest neighbor random walks. Of course, we already know (cf. (1.1.13)) that a nearest neighbor random on $\mathbb{Z}$ is recurrent if and only if it is symmetric. Thus, our interest here will be in higher dimensional analogs. In particular, in the hope that it will be convincing evidence that recurrence is subtle, we will show that the recurrence of the nearest neighbor, symmetric random walk on $\mathbb{Z}$ persists when $\mathbb{Z}$ is replaced by $\mathbb{Z}^2$ but disappears in $\mathbb{Z}^3$.

1.2.1. Random Walks on $\mathbb{Z}^d$: To describe the analog on $\mathbb{Z}^d$ of a nearest neighbor random walk on $\mathbb{Z}$, we begin by thinking of $\mathbf{N}_1 \equiv \{-1, 1\}$ as the set of *nearest neighbors* in $\mathbb{Z}$ of 0. It should then be clear why the set of nearest neighbors of the origin in $\mathbb{Z}^d$ consists of the $2d$ points in $\mathbb{Z}^d$ for which $(d-1)$ coordinates are 0 and the remaining coordinate is in $\mathbf{N}_1$. Next, we replace the $\mathbf{N}_1$-valued Bernoulli random variables in §1.1 by their d-dimensional analogs, namely: independent, identically distributed $\mathbf{N}_d$-valued random variables $\mathbf{B}_1, \dots, \mathbf{B}_n, \dots$.[8] Finally, a nearest neighbor random walk on $\mathbb{Z}^d$ is a family $\{\mathbf{X}_n : n \geq 0\}$ of the form

$$\mathbf{X}_0 = \mathbf{0} \quad \text{and} \quad \mathbf{X}_n = \sum_{m=1}^{n} \mathbf{B}_m \quad \text{for } n \geq 1.$$

The equivalent, stochastic process oriented description of $\{\mathbf{X}_n : n \geq 0\}$ is

$$\begin{aligned} &\mathbb{P}(\mathbf{X}_0 = \mathbf{0}) = 1 \quad \text{and, for } n \geq 1 \text{ and } \boldsymbol{\epsilon} \in \mathbf{N}_d, \\ &\mathbb{P}\big(\mathbf{X}_n - \mathbf{X}_{n-1} = \boldsymbol{\epsilon} \,\big|\, \mathbf{X}_0, \dots, \mathbf{X}_{n-1}\big) = p_{\boldsymbol{\epsilon}}, \end{aligned} \tag{1.2.1}$$

where $p_{\boldsymbol{\epsilon}} \equiv \mathbb{P}\big(\mathbf{B}_1 = \boldsymbol{\epsilon}\big)$. When $\mathbf{B}_1$ is uniformly distributed on $\mathbf{N}_d$, the random walk is said to be *symmetric*.

In keeping with the notation and terminology introduced above, we define the time $\rho_{\mathbf{0}}$ of first return to the origin equal to n if $n \geq 1$, $\mathbf{X}_n = \mathbf{0}$, and $\mathbf{X}_m \neq \mathbf{0}$ for $1 \leq m < n$, and we take $\rho_{\mathbf{0}} = \infty$ if no such $n \geq 1$ exists. Also, we will say that the walk is *recurrent* or *transient* according to whether $\mathbb{P}(\rho_{\mathbf{0}} < \infty)$ is 1 or strictly less than 1.

1.2.2. An Elementary Recurrence Criterion: Given $n \geq 1$, let $\rho_{\mathbf{0}}^{(n)}$ be the time of the nth return to $\mathbf{0}$. That is, $\rho_{\mathbf{0}}^{(1)} = \rho_{\mathbf{0}}$ and, for $n \geq 2$,

$$\rho_{\mathbf{0}}^{(n-1)} < \infty \implies \rho_{\mathbf{0}}^{(n)} = \inf\{m > \rho^{(n-1)} : \mathbf{X}_m = \mathbf{0}\}$$

and $\rho_{\mathbf{0}}^{(n-1)} = \infty \implies \rho_{\mathbf{0}}^{(n)} = \infty$. Equivalently, if $g : (\mathbf{N}_d)^{\mathbb{Z}^+} \longrightarrow \mathbb{Z}^+ \cup \{\infty\}$ is determined so that

$$g(\boldsymbol{\epsilon}_1, \dots \boldsymbol{\epsilon}_\ell, \dots) > n \quad \text{if } \sum_{\ell=1}^{m} \boldsymbol{\epsilon}_\ell \neq \mathbf{0} \quad \text{for } 1 \leq m \leq n,$$

then $\rho_{\mathbf{0}} = g(\mathbf{B}_1, \dots, \mathbf{B}_\ell, \dots)$, and $\rho_{\mathbf{0}}^{(n)} = m \implies \rho_{\mathbf{0}}^{(n+1)} = m + \rho_{\mathbf{0}} \circ \Sigma^m$ where $\rho_{\mathbf{0}} \circ \Sigma^m$ is equal to $g\big(\mathbf{B}_{m+1}, \dots, \mathbf{B}_{m+\ell}, \dots\big)$. In particular, this leads to

$$\begin{aligned} \mathbb{P}(\rho_{\mathbf{0}}^{(n+1)} < \infty) &= \sum_{m=1}^{\infty} \mathbb{P}\big(\rho_{\mathbf{0}}^{(n)} = m \;\&\; \rho_{\mathbf{0}} \circ \Sigma^m < \infty\big) \\ &= \mathbb{P}(\rho_{\mathbf{0}}^{(n)} < \infty)\mathbb{P}(\rho_{\mathbf{0}} < \infty), \end{aligned}$$

[8] The existence of the $\mathbf{B}_n$'s can be seen as a consequence of Theorem 6.3.2. Namely, let $\{U_n : n \in \mathbb{Z}^+\}$ be a family of mutually independent random variables which are uniformly distributed on $[0, 1)$. Next, let $(\mathbf{k}_1, \dots, \mathbf{k}_{2d})$ be an ordering of the elements of $\mathbf{N}_d$, set $\beta_0 = 0$ and $\beta_m = \sum_{\ell=1}^{m} \mathbb{P}(\mathbf{B}_1 = k_\ell)$ for $1 \leq m \leq 2d$, define $F : [0, 1) \longrightarrow \mathbf{N}_d$ so that $F \restriction [\beta_{m-1}, \beta_m) = k_m$, and set $\mathbf{B}_n = F(U_n)$.

since $\{\rho_{\mathbf{0}}^{(n)} = m\}$ depends only on $(\mathbf{B}_1, \dots, \mathbf{B}_m)$, and is therefore independent of $\rho_{\mathbf{0}} \circ \Sigma^m$, and the distribution of $\rho_{\mathbf{0}} \circ \Sigma^m$ is the same as that of $\rho_{\mathbf{0}}$. Thus, we have proved that

$$\mathbb{P}(\rho_{\mathbf{0}}^{(n)} < \infty) = \mathbb{P}(\rho_{\mathbf{0}} < \infty)^n \quad \text{for } n \geq 1. \tag{1.2.2}$$

One dividend of (1.2.2) is that it supports the epochal picture given above about the structure of recurrent walks. Namely, it says that *if the walk returns once to* $\mathbf{0}$ *with probability* 1, *then, with probability* 1, *it will do so infinitely often.* This observation has many applications. For example, it shows that if the mean value of ith coordinate of $\mathbf{B}_1$ is different from 0, then $\{\mathbf{X}_n : n \geq 0\}$ must be transient. To see this, use Y_n to denote the ith coordinate of $\mathbf{B}_n$, and observe that $\{Y_n - Y_{n-1} : n \geq 1\}$ is a sequence of mutually independent, identically distributed $\{-1, 0, 1\}$-valued random variables with mean value $\mu \neq 0$. But, by the Strong Law of Large Numbers (cf. Exercise 1.3.4 below), this means that $\frac{Y_n}{n} \longrightarrow \mu \neq 0$ with probability 1, which is possible only if $|\mathbf{X}_n| \geq |Y_n| \longrightarrow \infty$ with probability 1, and clearly this eliminates the possibility that, even with positive probability, $\mathbf{X}_n = \mathbf{0}$ infinitely often.

A second dividend of (1.2.2) is the following. Define

$$T_{\mathbf{0}} = \sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{0}\}}(\mathbf{X}_n)$$

to be the total time that $\{\mathbf{X}_n : n \geq 0\}$ spends at the origin. Since $\mathbf{X}_0 = \mathbf{0}$, $T_{\mathbf{0}} \geq 1$. Moreover, for $n \geq 1$, $T_{\mathbf{0}} > n \iff \rho_{\mathbf{0}}^{(n)} < \infty$. Hence, by (1.2.2),

$$\mathbb{E}[T_{\mathbf{0}}] = \sum_{n=0}^{\infty} \mathbb{P}(T_{\mathbf{0}} > n) = 1 + \sum_{n=1}^{\infty} \mathbb{P}(\rho_{\mathbf{0}}^{(n)} < \infty) = 1 + \sum_{n=1}^{\infty} \mathbb{P}(\rho_{\mathbf{0}} < \infty)^n,$$

and so

$$\mathbb{E}[T_{\mathbf{0}}] = \frac{1}{1 - \mathbb{P}(\rho_{\mathbf{0}} < \infty)} = \frac{1}{\mathbb{P}(\rho_{\mathbf{0}} = \infty)}. \tag{1.2.3}$$

Before applying (1.2.3) to the problem of recurrence, it is interesting to note that $T_{\mathbf{0}}$ is a random variable for which the following peculiar dichotomy holds:

$$\begin{aligned} &\mathbb{P}(T_{\mathbf{0}} < \infty) > 0 \implies \mathbb{E}[T_{\mathbf{0}}] < \infty \\ &\mathbb{E}[T_{\mathbf{0}}] = \infty \implies \mathbb{P}(T_{\mathbf{0}} = \infty) = 1. \end{aligned} \tag{1.2.4}$$

Indeed, if $\mathbb{P}(T_{\mathbf{0}} < \infty) > 0$, then, with positive probability, $\mathbf{X}_n$ cannot be $\mathbf{0}$ infinitely often and so, by (1.1.13), $\mathbb{P}(\rho_{\mathbf{0}} < \infty) < 1$, which, by (1.2.3), means that $\mathbb{E}[T_{\mathbf{0}}] < \infty$. On the other hand, if $\mathbb{E}[T_{\mathbf{0}}] = \infty$, then (1.2.3) implies that $\mathbb{P}(\rho_{\mathbf{0}} < \infty) = 1$ and therefore, by (1.2.2), that $\mathbb{P}(T_{\mathbf{0}} > n) = \mathbb{P}(\rho_{\mathbf{0}}^{(n)} < \infty) = 1$ for all $n \geq 1$. Hence (cf. (6.1.3)), $\mathbb{P}(T_{\mathbf{0}} = \infty) = 1$.

1.2.3. Recurrence of Symmetric Random Walk in $\mathbb{Z}^2$: The most frequent way that (1.2.3) gets applied to determine recurrence is in conjunction with the formula

$$\mathbb{E}[T_{\mathbf{0}}] = \sum_{n=0}^{\infty} \mathbb{P}(\mathbf{X}_n = \mathbf{0}). \tag{1.2.5}$$

Although the proof of (1.2.5) is essentially trivial (cf. Theorem 6.1.15):

$$\mathbb{E}[T_{\mathbf{0}}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{0}\}}(\mathbf{X}_n)\right] = \sum_{n=0}^{\infty} \mathbb{E}\left[\mathbf{1}_{\{\mathbf{0}\}}(\mathbf{X}_n)\right] = \sum_{n=0}^{\infty} \mathbb{P}(\mathbf{X}_n = \mathbf{0}),$$

in conjunction with (1.2.3) it becomes powerful. Namely, it says that

$$\{\mathbf{X}_n : n \geq 0\} \text{ is recurrent if and only if } \sum_{n=0}^{\infty} \mathbb{P}(\mathbf{X}_n = \mathbf{0}) = \infty, \tag{1.2.6}$$

and, since $\mathbb{P}(\mathbf{X}_n = \mathbf{0})$ is more amenable to estimation than quantities which involve knowing the trajectory at more than one time, this is valuable information.

In order to apply (1.2.6) to symmetric random walks, it is important to know that *when the walk is symmetric, then* $\mathbf{0}$ *is the most likely place for the walk to be at any even time.* To verify this, note that if $\mathbf{k} \in \mathbb{Z}^d$,

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}_{2n} = \mathbf{k}) = \sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell} \;\&\; \mathbf{X}_{2n} - \mathbf{X}_n = \mathbf{k} - \boldsymbol{\ell}) \\
&= \sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell}) \mathbb{P}(\mathbf{X}_{2n} - \mathbf{X}_n = \mathbf{k} - \boldsymbol{\ell}) = \sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell}) \mathbb{P}(\mathbf{X}_n = \mathbf{k} - \boldsymbol{\ell}) \\
&\leq \left(\sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell})^2\right)^{\frac{1}{2}} \left(\sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \mathbf{k} - \boldsymbol{\ell})^2\right)^{\frac{1}{2}} = \sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell})^2,
\end{aligned}$$

where, in the passage to the last line, we have applied Schwarz's inequality (cf. Exercise 1.3.1 below). Up to this point we have not used symmetry. However, if the walk is symmetric, then $\mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell}) = \mathbb{P}(\mathbf{X}_n = -\boldsymbol{\ell})$, and so the last line of the preceding can by continued as

$$\begin{aligned}
&\sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell}) \mathbb{P}(\mathbf{X}_n = -\boldsymbol{\ell}) \\
&\quad = \sum_{\boldsymbol{\ell} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_n = \boldsymbol{\ell}) \mathbb{P}(\mathbf{X}_{2n} - \mathbf{X}_n = -\boldsymbol{\ell}) = \mathbb{P}(\mathbf{X}_{2n} = \mathbf{0}).
\end{aligned}$$

Thus,

$$\{\mathbf{X}_n : n \geq 0\} \text{ symmetric} \implies \mathbb{P}(\mathbf{X}_{2n} = \mathbf{0}) = \max_{\mathbf{k} \in \mathbb{Z}^d} \mathbb{P}(\mathbf{X}_{2n} = \mathbf{k}). \tag{1.2.7}$$

To develop a feeling for how these considerations get applied, we begin by using them to give a second derivation of the recurrence of the nearest neighbor, symmetric random walk on $\mathbb{Z}$. For this purpose, note that, because $\mathbb{P}(|X_n| \le n) = 1$, (1.2.7) implies that

$$1 = \sum_{\ell=-2n}^{2n} \mathbb{P}(X_{2n} = \ell) \le (4n+1)\mathbb{P}(X_{2n} = 0),$$

and therefore, since the harmonic series diverges, that $\sum_{n=0}^{\infty} \mathbb{P}(X_n = 0) = \infty$.

The analysis for the symmetric, nearest neighbor random walk in $\mathbb{Z}^2$ requires an additional ingredient. Namely, the d-dimensional analog of the preceding line of reasoning would lead to $\mathbb{P}(\mathbf{X}_{2n} = \mathbf{0}) \ge (4n+1)^{-d}$, which is inconclusive except when $d = 1$. In order to do better, we need to use the fact that

$$\{\mathbf{X}_n : n \ge 0\} \text{ symmetric} \implies \mathbb{E}\big[|\mathbf{X}_n|^2\big] = n. \tag{1.2.8}$$

To prove (1.2.8), note that each coordinate of $\mathbf{B}_n$ is a random variable with mean value 0 and variance $\frac{1}{d}$. Hence, because the $\mathbf{B}_n$'s are mutually independent, the second moment of each coordinate of $\mathbf{X}_n$ is $\frac{n}{d}$.

Knowing (1.2.8), Markov's inequality (6.1.12) says that

$$\mathbb{P}\big(|\mathbf{X}_{2n}| \ge 2\sqrt{n}\big) \le \frac{1}{4n}\mathbb{E}\big[|\mathbf{X}_{2n}|^2\big] = \frac{1}{2},$$

which allows us to sharpen the preceding argument to give[9]

$$\begin{aligned}\frac{1}{2} &\le \mathbb{P}\big(|\mathbf{X}_{2n}| < 2\sqrt{n}\big) = \sum_{|\boldsymbol{\ell}|<2\sqrt{n}} \mathbb{P}\big(\mathbf{X}_{2n} = \boldsymbol{\ell}\big) \\ &\le (4\sqrt{n}+1)^d \mathbb{P}\big(\mathbf{X}_{2n} = \mathbf{0}\big) \le 2^{d-1}\big(4n^{\frac{d}{2}}+1\big)\mathbb{P}\big(\mathbf{X}_{2n} = \mathbf{0}\big).\end{aligned}$$

That is, we have now shown that

$$\mathbb{P}\big(\mathbf{X}_{2n} = \mathbf{0}\big) \ge 2^{-d}\big(4n^{\frac{d}{2}}+1\big)^{-1} \tag{1.2.9}$$

for the symmetric, nearest neighbor random walk on $\mathbb{Z}^d$. In particular, when $d = 2$, this proves that *the symmetric, nearest neighbor random walk on $\mathbb{Z}^2$ is recurrent.*

[9] For any $a, b \in [0,\infty)$ and $p \in [1,\infty)$, $(a+b)^p \le 2^{p-1}(a^p + b^p)$. This can be seen as an application of Jensen's inequality (cf. Exercise 5.6.2), which, in this case, is simply the statement that $x \in [0,\infty) \longmapsto x^p$ is convex.

1.2.4. Transience in $\mathbb{Z}^3$: Although (1.2.9) was sufficient to prove recurrence for the symmetric, nearest neighbor random walk in $\mathbb{Z}^2$, it only leaves open the possibility of transience in $\mathbb{Z}^d$ for $d \geq 3$. Thus, in order to nail down the question when $d \geq 3$, we will need to see how good an estimate (1.2.9) really is. In particular, it would suffice to prove that there is an upper bound of the same form.

To get an upper bound which complements the lower bound in (1.2.9), we first do so in the case when $d = 1$. For this purpose, let $0 \leq \ell \leq n$ be given, and observe that

$$\frac{\mathbb{P}(X_{2n} = 2\ell)}{\mathbb{P}(X_{2n} = 0)} = \frac{(n!)^2}{(n+\ell)!(n-\ell)!} = \frac{n(n-1)\cdots(n-\ell+1)}{(n+\ell)(n+\ell-1)\cdots(n+1)}$$
$$= \prod_{k=0}^{\ell-1}\left(1 - \frac{\ell}{n+\ell-k}\right) \geq \left(1 - \frac{\ell}{n+1}\right)^{\ell}.$$

Now recall that

$$\log(1-x) = -\sum_{m=1}^{\infty} \frac{x^m}{m} \quad \text{for } |x| < 1, \tag{1.2.10}$$

and therefore that $\log(1-x) \geq -\frac{3x}{2}$ for $0 \leq x \leq \frac{1}{2}$. Hence, the preceding shows that

$$\frac{\mathbb{P}(X_{2n} = 2\ell)}{\mathbb{P}(X_{2n} = 0)} \geq \exp\left(-\ell \log \frac{\ell}{n+1}\right) \geq e^{-\frac{3\ell^2}{2(n+1)}}$$

as long as $0 \leq \ell \leq \frac{n+1}{2}$. Because $\mathbb{P}(X_{2n} = -2\ell) = \mathbb{P}(X_{2n} = 2\ell)$, we can now say that

$$\mathbb{P}(X_{2n} = 0) \leq e^{\frac{3}{2}}\mathbb{P}(X_{2n} = 2\ell) \quad \text{for } |\ell| \leq \sqrt{n}.$$

But, because $\sum_\ell \mathbb{P}(X_{2n} = 2\ell) = 1$, this means that $\left(2\sqrt{n} - 1\right)\mathbb{P}(X_{2n} = 0) \leq e^{\frac{3}{2}}$, and so

$$\mathbb{P}(X_{2n} = 0) \leq e^{\frac{3}{2}}\left(2\sqrt{n} - 1\right)^{-1}, \quad n \geq 1, \tag{1.2.11}$$

when $\{X_n : n \geq 0\}$ is the symmetric, nearest neighbor random walk on $\mathbb{Z}$.

If, as they most definitely are not, the coordinates of the symmetric, nearest neighbor random walk were independent, then (1.2.11) would yield the sort of upper bound for which we are looking. Thus it is reasonable to examine to what extent we can relate the symmetric, nearest neighbor random walk on $\mathbb{Z}^d$ to d mutually independent symmetric, nearest neighbor random walks $\{X_{i,n} : n \geq 0\}$, $1 \leq i \leq d$, on $\mathbb{Z}$. To this end, refer to (1.2.1) which $p_\epsilon \equiv \frac{1}{2d}$, and think of choosing $\mathbf{X}_n - \mathbf{X}_{n-1}$ in two steps: first choose the coordinate which is to be non-zero and then choose whether it is to be $+1$ or -1. With this in mind, let $\{I_n : n \geq 0\}$ be a sequence of $\{1, \ldots, d\}$-valued, mutually

independent, uniformly distributed random variables which are independent of $\{X_{i,n} : 1 \le i \le d \ \& \ n \ge 0\}$, set, for $1 \le i \le d$, $N_{i,0} = 0$ and $N_{i,n} = \sum_{m=1}^{n} \mathbf{1}_{\{i\}}(I_m)$ when $n \ge 1$, and consider the sequence $\{\mathbf{Y}_n : n \ge 0\}$ given by

$$\mathbf{Y}_n = \big(X_{1,N_{1,n}}, \dots, X_{d,N_{d,n}}\big). \tag{1.2.12}$$

Without too much effort, one can check that $\{\mathbf{Y}_n : n \ge 0\}$ satisfies the conditions in (1.2.1) for the symmetric, nearest neighbor random walk on $\mathbb{Z}^d$ and therefore has the same distribution as $\{\mathbf{X}_n : n \ge 0\}$. In particular, by (1.2.11),

$$\begin{aligned}
\mathbb{P}(\mathbf{X}_{2n} = \mathbf{0}) &= \sum_{\mathbf{m}\in\mathbb{N}^d} \mathbb{P}\big(X_{i,2m_i} = 0 \ \& \ N_{i,2n} = 2m_i \text{ for } 1 \le i \le d\big) \\
&= \sum_{\substack{\mathbf{m}\in\mathbb{N}^d \\ m_1\wedge\cdots\wedge m_d \ge \frac{n}{d}}} \left(\prod_{i=1}^{d} \mathbb{P}\big(X_{i,2m_i} = 0\big)\right) \mathbb{P}\big(N_{i,2n} = 2m_i \text{ for } 1 \le i \le d\big) \\
&\quad + \sum_{\substack{\mathbf{m}\in\mathbb{N}^d \\ m_1\wedge\cdots\wedge m_d < \frac{n}{d}}} \left(\prod_{i=1}^{d} \mathbb{P}\big(X_{i,2m_i} = 0\big)\right) \mathbb{P}\big(N_{i,2n} = 2m_i \text{ for } 1 \le i \le d\big) \\
&\le e^{\frac{3d}{2}} \left(2\sqrt{\tfrac{n}{d}} - 1\right)^{-d} + \mathbb{P}\big(N_{i,2n} \le \tfrac{n}{d} \text{ for some } 1 \le i \le d\big).
\end{aligned}$$

Thus, we will have proved that there is a constant $A(d) < \infty$ such that

$$\mathbb{P}(\mathbf{X}_{2n} = \mathbf{0}) \le A(d) n^{-\frac{d}{2}}, \quad n \ge 1, \tag{1.2.13}$$

once we show that there is a constant $B(d) < \infty$ such that

$$\mathbb{P}\big(N_{i,2n} \le \tfrac{n}{d} \text{ for some } 1 \le i \le d\big) \le B(d) n^{-\frac{d}{2}}, \quad n \ge 1. \tag{1.2.14}$$

In particular, this will complete the proof that

$$d \ge 3 \implies \sum_{n=0}^{\infty} \mathbb{P}(\mathbf{X}_{2n} = \mathbf{0}) < \infty$$

and therefore that *the symmetric, nearest neighbor random walk in $\mathbb{Z}^d$ is transient when $d \ge 3$.*

To prove (1.2.14), first note that

$$\mathbb{P}\big(N_{i,2n} \le \tfrac{n}{d} \text{ for some } 1 \le i \le d\big) \le d\mathbb{P}\big(N_{1,2n} \le \tfrac{n}{d}\big).$$

Next, write $N_{1,n} = \sum_1^n Z_m$ where $Z_m = 1_{\{1\}}(I_m)$, and observe that $\{Z_m : m \geq 1\}$ is a sequence of $\{0,1\}$-valued Bernoulli random variables such that $\mathbb{P}(Z_m = 1) = p \equiv \frac{1}{d}$. In particular, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda \sum_1^n Z_m\right)\right] = \left(pe^{\lambda} + q\right)^n,$$

and so

$$\mathbb{E}\left[\exp\left(\lambda\left(np - \sum_1^n Z_m\right)\right)\right] = e^{n\psi(\lambda)} \text{ where } \psi(\lambda) \equiv \log\left(pe^{-\lambda q} + qe^{\lambda p}\right).$$

Since $\psi(0) = \psi'(0) = 0$, and

$$\psi''(\lambda) = \frac{pqe^{\lambda(p-q)}}{\left(qe^{\lambda p} + pe^{-\lambda q}\right)^2} = \frac{pq}{(qx_\lambda + px_\lambda^{-1})^2} \leq \frac{1}{4}$$

where $x_\lambda \equiv e^{\frac{1}{2}\lambda(p+q)}$, Taylor's formula allows us to conclude that

$$\mathbb{E}\left[\exp\left(\lambda\left(np - \sum_1^n Z_m\right)\right)\right] \leq e^{\frac{n\lambda^2}{8}}, \quad \lambda \in \mathbb{R}. \tag{1.2.15}$$

Starting from (1.2.15), there are many ways to arrive at (1.2.14). For example, for any $\lambda > 0$ and $R > 0$, Markov's inequality (6.1.12) plus (1.2.15) say that

$$\mathbb{P}\left(\sum_1^n Z_m \leq np - nR\right) = \mathbb{P}\left(\exp\left(\lambda\left(np - \sum_1^n Z_m\right)\right) \geq e^{n\lambda R}\right)$$
$$\leq e^{-\lambda nR + \frac{n\lambda^2}{8}},$$

which, when $\lambda = 4nR$, gives

$$\mathbb{P}\left(\sum_1^n Z_m \leq np - nR\right) \leq e^{-2nR^2}. \tag{1.2.16}$$

Returning to the notation used earlier and using the remark with which our discussion of (1.2.14) began, one see from (1.2.16) that

$$\mathbb{P}\left(N_{i,2n} \leq \tfrac{n}{d} \text{ for some } 1 \leq i \leq d\right) \leq de^{-\frac{2n^2}{d^2}},$$

which is obviously far more than is required by (1.2.14).

The argument which we have used in this subsection is an example of an extremely powerful method known as *coupling*. Loosely speaking, the coupling method entails writing a random variable about which one wants to know more (in our case $\mathbf{X}_{2n}$) as a function of random variables about which one knows quite a bit (in our case $\{(X_{i,m}, N_{i,m}) : 1 \leq i \leq d \ \& \ m \geq 0\}$). Of course, the power of the method reflects the power of its user: there are lots of ways in which to couple a random variable to other random variables, but most of them are useless.

1.3 Exercises

EXERCISE 1.3.1. *Schwarz's inequality* comes in many forms, the most elementary of which is the statement that, for any $\{a_n : n \in \mathbb{Z}\} \subseteq \mathbb{R}$ and $\{b_n : n \in \mathbb{Z}\} \subseteq \mathbb{R}$,

$$\sum_{n\in\mathbb{Z}} |a_n b_n| \leq \sqrt{\sum_{n\in\mathbb{Z}} a_n^2}\sqrt{\sum_{n\in\mathbb{Z}} b_n^2}.$$

Moreover, when the right hand side is finite, then

$$\left|\sum_{n\in\mathbb{Z}} a_n b_n\right| = \sqrt{\sum_{n\in\mathbb{Z}} a_n^2}\sqrt{\sum_{n\in\mathbb{Z}} b_n^2}$$

if and only if there is an $\alpha \in \mathbb{R}$ for which either $b_n = \alpha a_n$, $n \in \mathbb{Z}$, or $a_n = \alpha b_n$, $n \in \mathbb{Z}$. Here is an outline of one proof of these statements.

(**a**) Begin by showing that it suffices to treat the case in which $a_n = 0 = b_n$ for all but a finite number of n's.

(**b**) Given a real, quadratic polynomial $P(x) = Ax^2 + 2Bx + C$, use the quadratic formula to see that $P \geq 0$ everywhere if and only if $C \geq 0$ and $B^2 \leq AC$. Similarly, show that $P > 0$ everywhere if and only if $C > 0$ and $B^2 < AC$.

(**c**) Assuming that $a_n = 0 = b_n$ for all but a finite number of n's, set $P(x) = \sum_n (a_n x + b_n)^2$, and apply (**b**) to get the desired conclusions. Finally, use (**a**) to remove the restriction of the a_n's and b_n's.

EXERCISE 1.3.2. Let $\{Y_n : n \geq 1\}$ be a sequence of mutually independent, identically distributed random variables satisfying $\mathbb{E}[|Y_1|] < \infty$. Set $X_n = \sum_{m=1}^n Y_m$ for $n \geq 1$. *The Weak Law of Large Numbers* says that

$$\mathbb{P}\left(\left|\frac{X_n}{n} - \mathbb{E}[Y_1]\right| \geq \epsilon\right) \longrightarrow 0 \quad \text{for all } \epsilon > 0.$$

In fact,

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\frac{X_n}{n} - \mathbb{E}[Y_1]\right|\right] = 0, \tag{1.3.3}$$

from which the above follows as an application of Markov's inequality. Here are steps which lead to (1.3.3).

(**a**) First reduce to the case when $\mathbb{E}[Y_1] = 0$. Next, assume that $\mathbb{E}[Y_1^2] < \infty$, and show that

$$\mathbb{E}\left[\left|\frac{X_n}{n}\right|\right]^2 \leq \mathbb{E}\left[\left|\frac{X_n}{n}\right|^2\right] = \frac{\mathbb{E}[Y_1^2]}{n}.$$

Hence the result is proved when Y_1 has a finite second moment.

(**b**) Given $R > 0$, set $Y_n^{(R)} = Y_n \mathbf{1}_{[0,R)}(|Y_n|) - \mathbb{E}[Y_n,\ |Y_n| < R]$ and $X_n^{(R)} = \sum_{m=1}^{n} Y_m^{(R)}$. Note that, for any $R > 0$,

$$\mathbb{E}\left[\left|\frac{X_n}{n}\right|\right] \leq \mathbb{E}\left[\left|\frac{X_n^{(R)}}{n}\right|\right] + \mathbb{E}\left[\left|\frac{X_n - X_n^{(R)}}{n}\right|\right]$$

$$\leq \sqrt{\mathbb{E}\left[\left(\frac{X_n^{(R)}}{n}\right)^2\right]} + 2\mathbb{E}\big[|Y_1|,\ |Y_1| \geq R\big] \leq \frac{R}{n^{\frac{1}{2}}} + 2\mathbb{E}\big[|Y_1|,\ |Y_1| \geq R\big],$$

and use this, together with the Monotone Convergence Theorem, to complete the proof of (1.3.3).

EXERCISE 1.3.4. Refer to Exercise 1.3.2. *The Strong Law of Large Numbers* says that the statement in the Weak Law can be improved to the statement that $\frac{X_n}{n} \longrightarrow \mathbb{E}[Y_1]$ with probability 1. The proof of the Strong Law when one assumes only that $\mathbb{E}[|Y_1|] < \infty$ is a bit tricky. However, if one is willing to assume that $\mathbb{E}[Y_1^4] < \infty$, then a proof can be based on the same type argument which leads to the Weak Law.

Let $\{Y_n\}_1^\infty$ be a sequence of mutually independent random variables with the properties that $M = \sup_n \mathbb{E}\big[|Y_n|^4\big] < \infty$, and prove that, with probability 1, $\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \big(Y_m - \mathbb{E}[Y_m]\big) = 0$. Note that we have not assume yet that they are identically distributed, but when we add this assumption we get $\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} = \mathbb{E}[Y_1]$ with probability 1.

Here is an outline.

(**a**) Begin by reducing to the case when $\mathbb{E}[Y_n] = 0$ for all $n \in \mathbb{Z}^+$.

(**b**) After writing

$$\mathbb{E}\left[\left(\sum_1^n Y_k\right)^4\right] = \sum_{k_1,\dots,k_4=1}^{n} \mathbb{E}\big[Y_{k_1} \cdots Y_{k_4}\big]$$

and noting that the only terms which do not vanish are those for which each index is equal to at least one other index, conclude that

$$\mathbb{E}\left[\left(\sum_1^n Y_k\right)^4\right] = \sum_{k=1}^{n} \mathbb{E}\big[Y_k^4\big] + 6 \sum_{1\leq k<\ell\leq n} \mathbb{E}\big[Y_k^2\big]\mathbb{E}\big[Y_\ell^2\big].$$

Hence, since $\mathbb{E}[Y_k^2]^2 \leq \mathbb{E}[Y_k^4]$,

$$(*) \qquad \mathbb{E}\left[\left(\sum_1^n Y_k\right)^4\right] \leq 3Mn^2.$$

(**c**) Starting from (*), show that

$$\mathbb{P}\left(\left|\frac{\sum_1^n Y_k}{n}\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^4}\mathbb{E}\left[\left|\frac{\sum_1^n Y_k}{n}\right|^4\right] \leq \frac{3M}{\epsilon^4 n^2} \longrightarrow 0$$

for all $\epsilon > 0$. This is the Weak Law of Large Numbers for independent random variables with bounded fourth moments. Of course, the use of four moments here is somewhat ridiculous since the argument using only two moments is easier.

(**d**) Starting again from (*) and using (6.1.4), show that

$$\begin{aligned}\mathbb{P}\left(\sup_{n>m}\left|\frac{\sum_1^n Y_k}{n}\right| \geq \epsilon\right) &\leq \sum_{n=m+1}^{\infty} \mathbb{P}\left(\left|\sum_1^n Y_k\right| \geq n\epsilon\right) \\ &\leq \frac{4M}{\epsilon^4}\sum_{n=m+1}^{\infty}\frac{1}{n^2} \leq \frac{4M}{\epsilon^4 m} \longrightarrow 0 \quad \text{as } m \to \infty \text{ for all } \epsilon > 0.\end{aligned}$$

(**e**) Use the definition of convergence plus (6.1.4) to show that

$$\begin{aligned}\mathbb{P}\left(\frac{\sum_1^n Y_k}{n} \not\to 0\right) &= \mathbb{P}\left(\bigcup_{N=1}^{\infty}\bigcap_{m=1}^{\infty}\bigcup_{n>m}\left|\frac{\sum_1^n Y_k}{n}\right| \geq \frac{1}{N}\right) \\ &\leq \sum_{N=1}^{\infty}\mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n>m}\left|\frac{\sum_1^n Y_k}{n}\right| \geq \frac{1}{N}\right).\end{aligned}$$

Finally, apply the second line of (6.1.3) plus (**d**) above to justify

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n>m}\left|\frac{\sum_1^n Y_k}{n}\right| \geq \frac{1}{N}\right) = \lim_{m\to\infty}\mathbb{P}\left(\bigcup_{n>m}\left|\frac{\sum_1^n Y_k}{n}\right| \geq \frac{1}{N}\right) = 0$$

for each $N \in \mathbb{Z}^+$. Hence, with probability 1, $\frac{1}{n}\sum_1^n Y_k \longrightarrow 0$, which is the Strong Law of Large Numbers for independent random variables with bounded fourth moments.

EXERCISE 1.3.5. Readers who know DeMoivre's proof of the Central Limit Theorem will have realized that the estimate in (1.2.11) is a poor man's substitute for what one can get as a consequence of *Stirling's formula*

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{as } n \to \infty, \tag{1.3.6}$$

meaning that the ratio of the quantities on the two sides of "$\sim$" tends to 1. Indeed, given (1.3.6), show that

$$\mathbb{P}(X_{2n} = 0) \sim \sqrt{\frac{2}{\pi n}}.$$

Next, give a proof of (1.3.6) based on the following line of reasoning.

(**a**) Let $\tau_1, \ldots, \tau_n$ be a mutually independent, unit exponential random variables,[10] and show that for any $0 < R \leq \sqrt{n}$

$$1 - \frac{1}{R^2} \leq \mathbb{P}\left(-R \leq \frac{\tau_1 + \cdots + \tau_n - n}{\sqrt{n}} \leq R\right) = \frac{1}{(n-1)!} \int_{-\sqrt{n}R+n}^{\sqrt{n}R+n} t^{n-1} e^{-t}\, dt.$$

(**b**) Make a change of variables followed by elementary manipulations to show that

$$\begin{aligned}
\int_{-\sqrt{n}R+n}^{\sqrt{n}R+n} t^{n-1} e^{-t}\, dt &= \sqrt{n} e^{-n} \int_{-R}^{R} \bigl(n + \sqrt{n}\sigma\bigr)^{n-1} e^{-\sqrt{n}\sigma}\, d\sigma \\
&= n^{n-\frac{1}{2}} e^{-n} \int_{-R}^{R} \left(1 + \frac{\sigma}{\sqrt{n}}\right)^{n-1} e^{-\sqrt{n}\sigma}\, d\sigma \\
&= n^{n-\frac{1}{2}} e^{-n} \int_{-R}^{R} \exp\left(-\frac{\sigma^2}{2} + E_n(\sigma)\right) d\sigma,
\end{aligned}$$

where

$$E_n(\sigma) \equiv (n-1) \log\left(1 + \frac{\sigma}{\sqrt{n}}\right) - \sqrt{n}\sigma + \frac{\sigma^2}{2}.$$

(**c**) As an application of the Taylor's series for $\log(1+x)$ (cf. (1.2.10)), show that $E_n(\sigma) \longrightarrow 0$ uniformly for $|\sigma| \leq R$ when $n \to \infty$, and combine this with the results in (**a**) and (**b**) to arrive at

$$\overline{\lim_{n\to\infty}} \frac{n^{n+\frac{1}{2}} e^{-n}}{n!} \int_{-R}^{R} e^{-\frac{\sigma^2}{2}}\, d\sigma \leq 1$$

and

$$\underline{\lim_{n\to\infty}} \frac{n^{n+\frac{1}{2}} e^{-n}}{n!} \int_{-R}^{R} e^{-\frac{\sigma^2}{2}}\, d\sigma \geq 1 - \frac{1}{R^2}.$$

Because $\int_{-\infty}^{\infty} e^{-\frac{\sigma^2}{2}}\, d\sigma = \sqrt{2\pi}$, it is clear that (1.3.6) follows after one lets $R \nearrow \infty$.

EXERCISE 1.3.7. The argument in §1.2.3 is quite robust. Indeed, let $\{\mathbf{X}_n : n \geq 0\}$ be any symmetric random walk on $\mathbb{Z}^2$ whose jumps have finite second moment. That is, $\mathbf{X}_0 = \mathbf{0}$, $\{\mathbf{X}_n - \mathbf{X}_{n-1} : n \geq 1\}$ are mutually independent, identically distributed, symmetric ($\mathbf{X}_1$ has the same distribution as $-\mathbf{X}_1$), $\mathbb{Z}^2$-valued random variables with finite second moment. Show that $\{\mathbf{X}_n : n \geq 0\}$ is recurrent in the sense that $\mathbb{P}(\exists n \geq 1\ \mathbf{X}_n = \mathbf{0}) = 1$.

[10] A unit exponential random variable is a random variable τ for which $\mathbb{P}(\tau > t) = e^{-t \vee 0}$.

EXERCISE 1.3.8. Let $\{\mathbf{X}_n : n \geq 0\}$ be a random walk on $\mathbb{Z}^d$: $\mathbf{X}_0 = \mathbf{0}$, $\{\mathbf{X}_n - \mathbf{X}_{n-1} : n \geq 1\}$ are mutually independent, identically distributed, $\mathbb{Z}^d$-valued random variables. Further, for each $1 \leq i \leq d$, let $(\mathbf{X}_n)_i$ be the ith coordinate of $\mathbf{X}_n$, and assume that

$$\min_{1 \leq i \leq d} \mathbb{P}\big((\mathbf{X}_1)_i \neq 0\big) > 0 \quad \text{but} \quad \mathbb{P}\big(\exists\, i \neq j\ (\mathbf{X}_1)_i (\mathbf{X}_1)_j \neq 0\big) = 0.$$

If, for some $C < \infty$ and $(\alpha_1, \ldots, \alpha_d) \in [0, \infty)^d$ with $\sum_1^d \alpha_i > 1$, $\mathbb{P}\big((\mathbf{X}_n)_i = 0\big) \leq Cn^{-\alpha_i}$, $n \geq 1$, show that $\{\mathbf{X}_n : n \geq 0\}$ is transient in the sense that $\mathbb{P}(\exists n \geq 1\ \mathbf{X}_n = \mathbf{0}) < 1$.

EXERCISE 1.3.9. Let $\{\mathbf{X}_n : n \geq 0\}$ be a random walk on $\mathbb{Z}^d$, as in the preceding. Given $\mathbf{k} \in \mathbb{Z}^d$, set

$$T_{\mathbf{k}} = \sum_{n=0}^{\infty} \mathbf{1}_{\{\mathbf{k}\}}(\mathbf{X}_n) \quad \text{and} \quad \zeta_{\mathbf{k}} = \inf\{n \geq 0 : \mathbf{X}_n = \mathbf{k}\}.$$

Show that

$$\mathbb{E}[T_{\mathbf{k}}] = \mathbb{P}(\zeta_{\mathbf{k}} < \infty)\mathbb{E}[T_{\mathbf{0}}] = \frac{\mathbb{P}(\zeta_{\mathbf{k}} < \infty)}{\mathbb{P}(\rho_{\mathbf{0}} = \infty)}, \tag{1.3.10}$$

where $\rho_{\mathbf{0}} = \inf\{n \geq 1 : \mathbf{X}_n = \mathbf{0}\}$ is the time of first return to $\mathbf{0}$. In particular, if $\{\mathbf{X}_n : n \geq 0\}$ is transient in the sense described in the preceding exercise, show that

$$\mathbb{E}\left[\sum_{m=0}^{\infty} \mathbf{1}_{B(r)}(\mathbf{X}_n)\right] < \infty \quad \text{for all } r \in (0, \infty),$$

where $B(r) = \{\mathbf{k} : |\mathbf{k}| \leq r\}$; and from this conclude that $|\mathbf{X}_n| \longrightarrow \infty$ with probability 1. On the other hand, if $\{\mathbf{X}_n : n \geq 0\}$ is recurrent, show that $\mathbf{X}_n = \mathbf{0}$ infinitely often with probability 1. Hence, either $\{\mathbf{X}_n : n \geq 0\}$ is recurrent and $\mathbf{X}_n = \mathbf{0}$ infinitely often with probability 1 or it is transient and $|\mathbf{X}_n| \longrightarrow \infty$ with probability 1.

EXERCISE 1.3.11. Take $d = 1$ in the preceding, $X_0 = 0$, and $\{X_n - X_{n-1} : n \geq 1\}$ to be mutually independent, identically distributed random variables for which $0 < \mathbb{E}\big[|X_1|\big] < \infty$ and $\mathbb{E}[X_1] = 0$. By a slight variation on the argument given in §1.2.1, we will show here that this random walk is recurrent but that

$$\overline{\lim_{n \to \infty}}\, X_n = \infty \quad \text{and} \quad \underline{\lim_{n \to \infty}}\, X_n = -\infty \quad \text{with probability 1.}$$

(a) First show that it suffices to prove that $\sup_n X_n = \infty$ and that $\inf_n X_n = -\infty$. Next, use the Weak Law of Large Numbers (cf. Exercise 1.3.2) to show that

$$\lim_{n \to \infty} \max_{1 \leq m \leq n} \frac{\mathbb{E}\big[|X_n|\big]}{n} = 0.$$

(b) For $n \geq 1$, set $T_k^{(n)} = \sum_{m=0}^{n-1} \mathbf{1}_{\{k\}}(X_m)$, show that $\mathbb{E}[T_k^{(n)}] \leq \mathbb{E}[T_0^{(n)}]$ for all $k \in \mathbb{Z}$, and use this to arrive at

$$(4\mu(n)+1)\mathbb{E}[T_0^{(n)}] \geq \frac{n}{2} \quad \text{where } \mu(n) \equiv \max_{0 \leq m \leq n-1} \mathbb{E}[|X_m|].$$

Finally, apply part (a) to conclude that $\mathbb{E}[T_0] = \infty$. Hence, by (1.3.10), $\mathbb{P}(\rho_0 < \infty) = 1$, and so $\{X_n : n \geq 0\}$ is recurrent.

(c) To complete the program, proceed as in the derivation of (1.2.2) to pass from (b) to

$$(*) \qquad \mathbb{P}(\rho_0^{(m)} < \infty) = 1 \quad \text{for all } m \geq 1,$$

where $\rho_0^{(m)}$ is the time of the mth return to 0. Next, for $r \in \mathbb{Z}^+$, set $\eta_r = \inf\{n \geq 0 : X_n \geq r\}$, show that $\epsilon \equiv \mathbb{P}(\eta_1 > \rho_0) < 1$, and conclude that $\mathbb{P}(\eta_1 > \rho_0^{(m)}) \leq \epsilon^m$. Now, combine this with (*) to get $\mathbb{P}(\eta_1 < \infty) = 1$. Finally, argue that

$$\mathbb{P}(\eta_{r+1} < \infty) \geq \mathbb{P}(\eta_r < \infty)\mathbb{P}(\eta_1 < \infty)$$

and therefore that $\mathbb{P}(\eta_r < \infty) = 1$ for all $r \geq 1$. Since this means that, with probability 1, $\sup_n X_n \geq r$ for all $r \geq 1$, it follows that $\sup_n X_n = \infty$ with probability 1. To prove that $\inf_n X_n = -\infty$ with probability 1, simply replace $\{X_n : n \geq 0\}$ by $\{-X_n : n \geq 0\}$.

EXERCISE 1.3.12. [11] Here is an interesting application of one dimensional random walks to elementary *queuing theory*. Queuing theory deals with the distribution of the number of people waiting to be served (i.e., the length of the queue) when, during each time interval, the number of people who arrive and the number of people who are served are random. The queuing model which we will consider here is among the simplest. Namely, we will assume that, during the time interval $[n-1, n)$, the number of people who arrive minus the number who can be served is given by a $\mathbb{Z}$-valued random variable B_n. Further, we assume that the B_n's are mutually independent and identically distributed random variables satisfying $0 < \mathbb{E}[|B_1|] < \infty$. The associated queue is, apart from the fact that there are never a negative number of people waiting, the random walk $\{X_n : n \geq 0\}$ determined by the B_n's: $X_0 = 0$ and $X_n = \sum_{m=1}^n B_m$. To take into account the prohibition against having a queue of negative length, the queuing model $\{Q_n : n \geq 0\}$ is given by the prescription

$$Q_0 = 0 \quad \text{and} \quad Q_n = (Q_{n-1} + B_n)^+ \quad \text{for } n \geq 1.$$

(a) Show that

$$Q_n = X_n - \min_{0 \leq m \leq n} X_m = \max_{0 \leq m \leq n} (X_n - X_m),$$

and conclude that, for each $n \geq 0$, the distribution of Q_n is the same as that of $M_n \equiv \max_{0 \leq m \leq n} X_m$.

[11] So far as I know, this example was invented by Wm. Feller.

(**b**) Set $M_\infty \equiv \lim_{n\to\infty} M_n \in \mathbb{N}\cup\{\infty\}$, and, as a consequence of (**a**), arrive at

$$\lim_{n\to\infty} \mathbb{P}\big(Q_n = j\big) = \mathbb{P}\big(M_\infty = j\big) \quad \text{for } j \in \mathbb{N}.$$

(**c**) Set $\mu \equiv \mathbb{E}[B_1]$. The Weak Law of Large Numbers says that, for each $\epsilon > 0$, $\mathbb{P}\big(|X_n - n\mu| \geq n\epsilon\big) \longrightarrow 0$ as $n \to \infty$. In particular, when $\mu > 0$, show that $\mathbb{P}(M_\infty = \infty) = 1$. When $\mu = 0$, use Exercise 1.3.11 to reach the same conclusion. Hence, when $\mathbb{E}[B_1] \geq 0$, $\mathbb{P}(Q_n = j) \longrightarrow 0$ for all $j \in \mathbb{N}$. That is, when the expected number of arrivals is a least as large of the expected number of people served, then, with probability 1, the queue grows infinitely long.

(**d**) Now assume that $\mu \equiv \mathbb{E}[B_1] < 0$. Then the Strong Law of Large Numbers (cf. Exercise 1.3.4 for the case when B_1 has a finite fourth moment and Theorem 1.4.11 in [9] for the general case) says that $\frac{X_n}{n} \longrightarrow \mu$ with probability 1. In particular, conclude that $M_\infty < \infty$ with probability 1 and therefore that $\sum_{j\in\mathbb{N}} \nu_j = 1$ when $\nu_j \equiv \lim_{n\to\infty} \mathbb{P}(Q_n = j) = \mathbb{P}(M_\infty = j)$.

(**e**) Specialize to the case when the B_m's are $\{-1, 1\}$-valued Bernoulli random variables with $p \equiv \mathbb{P}(B_1 = 1) \in (0, 1)$, and set $q = 1 - p$. Use the calculations in (1.1.12) to show that

$$\lim_{n\to\infty} \mathbb{P}\big(Q_n = j\big) = \begin{cases} 0 & \text{if } p \geq q \\ \frac{q-p}{q}\left(\frac{p}{q}\right)^j & \text{if } p < q. \end{cases}$$

(**f**) Generalize (**e**) to the case when $B_m \in \{-1, 0, 1\}$, $p = \mathbb{P}(B_1 = 1)$, and $q = \mathbb{P}(B_1 = -1)$. The idea is that M_∞ in this case has the same distribution as $\sup_n Y_n$, where $\{Y_n : n \geq 0\}$ is the random walk corresponding to $\{-1, 1\}$-valued Bernoulli random variables which are 1 with probability $\frac{p}{p+q}$.

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 $\{1, \ldots, N\}$ 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 $(i_0, \ldots, i_n) \in \mathbb{S}^{n+1}$, and observe that

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

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,

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

Hence (2.1.4) results after one sums with respect to $(i_0, \ldots, i_{n-1})$. 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*: $\|\boldsymbol{\rho} + \boldsymbol{\rho}'\|_{\mathrm{v}} \leq \|\boldsymbol{\rho}\|_{\mathrm{v}} + \|\boldsymbol{\rho}'\|_{\mathrm{v}}$. 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:

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

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 $m \to \infty$,

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

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 $\{1, \ldots, N\}$ 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 $0 \leq m \leq n$,

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

since

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

In particular, if $\boldsymbol{\mu}$ is the initial distribution of $\{X_n : n\geq 0\}$, 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 $|\boldsymbol{\mu}\mathbf{f}|\leq\|\boldsymbol{\mu}\|_{\mathrm{v}}\|\mathbf{f}\|_{\mathrm{u}}$, since

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

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

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

2.1.4. The Markov Property: By definition, if $\boldsymbol{\mu}$ is the initial distribution of $\{X_n : n\geq 0\}$, 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 $m,n\geq 1$ 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: $\mathbf{P}\mathbf{1} = \mathbf{1}$. 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

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

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 $\boldsymbol{\mu}$,

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

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

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

Second, for each $k \geq 0$,

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

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

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

which is what we wanted.

To complete the proof of Theorem 2.2.5 from here, assume that $(\mathbf{A}_M)_{ij_0} \geq \epsilon$ 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 $\mathbf{A}_M$, $\|\boldsymbol{\mu}\mathbf{A}_n\mathbf{A}_M - \boldsymbol{\pi}\|_{\mathrm{v}} = \|(\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi})\mathbf{A}_M\|_{\mathrm{v}} \leq (1-\epsilon)\|\boldsymbol{\mu}\mathbf{A}_n - \boldsymbol{\pi}\|_{\mathrm{v}}$, which, in conjunction with (2.2.6), leads to

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

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 $|\boldsymbol{\mu}\mathbf{f}| \le \|\boldsymbol{\mu}\|_{\mathrm{v}}\|f\|_{\mathrm{u}}$)[5]

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

and (cf. Theorem 2.2.5)

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

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

2.3.1. The Mean Ergodic Theorem: Let $\{\mathbf{X}_n : n \ge 0\}$ 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,*

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

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

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

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,

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

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 $\|\bar{\mathbf{f}}\|_{\mathrm{u}} \le \|\mathbf{f}\|_{\mathrm{u}}$,

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

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 $n \geq 1$,

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

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

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

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 $(X_0, \ldots, X_n)$.

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

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

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

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

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

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

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 π: 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

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

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,

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

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 $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_k, \ldots)$ 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

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

and that

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

.

(**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 $\mathbb{E}\left[X_n \,\middle|\, X_0 = i\right] = i\gamma^n$. 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

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

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 $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_k, \ldots)$ 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.

CHAPTER **3**

More about the Ergodic Theory of Markov Chains

In Chapter 2, all of our considerations centered around one form or another of the Doeblin condition which says that there is a state which can be reached from any other state at a uniformly fast rate. Although there are lots of chains on an infinite state space which satisfy his condition, most do not. As a result, many chains on an infinite state space will not even admit a stationary probability distribution. Indeed, the fact that there are infinitely many states means that there is enough space for the chain to "get lost and disappear." There are two ways in which this can happen. Namely, the chain can disappear because, like the a nearest neighbor, non-symmetric random walk in $\mathbb{Z}$ (cf. (1.1.13)) or even the symmetric one in $\mathbb{Z}^3$ (cf. §1.2.4), it may have no recurrent states and, as a consequence, will spend a finite amount of time in any given state. A more subtle way for the chain to disappear is for it to be recurrent but not sufficiently recurrent for there to exist a stationary distribution. Such an example is the symmetric, nearest neighbor random walk in $\mathbb{Z}$ which is recurrent, but just barely so. In particular, although this random walk returns infinitely often to the place where is begins, it does so at too sparse a set of times. More precisely, by (1.2.7) and (1.2.13), if $\mathbf{P}$ is the transition probability matrix for the symmetric, nearest neighbor random walk on $\mathbb{Z}$, then

$$(\mathbf{P}^{2n})_{ij} \leq (\mathbf{P}^{2n})_{ii} = \mathbb{P}(X_{2n} = 0) \leq A(1)n^{-\frac{1}{2}} \longrightarrow 0,$$

and so, if $\boldsymbol{\mu}$ were a probability vector which was stationary for $\mathbf{P}$, then, by Lebesgue's Dominated Convergence Theorem, Theorem 6.1.11, we would have the contradiction

$$(\boldsymbol{\mu})_j = \lim_{n\to\infty} \sum_{i\in\mathbb{Z}} (\boldsymbol{\mu})_i (\mathbf{P}^{2n})_{ij} = 0 \quad \text{for all } j \in \mathbb{Z}.$$

In this chapter, we will see that ergodic properties can exist in the absence of Doeblin's condition. However, as we will see, what survives does so in a weaker form. Specifically, we will no longer be looking for convergence in the $\|\cdot\|_{\mathrm{v}}$-norm and instead will settle for pointwise convergence. That is, we will be looking for results of the form $(\boldsymbol{\mu}\mathbf{P})_j \longrightarrow (\boldsymbol{\pi})_j$ for each j rather than $\|\boldsymbol{\mu}\mathbf{P} - \boldsymbol{\pi}\|_{\mathrm{v}} \longrightarrow 0$.

3.1 Classification of States

In this section we deal with a topic which was hinted at but not explicitly discussed in Chapter 2. Namely, because we will no longer be making an assumption like Doeblin's, it will be necessary to take into account the possibility that the chain sees the state space as a union of disjoint parts. To be precise, given a pair (i, j) of states, recall that we write $i \to j$ and say that j is accessible from i if, with positive probability, the chain can go from state i to state j. That is, $(\mathbf{P}^n)_{ij} > 0$ for some $n \in \mathbb{N}$. Notice that accessibility is transitive in the sense that

$$i \to j \text{ and } j \to \ell \implies i \to \ell. \tag{3.1.1}$$

Indeed, if $(\mathbf{P}^m)_{ij} > 0$ and $(\mathbf{P}^n)_{j\ell} > 0$, then

$$(\mathbf{P}^{m+n})_{i\ell} = \sum_k (\mathbf{P}^m)_{ik}(\mathbf{P}^n)_{k\ell} \geq (\mathbf{P}^m)_{ij}(\mathbf{P}^n)_{j\ell} > 0.$$

If i and j are accessible from one another in the sense that $i \to j$ and $j \to i$, then we write $i \leftrightarrow j$ and say that i *communicates with* j. It should be clear that $\leftrightarrow$ is an equivalence relation. To wit, because $(\mathbf{P}^0)_{ii} = 1$, $i \leftrightarrow i$, and it is trivial that $j \leftrightarrow i$ if $i \leftrightarrow j$. Finally, if $i \leftrightarrow j$ and $j \leftrightarrow \ell$, then (3.1.1) makes it obvious that $i \leftrightarrow \ell$. Thus, "$\leftrightarrow$" leads to a partitioning of the state space into equivalence classes made up of communicating states. That is, for each state i, the communicating equivalence class $[i]$ of i is the set of states j such that $i \leftrightarrow j$; and, for every pair (i, j), either $[i] = [j]$ or $[i] \cap [j] = \emptyset$. In the case when every state communicates with every other state, we say that the chain is *irreducible*.

3.1.1. Classification, Recurrence, and Transience: In this subsection, we will show that recurrence and transience are *communicating class properties*. That is, either all members of a communicating equivalence class are recurrent or all members are transient.

Recall (cf. §2.3.2) that ρ_j is the time of first return to j and that we say j is *recurrent* or *transient* according to whether $\mathbb{P}(\rho_j < \infty | X_0 = j)$ is equal to or strictly less than 1.

3.1.2 THEOREM. *Assume that i is recurrent and that $j \neq i$. Then $i \to j$ if and only if $\mathbb{P}(\rho_j < \rho_i | X_0 = i) > 0$. Moreover, if $i \to j$, then $\mathbb{P}(\rho_k < \infty | X_0 = \ell) = 1$ for any $(k, \ell) \in \{i, j\}^2$. In particular, $i \to j$ implies that $i \leftrightarrow j$ and that j is recurrent.*

PROOF: Given $j \neq i$ and $n \geq 1$, set (cf. (2.3.5))

$$G_n(k_0, \ldots, k_n) = \big(F_{n-1,i}(k_0, \ldots, k_{n-1}) - F_{n,i}(k_0, \ldots, k_n)\big) F_{n,j}(k_0, \ldots, k_n).$$

If $\{\rho_j^{(m)} : m \geq 0\}$ are defined as in §2.3.2, then, by (2.1.2),

$$\begin{aligned}
\mathbb{P}\big(\rho_i^{(m+1)} < \rho_j \,\big|\, X_0 = i\big) &= \sum_{\ell=1}^{\infty} \mathbb{P}\big(\rho_i^{(m)} = \ell \;\&\; \rho_i^{(m+1)} < \rho_j \,\big|\, X_0 = i\big) \\
&= \sum_{\ell=1}^{\infty}\sum_{n=1}^{\infty} \mathbb{P}\big(\rho_i^{(m)} = \ell,\ \rho_i^{(m+1)} = \ell + n < \rho_j \,\big|\, X_0 = i\big) \\
&= \sum_{\ell,n=1}^{\infty} \mathbb{E}\Big[G_n(X_\ell, \dots, X_{\ell+n}),\ \rho_i^{(m)} = \ell < \rho_j \,\Big|\, X_0 = i\Big] \\
&= \sum_{\ell,n=1}^{\infty} \mathbb{E}\Big[G_n(X_0, \dots, X_n) \,\Big|\, X_0 = i\Big]\mathbb{P}\big(\rho_i^{(m)} = \ell < \rho_j \,\big|\, X_0 = i\big) \\
&= \sum_{\ell,n=1} \mathbb{P}\big(\rho_i = n < \rho_j \,\big|\, X_0 = i\big)\mathbb{P}\big(\rho_i^{(m)} = \ell < \rho_j \,\big|\, X_0 = i\big) \\
&= \mathbb{P}\big(\rho_i < \rho_j \,\big|\, X_0 = i\big)\mathbb{P}\big(\rho_i^{(m)} < \rho_j \,\big|\, X_0 = i\big),
\end{aligned}$$

and so

$$j \neq i \implies \mathbb{P}\big(\rho_i^{(m)} < \rho_j \,\big|\, X_0 = i\big) = \mathbb{P}\big(\rho_i < \rho_j \,\big|\, X_0 = i\big)^m. \tag{3.1.3}$$

Now suppose that $i \to j$ but $\mathbb{P}(\rho_j < \rho_i | X_0 = i) = 0$. Then $\mathbb{P}(\rho_i \leq \rho_j | X_0 = i) = 1$, and so, because $\mathbb{P}(\rho_j \neq \rho_i | X_0 = i) \geq \mathbb{P}(\rho_i < \infty | X_0 = i) = 1$, $\mathbb{P}(\rho_i < \rho_j | X_0 = i) = 1$. Hence, by (3.1.3), this means that $\mathbb{P}(\rho_i^{(m)} < \rho_j | X_0 = i) = 1$ for all $m \geq 1$, which, since $\rho^{(m)} \geq m$, leads to $\mathbb{P}(\rho_j = \infty | X_0 = i) = 1$ and therefore rules out $i \to j$. That is, we have now shown that $i \to j \implies \mathbb{P}(\rho_j < \rho_i | X_0 = i) > 0$, and the opposite implication needs no comment.

To prove that $i \to j \implies \mathbb{P}(\rho_i < \infty | X_0 = j) = 1$, first observe that

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

Thus, after combining this with $\mathbb{P}(\rho_i < \infty | X_0 = i) = 1$, we have

$$\mathbb{P}\big(\rho_j < \rho_i \,\big|\, X_0 = i\big) = \mathbb{P}\big(\rho_j < \rho_i \,\big|\, X_0 = i\big)\mathbb{P}\big(\rho_i < \infty \,\big|\, X_0 = j\big),$$

which, because $\mathbb{P}(\rho_j < \rho_i | X_0 = i) > 0$, is possible only if $\mathbb{P}(\rho_i < \infty | X_0 = j) = 1$. In particular, we have now proved that $j \to i$ and therefore that $i \leftrightarrow j$.

Similarly,

$$\begin{aligned}\mathbb{P}(\rho_j < \infty \,|\, X_0 = i) &= \mathbb{P}(\rho_j < \rho_i \,|\, X_0 = i) + \mathbb{P}(\rho_i < \rho_j < \infty \,|\, X_0 = i) \\ &= \mathbb{P}(\rho_j < \rho_i \,|\, X_0 = i) + \mathbb{P}(\rho_i < \rho_j \,|\, X_0 = i)\mathbb{P}(\rho_j < \infty \,|\, X_0 = i),\end{aligned}$$

and so

$$\mathbb{P}(\rho_j < \infty \,|\, X_0 = i)\mathbb{P}(\rho_j < \rho_i \,|\, X_0 = i) = \mathbb{P}(\rho_j < \infty \,|\, X_0 = i).$$

Hence, $i \rightarrow j \implies \mathbb{P}(\rho_j < \infty | X_0 = i) = 1$.

Finally,

$$\mathbb{P}(\rho_i < \rho_j < \infty \,|\, X_0 = j) = \mathbb{P}(\rho_j < \infty \,|\, X_0 = i)\mathbb{P}(\rho_i < \rho_j \,|\, X_0 = j).$$

Hence, because we now know that $\mathbb{P}(\rho_i < \infty | X_0 = j) = 1 = \mathbb{P}(\rho_j < \infty | X_0 = i)$ when $i \rightarrow j$, we see that $i \rightarrow j$ implies

$$\begin{aligned}\mathbb{P}(\rho_j < \infty \,|\, X_0 = j) &= \mathbb{P}(\rho_j < \rho_i \,|\, X_0 = j) + \mathbb{P}(\rho_i < \rho_j < \infty \,|\, X_0 = j) \\ &= \mathbb{P}(\rho_j < \rho_i \,|\, X_0 = j) + \mathbb{P}(\rho_j < \infty \,|\, X_0 = i)\mathbb{P}(\rho_i < \rho_j \,|\, X_0 = j) = 1,\end{aligned}$$

since $\mathbb{P}(\rho_i = \rho_j | X_0 = j) \leq \mathbb{P}(\rho_i = \infty | X_0 = j) = 0$. □

As an immediate consequence of Theorem 3.1.2 we have the following corollary.

3.1.4 Corollary. *If $i \leftrightarrow j$, then j is recurrent (transient) if and only if i is. Moreover, if i is recurrent, then $\mathbb{P}(\rho_j < \infty | X_0 = i)$ is either 1 or 0 according whether or not i communicates with j. In particular, if i is recurrent, then $(\mathbf{P}^n)_{ij} = 0$ for all $n \geq 0$ and all j which do not communicate with i.*

When a chain is irreducible, all or none of its states possess any particular communicating class property. Hence, when a chain is irreducible, we will say that it is *recurrent* or *transient* if any one, and therefore all, of its states is.

3.1.2. Criteria for Recurrence and Transience: There are many tests which can help determine whether a state is recurrent, but no one of them works in all circumstances. In this subsection, we will develop a few of the most common of these tests. Throughout, we will use $\mathbf{u}$ to denote the column vector determined by a function $u : \mathbb{S} \longrightarrow \mathbb{R}$.

We begin with a criterion for transience.

3.1.5 Theorem. *If u is a non-negative function on $\mathbb{S}$ with the property that $(\mathbf{Pu})_i \leq (\mathbf{u})_i$ for all $i \in \mathbb{S}$, then $(\mathbf{Pu})_j < (\mathbf{u})_j$ for some $j \in \mathbb{S}$ implies j is transient.*

PROOF: Set $\mathbf{f} = \mathbf{u} - \mathbf{P}\mathbf{u}$, and note that, for all $n \geq 1$,

$$u(j) \geq (\mathbf{u})_j - (\mathbf{P}^n\mathbf{u})_j = \sum_{m=0}^{n-1} \bigl((\mathbf{P}^m\mathbf{u})_j - (\mathbf{P}^{m+1}\mathbf{u})_j\bigr)$$
$$= \sum_{m=0}^{n-1} (\mathbf{P}^m\mathbf{f})_j \geq (\mathbf{f})_j \sum_{m=0}^{n-1} (\mathbf{P}^m)_{jj}.$$

Thus $\mathbb{E}[T_j | X_0 = j] = \sum_{m=0}^{\infty} (\mathbf{P}^m)_{jj} \leq \frac{u(j)}{(\mathbf{f})_j} < \infty$, which, by (2.3.7), means that j is transient. □

In order to prove our next criterion, we will need the following special case of a general result known as *Doob's Stopping Time Theorem.*

LEMMA 3.1.6. *Assume that $u : \mathbb{S} \longrightarrow \mathbb{R}$ is bounded below and that Γ is a non-empty subset of $\mathbb{S}$. If $(\mathbf{P}\mathbf{u})_i \leq u(i)$ for all $i \notin \Gamma$ and $\rho_\Gamma \equiv \inf\{n \geq 1 : X_n \in \Gamma\}$, then*

$$\mathbb{E}\bigl[u(X_{n \wedge \rho_\Gamma}) \,\big|\, X_0 = i\bigr] \leq u(i) \quad \text{for all } n \geq 0 \text{ and } i \in \mathbb{S}.$$

Moreover, if the inequality in the hypothesis is replaced by equality, then the inequality in the conclusion can be replaced by equality.

PROOF: Set $A_n = \{\rho_\Gamma > n\}$. Then, A_n is measurable with respect to $(X_0, \ldots, X_n)$, and so, by (2.1.1), for any i,

$$\mathbb{E}\bigl[u(X_{(n+1)\wedge\rho_\Gamma}) \,\big|\, X_0 = i\bigr] = \mathbb{E}\bigl[u(X_{n\wedge\rho_\Gamma}),\, A_n\complement \,\big|\, X_0 = i\bigr]$$
$$+ \sum_{k \notin S} \mathbb{E}\bigl[u(X_{n+1}),\, A_n \cap \{X_n = k\} \,\big|\, X_0 = i\bigr]$$
$$= \mathbb{E}\bigl[u(X_{n\wedge\rho_\Gamma}),\, A_n\complement \,\big|\, X_0 = i\bigr] + \sum_{k \notin S} \mathbb{E}\bigl[(\mathbf{P}\mathbf{u})_k,\, A_n \cap \{X_n = k\} \,\big|\, X_0 = i\bigr]$$
$$\leq \mathbb{E}\bigl[u(X_{n\wedge\rho_\Gamma}),\, A_n\complement \,\big|\, X_0 = i\bigr] + \mathbb{E}\bigl[u(X_{n\wedge\rho_\Gamma}),\, A_n \,\big|\, X_0 = i\bigr]$$
$$= \mathbb{E}\bigl[u(X_{n\wedge\rho_\Gamma}) \,\big|\, X_0 = i\bigr].$$

Clearly, the same argument works just as well in the case of equality. □

THEOREM 3.1.7. *Assume that j is recurrent, and set $C = \{i : i \leftrightarrow j\}$. If $u : \mathbb{S} \longrightarrow [0, \infty)$ is a bounded function and either $u(i) = (\mathbf{P}\mathbf{u})_i$ or $u(j) \geq u(i) \geq (\mathbf{P}\mathbf{u})_i$ for all $i \in C \setminus \{j\}$, then u is constant on C. On the other hand, if j is transient, then the function u given by*

$$u(i) = \begin{cases} 1 & \text{if } i = j \\ \mathbb{P}(\rho_j < \infty | X_0 = i) & \text{if } i \neq j \end{cases}$$

is a bounded, non-constant solution to $u(i) = (\mathbf{P}\mathbf{u})_i$ for all $i \neq j$.

PROOF: In proving the first part, we will assume, without loss in generality, that $C = \mathbb{S}$. Now suppose that j is recurrent and that $u(i) = (\mathbf{P}\mathbf{u})_i$ for $i \neq j$. By applying Lemma 3.1.6 with $\Gamma = \{j\}$, we see that, for $i \neq j$,

$$u(i) = u(j)\mathbb{P}\big(\rho_j \leq n \,\big|\, X_0 = i\big) + \mathbb{E}\big[u(X_n),\, \rho_j > n \,\big|\, X_0 = i\big].$$

Hence, since, by Theorem 3.1.2, $\mathbb{P}(\rho_j < \infty | X_0 = i) = 1$ and u is bounded, we get $u(i) = u(j)$ after letting $n \to \infty$. Next assume that $u(j) \geq u(i) \geq (\mathbf{P}\mathbf{u})_i$ for all $i \neq j$. Then, again by Lemma 3.1.6, we have

$$u(j) \geq u(i) \geq u(j)\mathbb{P}\big(\rho_j \leq n \,\big|\, X_0 = i\big) + \mathbb{E}\big[u(X_n),\, \rho_j > n \,\big|\, X_0 = i\big],$$

which leads to the required conclusion when $n \to \infty$.

To prove the second part, let u be given by the prescription described, and begin by observing that, because j is transient,

$$1 > \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = j\big) = \mathbf{P}_{jj} + \sum_{i \neq j} \mathbf{P}_{ji} u(i) \geq \mathbf{P}_{jj} + (1 - \mathbf{P}_{jj}) \inf_{i \neq j} u(i).$$

From this, one sees first that $\mathbf{P}_{jj} < 1$ and then that $\inf_{i \neq j} u(i) < 1 = u(j)$. That is, u is bounded and non-constant. At the same time, when $i \neq j$, by conditioning on what happens at time 1, we know that

$$u(i) = \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = i\big) = \mathbf{P}_{ij} + \sum_{k \neq j} \mathbf{P}_{ik} \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = k\big) = (\mathbf{P}\mathbf{u})_i. \quad \square$$

3.1.8 THEOREM. *Let $\{B_m : m \geq 0\}$ be a non-decreasing sequence of non-empty subsets of $\mathbb{S}$ with the property that*

$$\mathbb{P}\big(\exists n \in \mathbb{N}\ X_n \notin B_m \,\big|\, X_0 = j\big) = 1 \quad \textit{for some } j \in B_0 \textit{ and all } m \geq 0.$$

If there to exists a non-negative function u satisfying $(\mathbf{P}\mathbf{u})_i \leq u(i)$, $i \neq j$ and $a_m \equiv \inf_{i \notin B_m} u_i \longrightarrow \infty$ as $m \to \infty$, then j is recurrent.

PROOF: For each $m \geq 0$, set $\Gamma_m = \{j\} \cup B_m\complement$, and take $\rho_{\Gamma_m} \equiv \inf\{n \geq 1 : X_n \in \Gamma_m\} = \rho_j \wedge \tau_m$, where $\tau_m \equiv \inf\{n \geq 1 : X_n \notin B_m\}$. By Lemma 3.1.6,

$$u(j) \geq \mathbb{E}\big[u(X_{n \wedge \rho_{\Gamma_m}}) \,\big|\, X_0 = j\big] \geq a_m \mathbb{P}\big(\tau_m \leq n \wedge \rho_j \,\big|\, X_0 = j\big)$$

for all $n \geq 0$. Hence, because $\mathbb{P}(\tau_m < \infty | X_0 = j) = 1$, we conclude, after letting $n \to \infty$, that $u(j) \geq a_m \mathbb{P}(\tau_m \leq \rho_j | X_0 = j)$ for all $m \geq 0$, and therefore $\lim_{m \to \infty} \mathbb{P}(\tau_m \leq \rho_j | X_0 = j) = 0$. But this means that

$$\begin{aligned} \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = j\big) &\geq \mathbb{P}\big(\rho_j < \tau_m \,\big|\, X_0 = j\big) \\ &= 1 - \mathbb{P}\big(\tau_m \leq \rho_j \,\big|\, X_0 = j\big) \nearrow 1, \end{aligned}$$

and so j is recurrent. $\square$

COROLLARY 3.1.9. *Assume that* $\mathbf{P}$ *is irreducible, and let* $\{F_m : m \geq 0\}$ *be a non-decreasing sequence of non-empty, finite subsets of the state space. If* $j \in F_0$ *and there exists* u *a non-negative function which satisfies* $(\mathbf{Pu})_i \leq u(i)$, $i \neq j$ *and* $\inf_{i \notin F_m} u_i \longrightarrow \infty$, *then* j *is recurrent.*

PROOF: In view of Theorem 3.1.8, it suffices for us to check that $\mathbb{P}(\exists n \in \mathbb{N}\ X_n \notin F_m | X_0 = j) = 1$ for all $m \geq 0$. To this end, let $\tau_m = \inf\{n \geq 1 : X_n \notin F_m\}$. By irreducibility, $\mathbb{P}(\tau_m < \infty | X_0 = i) > 0$ for all m and i. Hence, because F_m is finite, for each m there exists a $\theta_m \in (0,1)$ and $N_m \geq 1$ such that $\mathbb{P}(\tau_m > N_m | X_0 = i) \leq \theta_m$ for all $i \in F_m$. But this means that $\mathbb{P}\big(\tau_m > (\ell+1)N_m \,\big|\, X_0 = j\big)$ equals

$$\begin{aligned}
&\sum_{i \in F_m} \mathbb{P}\big(\tau_m > (\ell+1)N_m \ \&\ X_{\ell N_m} = i \,\big|\, X_0 = j\big) \\
&= \sum_{i \in F_m} \mathbb{P}\big(\tau_m > N_m \,\big|\, X_0 = i\big)\mathbb{P}\big(\tau_m > \ell N_m \ \&\ X_{\ell N_m} = i \,\big|\, X_0 = j\big) \\
&\leq \theta_m \mathbb{P}\big(\tau_m > \ell N_m \,\big|\, X_0 = j\big).
\end{aligned}$$

Thus, $\mathbb{P}(\tau_m > \ell N_m | X_0 = j) \leq \theta_m^\ell$, and so $\mathbb{P}(\tau_m = \infty | X_0 = j) = 0$. □

Remark: The preceding criteria are examples, of which there are many others, which relate recurrence of $j \in \mathbb{S}$ to the existence or non-existence of certain types of functions which satisfy either $\mathbf{Pu} = u$ or $\mathbf{Pu} \leq u$ on $\mathbb{S} \setminus \{j\}$. All these criteria can be understood as mathematical implementations of the intuitive idea that

$$u(i) = (\mathbf{Pu})_i \text{ or } (\mathbf{Pu})_i \leq u(i) \text{ for } i \neq j,$$

implies that as long as $X_n \neq j$, $u(X_n)$ will be "nearly constant" or "nearly non-increasing" as n increases. The sense in which these "nearly's" should be interpreted is the subject of *martingale theory*, and our proofs of these criteria would have been simplified had we been able to call on martingale theory.

3.1.3. Periodicity: Periodicity is another important communicating class property. In order to describe this property, we must recall Euclid's concept of the *greatest common divisor* $\gcd(S)$ of a non-empty subset $S \subseteq \mathbb{Z}$. Namely, we say that $d \in \mathbb{Z}^+$ is a common divisor of S and write $d|S$ if $\frac{s}{d} \in \mathbb{Z}$ for every $s \in S$. Clearly, if $S = \{0\}$, then $d|S$ for every $d \in \mathbb{Z}^+$, and so we take $\gcd(S) = \infty$. On the other hand, if $S \neq \{0\}$, then no common divisor of S can be larger than $\min\big\{|s| : s \in S \setminus \{0\}\big\}$, and so we know that $\gcd(S) < \infty$.

Our interest in this concept comes from the role it plays in the ergodic theory of Markov chains. Namely, as we will see below, it allows us to distinguish between the chains for which powers of the transition probability matrix converge and those for which it is necessary to take averages. More precisely, given a state i, set

$$S(i) = \{n \geq 0 : (\mathbf{P}^n)_{ii} > 0\} \quad \text{and} \quad d(i) = \gcd\big(S(i)\big). \tag{3.1.10}$$

Then $d(i)$ is called the *period* of the state i, and, i is said to be *aperiodic* if $d(i) = 1$.

As we will see, averaging is required unless i is aperiodic. However, before we get into this connection with ergodic theory, we need to take care of a few mundane matters. In the first place, the period is a communicating class property:

$$i \leftrightarrow j \implies d(i) = d(j). \tag{3.1.11}$$

To see this, assume that $(\mathbf{P}^m)_{ij} > 0$ and $(\mathbf{P}^n)_{ji} > 0$, and let d be a common divisor of $S(i)$. Then for any $k \in S(j)$, $(\mathbf{P}^{m+k+n})_{ii} \geq (\mathbf{P}^m)_{ij}(\mathbf{P}^k)_{jj}(\mathbf{P}^n)_{ji} > 0$, and so $m+k+n \in S(i)$. Hence $d|\{m+k+n : k \in S(j)\}$. But, because $m+n \in S(i)$, and therefore $d|(m+n)$, this is possible only if d divides $S(j)$, and so we now know that $d(i) \leq d(j)$. After reversing the roles of i and j, one sees that $d(j) \leq d(i)$, which means that $d(i)$ must equal $d(j)$.

We next need the following elementary fact from number theory.

3.1.12 THEOREM. *Given $\emptyset \neq S \subseteq \mathbb{Z}$ with $S \neq \{0\}$, $\gcd(S) \leq \min\{|s| : s \in S \setminus \{0\}\}$ and equality holds if and only if $\{\gcd(S), -\gcd(S)\} \cap S \neq \emptyset$. More generally, there always exists an $M \in \mathbb{Z}^+$, $\{a_m\}_1^M \subseteq \mathbb{Z}$, and $\{s_m\}_1^M \subseteq S$ such that $\gcd(S) = \sum_1^M a_m s_m$. Finally, if $S \subseteq \mathbb{N}$ and $(s_1, s_2) \in S^2 \implies s_1 + s_2 \in S$, then there exists an $M \in \mathbb{Z}^+$ such that*

$$\{m\gcd(S) : m \geq M\} = \{s \in S : s \geq M\gcd(S)\}.$$

PROOF: The first assertion needs no comment. To prove the second assertion, let $\hat{S}$ be the smallest subset of $\mathbb{Z}$ which contains S and has the property that $(s_1, s_2) \in \hat{S}^2 \implies s_1 \pm s_2 \in \hat{S}$. As is easy to check, $\hat{S}$ coincides with the subset of $\mathbb{Z}$ whose elements can be expressed in the form $\sum_1^M a_m s_m$ for some $M \geq 1$, $\{a_m\}_1^M \subseteq \mathbb{Z}$, and $\{s_m\}_1^M \subseteq S$. In particular, this means that $\gcd(S)|\hat{S}$, and so $\gcd(S) \leq \gcd(\hat{S})$. On the other hand, because $S \subseteq \hat{S}$, $\gcd(\hat{S})|S$. Hence, $\gcd(S) = \gcd(\hat{S})$, and so, by the first part, we will be done once we show that $\gcd(\hat{S}) \in \hat{S}$. To this end, let $m = \min\{s \in \mathbb{Z}^+ : s \in \hat{S}\}$. We already know that $\gcd(\hat{S}) \leq m$. Thus, to prove the equality, we need only check that $m|\hat{S}$. But, by the Euclidean algorithm, for any $s \in \hat{S}$, we can write $s = am + r$ for some $(a, r) \in \mathbb{Z}^2$ with $0 \leq r < m$. In particular, $r = s - am \in \hat{S}$. Hence, if $r \neq 0$, then r would contradict the condition that m is the smallest positive element of $\hat{S}$.

To prove the final assertion, first note that it suffices to prove that there is an $M \in \mathbb{Z}^+$ such that $\{m\gcd(S) : m \geq M\} \subseteq S$. To this end, begin by checking that, under the stated hypothesis, $\hat{S} = \{s_2 - s_1 : (s_1, s_2) \in S \cup \{0\}\}$. Thus, $\gcd(S) = s_2 - s_1$ for some $s_1 \in S \cup \{0\}$ and $s_2 \in S \setminus \{0\}$. If $s_1 = 0$, then $m\gcd(S) = ms_2 \in S$ for all $m \in \mathbb{Z}^+$, and so we can take $M = 1$. If $s_1 \neq 0$, choose $a \in \mathbb{Z}^+$ so that $s_1 = a\gcd(S)$. Then, for any $(m, r) \in \mathbb{N}^2$ with $0 \leq r < a$,

$$(a^2 + ma + r)\gcd(S) = ms_1 + rs_2 + (a - r)s_1 = (m + a - r)s_1 + rs_2 \in S.$$

Hence, after another application of the Euclidean algorithm, we see that we can take $M = a^2$. □

As an immediate consequence of Theorem 3.1.12, we see that

$$(3.1.13) \quad d(i) < \infty \implies (\mathbf{P}^{nd(i)})_{ii} > 0 \quad \text{for all sufficiently large } n \in \mathbb{Z}^+.$$

In particular,[1]

$$(3.1.14) \quad i \text{ is aperiodic} \iff (\mathbf{P}^n)_{ii} > 0 \text{ for all sufficiently large } n \in \mathbb{Z}^+.$$

We close this subsection with an application of these considerations to the ergodic theory of Markov chains on a finite state space.

3.1.15 COROLLARY. *Suppose that* $\mathbf{P}$ *is an transition probability matrix on a finite state space* $\mathbb{S}$. *If there is an aperiodic state* $j_0 \in \mathbb{S}$ *such that* $i \to j_0$ *for every* $i \in \mathbb{S}$, *then there exists an* $M \in \mathbb{Z}^+$ *and an* $\epsilon > 0$ *such that* $(\mathbf{P}^M)_{ij_0} \geq \epsilon$ *for all* $i \in \mathbb{S}$. *In particular, (cf. (2.3.9))*

$$\left\|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}\right\|_{\mathrm{v}} \leq 2(1-\epsilon)^{[\frac{n}{M}]} \quad \textit{for all } n \in \mathbb{Z}^+ \textit{ and initial distributions } \boldsymbol{\mu}.$$

PROOF: Because j_0 is aperiodic, we know that there is an $M_0 \in \mathbb{N}$ such that $(\mathbf{P}^n)_{j_0 j_0} > 0$ for all $n \geq M_0$. Further, because $i \to j_0$, there exists an $m(i) \in \mathbb{Z}^+$ such that $(\mathbf{P}^{m(i)})_{ij_0} > 0$. Hence, $(\mathbf{P}^n)_{ij_0} > 0$ for all $n \geq m(i) + M_0$. Finally, take $M = M_0 + \max_{i \in \mathbb{S}} m(i)$, $\epsilon = \min_{i \in \mathbb{S}} (\mathbf{P}^M)_{ij_0}$, and apply (2.2.3). □

3.2 Ergodic Theory without Doeblin

In this section, we will see to what extent the results obtained in Chapter 2 for Markov chains which satisfy Doeblin's condition can be reproduced without Doeblin's condition. The progression which we will adopt here runs in the opposite direction from that in Chapter 2. That is, here we will start with the most general but weakest form of the result and afterwards will see what can be done to refine it.

3.2.1. Convergence of Matrices: Because we will be looking at power series involving matrices which may have infinitely many entries, it will be important for us to be precise about what is the class of matrices with which we are dealing and in what sense our series are converging. For our purposes, the most natural class will be of matrices $\mathbf{M}$ for which

$$(3.2.1) \qquad \|\mathbf{M}\|_{\mathrm{u,v}} \equiv \sup_{i \in \mathbb{S}} \sum_{j \in \mathbb{S}} |(\mathbf{M})_{ij}|$$

is finite, and the set of all such matrices will be denoted by $M_{\mathrm{u,v}}(\mathbb{S})$. An easy calculation shows that $M_{\mathrm{u,v}}(\mathbb{S})$ is a vector space over $\mathbb{R}$ and that $\|\cdot\|_{\mathrm{u,v}}$ is a good norm on $M_{\mathrm{u},v}(\mathbb{S})$. That is,

$$\|\mathbf{M}\|_{\mathrm{u,v}} = 0 \quad \text{if and only if} \quad \mathbf{M} = \mathbf{0},$$

[1] The "if" part of the following statement depends on the existence of infinitely many prime numbers.

$$\|\alpha\mathbf{M}\|_{\mathrm{u,v}} = |\alpha|\|\mathbf{M}\|_{\mathrm{u},v} \quad \text{for } \alpha \in \mathbb{R},$$

and

$$\|\mathbf{M} + \mathbf{M}'\|_{\mathrm{u,v}} \leq \|\mathbf{M}\|_{\mathrm{u,v}} + \|\mathbf{M}'\|_{\mathrm{u,v}}.$$

Slightly less obvious is the fact that

$$(\mathbf{M}, \mathbf{M}') \in M_{\mathrm{u,v}}(\mathbb{S})^2 \implies \begin{array}{l} \mathbf{MM}' \text{ exists and} \\ \|\mathbf{MM}'\|_{\mathrm{u,v}} \leq \|\mathbf{M}\|_{\mathrm{u,v}}\|\mathbf{M}'\|_{\mathrm{u,v}}. \end{array} \tag{3.2.2}$$

To see this, observe first that, since

$$\sum_k |(\mathbf{M})_{ik}||(\mathbf{M}')_{kj}| \leq \left(\sum_k |(\mathbf{M})_{ik}|\right) \sup_k |(\mathbf{M}')_{kj}| < \infty,$$

the sum in

$$(\mathbf{MM}')_{ij} = \sum_k (\mathbf{M})_{ik}(\mathbf{M}')_{kj}$$

is absolutely convergent. In addition, for each i,

$$\begin{aligned} \sum_j |(\mathbf{MM}')_{ij}| &\leq \sum_j \sum_k |(\mathbf{M})_{ik}||(\mathbf{M}')_{kj}| \\ &= \sum_k |(\mathbf{M})_{ik}| \left(\sum_j |(\mathbf{M}')_{kj}|\right) \leq \left(\sum_k |(\mathbf{M})_{ik}|\right) \|\mathbf{M}'\|_{\mathrm{u,v}}, \end{aligned}$$

and so the inequality in (3.2.2) follows.

We next want to show that the metric which $\|\cdot\|_{\mathrm{u,v}}$ determines on $M_{\mathrm{u,v}}(\mathbb{S})$ is complete. In order to do this, it will be useful to know that if $\{\mathbf{M}_n\}_0^\infty \subseteq M_{\mathrm{u,v}}(\mathbb{S})$, then

$$\begin{aligned} M_{ij} = \lim_{n\to\infty} (\mathbf{M}_n)_{ij} \text{ for each } (i,j) \\ \implies \sup_i \sum_j |M_{ij}| \leq \varliminf_{n\to\infty} \|\mathbf{M}_n\|_{\mathrm{u,v}}. \end{aligned} \tag{3.2.3}$$

Indeed, by Fatou's Lemma, Theorem 6.1.10,

$$\sum_{j\in\mathbb{S}} |M_{ij}| \leq \varliminf_{n\to\infty} \sum_{j\in\mathbb{S}} |(\mathbf{M}_n)_{ij}| \quad \text{for each } i \in \mathbb{S},$$

and so (3.2.3) is proved.

Knowing (3.2.3), the proof of completeness goes as follows. Assume that $\{\mathbf{M}_n\}_0^\infty \subseteq M_{\mathrm{u,v}}(\mathbb{S})$ is $\|\cdot\|_{\mathrm{u,v}}$-Cauchy convergent: $\lim_{m\to\infty} \sup_{n>m} \|\mathbf{M}_n - \mathbf{M}_m\|_{\mathrm{u,v}} = 0$. Obviously, for each $(i,j) \in \mathbb{S}^2$, $\{(\mathbf{M}_n)_{ij}\}_0^\infty$ is Cauchy convergent as a sequence in $\mathbb{R}$. Hence, there is a matrix $\mathbf{M}$ such that, for each (i,j), $(\mathbf{M})_{ij} = \lim_{n\to\infty}(\mathbf{M}_n)_{ij}$. Furthermore, by (3.2.3),

$$\|\mathbf{M} - \mathbf{M}_m\|_{\mathrm{u,v}} \leq \varliminf_{n\to\infty} \|\mathbf{M}_n - \mathbf{M}_m\|_{\mathrm{u,v}} \leq \sup_{n>m} \|\mathbf{M}_n - \mathbf{M}_m\|_{\mathrm{u,v}}.$$

Hence, $\|\mathbf{M} - \mathbf{M}_m\|_{\mathrm{u,v}} \longrightarrow 0$.

3.2.2. Abel Convergence: As we said in the introduction, we will begin with the weakest form of the convergence results at which we are aiming. That is, rather than attempting to prove the convergence of $(\mathbf{P}^n)_{ij}$ or even the Césaro means $\frac{1}{n}\sum_{m=0}^{n-1}(\mathbf{P}^m)_{ij}$ as $n \to \infty$, we will begin by studying the Abel sums $(1-s)\sum_{m=0}^{\infty} s^m(\mathbf{P}^m)_{ij}$ as $s \nearrow 1$.

We will say that a bounded sequence $\{x_n\}_0^\infty \subseteq \mathbb{R}$ is *Abel convergent* to x if

$$\lim_{s\nearrow 1}(1-s)\sum_{n=1}^{\infty} s^n x_n = x.$$

It should be clear that Abel convergence is weaker than (i.e., is implied by) ordinary convergence.[2] Indeed, if $x_n \longrightarrow x$, then, since $(1-s)\sum_0^\infty s^n = 1$, for any N:

$$\left|x - (1-s)\sum_{n=0}^{\infty} s^n x_n\right| = (1-s)\left|\sum_{n=0}^{\infty} s^n(x - x_n)\right|$$

$$\leq (1-s)\sum_{n=0}^{\infty} s^n|x - x_n| \leq N(1-s)\sup_{n\in\mathbb{N}}|x - x_n| + \sup_{n\geq N}|x - x_n|.$$

Hence,

$$\overline{\lim_{s\nearrow 1}}\left|x - (1-s)\sum_{1}^{\infty} s^n x_n\right| \leq \lim_{N\to\infty}\sup_{n\geq N}|x - x_n| = 0$$

if $x_n \longrightarrow x$. On the other hand, although $\{(-1)^n\}_1^\infty$ fails to converge to anything,

$$(1-s)\sum_{n=0}^{\infty} s^n(-1)^n = (1-s)\frac{1}{1+s} \longrightarrow 0 \quad \text{as } s \nearrow 1.$$

That is, Abel convergence does not, in general imply ordinary convergence.

With the preceding in mind, we set

$$\mathbf{R}(s) = (1-s)\sum_{n=0}^{\infty} s^n\mathbf{P}^n \quad \text{for } s \in [0,1). \tag{3.2.4}$$

However, before accepting this definition, it is necessary to check that the above series converges. To this end, first note that, because $\mathbf{P}^n$ is a transition probability matrix for each $n \geq 0$, $\|\mathbf{P}^n\|_{\mathrm{u,v}} = 1$. Hence, for $0 \leq m < n$,

$$\left\|(1-s)\sum_{\ell=0}^{n} s^\ell\mathbf{P}^\ell - (1-s)\sum_{\ell=0}^{m} s^\ell\mathbf{P}^\ell\right\|_{\mathrm{u,v}} \leq (1-s)\sum_{\ell=m}^{n} s^\ell\|\mathbf{P}^\ell\|_{\mathrm{u,v}} \leq s^m,$$

[2] In Exercise 3.3.1 below, it is shown that Abel convergence is also weaker than Césaro convergence.

and so, by the completeness proved above, the series in (3.2.4) converges with respect to the $\|\cdot\|_{u,v}$-norm.

Our goal here is to prove that

$$\lim_{s\nearrow 1}\big(\mathbf{R}(s)\big)_{ij} = \pi_{ij} \equiv \begin{cases} \big(\mathbb{E}[\rho_j | X_0 = j]\big)^{-1} & \text{if } i = j \\ \mathbb{P}(\rho_j < \infty | X_0 = i)\pi_{jj} & \text{if } i \neq j, \end{cases} \tag{3.2.5}$$

and the key to our doing so lies in the *renewal equation*

$$\begin{aligned} (\mathbf{P}^n)_{ij} &= \sum_{m=1}^{n} f(m)_{ij}(\mathbf{P}^{n-m})_{jj} \quad \text{for } n \geq 1, \\ &\text{where } f(m)_{ij} \equiv \mathbb{P}\big(\rho_j = m \,\big|\, X_0 = i\big), \end{aligned} \tag{3.2.6}$$

which is an elementary application of (2.1.13):

$$\begin{aligned} (\mathbf{P}^n)_{ij} &= \sum_{m=1}^{n} \mathbb{P}\big(X_n = j \;\&\; \rho_j = m \,\big|\, X_0 = i\big) \\ &= \sum_{m=1}^{n} \mathbb{P}\big(X_{n-m} = j \,\big|\, X_0 = j\big)\mathbb{P}\big(\rho_j = m \,\big|\, X_0 = i\big). \end{aligned}$$

Next, for $s \in [0,1)$, set

$$\hat{f}(s)_{ij} \equiv \sum_{m=1}^{\infty} s^m f(m)_{ij} = \mathbb{E}\big[s^{\rho_j} \,\big|\, X_0 = i\big],$$

and, starting from (3.2.6), conclude that

$$\begin{aligned} \big(\mathbf{R}(s)\big)_{ij} &= (1-s)\delta_{i,j} + (1-s)\sum_{n=1}^{\infty} s^n \left(\sum_{m=1}^{n} f(m)_{ij}(\mathbf{P}^{n-m})_{jj}\right) \\ &= (1-s)\delta_{i,j} + (1-s)\sum_{m=1}^{\infty} s^m f(m)_{ij}\left(\sum_{n=m}^{\infty} s^{n-m}(\mathbf{P}^{n-m})_{jj}\right) \\ &= (1-s)\delta_{i,j} + \hat{f}(s)_{ij}\big(\mathbf{R}(s)\big)_{jj}. \end{aligned}$$

That is,

$$\big(\mathbf{R}(s)\big)_{ij} = (1-s)\delta_{i,j} + \hat{f}(s)_{ij}\big(\mathbf{R}(s)\big)_{jj} \quad \text{for } s \in [0,1). \tag{3.2.7}$$

Given (3.2.7), (3.2.5) is easy. Namely,

$$\big(\mathbf{R}(s)\big)_{jj} = \frac{1-s}{1-\hat{f}(s)_{jj}}.$$

Hence, if j is transient, and therefore $\hat{f}(1)_{jj} < 1$,

$$\pi_{jj} = \lim_{s \nearrow 1} \big(\mathbf{R}(s)\big)_{jj} = 0 = \frac{1}{\mathbb{E}[\rho_j | X_0 = j]}.$$

On the other hand, if j is recurrent, then, since

$$\frac{1 - s^m}{1 - s} = \sum_{\ell=0}^{m-1} s^\ell \nearrow m,$$

the Monotone Convergence Theorem says that

$$\frac{1 - \hat{f}(s)_{jj}}{1 - s} = \sum_{m=1}^{\infty} \frac{1 - s^m}{1 - s} f(m)_{jj} \nearrow \sum_{m=1}^{\infty} m f(m)_{jj} = \mathbb{E}\big[\rho_j \,\big|\, X_0 = j\big]$$

as $s \nearrow 1$. At the same time, when $i \neq j$,

$$\big(\mathbf{R}(s)\big)_{ij} = \hat{f}(s)_{ij} \big(\mathbf{R}(s)\big)_{jj} \nearrow \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = i\big)\pi_{jj}.$$

3.2.3. Structure of Stationary Distributions: We will say that a probability vector $\boldsymbol{\mu}$ is $\mathbf{P}$-*stationary* and will write $\boldsymbol{\mu} \in \mathrm{Stat}(\mathbf{P})$ if $\boldsymbol{\mu} = \boldsymbol{\mu}\mathbf{P}$. Obviously, if $\boldsymbol{\mu}$ is stationary, then $\boldsymbol{\mu} = \boldsymbol{\mu}\mathbf{R}(s)$ for each $s \in [0, 1)$. Hence, by (3.2.5) and Lebesgue's Dominated Convergence Theorem,

$$(\boldsymbol{\mu})_j = \sum_i (\boldsymbol{\mu})_i \big(\mathbf{R}(s)\big)_{ij} \longrightarrow \sum_i (\boldsymbol{\mu})_i \pi_{ij}.$$

If j is transient, then $\pi_{ij} \equiv 0$. On the other hand, if j is recurrent, then, by Theorem 3.1.2, π_{ij} is either π_{jj} or 0, according to whether $i \leftrightarrow j$ or $i \not\leftrightarrow j$. Hence, in either case, we have that

$$\boldsymbol{\mu} \in \mathrm{Stat}(\mathbf{P}) \implies (\boldsymbol{\mu})_j = \left(\sum_{i \leftrightarrow j} (\boldsymbol{\mu})_i \right) \pi_{jj}. \tag{3.2.8}$$

We next want to show that

$$\pi_{jj} > 0 \text{ and } C = \{i : i \leftrightarrow j\} \implies \boldsymbol{\pi}^C \in \mathrm{Stat}(\mathbf{P}) \text{ when } (\boldsymbol{\pi}^C)_i \equiv \mathbf{1}_C(i)\pi_{ii}. \tag{3.2.9}$$

To do this, first note that $\pi_{jj} > 0$ only if j is recurrent. Thus, all $i \in C$ are recurrent and, for each $s \in (0, 1)$, $\big(\mathbf{R}(s)\big)_{k\ell} > 0 \iff (k, \ell) \in C^2$. In particular, $(\boldsymbol{\pi}^C)_i = \lim_{s \nearrow 1} \big(\mathbf{R}(s)\big)_{ji}$ for all i, and therefore, by Fatou's Lemma,

$$\sum_i (\boldsymbol{\pi}^C)_i \leq \varliminf_{s \nearrow 1} \sum_i \big(\mathbf{R}(s)\big)_{ji} = 1.$$

Similarly, for any i,

$$(\boldsymbol{\pi}^C\mathbf{P})_i = \sum_{k\in C} \pi_{kk}(\mathbf{P})_{ki} \leq \varliminf_{s\nearrow 1} \sum_{k\in C} \big(\mathbf{R}(s)\big)_{jk}(\mathbf{P})_{ki} = (\boldsymbol{\pi}^C)_i,$$

since

$$\sum_{k\in C} \big(\mathbf{R}(s)\big)_{jk}(\mathbf{P})_{ki} = \frac{(\mathbf{R}(s) - (1-s)\mathbf{I})_{ji}}{s} \longrightarrow \pi_{ji} = (\boldsymbol{\pi}^C)_i \quad \text{as } s \nearrow 1.$$

But if strict inequality were to hold for some i, then, by Fubini's Theorem, Theorem 6.1.15, we would have the contradiction

$$\sum_k (\boldsymbol{\pi}^C)_k = \sum_k (\boldsymbol{\pi}^C)_k \left(\sum_i (\mathbf{P})_{ki}\right) = \sum_i \left(\sum_k (\boldsymbol{\pi}^C)_k (\mathbf{P})_{ki}\right) < \sum_i (\boldsymbol{\pi}^C)_i.$$

Hence, we now know that $\boldsymbol{\pi}^C = \boldsymbol{\pi}^C\mathbf{P}$. Finally, to prove that $\boldsymbol{\pi}^C \in \mathrm{Stat}(\mathbf{P})$, we still have to check that $\sum_i (\boldsymbol{\pi}^C)_i = 1$. However, we have already checked that $\boldsymbol{\pi}^C = \boldsymbol{\pi}^C\mathbf{P}$, and so we know that $\boldsymbol{\pi}^C = \boldsymbol{\pi}^C\mathbf{R}(s)$. Therefore, since, as we already showed, $\sum_i (\boldsymbol{\pi}^C)_i \leq 1$, Lebesgue's Dominated Convergence Theorem justifies

$$0 < \pi_{jj} = \sum_i (\boldsymbol{\pi}^C)_i \big(\mathbf{R}(s)\big)_{ij} \longrightarrow \left(\sum_i (\boldsymbol{\pi}^C)_i\right) \pi_{jj},$$

which is possible only if $\sum_i (\boldsymbol{\pi}^C)_i = 1$.

Before summarizing the preceding as a theorem, we need to recall that a subset A of a linear space is said to be a *convex set* if $(1-\theta)a + \theta a' \in A$ for all $a, a' \in A$ and $\theta \in [0,1]$ and that $b \in A$ is an *extreme point* of the convex set A if $b = (1-\theta)a + \theta a'$ for some $\theta \in (0,1)$ and $a, a' \in A$ implies that $b = a' = a''$. In addition, we need the notion of positive recurrence. Namely, we say that $j \in \mathbb{S}$ is *positive recurrent* if $\mathbb{E}[\rho_j | X_0 = j] < \infty$. Obviously, only recurrent states can be positive recurrent. On the other hand, (1.1.13) together with (1.1.15) show that that there can exist *null recurrent* states, those which are recurrent but not positive recurrent.

3.2.10 Theorem. *$\mathrm{Stat}(\mathbf{P})$ is a convex subset of $\mathbb{R}^{\mathbb{S}}$. Moreover, $\mathrm{Stat}(\mathbf{P}) \neq \emptyset$ if and only if there is at least one positive recurrent state $j \in \mathbb{S}$. In fact, for any $\boldsymbol{\mu} \in \mathrm{Stat}(\mathbf{P})$, (3.2.8) holds, and $\boldsymbol{\mu}$ is an extreme point in $\mathrm{Stat}(\mathbf{P})$ if and only if there is a communicating class C of positive recurrent states for which (cf. (3.2.9)) $\boldsymbol{\mu} = \boldsymbol{\pi}^C$. In particular, $(\boldsymbol{\mu})_j = 0$ for any transient state j and, for any recurrent state j, either $(\boldsymbol{\mu})_i$ is strictly positive or it is 0 simultaneously for all states i's which communicate with j.*

Proof: The only statements not already covered are the characterization of the extreme points of $\mathrm{Stat}(\mathbf{P})$ and the final assertion in the case when j is recurrent.

In view of (3.2.8), the final assertion when j is recurrent comes down to showing that if j is positive recurrent and $i \leftrightarrow j$, then i is positive recurrent. To this end, suppose that j is positive recurrent, set $C = \{i : i \leftrightarrow j\}$, and let $i \in C$ be given. Then $\boldsymbol{\pi}^C \mathbf{P}^n = \boldsymbol{\pi}^C$ for all $n \geq 0$, and therefore, by choosing n so that $(\mathbf{P}^n)_{ji} > 0$, we see that $(\boldsymbol{\pi}^C)_i \geq (\boldsymbol{\pi}^C)_j (\mathbf{P}^n)_{ji} > 0$.

To handle the characterization of extreme points, first suppose that $\boldsymbol{\mu} \neq \boldsymbol{\pi}^C$ for any communicating class C of positive recurrent states. Then, by (3.2.8), there must exist non-communicating, positive recurrent states j and j' for which $(\boldsymbol{\mu})_j > 0 < (\boldsymbol{\mu})_{j'}$. But, again by (3.2.8), this means that $\boldsymbol{\mu} = \theta \boldsymbol{\pi}^C + (1-\theta)\boldsymbol{\nu}$, where $C = \{i : i \leftrightarrow j\}$, $\theta = \sum_{i \in C} (\boldsymbol{\mu})_i \in (0,1)$, and $(\boldsymbol{\nu})_i$ equals 0 or $(1-\theta)^{-1}(\boldsymbol{\mu})_i$ depending on whether i is or is not in C. Clearly $\boldsymbol{\nu} \in \text{Stat}(\mathbf{P})$ and, because $\boldsymbol{\nu}_{j'} > 0 = (\boldsymbol{\pi}^C)_{j'}$, $\boldsymbol{\nu} \neq \boldsymbol{\pi}^C$. Hence, $\boldsymbol{\mu}$ cannot be extreme. Equivalently, every extreme $\boldsymbol{\mu}$ is $\boldsymbol{\pi}^C$ for some communicating class C of positive recurrent states.

Conversely, given a communicating class C of positive recurrent states, suppose that $\boldsymbol{\pi}^C = (1-\theta)\boldsymbol{\mu} + \theta\boldsymbol{\nu}$ for some $\theta \in (0,1)$ and pair $(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \text{Stat}(\mathbf{P})^2$. Then $(\boldsymbol{\mu})_i = 0$ for all $i \notin C$, and so, by (3.2.8), we see first that $\boldsymbol{\mu} = \boldsymbol{\pi}^C$ and then, as a consequence, that $\boldsymbol{\nu} = \boldsymbol{\pi}^C$ as well. □

The last part of Theorem 3.2.10 shows that *positive recurrence is a communicating class property.* Indeed, if j is positive recurrent and $C = \{i : i \leftrightarrow j\}$, then $\boldsymbol{\pi}^C \in \text{Stat}(\mathbf{P})$, and so, since $(\boldsymbol{\pi}^C)_j > 0$, $\pi_{ii} = (\boldsymbol{\pi}^C)_i > 0$. In particular, if a chain is irreducible, we are justified in saying it is positive recurrent if any one of its states is. See Exercise 3.3.3 below for a criterion which guarantees positive recurrence.

3.2.4. A Small Improvement: The next step in our program will be the replacement of Abel convergence by Césaro convergence. That is, we will show that (3.2.5) can be replaced by

$$\lim_{n \to \infty} (\mathbf{A}_n)_{ij} = \pi_{ij}, \quad \text{where } \mathbf{A}_n = \frac{1}{n} \sum_{m=0}^{n-1} \mathbf{P}^m. \tag{3.2.11}$$

As is shown in Exercise 3.3.1, Césaro convergence does not, in general, follow from Abel convergence. In fact, general results which say when Abel convergence implies Césaro convergence can be quite delicate and, because the original one was proved by a man named Tauber, they are known as Tauberian theorems. Fortunately, the Tauberian theorem required here is quite straight-forward.

A key role in our proof will be played by the following easy estimate:

$$\begin{aligned} &\{a_m\}_0^\infty \subseteq [0,1] \;\&\; A_n = \frac{1}{n} \sum_0^{n-1} a_\ell \\ &\qquad \Longrightarrow |A_n - A_{n-m}| \leq \frac{m}{n} \quad \text{for } 0 \leq m < n. \end{aligned} \tag{3.2.12}$$

The proof is:

$$A_n - A_{n-m} = \frac{1}{n}\sum_{\ell=n-m}^{n-1} a_\ell - \frac{m}{n(n-m)}\sum_{\ell=0}^{n-m-1} a_\ell \begin{cases} \geq -\frac{m}{n} \\ \leq \frac{m}{n}. \end{cases}$$

3.2.13 LEMMA. *For all (i,j), $\overline{\lim}_{n\to\infty}(\mathbf{A}_n)_{ij} \leq e\pi_{jj}$. In addition, for any j and any subsequence $\{n_\ell : \ell \geq 0\} \subseteq \mathbb{N}$,*

$$\lim_{\ell\to\infty}(\mathbf{A}_{n_\ell})_{jj} = \alpha \implies \lim_{\ell\to\infty}(\mathbf{A}_{n_\ell})_{ij} = \mathbb{P}\big(\rho_j < \infty \,\big|\, X_0 = i\big)\alpha \quad \textit{for all } i.$$

PROOF: To prove the first part, observe that

$$(\mathbf{A}_n)_{ij} \leq \frac{1}{n}\big(1-\tfrac{1}{n}\big)^{-n}\sum_{m=0}^{n-1}\big(1-\tfrac{1}{n}\big)^m(\mathbf{P}^m)_{ij} \leq \big(1-\tfrac{1}{n}\big)^{-n}\big(\mathbf{R}(1-\tfrac{1}{n})\big)_{ij},$$

which, together with (1.2.10) and (3.2.5), shows that $\overline{\lim}_{n\to\infty}(\mathbf{A}_n)_{ij} \leq e\pi_{ij} \leq e\pi_{jj}$.

To handle the second part, use (3.2.6) to arrive at

$$(\mathbf{A}_n)_{ij} = \sum_{m=1}^{n-1} f(m)_{ij}\big(1-\tfrac{m}{n}\big)(\mathbf{A}_{n-m})_{jj} \quad \text{for } i \neq j.$$

Hence,

$$\begin{aligned}\big|(\mathbf{A}_n)_{ij} - \mathbb{P}\big(\rho_j < n \,\big|\, X_0 = i\big)\alpha\big| &\leq \sum_{m=1}^{n-1} f(m)_{ij}\left(\frac{m}{n} + \big|(\mathbf{A}_{n-m})_{jj} - \alpha\big|\right) \\ &\leq 2\sum_{m=1}^{n-1}\frac{m}{n}f(m)_{ij} + \big|(\mathbf{A}_n)_{jj} - \alpha\big|,\end{aligned}$$

where, in the second inequality, we have used (3.2.12) plus $\sum_{m=1}^{\infty} f(m)_{ij} \leq 1$. Finally, by Lebesgue's Dominated Convergence Theorem, $\sum_0^{n-1}\frac{m}{n}f(m)_{ij}$ tends to 0 as $n \to \infty$, and therefore, by applying the above with $n = n_\ell$ and letting $\ell \to \infty$, we get the desired conclusion. □

We can now complete the proof of (3.2.11). Namely, if $\pi_{jj} = 0$, then the first part of Lemma 3.2.13 guarantees that $\lim_{n\to\infty}(\mathbf{A}_n)_{ij} = 0 = \pi_{ij}$ for all i. Thus, assume that $\pi_{jj} > 0$. In this case Theorem 3.2.10 says that j must be positive recurrent and $\boldsymbol{\pi}^C \in \text{Stat}(\mathbf{P})$ when $C \equiv \{i : i \leftrightarrow j\}$. In particular, $\pi_{jj} = \sum_{i\in C}(\boldsymbol{\pi}^C)_i(\mathbf{A}_n)_{ij}$. At the same time, if $\alpha^+ = \overline{\lim}_{n\to\infty}(\mathbf{A}_n)_{jj}$ and the subsequence $\{n_\ell : \ell \geq 0\}$ is chosen so that $(\mathbf{A}_{n_\ell})_{jj} \longrightarrow \alpha^+$, then, by the second part of Lemma 3.2.13 and Corollary 3.1.4,

$$i \in C \implies \lim_{\ell\to\infty}(\mathbf{A}_{n_\ell})_{ij} = \alpha^+.$$

Hence, after putting these two remarks together, we arrive at

$$\pi_{jj} = \lim_{\ell\to\infty} \sum_{i\in C} (\boldsymbol{\pi}^C)_i (\mathbf{A}_{n_\ell})_{ij} = \alpha^+ \sum_{i\in C} (\boldsymbol{\pi}^C)_i = \alpha^+.$$

Similarly, if $\alpha^- = \underline{\lim}_{n\to\infty} (\mathbf{A}_n)_{jj}$, we can show that $\alpha^- = \pi_{jj}$, and so we now know that $\pi_{jj} > 0 \implies \lim_{n\to\infty} (\mathbf{A}_n)_{jj} = \pi_{jj}$, which, after another application of the second part of Lemma 3.2.13, means that we have proved (3.2.11).

3.2.5. The Mean Ergodic Theorem Again: Just as we were able to use Theorem 2.2.5 in §2.3.1 to prove Theorem 2.3.4, so here we can use (3.2.11) to prove the following version of the mean ergodic theorem.

3.2.14 THEOREM. *Let C a communicating class of positive recurrent states. If $\mathbb{P}(X_0 \in C) = 1$, then*

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2\right] = 0.$$

See Exercises 3.3.9 and 3.3.11 below for a more refined statement.

PROOF: Since $\mathbb{P}(X_m \in C \text{ for all } m \in \mathbb{N}) = 1$, without loss in generality we may and will assume that C is the whole state space. In keeping with this assumption, we will set $\boldsymbol{\pi} = \boldsymbol{\pi}^C$.

Next note that if $\mu_i = \mathbb{P}(X_0 = i)$, then

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2\right]$$
$$= \sum_{i\in\mathbb{S}} \mu_i \mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2 \,\middle|\, X_0 = i\right],$$

and so it suffices to prove that

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2 \,\middle|\, X_0 = i\right] = 0 \quad \text{for each } i \in \mathbb{S}.$$

But, because $\pi_{ii} > 0$ for all $i \in \mathbb{S}$ and

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2 \,\middle|\, X_0 = i\right]$$
$$\le \frac{1}{\pi_{ii}} \sum_{k\in\mathbb{S}} \pi_{kk} \mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2 \,\middle|\, X_0 = k\right],$$

it is enough to prove the result when π is the initial distribution of the Markov chain. Hence, from now on, we will be making this assumption along with $C = \mathbb{S}$.

Now let $\mathbf{f}$ the column vector whose ith component is $\mathbf{1}_{\{j\}}(i) - \pi_{jj}$. Then, just as in the proof of Theorem 2.3.4,

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1}\mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2\right] \le \frac{2}{n^2}\sum_{k=0}^{n-1}(n-k)\mathbb{E}\big[(\mathbf{f})_{X_k}(\mathbf{A}_{n-k}\mathbf{f})_{X_k}\big].$$

Since $\pi \in \mathrm{Stat}(\mathbf{P})$,

$$\mathbb{E}\big[(\mathbf{f})_{X_k}(\mathbf{A}_{n-k}\mathbf{f})_{X_k}\big] = \pi\big(f\mathbf{A}_{n-k}\mathbf{f}\big),$$

and therefore the preceding becomes

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1}\mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2\right] \le \frac{2}{n^2}\sum_{m=1}^{n} m\pi\big(f\mathbf{A}_m\mathbf{f}\big),$$

where $(f\mathbf{A}_m\mathbf{f})_i \equiv (\mathbf{f})_i(\mathbf{A}_m\mathbf{f})_i$.

Finally, by (3.2.11), for each $\epsilon > 0$ there exists an $N_\epsilon \in \mathbb{Z}^+$ such that

$$|\pi(f\mathbf{A}_n\mathbf{f})| \le \sum_i (\pi)_i|(\mathbf{A}_n)_{ij} - \pi_{jj}| < \epsilon \quad \text{for all } n \ge N_\epsilon.$$

Hence, we find that

$$\overline{\lim_{n\to\infty}}\ \mathbb{E}\left[\left(\frac{1}{n}\sum_{m=0}^{n-1}\mathbf{1}_{\{j\}}(X_m) - \pi_{jj}\right)^2\right]$$
$$\le \overline{\lim_{n\to\infty}}\frac{2}{n^2}\sum_{m=1}^{N_\epsilon} m\big|\pi\big(f\mathbf{A}_m\mathbf{f}\big)\big| + \overline{\lim_{n\to\infty}}\frac{2\epsilon}{n^2}\sum_{m=N_\epsilon+1}^{n} m \le \epsilon. \quad \square$$

3.2.6. A Refinement in The Aperiodic Case: Our goal in this subsection is to prove that (cf. (3.2.5))

(3.2.15) If j is transient or aperiodic, then $\lim_{n\to\infty}(\mathbf{P}^n)_{ij} = \pi_{ij}$ for all $i \in \mathbb{S}$.

Of course, the case when j is transient requires very little effort. Namely, by (2.3.7), we have that

$$j \text{ transient} \implies \sum_{n=0}^{\infty}(\mathbf{P}^n)_{ij} \le \mathbb{E}[T_j|X_0 = j] < \infty,$$

and therefore that $\lim_{n\to\infty}(\mathbf{P}^n)_{ij} = 0$. At the same time, because $\mathbb{P}(\rho_j = \infty | X_0 = j) > 0$ when j is transient, $\pi_{ij} \leq \pi_{jj} = 0$. Thus, from now on, we will concentrate on the case when j is recurrent and aperiodic.

Our first step is the observation that (cf. §2.3.2)

$$\text{if } j \text{ is aperiodic, then there exists an } N \in \mathbb{Z}^+ \text{ such that} \quad \max_{1\leq m\leq n} \mathbb{P}\big(\rho_j^{(m)} = n \,\big|\, X_0 = j\big) > 0 \quad \text{for all } n \geq N. \tag{3.2.16}$$

To check this, use (3.1.14) to produce an $N \in \mathbb{Z}^+$ such that $(\mathbf{P}^n)_{jj} > 0$ for all $n \geq N$. Then, since

$$(\mathbf{P}^n)_{jj} = \sum_{m=1}^{n} \mathbb{P}\big(\rho_j^{(m)} = n \,\big|\, X_0 = j\big) \quad \text{for } n \geq 1,$$

(3.2.16) is clear.

The second, and key, step is contained in the following lemma.

3.2.17 LEMMA. *Assume that j is aperiodic and recurrent, and set $\alpha_j^- = \underline{\lim}_{n\to\infty}(\mathbf{P}^n)_{jj}$ and $\alpha_j^+ = \overline{\lim}_{n\to\infty}(\mathbf{P}^n)_{jj}$. Then there exist subsequences $\{n_\ell^- : \ell \geq 1\}$ and $\{n_\ell^+ : \ell \geq 1\}$ such that*

$$\alpha_j^\pm = \lim_{\ell\to\infty} (\mathbf{P}^{n_\ell^\pm - r})_{jj} \quad \textit{for all } r \geq 0.$$

PROOF: Choose a subsequence $\{n_\ell : \ell \geq 1\}$ so that $(\mathbf{P}^{n_\ell})_{jj} \longrightarrow \alpha_j^+$, and, using (3.2.16), choose $N \geq 1$ so that $\max_{1\leq m\leq n} \mathbb{P}(\rho_j^{(m)} = n | X_0 = j) > 0$ for all $n \geq N$. Given $r \geq N$, choose $1 \leq m \leq r$ so that $\delta \equiv \mathbb{P}(\rho_j^{(m)} = r | X_0 = j) > 0$. Now for any $M \in \mathbb{Z}^+$, observe that, when $n_\ell \geq M + r$, $(\mathbf{P}^{n_\ell})_{jj}$ is equal to

$$\begin{aligned}
&\mathbb{P}\big(X_{n_\ell} = j \;\&\; \rho_j^{(m)} = r \,\big|\, X_0 = j\big) + \mathbb{P}\big(X_{n_\ell} = j \;\&\; \rho_j^{(m)} \neq r \,\big|\, X_0 = j\big) \\
&\quad = \delta(\mathbf{P}^{n_\ell - r})_{jj} + \mathbb{P}\big(X_{n_\ell} = j \;\&\; n_\ell - M \geq \rho_j^{(m)} \neq r \,\big|\, X_0 = j\big) \\
&\qquad\qquad + \mathbb{P}\big(X_{n_\ell} = j \;\&\; \rho_j^{(m)} > n_\ell - M \,\big|\, X_0 = j\big).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{P}\big(X_{n_\ell} = j \;\&\; n_\ell - M \geq \rho_j^{(m)} \neq r \,\big|\, X_0 = j\big) \\
&\quad = \sum_{\substack{k=1 \\ k\neq r}}^{n_\ell - M} \mathbb{P}(\rho_j^{(m)} = k | X_0 = j)(\mathbf{P}^{n_\ell - k})_{jj} \leq (1 - \delta) \sup_{n\geq M} (\mathbf{P}^n)_{jj},
\end{aligned}$$

while

$$\mathbb{P}\big(X_{n_\ell} = j \;\&\; \rho_j^{(m)} > n_\ell - M \,\big|\, X_0 = j\big) \leq \mathbb{P}(\rho_j^{(m)} > n_\ell - M | X_0 = j).$$

Hence, since j is recurrent and therefore $\mathbb{P}(\rho_j^{(m)} < \infty | X_0 = j) = 1$, we get

$$\alpha_j^+ \le \delta \varliminf_{\ell\to\infty} (\mathbf{P}^{n_\ell - r})_{jj} + (1-\delta) \sup_{n\ge M} (\mathbf{P}^n)_{jj}$$

after letting $\ell \to \infty$. Since this is true for all $M \ge 1$, it leads to

$$\alpha_j^+ \le \delta \varliminf_{\ell\to\infty} (\mathbf{P}^{n_\ell - r})_{jj} + (1-\delta)\alpha_j^+,$$

which implies $\varliminf_{\ell\to\infty}(\mathbf{P}^{n_\ell - r})_{jj} \ge \alpha_{jj}^+$. But obviously $\varlimsup_{\ell\to\infty}(\mathbf{P}^{n_\ell - r})_{jj} \le \alpha_j^+$, and so we have shown that $\lim_{\ell\to\infty}(\mathbf{P}^{n_\ell - r})_{jj} = \alpha_j^+$ for all $r \ge N$. Now choose L so that $n_L \ge N$, take $n_\ell^+ = n_{\ell+L} - N$, and conclude that $\lim_{\ell\to\infty}(\mathbf{P}^{n_\ell^+ - r})_{jj} = \alpha_j^+$ for all $r \ge 0$.

The construction of $\{n_\ell^- : \ell \ge 1\}$ is essentially the same and is left as an exercise. □

3.2.18 LEMMA. *If j is aperiodic and recurrent, then $\varlimsup_{n\to\infty}(\mathbf{P}^n)_{jj} \le \pi_{jj}$. Furthermore, if the subsequences $\{n_\ell^\pm : \ell \ge 1\}$ are the ones described in Lemma 3.2.17, then $\lim_{\ell\to\infty}(\mathbf{P}^{n_\ell^\pm})_{ij} = \alpha_{jj}^\pm$ for any i with $i \leftrightarrow j$.*

PROOF: To prove the second assertion, simply note that, by Lemma 3.2.17 and Lebesgue's Dominated Convergence Theorem,

$$(\mathbf{P}^{n_\ell^\pm})_{ij} = \sum_{r=1}^{n_\ell^\pm} \mathbb{P}(\rho_j = r | X_0 = i)(\mathbf{P}^{n_\ell^\pm - r})_{jj} \longrightarrow \mathbb{P}(\rho_j < \infty | X_0 = i)\alpha_j^\pm.$$

Turning to the first assertion, we again use the result in Lemma 3.2.17 to obtain

$$\alpha_j^+ \sum_{r=1}^{N} \mathbb{P}(\rho_j \ge r | X_0 = j) = \lim_{\ell\to\infty} \sum_{r=1}^{N} \mathbb{P}\big(\rho_j \ge r \,\big|\, X_0 = j\big)(\mathbf{P}^{n_\ell^+ - r})_{jj}.$$

for all $N \ge 1$. Thus, if we show that

$$(*) \qquad \sum_{r=1}^{N} \mathbb{P}\big(\rho_j \ge r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} \le 1, \quad n \ge N \ge 1,$$

then we will know that

$$\alpha_j^+ \mathbb{E}\big[\rho_j \,\big|\, X_0 = j\big] = \alpha_j^+ \sum_{r=1}^{\infty} \mathbb{P}\big(\rho_j \ge r \,\big|\, X_0 = j\big) \le 1,$$

which is equivalent to the first assertion.

To prove (*), note that, for any $n \geq 1$,

$$\begin{aligned}(\mathbf{P}^n)_{jj} &= \sum_{r=1}^{n} \mathbb{P}\big(\rho_j = r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} \\ &= \sum_{r=1}^{n} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} - \sum_{r=1}^{n} \mathbb{P}\big(\rho_j \geq r+1 \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} \\ &= \sum_{r=1}^{n} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} - \sum_{r=2}^{n+1} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n+1-r})_{jj},\end{aligned}$$

and so, since $\mathbb{P}(\rho_j \geq 1 | X_0 = j) = 1$,

$$\sum_{r=1}^{n+1} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n+1-r})_{jj} = \sum_{r=1}^{n} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj}$$

for all $n \geq 1$. But $\sum_{r=1}^{n} \mathbb{P}(\rho_j \geq r | X_0 = j)(\mathbf{P}^{n-r})_{jj} = 1$ when $n = 1$, and so we have now proved that

$$\sum_{r=1}^{N} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} \leq \sum_{r=1}^{n} \mathbb{P}\big(\rho_j \geq r \,\big|\, X_0 = j\big)(\mathbf{P}^{n-r})_{jj} = 1$$

for all $n \geq N \geq 1$. □

We can now complete the proof of (3.2.15) when j is recurrent and aperiodic. By the first part of Lemma 3.2.18, we know that $\lim_{n\to\infty}(\mathbf{P}^n)_{jj} = 0$ if $\pi_{jj} = 0$. Thus, if $\pi_{jj} = 0$, then, for any i,

$$\lim_{n\to\infty} (\mathbf{P}^n)_{ij} = \lim_{n\to\infty} \sum_{r=1}^{n} \mathbb{P}(\rho_j = r | X_0 = i)(\mathbf{P}^{n-r})_{jj} = 0 = \pi_{ij}.$$

In order to handle the case when j is positive recurrent, set $C = \{i : i \leftrightarrow j\}$, and take $\boldsymbol{\pi}^C$ accordingly. Then, $\boldsymbol{\pi}^C \in \text{Stat}(\mathbf{P})$. In particular, by the last part of Lemma 3.2.18 and Lebesgue's Dominated Convergence Theorem,

$$\pi_{jj} = \sum_{i\in C} (\boldsymbol{\pi}^C)_i (\mathbf{P}^{n_\ell^\pm})_{ij} \longrightarrow \alpha_j^\pm \sum_{i\in C} (\boldsymbol{\pi}^C)_i = \alpha_j^\pm,$$

and so $(\mathbf{P}^n)_{jj} \longrightarrow \pi_{jj}$. Finally, if $i \neq j$, then, again by the Lebesgue's theorem,

$$(\mathbf{P}^n)_{ij} = \sum_{r=1}^{n} \mathbb{P}(\rho_j = r | X_0 = i)(\mathbf{P}^{n-r})_{jj} \longrightarrow \mathbb{P}(\rho_j < \infty | X_0 = i)\pi_{jj} = \pi_{ij}.$$

3.2.7. Periodic Structure: The preceding result allows us to give a finer analysis even in the case when the period is not 1. Namely, consider a Markov

chain with transition probability matrix $\mathbf{P}$ which is irreducible and recurrent on $\mathbb{S}$, and assume that its period is $d \geq 2$. The basic result in this subsection is that there exists a partition of $\mathbb{S}$ into subsets $\mathbb{S}_r$, $0 \leq r < d$, with the properties that

(1) $(\mathbf{P}^{md+r})_{jk} > 0 \implies r(k) - r(j) = r \bmod d$,
(2) $r(k) - r(j) = r \bmod d \implies (\mathbf{P}^{md+r})_{jk} > 0 < (\mathbf{P}^{md-r})_{kj}$ for all sufficiently large $m \geq 1$,
(3) for each $0 \leq r < d$, the restriction of $\mathbf{P}^d$ to $\mathbb{S}_r$ is an aperiodic, recurrent, irreducible transition probability matrix,

where we have used $r(j)$ to denote the $0 \leq r < d$ for which $j \in \mathbb{S}_r$.

To prove that this decomposition exists, we begin by noting that, for $0 \leq r < d$,

$$(*) \qquad \exists\, m \geq 0\ (\mathbf{P}^{md+r})_{ij} > 0 \implies \exists\, n \geq 1\ (\mathbf{P}^{nd-r})_{ji} > 0.$$

Indeed, by irreducibility and the Euclidean algorithm, we know that there exists an $m' \geq 0$ and $0 \leq r' < d$ such that that $(\mathbf{P}^{m'd+r'})_{ji} > 0$. Furthermore, by (3.1.13), $(\mathbf{P}^{(m'+m'')d+r'})_{ji} \geq (\mathbf{P}^{m'd+r'})_{ji}(\mathbf{P}^{m''d})_{ii} > 0$ for all sufficiently large m''s, and so we may and will assume that $m' \geq 1$. But then $(\mathbf{P}^{(m+m')d+(r+r')})_{ii} > 0$, and so $d|(r+r')$, which, because $0 \leq r, r' < d$, means that $r' = 0$ if $r = 0$ and that $r' = d - r$ if $r \geq 1$.

Starting from (*), it is easy to see that, for each pair $(i,j) \in \mathbb{S}^2$, there is a unique $0 \leq r < d$ such that $(\mathbf{P}^{md+r})_{ij} > 0$ for some $m \geq 0$. Namely, suppose that $(\mathbf{P}^{md+r})_{ij} > 0 < (\mathbf{P}^{m'd+r'})_{ij}$ for some $m, m' \in \mathbb{N}$ and $0 \leq r, r' < d$. Then, by (*), there exists an $n \geq 1$ such that $(\mathbf{P}^{nd-r})_{ji} > 0$ and so $(\mathbf{P}^{(m'+n)d+(r'-r)})_{ii} > 0$. Since this means that $d|(r'-r)$, we have proved that $r = r'$.

Now, let i_0 be a fixed reference point in $\mathbb{S}$, and, for each $0 \leq r < d$, define $\mathbb{S}_r$ to be the set of j such that there exists an $m \geq 0$ for which $(\mathbf{P}^{md+r})_{i_0 j} > 0$. By the preceding, we know that the $\mathbb{S}_r$'s are mutually disjoint. In addition, by irreducibility and the Euclidean algorithm, $\mathbb{S} = \bigcup_{r=0}^{d-1} \mathbb{S}_r$. Turning to the proof of property (1), use (*) to choose $n \geq 0$ and $n' \geq 1$ so that $(\mathbf{P}^{nd+r(j)})_{i_0 j} > 0 < (\mathbf{P}^{n'd-r(k)})_{k i_0}$. Then $(\mathbf{P}^{(n+m+n')d+(r(j)+r-r(k))})_{i_0 i_0} > 0$, and so $d|(r(j)+r-r(k))$. Equivalently, $r(k) - r(j) = r \bmod d$. Conversely, if $r(k) - r(j) = r \bmod d$, choose $n \geq 0$ and $n' \geq 1$ so that $(\mathbf{P}^{nd+r(k)})_{i_0 k} > 0$ and $(\mathbf{P}^{n'd-r(j)})_{j i_0} > 0$. Then $(\mathbf{P}^{(n+m+n')d+r})_{jk} > 0$ for any $m \geq 1$ satisfying $(\mathbf{P}^{md})_{i_0 i_0} > 0$. Since, by (3.1.13), $(\mathbf{P}^{md})_{i_0 i_0} > 0$ for all sufficiently large m's, this completes the left hand inequality in (2), and the right hand inequality in (2) is proved in the same way. Finally, to check (3), note that, from (1), the restriction of $\mathbf{P}^d$ to $\mathbb{S}_r$ is a transition probability matrix, and by (2), it is both irreducible and aperiodic.

The existence of such a partition has several interesting consequences. In the first place, it says that the chain proceeds through the state space in a

cyclic way: if it starts from i, then after n steps it is in $\mathbb{S}_{r(i)+n}$, where the addition in the subscript should be interpreted modulo d. In fact, with this convention for addition, we have that

$$\mathbf{P}^n \mathbf{1}_{\mathbb{S}_r} = \mathbf{1}_{\mathbb{S}_{r+n}}. \tag{3.2.19}$$

To see this, simply observe that, on the one hand, $\mathbf{P}^n \mathbf{1}_{\mathbb{S}_r}(i) = 0$ unless $i \in \mathbb{S}_{r+n}$, while, on the other hand, $\sum_{r'=0}^{d-1} \mathbf{P}^n \mathbf{1}_{\mathbb{S}_{r'}} = 1$. Hence, $i \notin \mathbb{S}_{r+n} \implies (\mathbf{P}^n \mathbf{1}_{\mathbb{S}_r})_i = 0$, whereas $i \in \mathbb{S}_{r+n} \implies 1 = \sum_{r'=0}^{d-1} (\mathbf{P}\mathbf{1}_{\mathbb{S}_{r'}})_i = (\mathbf{P}\mathbf{1}_{\mathbb{S}_{r+n}})_i$.

Secondly, because, for each $0 \le r < d$, the restriction of $\mathbf{P}^d$ to $\mathbb{S}_r$ is an irreducible, recurrent, and aperiodic, transition probability matrix, we know that, for each $0 \le r < d$ and $j \in \mathbb{S}_r$, there exists a $\pi_{jj}^{(r)} \in [0,1]$ with the property that $(\mathbf{P}^{md})_{ij} \longrightarrow \pi_{jj}^{(r)}$ for all $(i,j) \in \mathbb{S}_r^2$. More generally,

$$\lim_{m\to\infty} (\mathbf{P}^{md+s})_{ij} = \begin{cases} \pi_{jj}^{(r(j))} & \text{if } r(j) - r(i) = s \bmod d \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.20}$$

In particular, if $(i,j) \in (\mathbb{S}_r)^2$, then

$$(\mathbf{A}_{nd})_{ij} = \frac{1}{nd} \sum_{m=0}^{n-1} \sum_{s=0}^{d-1} (\mathbf{P}^{md+s})_{ij} = \frac{1}{nd} \sum_{m=0}^{n-1} (\mathbf{P}^{md})_{ij} \longrightarrow \frac{\pi_{jj}^{(r)}}{d}.$$

Hence, since we already know that $(\mathbf{A}_n)_{ij} \longrightarrow \pi_{jj}$, it follows that

$$\pi_j^{(r)} = d\pi_{jj} \quad \text{for } 0 \le r < d \text{ and } j \in \mathbb{S}_r. \tag{3.2.21}$$

In the case when $\mathbf{P}$ is positive recurrent on $\mathbb{S}$, so is the restriction of $\mathbf{P}^d$ to each $\mathbb{S}_r$, and therefore $\sum_{j\in\mathbb{S}_r} \pi_{jj}^{(r)} = 1$. Thus, in the positive recurrent case, (3.2.21) leads to the interesting conclusion that $\boldsymbol{\pi}^{\mathbb{S}}$ assigns probability $\frac{1}{d}$ to each $\mathbb{S}_r$. See Exercise 3.3.12 and 5.6.7 below for other applications of these considerations.

3.3 Exercises

EXERCISE 3.3.1. Just as Césaro convergence is strictly weaker (i.e., it is implied by but does not imply) than ordinary convergence, so, in this exercise, we will show that Abel convergence is strictly weaker than Césaro convergence.

(**a**) Assume that the radius of convergence of $\{a_n\}_0^\infty \subseteq \mathbb{R}$ is less than or equal to 1. That is, $\overline{\lim}_{n\to\infty} |a_n|^{\frac{1}{n}} \le 1$. Set $R(s) = (1-s)\sum_0^\infty s^n a_n$ for $s \in [0,1)$ and $A_n = \frac{1}{n}\sum_0^{n-1} a_m$ for $n \ge 1$, and show that, for $s \in [0,1)$, $R(s) = (1-s)^2 \sum_1^\infty n s^{n-1} A_n$. Use this to conclude that

$$\lim_{n\to\infty} A_n = a \in \mathbb{R} \implies \lim_{s\nearrow 1} R(s) = a.$$

(**b**) Take $a_n = (-1)^{n+1} n$ for $n \ge 0$, check that the radius of convergence of $\{a_n\}_0^\infty$ is 1, and show that

$$\frac{1}{n}\sum_{m=0}^{n-1} a_m = \begin{cases} \frac{1}{2} & \text{if } n \text{ is even} \\ -\frac{1}{2} + \frac{1}{2n} & \text{if } n \text{ is odd} \end{cases} \quad \text{and} \quad (1-s)\sum_{m=0}^{\infty} s^m a_m = \frac{s(1-s)}{(1+s)^2}.$$

Hence, $\{a_n\}_0^\infty$ is Abel convergent to 0 but is Césaro divergent.

Exercise 3.3.2. Recall the queuing model in Exercise 1.3.12. Show that $\{Q_n : n \geq 0\}$ is an $\mathbb{N}$-valued Markov chain conditioned to start from 0, and write down the transition probability matrix for this chain. Further, for this chain: show that 0 is transient if $\mathbb{E}[B_1] > 0$, 0 is null recurrent if $\mathbb{E}[B_1] = 0$, and that 0 is positive recurrent if $\mathbb{E}[B_1] < 0$. In order to handle the case when $\mathbb{E}[B_1] = 0$, you might want to refer to Exercise 1.3.11.

Exercise 3.3.3. Here is a test for positive recurrence. Namely, given a transition probability matrix $\mathbf{P}$ and an element j of the state space $\mathbb{S}$, set $C = \{i \in \mathbb{S} : i \leftrightarrow j\}$. Assume u is a non-negative function on C with the property that

$$u(i) \geq (\mathbf{P}\mathbf{u})_i + \epsilon \quad \text{for all } i \in C \setminus \{j\}$$

for some $\epsilon > 0$.

(a) Begin by showing that

$$\mathbb{E}\big[u\big(X_{(n+1)\wedge\rho_j}\big) \,\big|\, X_0 = j\big] \leq \mathbb{E}\big[u\big(X_{n\wedge\rho_j}\big) \,\big|\, X_0 = j\big] - \epsilon\mathbb{P}\big(\rho_j > n \,\big|\, X_0 = j\big),$$

and use this to conclude that j is positive recurrent.

(b) Suppose that $\mathbb{S} = \mathbb{Z}$ and that $|i| \geq \sum_j |j|(\mathbf{P})_{ij} + \epsilon$ for all $i \in \mathbb{Z} \setminus \{0\}$. Show that 0 is positive recurrent for the chain determined by $\mathbf{P}$.

Exercise 3.3.4. Consider the nearest neighbor random walk on $\mathbb{Z}$ which moves forward with probability $p \in \left(\frac{1}{2}, 1\right)$ and backward with probability $q = 1 - p$. In other words, we are looking at the Markov chain on $\mathbb{Z}$ whose transition probability matrix $\mathbf{P}$ is given by $(\mathbf{P})_{ij} = p$ if $j = i + 1$, $(\mathbf{P})_{ij} = q$ if $j = i - 1$, and $(\mathbf{P})_{ij} = 0$ if $|j - i| \neq 1$. Obviously, this chain is irreducible, and the results in §§1.2.2–1.2.1 show that 0 is transient. Thus, the considerations in Exercise 2.4.10 apply.

(a) Let $\hat{\mathbf{P}}$ be constructed from $\mathbf{P}$ by the prescription in Exercise 2.4.10 when $j_0 = 0$, and, using (1.1.12), show that

$$(\hat{\mathbf{P}})_{ij} = \begin{cases} p & \text{if } i \leq 0 \;\&\; j = i + 1 \text{ or } i \geq 1 \;\&\; j = i - 1 \\ q & \text{if } i \leq 0 \;\&\; j = i - 1 \text{ or } i \geq 1 \;\&\; j = i + 1 \\ 0 & \text{otherwise.} \end{cases}$$

(b) On the basis of Exercise 2.4.10, we know that 0 is recurrent for the chain determined by $\hat{\mathbf{P}}$. Moreover, by part (b) of Exercise 3.3.3, one can check that it is positive recurrent. In fact, by combining part (b) of Exercise 2.4.10 with the computations in §1.1.4, show that

$$\mathbb{E}^{\hat{\mathbb{P}}}[\rho_0 | X_0 = 0] = \frac{2p}{p - q},$$

where the superscript $\hat{\mathbb{P}}$ is used to indicate that the expectation value is taken relative to the chain determined by $\hat{\mathbf{P}}$.

(c) Since $\mathbf{P}$ is irreducible, so is $\hat{\mathbf{P}}$. Hence, since 0 is positive recurrent for the chain determined by $\hat{\mathbf{P}}$, there is a unique stationary probability vector for $\hat{\mathbf{P}}$. Find this vector and use it to show that

$$\mathbb{E}^{\hat{\mathbb{P}}}[\rho_j | X_0 = j] = \begin{cases} \frac{2p}{p-q} & \text{if } j \in \{0,1\} \\ \frac{2pq^j}{p^j(p-q)} & \text{if } j \leq -1 \\ \frac{2p^j}{q^{j-1}(p-q)} & \text{if } j \geq 2. \end{cases}$$

EXERCISE 3.3.5. In section §1.2.4, we worked quite hard to prove that the nearest neighbor, symmetric random walk on $\mathbb{Z}^3$ is transient, and one might hope that the criteria provided in §3.1.2 would allow us to avoid working so hard. However, even if one knows which function u ought to be plugged into Theorem 3.1.5 or 3.1.7, the computation to show that it works is rather delicate. Namely, show that if $\alpha > 0$ is sufficiently large and

$$u(\mathbf{k}) = \left(\alpha^2 + \sum_{i=1}^{3} (\mathbf{k})_i^2\right)^{-\frac{1}{2}} \quad \text{for } \mathbf{k} \in \mathbb{Z}^3,$$

then $(\mathbf{Pu})_{\mathbf{k}} \leq u(\mathbf{k}) \leq u(\mathbf{0})$ when $\mathbf{u}$ is the column vector determined by the function u and $\mathbf{P}$ is the transition probability matrix for the nearest neighbor random walk on $\mathbb{Z}^3$. That is,

$$(\mathbf{P})_{\mathbf{k}\boldsymbol{\ell}} = \begin{cases} \frac{1}{6} & \text{if } \sum_{i=1}^{3} |(\mathbf{k})_i - (\boldsymbol{\ell})_i| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

What follows are some hints.

(a) Let $\mathbf{k} \in \mathbb{Z}^3$ be given, and set

$$M = 1 + \alpha^2 + \sum_{i=1}^{3} (\mathbf{k})_i^2 \quad \text{and} \quad x_i = \frac{(\mathbf{k})_i}{M} \quad \text{for } 1 \leq i \leq 3.$$

Show that $(\mathbf{Pu})_{\mathbf{k}} \leq u(\mathbf{k})$ if and only if

$$\left(1 - \frac{1}{M}\right)^{-\frac{1}{2}} \geq \frac{1}{3} \sum_{i=1}^{3} \frac{(1+2x_i)^{\frac{1}{2}} + (1-2x_i)^{\frac{1}{2}}}{2(1-4x_i^2)^{\frac{1}{2}}}.$$

(b) Show that $\left(1 - \frac{1}{M}\right)^{-\frac{1}{2}} \geq 1 + \frac{1}{2M}$ and that

$$\frac{(1+\xi)^{\frac{1}{2}} + (1-\xi)^{\frac{1}{2}}}{2} \leq 1 - \frac{\xi^2}{8} \quad \text{for } |\xi| < 1,$$

and conclude that $(\mathbf{Pu})_{\mathbf{k}} \leq u(\mathbf{k})$ if

$$1 + \frac{1}{2M} \geq \frac{1}{3} \sum_{i=1}^{3} \frac{1}{(1-4x_i^2)^{\frac{1}{2}}} - \frac{|\mathbf{x}|^2}{6},$$

where $|\mathbf{x}| = \sqrt{\sum_1^3 x_i^2}$ is the Euclidean length of $\mathbf{x} = (x_1, x_2, x_3)$.

(**c**) Show that there is a constant $C < \infty$ such that, as long as $\alpha \geq 1$,

$$\frac{1}{3}\sum_{i=1}^{3} \frac{1}{(1-4x_i^2)^{\frac{1}{2}}} \leq 1 + \frac{2|\mathbf{x}|^2}{3} + C|\mathbf{x}|^4,$$

and put this together with the preceding to conclude that we can take any $\alpha > 0$ with $\alpha^2 + 1 \geq 2C$.

An analogous computation shows that, for each $d \geq 3$ and all sufficiently large $\alpha > 0$, the function $\left(\alpha^2 + \sum_1^d (\mathbf{k})_i^2\right)^{-\frac{d-2}{2}}$ can be used to prove that the nearest neighbor, symmetric random walk is transient in $\mathbb{Z}^d$. The reason why one does not get a contradiction when $d = 2$ is that the non-constant, non-negative functions which satisfies $\mathbf{P}u \leq u$ when $d = 2$ are of the form $\log(\alpha^2 + (\mathbf{k})_1^2 + (\mathbf{k})_2^2)$ and therefore do not achieve their maximum value at $\mathbf{0}$.

EXERCISE 3.3.6. As we said at the beginning of this chapter, we would be content here with statements about the convergence of either $\{(\mathbf{P}^n)_{ij} : n \geq 0\}$ or $\{(\mathbf{A}_n)_{ij} : n \geq 0\}$ for each $(i,j) \in \mathbb{S}^2$. However, as we are about to see, there are circumstances in which the pointwise results we have just obtained self-improve.

(**a**) Assume that j is positive recurrent, and set $C = \{i : i \leftrightarrow j\}$. Given a probability vector $\boldsymbol{\mu}$ with the property that $\sum_{i \notin C}(\boldsymbol{\mu})_i = 0$, show that, in general, $(\boldsymbol{\mu}\mathbf{A}_n)_i \longrightarrow \pi_{ii}$ and, when j is aperiodic, $(\boldsymbol{\mu}\mathbf{P}^n)_i \longrightarrow \pi_{ii}$ for each $i \in C$.

(**b**) Here is an interesting fact about convergence of series. Namely, for each $m \in \mathbb{N}$, let $\{a_{m,n} : n \geq 0\}$ be a sequence of real numbers which converges to a real number b_m as $n \to \infty$. Further, assume that, for each $n \in \mathbb{N}$, the sequence $\{a_{m,n} : m \geq 0\}$ is absolutely summable. Finally, assume that

$$\sum_{m=0}^{\infty} |a_{m,n}| \longrightarrow \sum_{m=0}^{\infty} |b_m| < \infty \quad \text{as } n \to \infty.$$

Show that

$$\lim_{n\to\infty} \sum_{m=0}^{\infty} |a_{m,n} - b_m| = 0.$$

Hint: Using the triangle inequality, show that

$$\big||a_{m,n}| - |b_m| - |a_{m,n} - b_m|\big| \leq 2|b_m|,$$

and apply Lebesgue's Dominated Convergence Theorem to conclude that

$$\sum_{m=0}^{\infty} |a_{m,n} - b_m| \leq \left|\sum_{m=0}^{\infty} \big(|a_{m,n}| - |b_m|\big)\right| + \sum_{m=0}^{\infty} \big||a_{m,n}| - |b_m| - |a_{m,n} - b_m|\big| \longrightarrow 0$$

as $n \to \infty$.

(c) Return to the setting in (a), and use (b) together with the result in (a) to show that, in general, $\|\mu\mathbf{A}_n - \boldsymbol{\pi}^C\|_{\mathrm{v}} \longrightarrow 0$, and $\|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}^C\|_{\mathrm{v}} \longrightarrow 0$ when j is aperiodic. In particular, for each probability vector $\boldsymbol{\mu}$ with $\sum_{i\in C}(\boldsymbol{\mu})_i = 1$,

$$\lim_{n\to\infty} \sup\{|\boldsymbol{\mu}\mathbf{P}^n\mathbf{f} - \boldsymbol{\pi}^C\mathbf{f}| : \|f\|_{\mathrm{u}} \le 1\} = 0,$$

where $\mathbf{f}$ is the column vector determined by a function f. Of course, this is still far less than what we had under Doeblin's condition since his condition provided us with a rate of convergence which was independent of $\boldsymbol{\mu}$. In general, no such uniform rate of convergence will exist.

EXERCISE 3.3.7. Here is an important interpretation of $\boldsymbol{\pi}^C$ when C is a positive recurrent communicating class. Namely, let i be a recurrent state and, for $k \in \mathbb{S}$, let μ_k be the expected number of times the chain visits k before returning to i given that the chain started from i:

$$\mu_k = \mathbb{E}\left[\sum_{m=0}^{\rho_i - 1} \mathbf{1}_{\{k\}}(X_m) \,\middle|\, X_0 = i\right] \in [0, \infty].$$

Determine the row vector $\boldsymbol{\mu} \in [0,\infty]^{\mathbb{S}}$ by $(\boldsymbol{\mu})_k = \mu_k$.

(a) Show that, for all $j \in \mathbb{S}$,

$$(\boldsymbol{\mu}\mathbf{P})_j = \mathbb{E}\left[\sum_{m=1}^{\rho_i} \mathbf{1}_{\{j\}}(X_m) \,\middle|\, X_0 = i\right] = \mu_j.$$

Thus, without any further assumptions about i, $\boldsymbol{\mu}$ is $\mathbf{P}$-stationary in the sense that $\boldsymbol{\mu} = \boldsymbol{\mu}\mathbf{P}$.

(b) Clearly $\mu_i = 1$ and $\sum_j \mu_j = \infty$ unless i is positive recurrent. Nonetheless, show that $\mu_j = 0$ unless $i \leftrightarrow j$ and that $\mu_j \in (0,\infty)$ if $i \leftrightarrow j$.

Hint: Show that

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

(c) If i is positive recurrent, show that

$$\bar{\boldsymbol{\mu}} \equiv \frac{\boldsymbol{\mu}}{\sum_k \mu_k} = \boldsymbol{\pi}^C.$$

Equivalently, when i is positive recurrent,

$$(\boldsymbol{\pi}^C)_j = \frac{\mathbb{E}\left[\sum_{m=0}^{\rho_i - 1} \mathbf{1}_{\{j\}}(X_m) \,\middle|\, X_0 = i\right]}{\mathbb{E}[\rho_i \,|\, X_0 = i]}.$$

In words, $(\boldsymbol{\pi}^C)_j$ *is the relative expected amount of time the chains spends at* j *before returning to* i.

EXERCISE 3.3.8. We continue with the program initiated in Exercise 3.3.7 but assume now that the reference point i is null recurrent. In this case, $\sum_j (\boldsymbol{\mu})_j = \infty$ when $\boldsymbol{\mu} \in [0,\infty)^{\mathbb{S}}$ is the $\mathbf{P}$-stationary measure introduced in Exercise 3.3.8. In this exercise we will show that, up to a multiplicative constant, $\boldsymbol{\mu}$ is the only $\mathbf{P}$-stationary $\boldsymbol{\nu} \in [0,\infty)^{\mathbb{S}}$ with the property that $(\boldsymbol{\nu})_j = 0$ unless $i \to j$ (and therefore $i \leftrightarrow j$). Equivalently, given such a $\boldsymbol{\nu}$, $\boldsymbol{\nu} = (\boldsymbol{\nu})_i \boldsymbol{\mu}$.

(**a**) Assume that $\boldsymbol{\nu} \in [0,\infty)^{\mathbb{S}}$ satisfies $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}$. If $(\boldsymbol{\nu})_i = 1$, show that, for all $j \in \mathbb{S}$ and $n \geq 0$,

$$(\boldsymbol{\nu})_j = \sum_{k \neq i} (\boldsymbol{\nu})_k \mathbb{P}\big(X_n = j \;\&\; \rho_i > n \,\big|\, X_0 = k\big) + \mathbb{E}\left[\sum_{m=0}^{n \wedge (\rho_i - 1)} \mathbf{1}_{\{j\}}(X_m) \,\middle|\, X_0 = i\right].$$

Hint: Work by induction on $n \geq 0$. When $n = 0$ there is nothing to do. To carry out the inductive step, use (2.1.1) and Fubini's Theorem to show that

$$\begin{aligned}\sum_{k \neq i} (\boldsymbol{\nu})_k \mathbb{P}\big(X_n = j \;\&\; \rho_i > n \,\big|\, X_0 = k\big) &= \sum_{k \neq i} (\boldsymbol{\nu}\mathbf{P})_k \mathbb{P}\big(X_n = j \;\&\; \rho_i > n \,\big|\, X_0 = k\big) \\ &= \sum_{\ell} (\boldsymbol{\nu})_\ell \mathbb{P}\big(X_{n+1} = j \;\&\; \rho_i > n+1 \,\big|\, X_0 = \ell\big).\end{aligned}$$

(**b**) Assuming that $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}$ and $(\boldsymbol{\nu})_j = 0$ unless $i \to j$, show that $\boldsymbol{\nu} = (\boldsymbol{\nu})_i \boldsymbol{\mu}$. **Hint**: First show that $\boldsymbol{\nu} = \mathbf{0}$ if $(\boldsymbol{\nu})_i = 0$, and thereby reduce to the case when $(\boldsymbol{\nu})_i = 1$. Starting from the result in (**a**), apply the Monotone Convergence Theorem to see that the right hand side tends to $(\boldsymbol{\mu})_j$ as $n \to \infty$.

EXERCISE 3.3.9. Let C be a communicating class of positive recurrent states. The reason why Theorem 3.2.14 is called a "mean ergodic theorem" is that the asserted convergence is taking place in the sense of mean square convergence. Of course, mean square convergence implies convergence in probability, but, in general, it cannot be used to get convergence with probability 1. Nonetheless, as we will show here, when $\mathbb{P}(X_0 \in C) = 1$ and $j \in C$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) = \pi_{jj} \quad \text{with probability 1.} \tag{3.3.10}$$

Observe that, in § 2.3.3, we proved the individual ergodic theorem (2.3.10) under Doeblin's condition holds, and (3.3.10) says that the same sort of individual ergodic theorem holds even when Doeblin's condition is not present. In fact, there is a very general result, of which (3.3.10) is a very special case,

which was proved originally by G.D. Birkhoff. However, we will not follow Birkhoff and instead, as we did in § 2.3.3, we base our proof on the Strong Law of Large Numbers, although' this time we need the full statement which holds (cf. Theorem 1.4.11 in [9])) for averages of mutually independent, identically distributed, integrable random variables.

(**a**) Show that it suffices to prove the result when $\mathbb{P}(X_0 = i) = 1$ for some $i \in C$.

(**b**) Set $\rho_i^{(0)} = 0$, and use $\rho_i^{(m)}$ to denote the time of the mth return to i. If

$$\tau_m = \rho_i^{(m)} - \rho_i^{(m-1)} \quad \text{and} \quad Y_m = \sum_{\ell=\rho_i^{(m-1)}}^{\rho_i^{(m)}-1} \mathbf{1}_{\{j\}}(X_\ell),$$

show that, conditional on $X_0 = i$, both $\{\tau_m : m \geq 1\}$ and $\{Y_m : m \geq 1\}$ are sequences of mutually independent, identically distributed, non-negative, integrable random variables. In particular, as an application of the Strong Law of Large Numbers and (3.2.5) plus the result in Exercise 3.3.7, conclude that, conditional on $X_0 = i$,

$$(*) \qquad \lim_{m\to\infty} \frac{\rho_i^{(m)}}{m} = \frac{1}{\pi_{ii}} \quad \text{and} \quad \lim_{m\to\infty} \frac{1}{m} \sum_{\ell=0}^{\rho_i^{(m)}-1} \mathbf{1}_{\{j\}}(X_m) = \pi_{ii}\pi_{ij}$$

with probability 1. Hence, $\lim_{m\to\infty} \frac{1}{\rho_i^{(m)}} \sum_{\ell=0}^{\rho_i^{(m)}-1} \mathbf{1}_{\{j\}}(X_m) = \pi_{ij} = \pi_{ij}$ with probability 1.

(**c**) In view of the results in (**a**) and (**b**), we will be done once we check that, conditional on $X_0 = i$,

$$\lim_{n\to\infty} \left| \frac{1}{n} \sum_{0}^{n-1} \mathbf{1}_{\{j\}}(X_\ell) - \frac{1}{\rho_i^{(m_n)}} \sum_{0}^{\rho_i^{(m_n)}-1} \mathbf{1}_{\{j\}}(X_\ell) \right| = 0$$

with probability 1, where m_n is the $\mathbb{Z}^+$-valued random variable determined so that $\rho_i^{(m_n-1)} \leq n < \rho_i^{(m_n)}$. To this end, first show that

$$\left| \frac{1}{n} \sum_{0}^{n-1} \mathbf{1}_{\{j\}}(X_\ell) - \frac{1}{\rho_i^{(m_n)}} \sum_{0}^{\rho_i^{(m_n)}-1} \mathbf{1}_{\{j\}}(X_\ell) \right| \leq \frac{\tau_{m_n}}{m_n}.$$

Next, from the first part of (*), show that $\mathbb{P}\big(\lim_{n\to\infty} m_n = \infty \big| X_0 = i\big) = 1$. Finally, check that, for any $\epsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(\sup_{m\geq M} \frac{\tau_m}{m} \geq \epsilon \,\middle|\, X_0 = i \right) \\ &\quad \leq \sum_{m=M}^{\infty} \mathbb{P}\big(\rho_i \geq m\epsilon \,\big|\, X_0 = i\big) \leq \frac{1}{\epsilon}\mathbb{E}\big[\rho_i,\, \rho_i \geq M\epsilon \,|\, X_0 = i\big], \end{aligned}$$

and use this to complete the proof of (3.3.10).

(**d**) Introduce the *empirical measure* $\mathbf{L}_n$, which is the random probability vector measuring the average time spent at points. That is, $(\mathbf{L}_n)_i = \frac{1}{n}\sum_0^{n-1} \mathbf{1}_{\{i\}}(X_m)$. By combining the result proved here with the one in (**b**) of Exercise 3.3.6, conclude that $\lim_{n\to\infty} \|\mathbf{L}_n - \boldsymbol{\pi}^C\|_v = 0$ with probability 1 when $\mathbb{P}(X_0 \in C) = 1$.

Exercise 3.3.11. Although the statement in Exercise 3.3.9 applies only to positive recurrent states, it turns out that there is a corresponding limit theorem for states which are not positive recurrent. Namely, show that if j is not positive recurrent, then, no matter what the initial distribution,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} \mathbf{1}_{\{j\}}(X_m) = 0\right) = 1.$$

When j is transient, $\mathbb{E}\left[\sum_0^\infty \mathbf{1}_{\{j\}}(X_m)\right] < \infty$ and therefore the result is trivial. To handle j are null recurrent, begin by noting that it suffices to handle the case when $\mathbb{P}(X_0 = j) = 1$. Next, note that, conditional on $X_0 = j$, $\{\rho_j^{(m+1)} - \rho_j^{(m)} : m \geq 0\}$ is a sequence of independent, identically distributed, $\mathbb{Z}^+$-valued random variables, and apply the Strong Law of Large Numbers to see that, for any $R \geq 1$,

$$\begin{aligned}\mathbb{P}\left(\varliminf_{m\to\infty} \frac{\rho_j^{(m)}}{m} \geq H(R) \,\middle|\, X_0 = j\right) \\ \geq \mathbb{P}\left(\varliminf_{m\to\infty} \frac{\rho_j^{(m)} \wedge R}{m} \geq H(R) \,\middle|\, X_0 = j\right) = 1,\end{aligned}$$

where $H(r) \equiv \frac{1}{2}\mathbb{E}[\rho_j \wedge R | X_0 = j] \nearrow \infty$ as $R \nearrow \infty$. Hence, given $X_0 = j$, $\frac{\rho_j^{(m)}}{m} \longrightarrow \infty$ with probability 1. Finally, check that, for any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{n\geq N} \frac{1}{n}\sum_0^{n-1} \mathbf{1}_{\{j\}}(X_m) \geq \epsilon \,\middle|\, X_0 = j\right) \leq \mathbb{P}\left(\sup_{n\geq N} \frac{\rho_j^{([n\epsilon])}}{n} \leq \frac{1}{\epsilon} \,\middle|\, X_0 = j\right) \\ \leq \mathbb{P}\left(\sup_{m\geq N\epsilon} \frac{\rho_j^{(m)}}{m} \leq \frac{1}{\epsilon} \,\middle|\, X_0 = j\right),\end{aligned}$$

and combine this with the preceding to reach the desired conclusion.

Exercise 3.3.12. When j is aperiodic for $\mathbf{P}$, we know that $\lim_{n\to\infty}(\mathbf{P}^n)_{ij}$ exists for all $i \in \mathbb{S}$ and is 0 unless j is positive recurrent. When $d(j) > 1$ and j is positive recurrent, show that $\lim_{n\to\infty}(\mathbf{P}^n)_{jj}$ will fail to exist. On the other hand, even if $d(j) > 1$, show that $\lim_{n\to\infty}(\mathbf{P}^n)_{ij} = 0$ for any $j \in \mathbb{S}$ which is not positive recurrent.

CHAPTER 4

Markov Processes in Continuous Time

Up until now we have been dealing with Markov processes for which time has a discrete parameter $n \in \mathbb{N}$. In this chapter we will introduce Markov processes for which time has a continuous parameter $t \in [0, \infty)$, even though our processes will continue to take their values in a countable state space $\mathbb{S}$.

4.1 Poisson Processes

Just as Markov processes with independent, identically distributed increments (a.k.a. random walks) on $\mathbb{Z}^d$ are the simplest discrete parameter Markov processes, so the simplest continuous time Markov processes are those whose increments are mutually independent and *homogeneous* in the sense that the distribution of an increment depends only on the length of the time interval over which the increment is taken. More precisely, we will be dealing in this section with $\mathbb{Z}^d$-valued stochastic processes $\{X(t) : t \geq 0\}$ with the property that $\mathbb{P}(X(0) = \mathbf{0}) = 1$ and

$$\begin{aligned}\mathbb{P}\big(X(t_1) - X(t_0) = j_1, \ldots, X(t_n) - X(t_{n-1}) = j_n\big) \\ = \prod_{m=1}^{n} \mathbb{P}\big(X(t_m - t_{m-1}) = j_m\big)\end{aligned}$$

for all $n \geq 1$, $0 \leq t_0 < \cdots < t_n$, and $(j_1, \ldots, j_n) \in (\mathbb{Z}^d)^n$.

4.1.1. The Simple Poisson Process: The *simple Poisson process* is the $\mathbb{N}$-valued stochastic process $\{N(t) : t \geq 0\}$ which starts from 0 (i.e., $N(0) = 0$), sits at 0 for a unit exponential holding time E_1 (i.e., $N(t) = 0$ for $0 \leq t < E_1$ and $\mathbb{P}(E_1 > t) = e^{-t}$), at time E_1 moves to 1 where it remains for an independent unit exponential holding time E_2, moves to 2 at time $E_1 + E_2$, etc. More precisely, if $\{E_n : n \geq 1\}$ is a sequence of mutually independent, unit exponential random variables and

$$J_n = \begin{cases} 0 & \text{when } n = 0 \\ \sum_{m=1}^{n} E_m & \text{when } n \geq 1, \end{cases}$$

then the stochastic process $\{N(t) : t \geq 0\}$ given by

$$N(t) = \max\{n \geq 0 : J_n \leq t\} \tag{4.1.1}$$

is a simple Poisson process. When thinking about the meaning of (4.1.1), keep in mind that, because,

$$\text{with probability 1, } E_n > 0 \text{ for all } n \geq 1 \text{ and } \sum_{m=1}^{\infty} E_m = \infty,$$

with probability 1 the path $t \in [0,\infty) \longmapsto N(t)$ is piecewise constant, right continuous, and, when it jumps, it jumps by $+1$: $N(t) - N(t-) \in \{0,1\}$ for all $t > 0$, where $N(t-) \equiv \lim_{s \nearrow t} N(s)$ is the left limit of $N(\,\cdot\,)$ at t.

We now want to show that $\{N(t) : t \geq 0\}$ moves along in independent, homogeneous increments.[1] That is, we want to show that, for each $s, t \in [0,\infty)$, $N(t) - N(s)$ is independent of (cf. §6.1.3) $\sigma\big(\{N(\tau) : \tau \in [0,s]\}\big)$ and has the same distribution as $N(t)$:

$$(4.1.2) \quad \mathbb{P}\big(N(s+t) - N(s) = n \,\big|\, N(\tau),\, \tau \in [0,s]\big) = \mathbb{P}\big(N(t) = n\big), \quad n \in \mathbb{N}.$$

Equivalently, what we have to check is that when $(s,t) \in [0,\infty)^2$ and $A \in \sigma\big(\{N(\tau) : \tau \in [0,s]\}\big)$, $\mathbb{P}\big(\{N(s+t) - N(s) \geq n\} \cap A\big) = \mathbb{P}(N(t) \geq n)\mathbb{P}(A)$ for all $n \in \mathbb{N}$. Since this is trivial when $n = 0$, we will assume that $n \in \mathbb{Z}^+$. In addition, since we can always write A as the disjoint union of the sets $A \cap \{N(s) = m\}$, $m \in \mathbb{N}$, and each of these is again in $\sigma\big(\{N(\tau) : \tau \in [0,s]\}\big)$, we may and will assume that, for some m, $N(s) = m$ on A. But in that case we can write $A = \{J_{m+1} > s\} \cap B$ where $B \in \sigma\big(\{E_1, \ldots, E_m\}\big)$ and $J_m \leq s$ on B. Hence, since $N(s+t) \geq m+n \iff J_{m+n} \leq s+t$, and $\sigma(\{J_\ell - J_m : \ell > m\})$ is independent of $\sigma\big(\{E_k : k \leq m\}\big)$, an application of (6.4.2) shows that

$$\begin{aligned}
&\mathbb{P}\big(\{N(s+t) - N(s) \geq n\} \cap A\big) = \mathbb{P}\big(\{J_{m+n} \leq s+t\} \cap \{J_{m+1} > s\} \cap B\big) \\
&\quad = \mathbb{P}\big(\{J_{m+n} - J_m \leq s+t-J_m\} \cap \{J_{m+1} - J_m > s - J_m\} \cap B\big) \\
&\quad = \mathbb{E}\big[v(J_m),\, B\big],
\end{aligned}$$

where, for $\xi \in [0,s]$,

$$\begin{aligned}
&v(\xi) \equiv \mathbb{P}\big(\{J_{m+n} - J_m \leq s+t-\xi\} \cap \{J_{m+1} - J_m > s - \xi\}\big) \\
&= \mathbb{P}\big(\{J_n \leq s+t-\xi\} \cap \{E_1 > s - \xi\}\big) = \mathbb{P}\big(\{J_n \leq s+t-\xi\} \cap \{E_n > s-\xi\}\big) \\
&= \mathbb{P}\big(\{J_{n-1} + E_n \leq s+t-\xi\} \cap \{E_n > s-\xi\}\big) = \mathbb{E}\big[w(\xi, E_n),\, E_n > s - \xi\big]
\end{aligned}$$

when $w(\xi,\eta) \equiv \mathbb{P}(J_{n-1} \leq s+t-\xi-\eta)$ for $\xi \in [0,s]$ and $\eta \in [s-\xi, s+t-\xi]$.

Up to this point we have not used any property of exponential random variables other than that they of positive. However, in our next step we will

[1] Actually, our primary goal here is to develop a line of reasoning which will serve us well later on. A more straight-forward, but less revealing, proof that the simple Poisson process has independent, homogeneous increments is given in Exercise 4.5.1 below.

use their characteristic property, namely, the fact that an exponential random variable E "has no memory." That is, $\mathbb{P}(E > a + b | E > a) = \mathbb{P}(E > b)$, from which it is an easy step to

$$\mathbb{E}\big[f(E),\, E > a\big] = e^{-a}\mathbb{E}\big[f(a+E)\big] \tag{4.1.3}$$

for any non-negative, $\mathcal{B}_{[0,\infty)}$-measurable function f. In particular, this means that

$$\begin{aligned} v(\xi) &= \mathbb{E}\big[w(\xi, E_n),\, E_n > s - \xi\big] = e^{-(s-\xi)}\mathbb{E}\big[w(\xi, E_n + s - \xi)\big] \\ &= e^{-(s-\xi)}\mathbb{P}(J_{n-1} \le t - E_n) = e^{-(s-\xi)}\mathbb{P}(J_n \le t) = e^{-(s-\xi)}\mathbb{P}\big(N(t) \ge n\big). \end{aligned}$$

Hence, we have now shown that

$$\mathbb{P}\big(\{N(s+t) - N(s) \ge n\} \cap A\big) = \mathbb{E}\big[e^{-(s-J_m)},\, B\big]\mathbb{P}\big(N(t) \ge n\big).$$

Finally, since

$$\mathbb{P}(A) = \mathbb{P}\big(\{J_{m+1} > s\} \cap B\big) = \mathbb{P}\big(\{E_{m+1} > s - J_m\} \cap B\big) = \mathbb{E}\big[e^{-(s-J_m)},\, B\big],$$

the proof of (4.1.2) is complete. Hence, we now know that *the simple Poisson process* $\{N(t) : t \ge 0\}$ *has homogeneous and mutually independent increments.*

Before moving on, we must still find out what is the distribution of $N(t)$. But, the sum of n mutually independent, unit exponential random variables has a $\Gamma(n)$-distribution, and so

$$\begin{aligned} \mathbb{P}\big(N(t) = n\big) &= \mathbb{P}\big(J_n \le t < J_{n+1}\big) = \mathbb{P}(J_n \le t) - \mathbb{P}(J_{n+1} \le t) \\ &= \frac{1}{(n-1)!}\int_0^t \tau^{n-1}\, d\tau - \frac{1}{n!}\int_0^t \tau^n\, d\tau = \frac{t^n}{n!}. \end{aligned}$$

That is, $N(t)$ *is a Poisson random variable with mean* t. More generally, when we combine this with (4.1.2), we get that, for all $s,\, t \in [0,\infty)$ and $n \in \mathbb{N}$,

$$\mathbb{P}\big(N(s+t) - N(s) = n \,\big|\, N(\tau),\, \tau \in [0,s]\big) = e^{-t}\frac{t^n}{n!}. \tag{4.1.4}$$

Alternatively, again starting from (4.1.2), we can now give the following Markovian description of $\{N(t) : t \ge 0\}$:

$$\mathbb{P}\big(N(s+t) = n \,\big|\, N(\tau),\, \tau \in [0,s]\big) = e^{-t}\frac{t^{n-N(s)}}{(n-N(s))!}\mathbf{1}_{[0,n]}\big(N(s)\big). \tag{4.1.5}$$

4.1.2. Compound Poisson Processes on $\mathbb{Z}^d$: Having constructed the simplest Poisson process, we can easily construct a rich class of processes which are the continuous time analogs of random walks on $\mathbb{Z}^d$. Namely, suppose that

$\boldsymbol{\mu}$ is a probability vector on $\mathbb{Z}^d$ which gives 0 mass to the origin $\mathbf{0}$. Then the *Poisson process with jump distribution* $\boldsymbol{\mu}$ *and rate* $R \in (0,\infty)$ is the stochastic process $\{\mathbf{X}(t) : t \geq 0\}$ which starts at $\mathbf{0}$, sits there for an exponential holding time having mean value R^{-1}, at which time it jumps by the amount $\mathbf{k} \in \mathbb{Z}^d$ with probability $(\boldsymbol{\mu})_{\mathbf{k}}$, sits where it lands for another, independent holding time with mean R^{-1}, jumps again a random amount with distribution $\boldsymbol{\mu}$, etc. Thus, the simple Poisson process is the case when $d = 1$, $(\boldsymbol{\mu})_1 = 1$, and $R = 1$. In particular, the amount by which the simple Poisson process jumps is deterministic whereas the amount by which a compound Poisson process jumps will, in general, be random.

To construct a compound Poisson process, let $\{\mathbf{B}_n : n \geq 1\}$ be a sequence of mutually independent $\mathbb{Z}^d$-valued random variables with distribution $\boldsymbol{\mu}$, introduce the random walk $\mathbf{X}_0 = \mathbf{0}$ and $\mathbf{X}_n = \sum_{m=1}^{n} \mathbf{B}_m$ for $n \geq 1$, and define $\{\mathbf{X}(t) : t \geq 0\}$ so that $\mathbf{X}(t) = \mathbf{X}_{N(Rt)}$, where $\{N(t) : t \geq 0\}$ is a simple Poisson process which is independent of the $\mathbf{B}_m$'s. The existence of all these random variables is guaranteed (cf. the footnote in § 1.2.1) by Theorem 6.3.2. Obviously, $\mathbf{X}(0) = \mathbf{0}$ and $t \in [0,\infty) \longmapsto \mathbf{X}(t) \in \mathbb{Z}^d$ is a piecewise constant, right continuous $\mathbb{Z}^d$-valued path. In addition, because $(\boldsymbol{\mu})_{\mathbf{0}} = 0$, it is clear[2] that the number of jumps that $t \rightsquigarrow \mathbf{X}(t)$ makes during a time interval $(s,t]$ is precisely $N(Rt) - N(Rs)$ and that $\mathbf{X}_n - \mathbf{X}_{n-1}$ is the amount of the nth jump of $t \rightsquigarrow \mathbf{X}(t)$. Thus, if $J_0 \equiv 0$ and, for $n \geq 1$, J_n is the time of the nth jump of $t \rightsquigarrow \mathbf{X}(t)$, then $N(Rt) = n \iff J_n \leq Rt < J_{n+1}$, and $\mathbf{X}(J_n) - \mathbf{X}(J_{n-1}) = \mathbf{X}_n - \mathbf{X}_{n-1}$. Equivalently, if $\{E_n : n \geq 1\}$ denotes the sequence of unit exponential random variables out of which $\{N(t) : t \geq 0\}$ is built, then $J_n - J_{n-1} = \frac{E_n}{R}$, $\mathbf{X}(t) - \mathbf{X}(t-) = \mathbf{0}$ for $t \in (J_{n-1}, J_n)$, and $\mathbf{X}(J_n) - \mathbf{X}(J_{n-1}) = \mathbf{B}_n$. Hence, $\{\mathbf{X}(t) : t \geq 0\}$ is indeed a compound Poisson process with jump distribution $\boldsymbol{\mu}$ and rate R.

We next want to show that a compound process moves along in homogeneous, mutually independent increments:

$$(4.1.6)\quad \mathbb{P}\big(\mathbf{X}(s+t) - \mathbf{X}(s) = \mathbf{k} \,\big|\, \mathbf{X}(\tau),\, \tau \in [0,s]\big) = \mathbb{P}(\mathbf{X}(t) = \mathbf{k}), \quad \mathbf{k} \in \mathbb{Z}^d.$$

For this purpose, we use the representation $\mathbf{X}(t) = \mathbf{X}_{N(Rt)}$ introduced above. Given $A \in \sigma\big(\{\mathbf{X}(\tau) : \tau \in [0,s]\}\big)$, we need to show that

$$\mathbb{P}\big(\{\mathbf{X}(s+t) - \mathbf{X}(s) = \mathbf{k}\} \cap A\big) = \mathbb{P}\big(\{\mathbf{X}(s+t) - \mathbf{X}(s) = \mathbf{k}\}\big)\mathbb{P}(A),$$

and, just as in the derivation of (4.1.2), we will, without loss in generality, assume that, for some $m \in \mathbb{N}$, $N(Rs) = m$ on A. But then A is independent

[2] This is the reason for our having assumed that $(\boldsymbol{\mu})_{\mathbf{0}} = 0$. However, one should realize that this assumption causes no loss in generality. Namely, if $(\boldsymbol{\mu})_{\mathbf{0}} = 1$, then the resulting compound process would be trivial: it would never move. On the other hand, if $(\boldsymbol{\mu})_{\mathbf{0}} \in (0,1)$, then we could replace $\boldsymbol{\mu}$ by $\bar{\boldsymbol{\mu}}$, where $(\bar{\boldsymbol{\mu}})_{\mathbf{0}} = 0$ and $(\bar{\boldsymbol{\mu}})_{\mathbf{k}} = (1 - (\boldsymbol{\mu})_{\mathbf{0}})^{-1}(\boldsymbol{\mu})_{\mathbf{k}}$ when $\mathbf{k} \neq \mathbf{0}$, and R by $\bar{R} = (1 - (\boldsymbol{\mu})_{\mathbf{0}})R$. The compound Poisson process corresponding to $\bar{\boldsymbol{\mu}}$ and $\bar{R}$ would have exactly the same distribution of the one corresponding to $\boldsymbol{\mu}$ and R.

of $\sigma\big(\{\mathbf{X}_{m+n} - \mathbf{X}_m : n \geq 0\} \cup \{N\big(R(s+t)\big) - N(Rs)\}\big)$, and so

$$\begin{aligned}
&\mathbb{P}\big(\{\mathbf{X}(s+t) - \mathbf{X}(s) = \mathbf{k}\} \cap A\big) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\big(\{\mathbf{X}(s+t) - \mathbf{X}(s) = \mathbf{k} \;\&\; N\big(R(s+t)\big) - N(Rs) = n\} \cap A\big) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\big(\{\mathbf{X}_{m+n} - \mathbf{X}_m = \mathbf{k} \;\&\; N\big(R(s+t)\big) - N(Rs) = n\} \cap A\big) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\big(\mathbf{X}_n = \mathbf{k}\big)\mathbb{P}\big(N(Rt) = n\big)\mathbb{P}(A) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\big(\mathbf{X}_n = \mathbf{k} \;\&\; N(Rt) = n\big)\mathbb{P}(A) = \mathbb{P}\big(\mathbf{X}(t) = \mathbf{k}\big)\mathbb{P}(A).
\end{aligned}$$

Hence, (4.1.6) is proved.

Finally, to compute the distribution of $\mathbf{X}(t)$, begin by recalling that the distribution of the sum of n independent, identically distributed random variables is the *n-fold convolution* of their distribution. Thus, $\mathbb{P}(\mathbf{X}_n = \mathbf{k}) = (\boldsymbol{\mu}^{\star n})_{\mathbf{k}}$, where $(\boldsymbol{\mu}^{\star 0})_{\mathbf{k}} = \delta_{\mathbf{0},\mathbf{k}}$ is the point mass at $\mathbf{0}$ and

$$\big(\boldsymbol{\mu}^{\star n}\big)_{\mathbf{k}} = \sum_{\mathbf{j} \in \mathbb{Z}^d} \big(\boldsymbol{\mu}^{\star(n-1)}\big)_{\mathbf{k}-\mathbf{j}} (\boldsymbol{\mu})_{\mathbf{j}}$$

for $n \geq 1$. Hence,

$$\mathbb{P}\big(\mathbf{X}(s+t) = \mathbf{k}\big) = \sum_{n=0}^{\infty} \mathbb{P}\big(\mathbf{X}_n = \mathbf{k} \;\&\; N(Rt) = n\big) = e^{-Rt} \sum_{n=0}^{\infty} \frac{(Rt)^n}{n!} (\boldsymbol{\mu}^{\star n})_{\mathbf{k}}.$$

Putting this together with (4.1.6), we now see that for $A \in \sigma\big(\{N(\tau) : \tau \in [0, s]\}\big)$,

$$\begin{aligned}
\mathbb{P}\big(\{\mathbf{X}(s+t) = \mathbf{k}\} \cap A\big) &= \sum_{\mathbf{j} \in \mathbb{Z}^d} \mathbb{P}\big(\{\mathbf{X}(s+t) = \mathbf{k}\} \cap A \cap \{\mathbf{X}(s) = \mathbf{j}\}\big) \\
&= \sum_{\mathbf{j} \in \mathbb{Z}^d} \mathbb{P}\big(\{\mathbf{X}(s+t) - \mathbf{X}(s) = \mathbf{k} - \mathbf{j}\} \cap A \cap \{\mathbf{X}(s) = \mathbf{j}\}\big) \\
&= \sum_{\mathbf{j} \in \mathbb{Z}^d} \big(\mathbf{P}(t)\big)_{\mathbf{jk}} \mathbb{P}\big(A \cap \{\mathbf{X}(s) = \mathbf{j}\}\big) = \mathbb{E}\big[\big(\mathbf{P}(t)\big)_{\mathbf{X}(s)\mathbf{k}}, A\big],
\end{aligned}$$

where

$$\big(\mathbf{P}(t)\big)_{\mathbf{k}\boldsymbol{\ell}} \equiv e^{-Rt} \sum_{m=0}^{\infty} \frac{(Rt)^m}{m!} (\boldsymbol{\mu}^{\star m})_{\boldsymbol{\ell}-\mathbf{k}}. \tag{4.1.7}$$

Equivalently, we have proved that $\{\mathbf{X}(t) : t \geq 0\}$ *is a continuous time Markov process with transition probability* $t \rightsquigarrow \mathbf{P}(t)$ in the sense that

$$\mathbb{P}\big(\mathbf{X}(s+t) = \mathbf{k} \,\big|\, \mathbf{X}(\sigma),\ \sigma \in [0,s]\big) = \big(\mathbf{P}(t)\big)_{\mathbf{X}(s)\mathbf{k}}. \tag{4.1.8}$$

Observe that, as a consequence of (4.1.8), we find that $\{\mathbf{P}(t) : t \geq 0\}$ is a *semigroup*. That is, it satisfies the *Chapman–Kolmogorov* equation

$$\mathbf{P}(s+t) = \mathbf{P}(s)\mathbf{P}(t), \quad s,t \in [0,\infty). \tag{4.1.9}$$

Indeed,

$$\begin{aligned}\big(\mathbf{P}(s+t)\big)_{\mathbf{0k}} &= \sum_{\mathbf{j}\in\mathbb{Z}^d} \mathbb{P}\big(\mathbf{X}(s+t) = \mathbf{k} \,\&\, \mathbf{X}(s) = \mathbf{j}\big) \\ &= \sum_{\mathbf{j}\in\mathbb{Z}^d} \big(\mathbf{P}(t)\big)_{\mathbf{j}\boldsymbol{\ell}} \big(\mathbf{P}(s)\big)_{\mathbf{0j}} = \big(\mathbf{P}(s)\mathbf{P}(t)\big)_{\mathbf{0k}},\end{aligned}$$

from which the asserted matrix equality follows immediately when one remembers that $\big(\mathbf{P}(\tau)\big)_{\mathbf{k}\boldsymbol{\ell}} = \big(\mathbf{P}(\tau)\big)_{\mathbf{0}(\boldsymbol{\ell}-\mathbf{k})}$.

4.2 Markov Processes with Bounded Rates

There are two directions in which one can generalize the preceding without destroying the Markov property. For one thing, one can make the distribution of jumps depend on where the process is at the time it makes the jump. This change comes down to replacing the random walk in the compound Poisson process by more a general Markov chain. The second way to increase the randomness is to make the rate of jumping depend on the place where the process is waiting before it jumps. That is, instead of the holding times all having the same mean value, the holding time at a particular state will depend on that state.

4.2.1. Basic Construction: Let $\mathbb{S}$ be a countable state space and $\mathbf{P}$ a transition probability matrix with the property that $(\mathbf{P})_{ii} = 0$ for all $i \in \mathbb{S}$.[3] Further, let $\mathfrak{R} \equiv \{R_i : i \in \mathbb{S}\} \subseteq [0,\infty)$ be a family of *rates*. Then a continuous time *Markov process* on $\mathbb{S}$ with rates $\mathfrak{R}$ and transition probability matrix $\mathbf{P}$ is an $\mathbb{S}$-valued family $\{X(t) : t \geq 0\}$ of random variables with the properties that

(a) $t \rightsquigarrow X(t)$ is piecewise constant and right continuous,

(b) If $J_0 \equiv 0$ and, for $n \geq 1$, J_n is the time of the nth jump of $t \rightsquigarrow X(t)$, then (4.2.1)

$$\begin{aligned}&\mathbb{P}\big(J_n > J_{n-1} + t \,\&\, X(J_n) = j \,\big|\, X(\tau),\ \tau \in [0, J_n)\big) \\ &\quad = e^{-tR_{X(J_{n-1})}} (\mathbf{P})_{X(J_{n-1})j} \text{ on } \{J_{n-1} < \infty\}.\end{aligned}$$

[3] We make this assumption for the same reason as we assumed in §4.1.2 that $(\boldsymbol{\mu})_{\mathbf{0}} = 0$, and, just as it resulted in no loss in generality there, so it does not reduce the generality here.

Our first task is to see that, together with the initial distribution, the preceding completely determines the distribution of $\{X(t) : t \geq 0\}$, and, for reasons which will become clear soon, we will restrict our attention throughout this section to the case when the rates $\mathfrak{R}$ are *bounded* in the sense that $\sup_i R_i < \infty$. In fact, for the moment, we will also assume that the rates $\mathfrak{R}$ are *non-degenerate* in the sense that $\mathfrak{R} \subseteq (0, \infty)$. Since it is clear that non-degeneracy implies that $\mathbb{P}(J_n < \infty) = 1$ for each $n \geq 0$, we may (cf. (6.1.5)) and will assume that $J_n < \infty$ for all $n \geq 0$. Now set $X_n = X(J_n)$ and $E_n = \frac{J_n - J_{n-1}}{R_{n-1}}$ for $n \geq 1$, and observe that, by (**b**) in (4.2.1),

$$\mathbb{P}\big(E_n > t \;\&\; X_n = j \,\big|\, \{E_1, \ldots, E_{n-1}\} \cup \{X_0, \ldots, X_{n-1}\}\big) = e^{-t}(\mathbf{P})_{X_n j}.$$

Hence, $\{X_n : n \geq 0\}$ is a Markov chain with transition probability matrix $\mathbf{P}$ and the same initial distribution as the distribution of $X(0)$, $\{E_n : n \geq 1\}$ is a sequence of mutually independent, unit exponential random variables, and $\sigma(\{X_n : n \geq 0\})$ is independent of $\sigma(\{E_n : n \geq 1\})$. Thus, the joint distribution of $\{X_n : n \geq 0\}$ and $\{E_n : n \geq 1\}$ is uniquely determined. Moreover, $\{X(t) : t \geq 0\}$ can be recovered from $\{X_n : n \geq 0\} \cup \{E_n : n \geq 1\}$. Namely, given $(e_1, \ldots, e_n, \ldots) \in (0, \infty)^{\mathbb{Z}^+}$ and $(j_0, \ldots, j_n, \ldots) \in \mathbb{S}^{\mathbb{N}}$, define

$$\begin{aligned} &\Phi^{(\mathfrak{R},\mathbf{P})}\big(t; (e_1, \ldots, e_n, \ldots), (j_0, \ldots, j_n, \ldots)\big) = j_n \text{ for } \xi_n \leq t < \xi_{n+1} \\ &\qquad \text{where } \xi_0 = 0 \text{ and } \xi_n = \sum_{m=1}^{n} R_{j_{m-1}}^{-1} e_m \text{ when } n \geq 1. \end{aligned} \tag{4.2.2}$$

Clearly,

$$\begin{aligned} X(t) = \Phi^{(\mathfrak{R},\mathbf{P})}\big(t; (E_1, \ldots, E_n, \ldots), (X_0, \ldots, X_n, \ldots)\big) \\ \text{for } 0 \leq t < \sum_{m=1}^{\infty} R_{j_{m-1}}^{-1} E_m. \end{aligned} \tag{4.2.3}$$

Thus, we will know that the distribution of $\{X(t) : t \geq 0\}$ is uniquely determined once we check that $\sum_{m=1}^{\infty} R_{X_{m-1}}^{-1} E_n = \infty$ with probability 1. But this is precisely why we made the assumption that $\mathfrak{R}$ is bounded. Namely, by either the Strong Law of Large Numbers or explicit calculation (e.g. of $\mathbb{E}\left[\exp\left(-\sum_1^{\infty} E_m\right)\right]$), we know that $\sum_1^{\infty} E_m = \infty$ with probability 1. Hence, the boundedness of $\mathfrak{R}$ is more than enough to guarantee that $\sum_1^{\infty} R_{X_{m-1}}^{-1} E_m = \infty$ with probability 1.

To handle[4] the *degenerate* case, the one when some of the R_i's may be 0, set $\mathbb{S}_0 = \{i : R_i = 0\}$ and determine $\bar{\mathfrak{R}}$ so that $\bar{R}_i = R_i$ if $i \notin \mathbb{S}_0$ and $\bar{R}_i = 1$ if $i \in \mathbb{S}_0$. Our goal is to show that the distribution of $\{X(t) : t \geq 0\}$ is the same as that of $\{\bar{X}(t \wedge \zeta : t \geq 0\}$, where $\{\bar{X}(t) : t \geq 0\}$ is a process with rates

[4] The discussion with follows is a little technical and need not be fully assimilated in order to proceed.

$\bar{\mathfrak{R}}$ and transition probability matrix $\mathbf{P}$ and $\zeta \equiv \inf\{t \geq 0 : \bar{X}(t) \in \mathbb{S}_0\}$ is the first time $t \rightsquigarrow \bar{X}(t)$ hits $\mathbb{S}_0$. That is, we are claiming that the distribution of the process $\{X(t) : t \geq 0\}$ is that of the process $\{\bar{X}(t) : t \geq 0\}$ *stopped* when it hits $\mathbb{S}_0$.

To verify the preceding assertion, let $\{X_n^{(i)} : n \geq 0 \ \& \ i \in \mathbb{S}_0\}$ be a family of $\mathbb{S}$-valued random variables and $\{\bar{E}_n : n \geq 1\}$ a family of $(0, \infty)$-valued random variables with the properties that

(1) $\sigma\big(\{X_n^{(i)} : n \geq 0 \ \& \ i \in \mathbb{S}_0\}\big)$, $\sigma\big(\{\bar{E}_n : n \geq 1\}\big)$, and $\sigma\big(\{X(t) : t \geq 0\}\big)$ are mutually independent,
(2) for each $i \in \mathbb{S}_0$, $\{X_n^{(i)} : n \geq 0\}$ is a Markov chain starting from i with transition probability matrix $\mathbf{P}$,
(3) $\{\bar{E}_n : n \geq 1\}$ are mutually independent, unit exponential random variables.

Then, for each $i \in \mathbb{S}$, the process $\{\bar{X}^{(i)}(t) : t \geq 0\}$ given by

$$\bar{X}^{(i)}(t) = \Phi^{(\bar{\mathfrak{R}}, \mathbf{P})}\big(t; (\bar{E}_1, \ldots, \bar{E}_n, \ldots), (X_0^{(i)}, \ldots, X_n^{(i)}, \ldots)\big)$$

starts at i and is associated with the rates $\bar{\mathfrak{R}}$ and transition probability matrix $\mathbf{P}$. Finally, define $\{\bar{X}(t) : t \geq 0\}$ so that

$$\bar{X}(t) = \begin{cases} X(t) & \text{if } t < \zeta \\ \bar{X}^{(X(\zeta))}(t - \zeta) & \text{if } t \geq \zeta. \end{cases}$$

Obviously, $X(t) = \bar{X}(t \wedge \zeta)$. Hence, we will be done once we show that the distribution of $\{\bar{X}(t) : t \geq 0\}$ is that of a process associated with the rates $\bar{\mathfrak{R}}$ and transition probability matrix $\mathbf{P}$. To see this, set $\bar{J}_0 \equiv 0$ and, for $m \geq 1$, let $\bar{J}_m$ be the time of the mth jump of $t \rightsquigarrow \bar{X}(t)$. Next, suppose that $A \in \sigma\big(\{\bar{X}(\tau) : \tau \in [0, J_n)\}\big)$. We need to show that

$$(*) \quad \mathbb{P}\big(\{\bar{J}_n > \bar{J}_{n-1} + t \ \& \ \bar{X}(\bar{J}_n) = j\} \cap A\big) = \mathbb{E}\Big[e^{-t\bar{R}_{\bar{X}(\bar{J}_{n-1})}}(\mathbf{P})_{\bar{X}(\bar{J}_{n-1})j}, \ A\Big],$$

and because we can always write A is the disjoint union of sets of the form $\{\bar{X}(\bar{J}_m) = j_m \text{ for } 0 \leq m < n\}$, we may and will assume that A itself is of this form. If $R_{j_m} > 0$ for each $0 \leq m < n$, then $A \in \sigma\big(\{X(\sigma) : \sigma \in [0, J_n)\}\big)$, and $\big(\bar{J}_\ell, \bar{X}(\bar{J}_\ell)\big) = \big(J_\ell, X(J_\ell)\big)$ for $0 \leq \ell \leq n$ on A. Thus (*) holds in this case. If $j_\ell \in \mathbb{S}_0$ for some $0 \leq \ell < n$, use m to denote the first such ℓ, set $i = j_m$, and let $\{\bar{J}_\ell^{(i)} : \ell \geq 0\}$ be the jump times of $t \rightsquigarrow \bar{X}^{(i)}(t)$. Then we can write $A = B \cap C$, where $B = \{X(J_\ell) = j_\ell \text{ for } 0 \leq \ell \leq m\}$ and

$$C = \begin{cases} \{\bar{X}^{(i)}(\bar{J}_{\ell-m}^{(i)}) = j_\ell \text{ for } m < \ell < n - 1\} & \text{if } 0 \leq m < n - 1 \\ \Omega & \text{if } m = n - 1. \end{cases}$$

In addition, on A, $\big(\bar{J}_\ell, \bar{X}(\bar{J}_\ell)\big) = \big(\bar{J}^{(i)}_{\ell-m}, \bar{X}^{(i)}(\bar{J}^{(i)}_{\ell-m})\big)$ for $\ell > m$. Hence,

$$\begin{aligned}
&\mathbb{P}\Big(\big\{\bar{J}_n > \bar{J}_{n-1} + t \;\&\; \bar{X}(\bar{J}_n) = j\big\} \cap A\Big) \\
&\quad = \mathbb{P}\Big(\big\{\bar{J}^{(i)}_{n-m} > \bar{J}^{(i)}_{n-m-1} + t \;\&\; \bar{X}^{(i)}(\bar{J}^{(i)}_{n-m}) = j\big\} \cap C\Big)\mathbb{P}(B) \\
&\quad = \mathbb{E}\Big[\exp\big(-t\bar{R}_{\bar{X}^{(i)}(\bar{J}^{(i)}_{n-m-1})}\big)(\mathbf{P})_{\bar{X}^{(i)}(\bar{J}^{(i)}_{n-m-1})\,j},\, C\Big]\mathbb{P}(B) \\
&\quad = \mathbb{E}\Big[e^{-t\bar{R}_{\bar{X}(\bar{J}_{n-1})}}(\mathbf{P})_{\bar{X}(\bar{J}_{n-1})\,j},\, A\Big],
\end{aligned}$$

and so (*) holds in this case also.

In view of the preceding, we know that the distribution of $\{X(t) : t \geq 0\}$ is uniquely determined by (4.2.1) plus its initial distribution, and, in fact, that its distribution is the same as the process determined by (4.2.3). Of course, in the degenerate case, although (4.2.3) holds, the exponential random variables $\{E_n : n \geq 1\}$ and the Markov chain $\{X_n : n \geq 0\}$ will *not* be functions of the process $\{X(t) : t \geq 0\}$. However, now that we have proved the uniqueness of their distribution, there is no need to let this bother us, and so we may and will always assume that our process is given by (4.2.3).

4.2.2. The Markov Property: We continue with the assumption that the rates are bounded, and we now want to check that the process $\{X(t) : t \geq 0\}$ described above possesses the Markov property:

$$\begin{aligned}
&\mathbb{P}\big(X(s+t) = j \,\big|\, X(\tau),\, \tau \in [0,s]\big) = \big(\mathbf{P}(t)\big)_{X(s)j} \\
&\qquad \text{where } \big(\mathbf{P}(t)\big)_{ij} \equiv \mathbb{P}\big(X(t) = j \,\big|\, X(0) = i\big).
\end{aligned} \tag{4.2.4}$$

For this purpose, it will be useful to have checked that

$$\begin{aligned}
&\xi_m \leq s < \xi_{m+1} \implies \Phi^{\mathfrak{R},\mathbf{P}}\big(s+t; (e_1, \ldots, e_n, \ldots), (j_0, \ldots, j_n, \ldots)\big) = \\
&\Phi^{\mathfrak{R},\mathbf{P}}\big(t; (e_{m+1} - R_{j_m}(s - \xi_m), e_{m+2} \ldots, e_{m+n}, \ldots), (j_m, \ldots, j_{m+n}, \ldots)\big)
\end{aligned} \tag{4.2.5}$$

Now, let $A \in \sigma\big(\{X(\tau) : \tau \in [0,s]\}\big)$ be given, and assume that $X(s) = i$ on A. What we need to do is verify that

$$(*) \qquad \mathbb{P}\big(\{X(s+t) = j\} \cap A\big) = \big(\mathbf{P}(t)\big)_{ij}\mathbb{P}(A).$$

To this end, set $A_m \equiv A \cap \{N(s) = m\} = \{E_{m+1} > R_i(s - J_m)\} \cap B_m$, where $\{J_m \leq s\} \supseteq B_m \in \sigma\big(\{E_1, \ldots, E_m\} \cup \{X_0, \ldots, X_m\}\big)$. Then

$$\begin{aligned}
\mathbb{P}\big(\{X(s+t) = j\} \cap A\big) &= \sum_{m=0}^{\infty} \mathbb{P}\big(\{X(s+t) = j\} \cap A_m\big) \\
&= \sum_{m=0}^{\infty} \mathbb{P}\big(\{X(s+t) = j \;\&\; E_{m+1} > R_i(s - J_m)\} \cap B_m\big)
\end{aligned}$$

and, by (4.2.3), (4.2.5), and (4.1.3),

$$\begin{aligned}
&\mathbb{P}\big(\{X(s+t)=j \;\&\; E_{m+1} > R_i(s-J_m)\}\cap B_m\big)\\
&=\mathbb{P}\Big(\big\{\Phi^{\mathfrak{R},\mathbf{P}}(t;(E_{m+1}-R_i(s-J_m),E_{m+2},\dots,E_{m+n},\dots),\\
&\qquad\qquad (i,X_{m+1},\dots,X_{m+n},\dots)\big)=j\big\}\cap\{E_{m+1}>R_i(s-J_m)\}\cap B_m\Big)\\
&=\mathbb{P}\big(X(t)=j\,\big|\,X(0)=i\big)\mathbb{E}\Big[e^{-R_i(s-J_m)},\,B_m\Big]=\big(\mathbf{P}(t)\big)_{ij}\mathbb{P}(A_m),
\end{aligned}$$

from which (*) is now an immediate consequence.

Just as (4.1.8) implied the semigroup property (cf. (4.1.9)) for the transition probability matrices there, so (4.2.4) implies the family $\{\mathbf{P}(t) : t \geq 0\}$ is a semigroup. In fact, the proof is the same as it was there:

$$\begin{aligned}
\big(\mathbf{P}(t)\big)_{ij}&=\sum_{k\in\mathbb{S}}\mathbb{P}\big(X(s+t)=j\;\&\;X(s)=k\,\big|\,X(0)=i\big)\\
&=\sum_{k\in\mathbb{S}}\big(\mathbf{P}(t)\big)_{kj}\mathbb{P}\big(X(s)=k\big)=\sum_{k\in\mathbb{S}}\big(\mathbf{P}(t)\big)_{kj}\big(\mathbf{P}(s)\big)_{ik}=\big(\mathbf{P}(s)\mathbf{P}(t)\big)_{ij}.
\end{aligned}$$

In addition, it is important to realize that, at least in principle, *the distribution of a Markov process is uniquely determined by its initial distribution together with the semigroup of transition probability matrices* $\{\mathbf{P}(t) : t > 0\}$. To be precise, suppose that $\{X(t) : t \geq 0\}$ is a family of $\mathbb{S}$-valued random variables for which (4.2.4) holds, and let $\boldsymbol{\mu}$ be its initial distribution (i.e., the distribution of $X(0)$.) Then, for all $n \geq 1$, $0 = t_0 < t_1 < \cdots < t_n$, and $j_0,\dots,j_n \in \mathbb{S}$,

$$\begin{aligned}
&\mathbb{P}\big(X(t_m)=j_m \text{ for } 0\leq m\leq n\big)\\
&\qquad=(\boldsymbol{\mu})_{j_0}\big(\mathbf{P}(t_1-t_0)\big)_{j_0j_1}\cdots\big(\mathbf{P}(t_n-t_{n-1})\big)_{j_{n-1}j_n}.
\end{aligned}\tag{4.2.6}$$

To verify this, first note that, by (4.2.4),

$$\mathbb{P}\big(X(t_0)=j_0\;\&\;X(t_1)=j_1\big)=\big(\mathbf{P}(t_1)\big)_{j_0j_1}(\boldsymbol{\mu})_{j_0}=(\boldsymbol{\mu})_{j_0}\big(\mathbf{P}(t_1-t_0)\big)_{j_0j_1}.$$

Thus (4.2.6) holds for $n = 1$. Now let $n \geq 2$ be given, assume that (4.2.6) holds for $n-1$, and set $A = \{X(t_m) = j_m : 0 \leq m \leq n-1\}$. Then, by (4.2.4),

$$\begin{aligned}
&\mathbb{P}\big(X(t_m)=j_m \text{ for } 0\leq m\leq n\big)\\
&\quad=\mathbb{P}\big(\{X(t_n)=j_n\}\cap A\big)=\big(\mathbf{P}(t_n-t_{n-1})\big)_{j_{n-1}j_n}\mathbb{P}(A),
\end{aligned}$$

and so (4.2.6) follows from the induction hypothesis. Finally, by the application of Theorem 6.1.6 given at the end of §6.1.4, the distribution of $\{X(t) : t \geq 0\}$ is completely determined by the probabilities it assigns to sets of the form $\{X(t_m) = j_m \text{ for } 0 \leq m \leq n\}$, and so our uniqueness assertion has now been justified.

4.2.3. The Q-Matrix and Kolmogorov's Backward Equation: As we saw at the end of the preceding section, apart from its initial distribution, the distribution of a Markov process is completely determined by the semigroup $\{\mathbf{P}(t) : t \geq 0\}$. Thus, it is important to develop methods for calculating the transition probabilities $\mathbf{P}(t)$ directly from the data contained in the rates $\mathfrak{R}$ and the transition probability $\mathbf{P}$.

Based on ones experience with real valued functions, one should suspect that (4.1.9) means that $\mathbf{P}(t)$ must to be expressible as $e^{t\mathbf{Q}}$ for some $\mathbf{Q}$. In fact, $\mathbf{Q}$ ought to be obtainable by differentiating $t \rightsquigarrow \mathbf{P}(t)$ at $t = 0$.

As the first step in our program to give the substance to the preceding speculations, we will prove that

$$(*) \qquad \big(\mathbf{P}(t)\big)_{ij} = \delta_{i,j} e^{-tR_i} + R_i \int_0^t e^{-\tau R_i} \big(\mathbf{P}\mathbf{P}(t-\tau)\big)_{ij}\, d\tau.$$

Clearly, there is nothing to do when $R_i = 0$, and so we now assume that $R_i > 0$. Because

$$\big(\mathbf{P}(t)\big)_{ij} = \delta_{ij}\mathbb{P}\big(E_1 > tR_i \,\big|\, X(0) = i\big) + \mathbb{P}\big(E_1 \leq tR_i \;\&\; X(t) = j \,\big|\, X(0) = i\big),$$

and (cf. (4.2.3) and (4.2.5))

$$\begin{aligned}
&\mathbb{P}\big(E_1 \leq tR_i \;\&\; X(t) = j \,\big|\, X(0) = i\big) \\
&= \mathbb{P}\Big(\Phi^{\mathfrak{R},\mathbf{P}}\big(t - R_i^{-1}E_1; (E_2, \dots, E_n, \dots), (X_1, \dots, X_n, \dots)\big) = j \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \;\&\; E_1 \leq tR_i \,\Big|\, X_0 = i\Big) \\
&= \mathbb{E}\Big[\big(\mathbf{P}(t - R_i^{-1}E_1)\big)_{X_1 j},\, E_1 \leq R_i t \,\Big|\, X_0 = i\Big] \\
&= R_i \int_0^t e^{-\tau R_i} \sum_{k \in \mathbb{S}} (\mathbf{P})_{ik} \big(\mathbf{P}(t-\tau)\big)_{kj}\, d\tau.
\end{aligned}$$

we have completed the proof of (*).

The expression in (*) is an integrated version of a renowned equation due to Kolmogorov. Namely, when one differentiates (*) with respect to t, one arrives at *Kolmogorov's backward equation* :

$$\frac{d}{dt}\mathbf{P}(t)_{ij} = -R_i \mathbf{P}(t)_{ij} + R_i \big(\mathbf{P}\mathbf{P}(t)\big)_{ij},$$

which, when written in matrix notation, becomes

$$(4.2.7) \qquad \frac{d}{dt}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t) \quad \text{with } \mathbf{P}(0) = \mathbf{I} \quad \text{when } \mathbf{Q} = \mathbf{R}(\mathbf{P} - \mathbf{I}),$$

where $\mathbf{R}$ is the diagonal matrix whose ith diagonal entry is R_i. The reason for the adjective "backward" is that (4.2.7) describes the evolution of $t \rightsquigarrow \big(\mathbf{P}(t)\big)_{ij}$

as a function of time t and its *backward variable* i, so called because, if one adopts the perspective of someone traveling along the path $t \rightsquigarrow X(t)$, then i, being the place where he started, is the variable he sees when he is "looking backward." Unfortunately, the terminology for the matrix $\mathbf{Q}$ is much less inspired. Namely, probabilists call any matrix whose off-diagonal entries are non-negative and whose rows sum to 0 a *Q-matrix*.

4.2.4. Kolmogorov's Forward Equation: Recall the norm $\|\cdot\|_{\mathrm{u,v}}$ introduced at the beginning of §3.2.1.

Starting from (4.2.7), we have

$$\mathbf{P}(t) = \mathbf{I} + \int_0^t \mathbf{Q}\mathbf{P}(\tau)\,d\tau = \mathbf{I} + t\mathbf{Q} + \int_0^t (t-\tau)\mathbf{Q}^2\mathbf{P}(\tau)\,d\tau,$$

and so

$$\bigl\|\mathbf{P}(t) - \mathbf{I} - t\mathbf{Q}\bigr\|_{\mathrm{u,v}} \le \frac{\|\mathbf{Q}\|_{\mathrm{u,v}}^2 t^2}{2}. \tag{4.2.8}$$

Because, by the semigroup property (cf. (4.1.9)),

$$\mathbf{P}(t+h) - \mathbf{P}(t) - h\mathbf{Q}\mathbf{P}(t) = \mathbf{P}(t)\bigl(\mathbf{P}(h) - \mathbf{I} - h\mathbf{Q}\bigr),$$

we can pass from the above to *Kolmogorov's forward equation*

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q} \quad \text{with } \mathbf{P}(0) = \mathbf{I}, \tag{4.2.9}$$

so called because it describes the evolution of $t \rightsquigarrow \mathbf{P}(t)$ as a function of the *forward variable* j: the variable which the traveler sees when he "looks forward" in time.

4.2.5. Solving Kolmogorov's Equation: As we mentioned earlier, (4.1.9) suggests that $\mathbf{P}(t) = e^{t\dot{\mathbf{P}}(0)}$, and, thanks to (4.2.7), we now know that $\dot{\mathbf{P}}(0) = \mathbf{Q}$. That is, we are guessing that $\mathbf{P}(t) = e^{t\mathbf{Q}}$, where the meaning of the exponential is given by the power series

$$e^{\mathbf{M}} \equiv \sum_{m=0}^{\infty} \frac{\mathbf{M}^m}{m!} \quad \text{for } \mathbf{M} \in M_{\mathrm{u,v}}(\mathbb{S}). \tag{4.2.10}$$

Because $\|\mathbf{M}^m\|_{\mathrm{u,v}} \le \|\mathbf{M}\|_{\mathrm{u,v}}^m$, there is no question that the preceding series converges for every $\mathbf{M} \in M_{\mathrm{u,v}}(\mathbb{S})$. In fact, because

$$\left\| e^{\mathbf{M}} - \sum_{m=0}^{n-1} \frac{\mathbf{M}^m}{m!} \right\|_{\mathrm{u,v}} \le \sum_{m=n}^{\infty} \frac{\|\mathbf{M}\|_{\mathrm{u,v}}^m}{m!},$$

$$\left\| e^{\mathbf{M}} - \sum_{m=0}^{n-1} \frac{\mathbf{M}^m}{m!} \right\|_{\mathrm{u,v}} \leq \frac{\|\mathbf{M}\|_{\mathrm{u,v}}^n}{n!} e^{\|\mathbf{M}\|_{\mathrm{u,v}}}. \tag{4.2.11}$$

Also, it is easy (cf. Exercise 4.5.2 below) to check that

$$e^{\mathbf{M}_1+\mathbf{M}_2} = e^{\mathbf{M}_1} e^{\mathbf{M}_2} \quad \text{for } \mathbf{M}_1,\ \mathbf{M}_2 \in M_{\mathrm{u,v}}(\mathbb{S}) \text{ which commute.} \tag{4.2.12}$$

In particular, this means that $t \rightsquigarrow e^{t\mathbf{Q}}$ has the semigroup property: $e^{(s+t)\mathbf{Q}} = e^{s\mathbf{Q}} e^{t\mathbf{Q}}$. Finally, because

$$\left\| e^{h\mathbf{Q}} - \mathbf{I} - h\mathbf{Q} \right\|_{\mathrm{u,v}} \leq \frac{\|\mathbf{Q}\|_{\mathrm{u,v}}^2 h^2}{2} e^{h\|\mathbf{Q}\|_{\mathrm{u,v}}},$$

we see that

$$\left\| \frac{e^{(t+h)\mathbf{Q}} - e^{t\mathbf{Q}}}{h} - \mathbf{Q}e^{t\mathbf{Q}} \right\|_{\mathrm{u,v}} = \left\| e^{t\mathbf{Q}} \left(\frac{e^{h\mathbf{Q}} - \mathbf{I}}{h} - \mathbf{Q} \right) \right\|_{\mathrm{u,v}} \leq \frac{\|\mathbf{Q}\|_{\mathrm{u,v}}^2 h e^{(t+h)\|\mathbf{Q}\|_{\mathrm{u,v}}}}{2}$$

and so $\frac{d}{dt} e^{t\mathbf{Q}} = e^{t\mathbf{Q}}\mathbf{Q}$.

With these preparations, it is an easy matter to complete the identification

$$\mathbf{P}(t) = e^{t\mathbf{Q}} \equiv \sum_{m=0}^{\infty} \frac{t^m \mathbf{Q}^m}{m!} \quad \text{for all } t \in [0, \infty). \tag{4.2.13}$$

Namely, by the preceding and (4.2.7), we see that, for any $t > 0$ and $\tau \in [0, t]$,

$$\frac{d}{d\tau} e^{(t-\tau)\mathbf{Q}} \mathbf{P}(\tau) = e^{(t-\tau)\mathbf{Q}} (\mathbf{Q} - \mathbf{Q}) \mathbf{P}(\tau) = \mathbf{0},$$

and so $\tau \in [0, t] \longmapsto e^{(t-\tau)\mathbf{Q}} \mathbf{P}(\tau) \in M_{\mathrm{u,v}}(\mathbb{S})$ is constant.

Before closing this discussion, there is an important point which should be addressed. Namely, we have proved, via (4.2.13), that, for each $t \geq 0$, $e^{t\mathbf{Q}}$ is a transition probability matrix, although this fact is not at all evident from the power series expression for $e^{t\mathbf{Q}}$. In particular, without further considerations, it is far from clear why the entries of $e^{t\mathbf{Q}}$ are non-negative or why $\|e^{t\mathbf{Q}}\|_{\mathrm{u,v}} = 1$, independent of $\|\mathbf{Q}\|_{\mathrm{u,v}}$. Thus, we will now provide another way to see these properties, one that does not require the identification of $e^{t\mathbf{Q}}$ as $\mathbf{P}(t)$. Namely, set

$$\hat{\mathbf{P}}_{ij} = \begin{cases} \frac{Q_{ij}}{M} & \text{if } i \neq j \\ 1 + \frac{Q_{ii}}{M} & \text{if } i = j \end{cases} \quad \text{where } M \equiv \sup_i R_i = \sup_i \bigl(-(\mathbf{Q})_{ii}\bigr).$$

Then $\hat{\mathbf{P}}$ is a transition probability matrix, and $\mathbf{Q} = M(\hat{\mathbf{P}} - \mathbf{I})$. Hence, because $\mathbf{I}$ commutes with $\hat{\mathbf{P}}$,

$$e^{t\mathbf{Q}} = e^{-tM\mathbf{I}} e^{tM\hat{\mathbf{P}}} = e^{-tM} \sum_{m=0}^{\infty} \frac{(tM)^m}{m!} \hat{\mathbf{P}}^m.$$

The non-negativity of the entries of $e^{t\mathbf{Q}}$ is obvious from this representation as is the fact that

$$\sum_{j\in\mathbb{S}}\left(e^{t\mathbf{Q}}\right)_{ij} = e^{-tM}\sum_{m=1}^{\infty}\frac{(tM)^m}{m!}\left(\sum_{j\in\mathbb{S}}(\hat{\mathbf{P}}^m)_{ij}\right) = 1.$$

In order to appreciate what is going on here, it is well to notice that this line of reasoning works only because $t \geq 0$: when $t < 0$, the entries of $e^{t\mathbf{Q}}$ need not be non-negative and $\|e^{t\mathbf{Q}}\|_{\mathrm{u,v}}$ may be as large as $e^{|t|\|\mathbf{Q}\|_{\mathrm{u,v}}}$.

4.2.6. A Markov Process from its Infinitesimal Characteristics: In some applications the most natural way to describe a Markov process $\{X(t) : t \geq 0\}$ in terms of its rates $\mathfrak{R}$ and the underlying transition probability $\mathbf{P}$ is to specify its initial distribution and say that, given its past $\sigma\big(\{X(\tau) : \tau \in [0,t]\}\big)$ up until time $t \geq 0$, the probability of its having moved away from $X(t)$ at time $t+h$ (where $h > 0$ is small) will be approximately $hR_{X(t)}$, and given $\sigma\big(\{X(\tau) : \tau \in [0,t]\} \cup \{X(t+h) \neq X(t)\}\big)$ (i.e., both its past and that it has moved) the probability that $X(t+h) = j$ is approximately $(\mathbf{P})_{X(t)j}$. In such a description, $\mathfrak{R}$ and $\mathbf{P}$ become the *infinitesimal characteristics* of the process and the question is whether the distribution of the process can be reconstructed from this description. However, before attempting such a reconstruction, we will make the description more quantitative by insisting that there exist an $\epsilon : (0,\infty) \longrightarrow (0,\infty)$ which tends to 0 at 0 and for which

$$\left|\mathbb{P}\big(X(t+h) \neq X(t) \,\big|\, X(\tau),\, \tau \in [0,t]\big) - hR_{X(t)}\right| \leq h\epsilon(h)$$

and

$$\left|\mathbb{P}\big(X(t+h) = j \,\big|\, X(\tau),\, \tau \in [0,t],\ \&\ X(t+h) \neq X(t)\big) - (\mathbf{P})_{X(t)j}\right| \leq \epsilon(h).$$

Clearly, the first of these is equivalent to

$$\left|\mathbb{P}\big(X(t+h) = X(t) \,\big|\, X(\tau),\, \tau \in [0,t]\big) - 1 - h(\mathbf{Q})_{X(t)X(t)}\right| \leq h\epsilon(h),$$

whereas the two together lead to

$$\left|\mathbb{P}\big(X(t+h) = j \,\big|\, X(\tau),\, \tau \in [0,t]\big) - h(\mathbf{Q})_{X(t)j}\right| \leq h\epsilon(h)\big(R_{X(t)} + \epsilon(h)\big)$$

when $j \neq X(t)$. Hence, because we are assuming that the rates are bounded, the above description implies that

$$(*) \qquad \left|\mathbb{P}\big(X(t+h) = j \,\big|\, X(\tau),\, \tau \in [0,t]\big) - \delta_{X(t),j} - h(\mathbf{Q})_{X(t)j}\right| \leq h\epsilon'(h),$$

where, again, $\epsilon'(h) \longrightarrow 0$ when $h \searrow 0$.

We will now show that (*) is sufficient to show that (4.2.4) is satisfied with the $\mathbf{P}(t) = e^{t\mathbf{Q}}$ and therefore, by the result discussed toward the end of §4.2.2, that $\{X(t) : t \geq 0\}$ is a Markov process corresponding to rates $\mathfrak{R}$ and transition probability matrix $\mathbf{P}$. To this end, given $s \geq 0$ and $A \in \sigma\big(\{X(\tau) : \tau \in [0,s]\}\big)$, determine the row vectors $\{\boldsymbol{\mu}(t) : t \geq 0\}$ so that $\big(\boldsymbol{\mu}(t)\big)_j = \mathbb{P}\big(\{X(s+t) = j\} \cap A\big)$ for each $j \in \mathbb{S}$. Then, because

$$\big(\boldsymbol{\mu}(t+h)\big)_j = \mathbb{E}\Big[\mathbb{P}\big(X(t+h) = j \,\big|\, X(\tau),\, \tau \in [0,t]\big),\, A\Big]$$

and $\big(\boldsymbol{\mu}(t)\big)_j = \mathbb{E}\big[\delta_{X(t),j},\, A\big]$, (*) tells us that

$$\begin{aligned} h\epsilon'(h) &\geq \Big|\big(\boldsymbol{\mu}(t+h)\big)_j - \big(\boldsymbol{\mu}(t)\big)_j - h\mathbb{E}\big[(\mathbf{Q})_{X(s+t)j},\, A\big]\Big| \\ &= \Big|\big(\boldsymbol{\mu}(t+h)\big)_j - \big(\boldsymbol{\mu}(t)\big)_j - h\big(\boldsymbol{\mu}(t)\mathbf{Q}\big)_j\Big| \quad \text{for } t \geq 0 \text{ and } h > 0. \end{aligned}$$

Hence, $\frac{d}{dt}\boldsymbol{\mu}(t) = \boldsymbol{\mu}(t)\mathbf{Q}$ for $t > 0$, and so

$$\frac{d}{d\tau}\boldsymbol{\mu}(\tau)e^{(t-\tau)\mathbf{Q}} = \big(\boldsymbol{\mu}(\tau)\mathbf{Q} - \boldsymbol{\mu}(\tau)\mathbf{Q}\big)e^{(t-\tau)\mathbf{Q}} = \mathbf{0} \quad \text{for } \tau \in (0,t),$$

from which we conclude that

$$\mathbb{P}\big(\{X(s+t) = j\} \cap A\big) = \boldsymbol{\mu}(t) = \boldsymbol{\mu}(0)e^{t\mathbf{Q}} = \boldsymbol{\mu}(0)\mathbf{P}(t),$$

which is equivalent to

$$\mathbb{P}\big(\{X(t) = j\} \cap A\big) = \mathbb{E}\big[\big(\mathbf{P}(t)\big)_{X(s)j},\, A\big].$$

That is, we have now shown that (*) implies (4.2.4), and therefore, by the final part of §4.2.2, we see that the distribution of $\{X(t) : t \geq 0\}$ is that of a Markov process corresponding to rates $\mathfrak{R}$ and transition probability $\mathbf{P}$.

4.3 Unbounded Rates

Thus far we have been assuming that the rates $\mathfrak{R} = \{R_i : i \in \mathbb{S}\}$ are bounded, and the essential application which we made of this assumption came in §4.2.1. Namely, a bound on the rates guaranteed that the $J_n \nearrow \infty$ with probability 1 and therefore that our Markov process was completely determined for all time. As we will see below (cf. Exercises 4.5.5 and 4.5.6), when the rates are unbounded, $J_\infty \equiv \lim_{n\to\infty} J_n$ may be finite with positive probability, in which case our description fails to tell the process what to do during the time interval $[J_\infty, \infty)$. In this section, we will give conditions, other than a bound on $\mathfrak{R}$, which guarantee that $J_\infty = \infty$ with probability 1. In addition, we will give a very cursory discussion of what one can do to salvage the situation when $J_\infty < \infty$ with positive probability.

4.3.1. Explosion: The setting here is the same as the one in §4.2.1, only now we are no longer assuming that the rates are bounded. Thus, if $t \rightsquigarrow X(t)$ is given by the prescription in (4.2.3), J_n denotes the time of the nth jump (i.e., the nth t for which $X(t) \neq X(t-)$), and $J_\infty \equiv \lim_{n\to\infty} J_n$, then the distribution of $t \rightsquigarrow X(t)$ is uniquely determined only until time J_∞.

Our first step is to show that, with probability 1, J_∞ coincides with the time when the process explodes out of the state space. To be precise, choose an *exhaustion* $\{F_N : N \geq 1\}$ of $\mathbb{S}$ by non-empty, finite subsets. That is, for each N, F_N is a non-empty, finite subset of $\mathbb{S}$, $F_N \subseteq F_{N+1}$, and $\mathbb{S} = \bigcup_N F_N$. Next, take $\mathfrak{R}^{(N)}$ to be the set of rates given by

$$R_i^{(N)} = \begin{cases} R_i & \text{if } i \in F_N \\ 0 & \text{if } i \notin F_N. \end{cases}$$

Then, for each $N \geq 1$, (4.2.3) with $\mathfrak{R}^{(N)}$ replacing $\mathfrak{R}$ determines a Markov process $\{X^{(N)}(t) : t \geq 0\}$. Moreover, if $\zeta_N \equiv \inf\{t \geq 0 : X^{(N)}(t) \notin F_N\}$, then $X^{(N+1)}(t) = X^{(N)}(t)$ for $t \in [0, \zeta_N] \cap [0, \infty)$. Hence, $\zeta_N \leq \zeta_{N+1}$, and so the *explosion time* $\mathfrak{e} \equiv \lim_{N\to\infty} \zeta_N$ exists (in $[0, \infty]$).

4.3.1 THEOREM. *$\mathfrak{e} = J_\infty$ with probability 1, and, for each $N \geq 1$, $X^{(N)}(t) = X(t \wedge \zeta_N)$ for $t \in [0, \infty)$. In particular, if $\mathbb{P}(\mathfrak{e} = \infty) = 1$, then the distribution of $X(0)$ together with the description in (4.2.1) uniquely determine the distribution of a process $\{X(t) : t \geq 0\}$.*

PROOF: Because $\{\mathfrak{e} \neq J_\infty\}$ can be written as the union of the sets $\{\mathfrak{e} > T \geq J_\infty\} \cup \{J_\infty > T \geq \mathfrak{e}\}$ as T runs over the positive rationals, in order to prove that $\mathfrak{e} = J_\infty$ with probability 1, (6.1.5) says that it suffices for us to show that $\mathbb{P}(\mathfrak{e} > T \geq J_\infty) = 0 = \mathbb{P}(J_\infty > T \geq \mathfrak{e})$ for each $T > 0$. To this end, first suppose that $\mathbb{P}(\mathfrak{e} > T \geq J_\infty) > 0$ for some T. Then there exists an N such that $\mathbb{P}(\zeta_N > T \geq J_\infty) > 0$. On the other hand, $\zeta_N > T \geq J_\infty \implies T \geq J_\infty \geq r_N^{-1} \sum_1^\infty E_m$, where $r_N = \sup_{j\in F_N} R_j$, and therefore we are led to the contradiction

$$0 < \mathbb{P}(\zeta_N > T \geq J_\infty) \leq \mathbb{P}\left(\sum_{m=1}^{\infty} E_m \leq r_N T\right) = 0.$$

Next suppose that $\mathbb{P}(J_\infty > T \geq \mathfrak{e}) > 0$. Then there exists an $n \geq 1$ such that $\mathbb{P}(J_n > T \geq \mathfrak{e}) > 0$. On the other hand, if $\boldsymbol{\mu}$ is the distribution of $X(0)$, then $\sum_{m=0}^{n} \sum_{j\notin F_N} (\boldsymbol{\mu}\mathbf{P}^m)_j \longrightarrow 0$ as $N \to \infty$, and so,

$$\begin{aligned} \mathbb{P}\big(J_n > T \geq \mathfrak{e}\big) &\leq \mathbb{P}\big(J_n > T \geq \zeta_N\big) \leq \mathbb{P}\big(\exists\, 0 \leq m \leq n \; X_m \notin F_N\big) \\ &\leq \sum_{m=0}^{n} \sum_{j\notin F_N} (\boldsymbol{\mu}\mathbf{P}^m)_j \longrightarrow 0 \quad \text{as } N \to \infty. \end{aligned}$$

That is, we know that $\mathbb{P}(J_n > T \geq \mathfrak{e}) = 0$.

Given the preceding, it should be clear that $X(t) = X^{(N)}(t)$ as long as $t \in [0, \zeta_N)$ and that $X(\zeta_N) = X^{(N)}(\zeta_N)$ if $\zeta_N < \infty$. Hence, $X(t \wedge \zeta_N) = X^{(N)}(t)$ for all $N \in \mathbb{N}$ and $t \geq 0$. Finally, because (4.2.1), with $\mathfrak{R}^{(N)}$ replacing $\mathfrak{R}$, together with the initial distribution uniquely determine the distribution of $\{X^{(N)}(t) : t \geq 0\}$, it follows that, for each $N \in \mathbb{N}$, the distribution of $\{X(t \wedge \zeta_N) : t \geq 0\}$ is uniquely determined by the initial distribution and (4.2.1), and therefore, when $\mathbb{P}(\mathfrak{e} = \infty) = 1$, so is the distribution of $\{X(t) : t \geq 0\}$. □

4.3.2 COROLLARY. *If* $\mathbb{P}(\mathfrak{e} = \infty | X(0) = i) = 1$ *for all* $i \in \mathbb{S}$ *and* $\big(\mathbf{P}(t)\big)_{ij} \equiv \mathbb{P}(X(t) = j | X(0) = i)$, *then, for each initial distribution* $\boldsymbol{\mu}$ *and* $T > 0$,

$$\lim_{N \to \infty} \sup_{t \in (0,T]} \|\boldsymbol{\mu}\mathbf{P}^{(N)}(t) - \boldsymbol{\mu}\mathbf{P}(t)\|_{\mathrm{v}} = 0, \tag{4.3.3}$$

where $\{\mathbf{P}^{(N)} : t > 0\}$ *is the semigroup determined by* $\mathfrak{R}^{(N)}$ *and* $\mathbf{P}$. *Moreover,* $\{X(t) : t \geq 0\}$ *satisfies the Markov property in* (4.2.4). *Finally,* $\{\mathbf{P}(t) : t \geq 0\}$ *is a semigroup which satisfies Kolmogorov's backward equation in the sense that, for each* $(i, j) \in \mathbb{S}^2$,

$$\big(\mathbf{P}(t)\big)_{ij} = \delta_{i,j} + \int_0^t \big(\mathbf{Q}\mathbf{P}(\tau)\big)_{ij}\, d\tau. \tag{4.3.4}$$

PROOF: First note that

$$\mathbb{P}(\mathfrak{e} = \infty) = \sum_{i \in \mathbb{S}} (\boldsymbol{\mu})_i \mathbb{P}\big(\mathfrak{e} = \infty \,\big|\, X(0) = i\big) = 1,$$

and therefore that $\lim_{N \to \infty} \mathbb{P}(\zeta_N \leq T) = 0$ for each $T \in [0, \infty)$. At the same time, by Theorem 4.3.1,

$$\big|\big(\boldsymbol{\mu}\mathbf{P}^{(N)}(t)\big)_j - \big(\boldsymbol{\mu}\mathbf{P}(t)\big)_j\big| \leq \mathbb{P}\big(X(t) = j \;\&\; \zeta_N \leq T\big)$$

for all $0 \leq t \leq T$ and $j \in \mathbb{S}$. Thus,

$$\sup_{t \in (0,T]} \big\|\boldsymbol{\mu}\mathbf{P}^{(N)}(t) - \boldsymbol{\mu}\mathbf{P}(t)\big\|_{\mathrm{v}} \leq \mathbb{P}\big(\zeta_N \leq T\big) \longrightarrow 0 \quad \text{as } N \to \infty.$$

To prove the Markov property (4.2.4), let $A \in \sigma\big(\{X(\tau) : \tau \in [0, s]\}\big)$ be given, and assume that $X(s) = i$ on A. Then, since $A \cap \{\zeta_N > s\} \in \sigma\big(\{X^{(N)}(\tau) : \tau \in [0, s]\}\big)$, the Markov property for $\{X^{(N)}(t) : t \geq 0\}$ plus the result in Theorem 4.3.1 justify

$$\begin{aligned}\mathbb{P}\big(\{X(s+t) = j\} \cap A\big) &= \lim_{N \to \infty} \mathbb{P}\big(\{X^{(N)}(s+t) = j\} \cap A \cap \{\zeta_N > s\}\big) \\ &= \lim_{N \to \infty} \big(\mathbf{P}^{(N)}(t)\big)_{ij} \mathbb{P}\big(A \cap \{\zeta_N > s\}\big) = \big(\mathbf{P}(t)\big)_{ij} \mathbb{P}(A).\end{aligned}$$

Of course, once we know that (4.2.4) holds, then, by exactly the same argument with which we derived (4.1.9) in the bounded case, we know that the $\{\mathbf{P}(t) : t \geq 0\}$ here is also a semigroup.

To check (4.3.4), note that, by (4.2.7) applied to $\{\mathbf{P}^{(N)}(t) : t \geq 0\}$,

$$\big(\mathbf{P}^{(N)}(t)\big)_{ij} = \delta_{i,j} + \int_0^t \big(\mathbf{Q}\mathbf{P}^{(N)}(\tau)\big)_{ij}\, d\tau$$

as soon as N is large enough that $i \in F_N$. Hence, since $1 \geq \big(\mathbf{P}^{(N)}(\tau)\big)_{kj} \longrightarrow \big(\mathbf{P}(\tau)\big)_{kj}$ while $\sum_k |(\mathbf{Q})_{ik}| = 2R_i < \infty$, (4.3.4) follows by Lebesgue's Dominated Convergence Theorem. □

4.3.2. Criteria for Non-explosion or Explosion: In this subsection we will first develop two criteria which guarantee *non-explosion*: $\mathfrak{e} = \infty$ with probability 1. We will also give a condition which guarantees explosion with probability 1.

4.3.5 THEOREM. *Let* $\mathbf{P}$ *be a transition probability matrix satisfying* $(\mathbf{P})_{ii} = 0$ *for all* $i \in \mathbb{S}$, *and let* $\boldsymbol{\mu}$ *be a probability vector with the property that* $(\boldsymbol{\mu})_i = 0$ *unless* i *is recurrent for* $\mathbf{P}$. *Then for every choice of rates* $\mathfrak{R}$, *there is no explosion of the process described in (4.2.1) with initial distribution* $\boldsymbol{\mu}$.

PROOF: First observe that it is enough to handle the case when the rates are non-degenerate. Indeed, because what we are trying to check is that $J_\infty = \sum_1^\infty R_{X_m}^{-1} E_m = \infty$ with probability 1, it is clear that making the rates smaller can only make explosion less likely. In addition, because $\mathbb{P}(\mathfrak{e} = \infty) = \sum_{i \in \mathbb{S}} (\boldsymbol{\mu})_i \mathbb{P}\big(\mathfrak{e} = \infty | X(0) = i\big)$, it suffices for us to show that $\mathbb{P}\big(\mathfrak{e} = \infty | X(0) = i\big) = 1$ whenever i is recurrent for $\mathbf{P}$.

Because we are assuming that the rates are non-degenerate, (4.2.3) guarantees that the points visited by $\{X(t) : t \in [0, J_\infty)\}$ will be the same as the points visited by $\{X_n : n \geq 0\}$. In particular, for any $N \geq 1$,

$$\begin{aligned}&\mathbf{P}\big((\sigma^{(N)})_i < \zeta_N \,\big|\, X(0) = i\big)\\ &\quad = \theta_N \equiv \mathbb{P}\big(\exists\, n \geq 1 \; X_n = i \text{ and } X_m \in F_N \text{ for } 0 \leq m \leq n \,\big|\, X_0 = i\big),\end{aligned}$$

where $(\sigma^{(N)})_i$ is the first time after $J_1^{(N)}$ that $t \rightsquigarrow X^{(N)}(t)$ returns to i. At the same time, essentially the same argument as was used to prove Theorem 2.3.6 shows that $\mathbb{P}\big((\sigma^{(N)})_i^{(m)} < \zeta_N \,\big|\, X(0) = i\big) = \theta_N^m$, where $(\sigma^{(N)})_i^{(m)}$ is defined inductively so that $(\sigma^{(N)})_i^{(1)} = (\sigma^{(N)})_i$ and

$$(\sigma^{(N)})_i^{(m+1)} = \begin{cases} \inf\big\{t \geq J_{\ell+1}^{(N)} : X^{(N)} = i\big\} & \text{if } (\sigma^{(N)})_i^{(m)} = J_\ell^{(N)} < \infty \\ \infty & \text{if } (\sigma^{(N)})_i^{(m)} = \infty \end{cases}$$

and $\{J_n^{(N)} : n \geq 0\}$ are the jump times of $t \rightsquigarrow X^{(N)}(t)$. Thus, if i is recurrent for $\mathbf{P}$ and therefore $\theta_N \longrightarrow 1$ as $N \to \infty$, we know that

$$\lim_{N \to \infty} \mathbb{P}\big((\sigma^{(N)})_i^{(m)} < \zeta_N \,\big|\, X(0) = i\big) = 1 \quad \text{for each } m \geq 1.$$

On the other hand, since $(\sigma^{(N)})_i^{(1)} \geq J_1^{(N)}$ and

$$(\sigma^{(N)})_i^{(m)} = J_\ell^{(N)} \implies (\sigma^{(N)})_i^{(m+1)} - (\sigma^{(N)})_i^{(m)} \geq J_{\ell+1}^{(N)} - J_\ell^{(N)} = \frac{E_{\ell+1}}{R_i},$$

$$\mathbb{P}\big((\sigma^{(N)})_i^{(m)} \leq T \,\big|\, X(0) = i\big) \leq \mathbb{P}\left(\sum_{\ell=1}^{m} E_\ell \leq R_i T\right) \leq \frac{(TR_i)^m}{m!}.$$

But, for all $m \geq 1$ and $N \geq 1$,

$$\mathbb{P}\big(\zeta_N \leq T \,\big|\, X(0) = i\big) \leq \mathbb{P}\big(\zeta_N \leq (\sigma^{(N)})_i^{(m)} \,\big|\, X(0) = i\big) + \mathbb{P}\big((\sigma^{(N)})_i^{(m)} \leq T \,\big|\, X(0) = i\big),$$

and so, after first letting $N \to \infty$ and then $m \to \infty$, we see that, when i is recurrent for $\mathbf{P}$, $\mathbb{P}\big(\mathfrak{e} \leq T \,\big|\, X(0) = i\big) = 0$. □

We group our second non-explosion criterion together with our criterion for explosion as two parts of the same theorem. In this theorem, the process is determined by (4.2.1) with rates $\mathfrak{R}$ and transition probability $\mathbf{P}$.

4.3.6 THEOREM. *If there exists a non-negative function u on $\mathbb{S}$ with the properties that $U_N \equiv \inf_{j \notin F_N} u(j) \longrightarrow \infty$ as $N \to \infty$ and, for some $\alpha \in [0, \infty)$,*

$$\sum_{j \in \mathbb{S}} (\mathbf{P})_{ij} u(j) \leq \left(1 + \frac{\alpha}{R_i}\right) u(i) \quad \textit{whenever } i \in \mathbb{S} \textit{ and } R_i > 0,$$

then $\mathbb{P}(\mathfrak{e} = \infty | X(0) = i) = 1$ for all $i \in \mathbb{S}$. On the other hand, if, for some $i \in \mathbb{S}$, $R_j > 0$ whenever $i \to j$ and there exists a non-negative function u on $\mathbb{S}$ with the property that, for some $\epsilon > 0$,

$$\sum_{\{k:\, i \to k\}} (\mathbf{P})_{jk} u(k) \leq u(j) - \frac{\epsilon}{R_j} \quad \textit{whenever } i \to j,$$

then $\mathbb{P}\big(\mathfrak{e} = \infty | X(0) = i\big) = 0$.

PROOF: To prove the first part, for each $N \geq 1$, set $u^{(N)}(j) = u(j)$ when $j \in F_N$ and $u^{(N)}(j) = U_N$ when $j \notin F_N$. It is an easy matter to check that if $\mathbf{Q}^{(N)} = \mathbf{R}^{(N)}(\mathbf{P} - \mathbf{I})$, where $(\mathbf{R}^{(N)})_{ij} = R_i^{(N)} \delta_{i,j}$, and $\mathbf{u}^{(N)}$ is the column vector determined by $(\mathbf{u}^{(N)})_i = u^{(N)}(i)$, then, for all $i \in \mathbb{S}$, $\big(\mathbf{Q}^{(N)}\mathbf{u}^{(N)}\big)_i \leq \alpha u^{(N)}(i)$. Hence, by Kolmogorov's forward equation (4.2.9) applied to $\{\mathbf{P}^{(N)}(t)\, t \geq 0\}$,

$$\frac{d}{dt}\big(\mathbf{P}^{(N)}(t)\mathbf{u}^{(N)}\big)_i \leq \alpha\big(\mathbf{P}^{(N)}(t)\mathbf{u}^{(N)}\big)_i,$$

and so $\big(\mathbf{P}^{(N)}(T)\mathbf{u}^{(N)}\big)_i \leq e^{\alpha T} u^{(N)}(i)$. But, since

$$u^{(N)}\big(X^{(N)}(T)\big) = u^{(N)}\big(X^{(N)}(\zeta_N)\big) \geq U_N \quad \text{if } \zeta_N \leq T,$$

this means that

$$\mathbf{P}\big(\zeta_N \le T \,\big|\, X(0)=i\big) \le \frac{1}{U_N}\mathbb{E}\big[u^{(N)}\big(X^{(N)}(T)\big) \,\big|\, X(0)=i\big]$$
$$= \frac{\big(\mathbf{P}^{(N)}(T)\mathbf{u}^{(N)}\big)_i}{U_N} \le \frac{e^{\alpha T}u^{(N)}(i)}{U_N} \le \frac{e^{\alpha T}u(i)}{U_N} \quad \text{if } i \in F_N,$$

and so, by the Monotone Convergence Theorem, we have now proved that $\mathbf{P}(\mathfrak{e} \le T | X(0)=i) \le \lim_{N\to\infty}\mathbb{P}(\zeta_N \le T | X(0)=i) = 0$.

In proving the second assertion, we may and will assume that $i \to j$ for all $j \in \mathbb{S}$. Now take $u^{(N)}(j) = u(j)$ if $j \in F_N$ and $u^{(N)}(j) = 0$ if $j \notin F_N$, and use $\mathbf{u}^{(N)}$ to denote the associated column vector. Clearly $\big(\mathbf{Q}^{(N)}\mathbf{u}^{(N)}\big)_i$ is either less than or equal to $-\epsilon$ or equal to 0 according to whether $i \in F_N$ or $i \notin F_N$. Hence, again by Kolmogorov's forward equation,

$$\frac{d}{dt}\big(\mathbf{P}^{(N)}(t)\mathbf{u}^{(N)}\big)_i \le -\epsilon \sum_{j\in F_N}\big(\mathbf{P}^{(N)}(t)\big)_{ij},$$

and so

$$\mathbb{E}\big[u^{(N)}\big(X^{(N)}(T)\big) \,\big|\, X^{(N)}(0)=i\big] - u^{(N)}(i)$$
$$\le -\epsilon\mathbb{E}\left[\int_0^T \mathbf{1}_{F_N}\big(X^{(N)}(t)\big)\,dt \,\middle|\, X^{(N)}(0)=i\right] = -\epsilon\mathbb{E}\big[T\wedge\zeta_N \,\big|\, X^{(N)}(0)=i\big].$$

But, since $u^{(N)} \ge 0$, this means that $\mathbb{E}[\zeta_N | X^{(N)}(0)=i] \le \frac{u(i)}{\epsilon}$ for all N, and so $\mathbb{E}[\mathfrak{e} | X(0)=i] \le \frac{u(i)}{\epsilon} < \infty$. □

4.3.3. What to Do When Explosion Occurs: Although I do not intend to repent, I feel compelled to admit that we have ignored what, from the mathematical standpoint, is the most interesting aspect of the theory under consideration. Namely, so far we have said nothing about the options one has when explosion occurs with positive probability, and in this subsection we will discuss only the most banal of the many choices available.

If one thinks of $\{\mathfrak{e} < \infty\}$ as the event that the process escapes its state space in a finite amount of time, then continuation of the process after $\mathfrak{e}$ might entail the introduction of at least one *new state*. Indeed, the situation here is very similar to the one encountered by topologists when they want to "compactify" a space. The simplest compactification of a separable, locally compact space is the *one point compactification*, the compactification which recognizes escape to infinity but ignores all details about the route taken. The analog of one point compactification in the present context is, at the time of explosion, to send the process to an absorbing point $\Delta \notin \mathbb{S}$. That is, one defines $t \in [0,\infty) \longmapsto X(t) \in \mathbb{S}\cup\{\Delta\}$ by the prescription in (4.2.3) (remember that, by Theorem 4.3.1, $\mathfrak{e} = J_\infty$ with probability 1) as long as $t \in [0, J_\infty)$ and

takes $X(t) = \Delta$ for $t \in [J_\infty, \infty)$. For various reasons, none of which I will explain, this extension is called that the *minimal extension* of the process. The minimal extension has the virtue that it always works. On the other hand, it, like the one point compactification, has the disadvantage that it completely masks all the fine structure of any particular case. For example, when $\mathbb{S} = \mathbb{Z}$, explosion can occur because, given that $\mathfrak{e} < \infty$, $\lim_{t \nearrow \mathfrak{e}} X(t) = +\infty$ with probability 1 or because, although $\lim_{t \nearrow \mathfrak{e}} |X(t)| = \infty$ with probability 1, both $\lim_{t \nearrow \mathfrak{e}} X(t) = +\infty$ and $\lim_{t \nearrow \mathfrak{e}} X(t) = -\infty$ occur with positive probability. In the latter case, one might want to record which of the two possibilities occurred, and this could be done by introducing two new absorbing states, Δ_+ for those paths which escape via $+\infty$ and Δ_- for those which escape via $-\infty$.

Alternatively, rather than thinking of the explosion time as a time to banish the process from $\mathbb{S}$, one can turn it into a time of renewal by redistributing the process over $\mathbb{S}$ at time $\mathfrak{e}$ and running it again until it explodes, etc.

Obviously, there is an infinity of possibilities. Suffice it to say that the preceding discussion hardly scratches the surface.

4.4 Ergodic Properties

In this section we will examine the ergodic behavior of the Markov processes which we have been discussing in this chapter. Our running assumption will be that the process with which we are dealing is a continuous time Markov process of the sort described in §4.2.1 under the condition that there is no explosion.

4.4.1. Classification of States: Just as in §3.1, we begin by classifying the states of $\mathbb{S}$.

In order to make our first observation, write $\mathbf{Q} = \mathbf{R}(\mathbf{P} - \mathbf{I})$, where $\mathbf{R}$ is the diagonal matrix of rates from $\mathfrak{R}$ and $\mathbf{P}$ is a transition probability matrix whose diagonal entries are 0. Next, define $\mathbf{P}^{\mathfrak{R}}$ to be the transition probability matrix given by[5]

$$(\mathbf{P}^{\mathfrak{R}})_{ij} = \begin{cases} (\mathbf{P})_{ij} & \text{if } R_i > 0 \\ \delta_{ij} & \text{if } R_i = 0. \end{cases} \tag{4.4.1}$$

Obviously, it is again true that $\mathbf{Q} = \mathbf{R}(\mathbf{P}^{\mathfrak{R}} - \mathbf{I})$. In addition,

$$\begin{aligned} i \to j \text{ relative to } \mathbf{P}^{\mathfrak{R}} &\iff \exists\, n \geq 0\ (\mathbf{Q}^n)_{ij} > 0 \\ &\iff \big(\mathbf{P}(t)\big)_{ij} > 0 \text{ for all } t > 0. \end{aligned} \tag{4.4.2}$$

Because of (4.2.13), this is obvious when $\mathfrak{R}$ is bounded, and, in general, it follows from the bounded case plus (cf. the notation in §4.3.1) $\lim_{N \to \infty} \big(\mathbf{P}^{(N)}(t)\big)_{ij} = \big(\mathbf{P}(t)\big)_{ij}$.

[5] It is worth noting that, as distinguished from $\mathbf{P}$, $\mathbf{P}^{\mathfrak{R}}$ is completely determined by $\mathbf{Q}$.

On the basis of (4.4.2), we will write $i \overset{\mathbf{Q}}{\to} j$ when any one of the equivalent conditions in (4.4.2) holds, and we will say that i **Q**-*communicates* with j and will write $i \overset{\mathbf{Q}}{\leftrightarrow} j$ when $i \overset{\mathbf{Q}}{\to} j$ and $j \overset{\mathbf{Q}}{\to} i$. In this connection, $\mathbb{S}$ will be said to be **Q**-*irreducible* if all states communicate with all other states.

We turn next to the question of recurrence and transience. Because

$$R_i = 0 \implies \mathbb{P}\big(X(t) = i \text{ for all } t \geq 0 \,\big|\, X(0) = i\big) = 1,$$

it is obvious that i should be considered recurrent when $R_i = 0$. On the other hand, since in the continuous time setting there is no "first positive time," it is less obvious what should be meant by recurrence of i when $R_i > 0$. However, if one adopts the attitude that time 1 in the discrete time context represents the first time that the chain can move, then it becomes clear that the role of 1 there should be played here by the first jump time J_1, and therefore that the role of ρ_i is played by

$$\sigma_i = \inf\{t \geq J_1 : X(t) = i\}. \tag{4.4.3}$$

Hence, we will say that i is **Q**-*recurrent* if either $R_i = 0$ or $\mathbb{P}(\sigma_i < \infty | X(0) = i) = 1$, and we will say that it is **Q**-*transient* if it is not **Q**-recurrent.

Our next observation is that i *is* **Q**-*recurrent if and only if it is recurrent with respect to the transition probability matrix* $\mathbf{P}^{\mathfrak{R}}$ in (4.4.1). Indeed, as is evident from the discussion in §4.2.1, the points visited by $t \rightsquigarrow X(t)$ are exactly the same as those visited by $n \rightsquigarrow X_n = X(J_n)$, and $\{X_n : n \geq 0\}$ is a Markov chain with transition probability matrix $\mathbf{P}^{\mathfrak{R}}$. As a consequence of this observation, we see that both Theorem 3.1.2 and Corollary 3.1.4 apply to the present setting. In particular, **Q**-*recurrence and* **Q**-*transience are* **Q**-*communicating class properties.*

We next develop the analog for continuous time of the relations given in (2.3.7). Namely, set $\sigma_j^{(1)} = \sigma_j$ and, for $m \geq 1$, $\sigma_j^{(m+1)} = \infty$ if $\sigma_j^{(m)} = \infty$ and $\sigma_j^{(m+1)} = \inf\{t \geq J_{\ell+1} : X(t) = j\}$ if $\sigma_j^{(m)} = J_\ell < \infty$. Then, just as in the proof of Theorem 2.3.6, one sees that

$$\mathbb{P}\big(\sigma_i^{(m)} < \infty \,\big|\, X(0) = i\big) = \mathbb{P}\big(\sigma_i < \infty \,\big|\, X(0) = i\big)^m.$$

In addition, one can easily check that

$$\begin{aligned}
&\mathbb{E}\left[\int_{\sigma_i^{(m)}}^{\sigma_i^{(m+1)}} \mathbf{1}_{\{i\}}\big(X(t)\big)\, dt,\, \sigma_i^{(m)} < \infty \,\middle|\, X(0) = i\right] \\
&\quad = \mathbb{E}\left[\int_0^{\sigma_i} \mathbf{1}_{\{i\}}\big(X(t)\big)\, dt \,\middle|\, X(0) = i\right] \mathbb{P}\big(\sigma_i^{(m)} < \infty \,\big|\, X(0) = i\big) \\
&\quad = \mathbb{E}\big[J_1 \,\big|\, X(0) = i\big] \mathbb{P}\big(\sigma_i^{(m)} < \infty \,\big|\, X(0) = i\big) = \frac{\mathbb{P}(\sigma_i < \infty | X(0) = i)^m}{R_i}
\end{aligned}$$

and

$$\mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{j\}}\big(X(t)\big)\,dt \,\middle|\, X(0)=i\right] = \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{j\}}\big(X(t)\big)\,dt \,\middle|\, X(0)=j\right]\mathbb{P}\big(\sigma_j<\infty \,\big|\, X(0)=i\big).$$

Hence, by exactly the argument with which we passed from Theorem 2.3.6 to (2.3.7), we now know that

$$\begin{aligned}
&\mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{j\}}\big(X(t)\big)\,dt \,\middle|\, X(0)=i\right] = \frac{1}{R_j}\left(\delta_{i,j}+\frac{\mathbb{P}(\sigma_j<\infty|X(0)=i)}{\mathbb{P}(\sigma_j=\infty|X(0)=j)}\right)\\
&\mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{i\}}\big(X(t)\big)\,dt \,\middle|\, X(0)=i\right]=\infty\\
&\qquad\qquad\Longleftrightarrow\ \mathbb{P}\left(\int_0^\infty \mathbf{1}_{\{i\}}\big(X(t)\big)\,dt=\infty \,\middle|\, X(0)=i\right)=1
\end{aligned} \tag{4.4.4}$$

$$\begin{aligned}
&\mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{i\}}\big(X(t)\big)\,dt \,\middle|\, X(0)=i\right]<\infty\\
&\qquad\qquad\Longleftrightarrow\ \mathbb{P}\left(\int_0^\infty \mathbf{1}_{\{i\}}\big(X(t)\big)\,dt<\infty \,\middle|\, X(0)=i\right)=1.
\end{aligned}$$

Our final goal in this subsection is to prove the following statement.

4.4.5 THEOREM. *For any given state $i\in\mathbb{S}$, the following are equivalent:*

(1) *i is $\mathbf{Q}$-recurrent.*
(2) *There is a $t\in(0,\infty)$ such that i is recurrent relative to the transition probability matrix $\mathbf{P}(t)$.*
(3) *i is recurrent relative to $\mathbf{P}(t)$ for all $t\in(0,\infty)$.*

PROOF: We will prove this equivalence by checking that the same statement holds when "recurrent" is replaced throughout by "transient." To this end, first observe that

$$\mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{i\}}\big(X(t)\big)\,dt \,\middle|\, X(0)=i\right]=\int_0^\infty \big(\mathbf{P}(t)\big)_{ii}\,dt,$$

and therefore, from the first line of (4.4.4), that

$$i \text{ is } \mathbf{Q}\text{-transient} \iff \int_0^\infty \big(\mathbf{P}(t)\big)_{ii}\,dt<\infty. \tag{4.4.6}$$

Next, notice that for $0\le s<t$,

$$\big(\mathbf{P}(t)\big)_{ii}\ge\big(\mathbf{P}(t-s)\big)_{ii}\big(\mathbf{P}(s)\big)_{ii}\ge e^{-(t-s)R_i}\big(\mathbf{P}(s)\big)_{ii} \tag{4.4.7}$$

since $\big(\mathbf{P}(h)\big)_{ii} \geq \mathbb{P}(J_1 > h \big| X(0) = i) = e^{-hR_i}$. Hence, for any $t > 0$ and $n \in \mathbb{N}$,

$$\big(\mathbf{P}(t)^{n+1}\big)_{ii} = \big(\mathbf{P}\big((n+1)t\big)\big)_{ii} \geq e^{-tR_i}\big(\mathbf{P}(\tau)\big)_{ii} \quad \text{for all } \tau \in [nt, (n+1)t]$$

and

$$e^{-tR_i}\big(\mathbf{P}(t)^n\big)_{ii} = e^{-tR_i}\big(\mathbf{P}(nt)\big)_{ii} \leq \big(\mathbf{P}(\tau)\big)_{ii} \quad \text{for all } \tau \in [nt, (n+1)t].$$

Since this means that

$$te^{-tR_i}\sum_{n=0}^{\infty}\big(\mathbf{P}(t)^n\big)_{ii} \leq \int_0^{\infty}\big(\mathbf{P}(\tau)\big)_{ii}\,d\tau \leq te^{tR_i}\sum_{n=0}^{\infty}\big(\mathbf{P}(t)^{n+1}\big)_{ii},$$

the asserted implications are now immediate from (4.4.6). □

4.4.2. Stationary Measures and Limit Theorems: In this subsection, we will complete our program by proving the following basic result.

4.4.8 THEOREM. *For each $j \in \mathbb{S}$*

$$\hat{\pi}_{jj} \equiv \lim_{t\to\infty}\big(\mathbf{P}(t)\big)_{jj} \text{ exists}$$

and

$$\lim_{t\to\infty}\big(\mathbf{P}(t)\big)_{ij} = \hat{\pi}_{ij} \equiv \mathbb{P}\big(\sigma_j < \infty \big| X(0) = i\big)\hat{\pi}_{jj} \text{ for } i \neq j.$$

Moreover, if $\hat{\pi}_{jj} > 0$, then $\hat{\pi}_{ii} > 0$ for all $i \in C \equiv \{i : i \overset{\mathbf{Q}}{\leftrightarrow} j\}$, and when the row vector $\hat{\boldsymbol{\pi}}^C$ is determined by $(\hat{\boldsymbol{\pi}}^C)_i = 1_C(i)\hat{\pi}_{ii}$, then, for each $s > 0$, $\hat{\boldsymbol{\pi}}^C$ is the unique probability vector $\boldsymbol{\mu} \in \mathrm{Stat}\big(\mathbf{P}(s)\big)$ for which $(\boldsymbol{\mu})_k = 0$ when $k \notin C$. In fact, if $\boldsymbol{\mu} \in \mathrm{Stat}\big(\mathbf{P}(s)\big)$ for some $s > 0$, then, for each $j \in C$,

$$(\boldsymbol{\mu})_j = \left(\sum_{i \overset{\mathbf{Q}}{\leftrightarrow} j}(\boldsymbol{\mu})_i\right)\hat{\pi}_{jj}.$$

PROOF: We begin with the following continuous-time version of the renewal equation (cf. (3.2.6)):

$$\big(\mathbf{P}(t)\big)_{ij} = e^{-tR_i}\delta_{i,j} + \mathbb{E}\big[\big(\mathbf{P}(t-\sigma_j)\big)_{jj},\, \sigma_j \leq t \,\big|\, X(0) = i\big]. \tag{4.4.9}$$

The proof of (4.4.9) runs as follows. First, write $\big(\mathbf{P}(t)\big)_{ij}$ as

$$\mathbb{P}\big(X(t) = j \;\&\; J_1 > t \,\big|\, X(0) = i\big) + \mathbb{P}\big(X(t) = j \;\&\; J_1 \leq t \,\big|\, X(0) = i\big).$$

Clearly, the first term on the right is 0 unless $i = j$, in which case it is equal e^{-tR_i}. To handle the second term, write it as

$$\sum_{m=1}^{\infty} \mathbb{P}\big(X(t) = j \;\&\; \sigma_j = J_m \le t \,\big|\, X(0) = i\big),$$

and observe that (cf. (4.2.3) and (4.2.5))

$$\begin{aligned}
&\mathbb{P}\big(X(t) = j \;\&\; \sigma_j = J_m \le t \,\big|\, X(0) = i\big) \\
&= \mathbb{P}\Big(\Phi^{\mathfrak{R},\mathbf{P}}\big(t - J_m; (E_{m+1}, \dots, E_{m+n}, \dots), (j, X_{m+1}, \dots, X_{m+n}, \dots)\big) = j \\
&\qquad\qquad \;\&\; \sigma_j = J_m \le t \,\Big|\, X_0 = i\Big) \\
&= \mathbb{E}\Big[\big(\mathbf{P}(t - J_m)\big)_{jj}, \sigma_j = J_m \le t \,\Big|\, X(0) = i\Big].
\end{aligned}$$

Hence, after summing this over $m \ge 1$ and combining this result with the preceding, one arrives at (4.4.9).

Knowing (4.4.9), we see that the first part of the theorem will be proved once we treat the case when $i = j$. To this end, we begin by observing that, because, by (4.4.7), $\big(\mathbf{P}(s)\big)_{ii} \ge e^{-sR_i} > 0$ for all $s > 0$ and $i \in \mathbb{S}$, each i is aperiodic relative to $\mathbf{P}(s)$, and so, by (3.2.15), we know that $\pi(s)_{ii} \equiv \lim_{n\to\infty}\big(\mathbf{P}(s)^n\big)_{ii}$ exists for all $s > 0$ and $i \in \mathbb{S}$. We want to show that $\pi(1)_{ii} = \lim_{t\to\infty}\big(\mathbf{P}(t)\big)_{ii}$, and when $\pi(1)_{ii} = 0$, this is easy. Indeed, by (4.4.7),

$$\overline{\lim_{t\to\infty}}\big(\mathbf{P}(t)\big)_{ii} \le e^{R_i}\,\overline{\lim_{t\to\infty}}\big(\mathbf{P}([t]+1)\big)_{ii} = e^{R_j}\pi(1)_{ii},$$

where $[t]$ denotes the integer part of t.

In view of the preceding, what remains to be proved in the first assertion is that $\lim_{t\to\infty}\big(\mathbf{P}(t)\big)_{ii} = \pi(1)_{ii}$ when $\pi(1)_{ii} > 0$, and the key step in our proof will be the demonstration that $\pi(s)_{ii} = \pi(1)_{ii}$ for all $s > 0$, a fact which we already know when $\pi(1)_{ii} = 0$. Thus, assume that $\pi(1)_{ii} > 0$, and let C be the $\mathbf{Q}$-communicating class of i relative to $\mathbf{P}(1)$. By (4.4.2), C is also the communicating class of i relative to $\mathbf{P}(s)$ for every $s > 0$. In addition, because $\pi(1)_{ii} > 0$, i is recurrent, in fact positive recurrent, relative to $\mathbf{P}(1)$, and therefore, by Theorem 4.4.5, it is also recurrent relative to $\mathbf{P}(s)$ for all $s > 0$. Now determine the row vector $\boldsymbol{\pi}(1)^C$ so that $(\boldsymbol{\pi}(1)^C)_j = \mathbf{1}_C(j)\pi(1)_{jj}$. Then, because $\pi(1)_{ii} > 0$, we know, by Theorem 3.2.10, that $\boldsymbol{\pi}(1)^C$ is the one and only $\boldsymbol{\mu} \in \mathrm{Stat}\big(\mathbf{P}(1)\big)$ which vanishes off of C. Next, given $s > 0$, consider $\boldsymbol{\mu} = \boldsymbol{\pi}(1)^C\mathbf{P}(s)$. Then $\boldsymbol{\mu}$ is a probability vector and, because $\big(\mathbf{P}(1)\big)_{jk} = 0$ when $j \in C$ and $k \notin C$, $\boldsymbol{\mu}$ vanishes off of C. Also, because

$$\begin{aligned}
\boldsymbol{\mu}\mathbf{P}(1) &= \boldsymbol{\pi}(1)^C\mathbf{P}(s)\mathbf{P}(1) = \boldsymbol{\pi}(1)^C\mathbf{P}(s+1) \\
&= \boldsymbol{\pi}(1)^C\mathbf{P}(1)\mathbf{P}(s) = \boldsymbol{\pi}(1)^C\mathbf{P}(s) = \boldsymbol{\mu},
\end{aligned}$$

$\boldsymbol{\mu}$ is stationary for $\mathbf{P}(1)$. Hence, by uniqueness, we conclude that $\boldsymbol{\pi}(1)^C\mathbf{P}(s) = \boldsymbol{\pi}(1)^C$, and therefore that $\boldsymbol{\pi}(1)^C$ is a stationary measure for $\mathbf{P}(s)$ which vanishes off of C. But, by (3.2.8) and the fact that C is the communicating class of i relative to $\mathbf{P}(s)$, this means that

$$\pi(1)_{ii} = \bigl(\boldsymbol{\pi}(1)^C\bigr)_i = \left(\sum_{j\in C}\bigl(\boldsymbol{\pi}(1)^C\bigr)_j\right)\pi(s)_{ii} = \pi(s)_{ii}.$$

To complete the proof of the first part from here, note that, by (4.4.7),

$$e^{-sR_j}\bigl(\mathbf{P}(ns)\bigr)_{jj} \le \bigl(\mathbf{P}(t)\bigr)_{jj} \le e^{sR_j}\bigl(\mathbf{P}((n+1)s)\bigr)_{jj} \quad \text{when } ns \le t \le (n+1)s,$$

and so, since $\pi(s)_{jj} = \pi(1)_{jj}$ and $\mathbf{P}(nt) = \mathbf{P}(t)^n$,

$$e^{-sR_j}\pi(1)_{jj} \le \varliminf_{t\to\infty}\bigl(\mathbf{P}(t)\bigr)_{jj} \le \varlimsup_{t\to\infty}\bigl(\mathbf{P}(t)\bigr)_{jj} \le e^{sR_j}\pi(1)_{jj}.$$

Now let $s \searrow 0$, and simply define $\hat{\pi}_{jj} = \pi(1)_{jj}$.

Given the first part, the proof of the second part is easy. Namely, if $\hat{\pi}_{jj} > 0$ and $C = \{i : i \overset{\mathbf{Q}}{\leftrightarrow} j\}$, then, for each $s > 0$, we know that C is the communicating class of j relative to $\mathbf{P}(s)$ and that $\hat{\boldsymbol{\pi}}^C = \boldsymbol{\pi}(s)^C \in \mathrm{Stat}\bigl(\mathbf{P}(s)\bigr)$. Conversely, by (3.2.8), we know that if $\boldsymbol{\mu} \in \mathrm{Stat}\bigl(\mathbf{P}(s)\bigr)$ for some $s > 0$, then

$$(\boldsymbol{\mu})_j = \left(\sum_{\{i:i\overset{\mathbf{Q}}{\leftrightarrow}j\}}(\boldsymbol{\mu})_i\right)\pi(s)_{jj} = \left(\sum_{\{i:i\overset{\mathbf{Q}}{\leftrightarrow}j\}}(\boldsymbol{\mu})_i\right)\hat{\pi}_{jj}. \quad \square$$

With the preceding result in mind, we will say that i is $\mathbf{Q}$-*positive recurrent* if $\hat{\pi}_{ii} > 0$ and will say that i is $\mathbf{Q}$-*null recurrent* if i is $\mathbf{Q}$-recurrent but not $\mathbf{Q}$-positive recurrent. From the preceding, we already know that *Q-positive recurrence is a Q-communicating class property.*

The following corollary comes at very little additional cost.

4.4.10 COROLLARY. **Mean Ergodic Theorem** *Assume that j is $\mathbf{Q}$-positive recurrent and that $\mathbb{P}\bigl(X(0)\overset{\mathbf{Q}}{\leftrightarrow}j\bigr) = 1$. Then,*

$$\lim_{T\to\infty}\mathbb{E}\left[\left(\frac{1}{T}\int_0^T \mathbf{1}_{\{j\}}\bigl(X(t)\bigr)\,dt - \hat{\pi}_{jj}\right)^2\right] = 0.$$

See Exercise 4.5.10 for the more refined version.

PROOF: The proof is really just an obvious transcription to the continuous setting of the argument used to prove Theorem 3.2.14. Namely, set $C = \{i : j \overset{\mathcal{Q}}{\leftrightarrow} i\}$. By precisely the argument given there, it suffices to handle the case when $\hat{\boldsymbol{\pi}}^C$ is the initial distribution. Thus, if $f = \mathbf{1}_{\{j\}} - \hat{\pi}_{jj}$, what we need to show is that

$$\lim_{T\to\infty} \mathbb{E}\left[\left(\frac{1}{T}\int_0^T f\big(X(t)\big)\,dt\right)^2\right] = 0,$$

when $\hat{\boldsymbol{\pi}}^C$ is the distribution of $X(0)$. But, because $\hat{\boldsymbol{\pi}}^C$ is $\mathbf{P}(t)$-stationary,

$$\mathbb{E}\left[\left(\frac{1}{T}\int_0^T f\big(X(t)\big)\,dt\right)^2\right] = \frac{2}{T^2}\int_0^T \left(\int_0^t \mathbb{E}\Big[f\big(X(s)\big)f\big(X(t)\big)\Big]\,ds\right) dt$$

$$= \frac{2}{T^2}\int_0^T \left(\int_0^t \alpha(t-s)\,ds\right) dt = \frac{2}{T}\int_0^T \big(1 - \tfrac{t}{T}\big)\,\alpha(t)\,dt,$$

where $\alpha(t) \equiv \hat{\boldsymbol{\pi}}^C\big(f\mathbf{P}(t)\mathbf{f}\big)$, $\mathbf{f}$ being the column vector corresponding to the function f and $f\mathbf{P}(t)\mathbf{f}$ being the column vector determined by $(f\mathbf{P}(t)\mathbf{f})_i = f(i)(\mathbf{P}(t)\mathbf{f})_i$. Finally, since, for each $i \in C$, $\lim_{t\to\infty}(\mathbf{P}(t)\mathbf{f})_i = 0$, $\lim_{t\to\infty}\alpha(t) = 0$, and therefore

$$\frac{2}{T}\int_0^T \big(1 - \tfrac{t}{T}\big)\,\alpha(t)\,dt \le \frac{2}{T}\int_0^T |\alpha(t)|\,dt \longrightarrow 0 \quad \text{as } T \to \infty. \quad \square$$

4.4.3. Interpreting $\hat{\pi}_{ii}$: Although we have shown that the limit $\hat{\pi}_{ij} = \lim_{t\to\infty}\big(\mathbf{P}(t)\big)_{ij}$ exists, we have yet to give an expression, analogous to the one in (3.2.5), for $\hat{\pi}_{ii}$. However, as we are about to see, such an expression is readily available from (4.4.9). Namely, for each $\alpha > 0$, set

$$L(\alpha)_{ij} = \alpha\mathbb{E}\left[\int_0^\infty e^{-\alpha t}\mathbf{1}_{\{j\}}\big(X(t)\big)\,dt \,\middle|\, X(0) = i\right] = \alpha\int_0^\infty e^{-\alpha t}\big(\mathbf{P}(t)\big)_{ij}\,dt.$$

Because $\hat{\pi}_{ii} = \lim_{t\to\infty}\big(\mathbf{P}(t)\big)_{ii}$, the second of these representations makes it is easy to identify $\hat{\pi}_{ii}$ as $\lim_{\alpha\searrow 0} L(\alpha)_{ii}$. At the same time, from (4.4.9), we see that

$$L(\alpha)_{ii} = \frac{\alpha}{\alpha + R_i} + \mathbb{E}\big[e^{-\alpha\sigma_i} \,\big|\, X(0) = i\big]L(\alpha)_{ii}.$$

Hence, we have now shown that

$$\hat{\pi}_{ii} = \begin{cases} 1 & \text{if } R_i = 0 \\ \frac{1}{R_i\mathbb{E}[\sigma_i | X(0)=i]} & \text{if } R_i > 0, \end{cases}$$

which, in conjunction with the second equality in Theorem 4.4.8, proves that

$$\hat{\pi}_{ij} = \begin{cases} \delta_{i,j} + \mathbb{P}(\sigma_j < \infty | X(0) = i) & \text{if } R_j = 0 \\ \frac{\mathbb{P}(\sigma_j < \infty \,|\, X(0)=i)}{R_j\mathbb{E}[\sigma_j | X(0)=j]} & \text{if } R_j > 0. \end{cases} \tag{4.4.11}$$

Of course, as an immediate corollary of (4.4.11), we now know that i *is positive recurrent if and only if either* $R_i = 0$ *or* $\mathbb{E}[\sigma_i | X(0) = i] < \infty$.

4.5 Exercises

EXERCISE 4.5.1. The purpose of this exercise is to give another derivation of (4.1.5). Thus, let $\{E_n : n \geq 1\}$ be a sequence of mutually independent, unit exponential random variables, and define $\{J_n : n \geq 0\}$ and $\{N(t) : t \geq 0\}$ accordingly, as in §4.1.1. Given $0 = t_0 < \cdots < t_\ell$ and $0 \leq n_1 \leq \cdots \leq n_\ell$, use the change of variables formula for multi-dimensional integrals to justify:

$$\begin{aligned}
&\mathbf{P}\big(N(t_1) = n_1, \ldots, N(t_\ell) = n_\ell\big) \\
&\quad = \mathbf{P}\big(J_{n_1} \leq t_1 < J_{n_1+1}, \ldots, J_{n_\ell} \leq t_\ell < J_{n_\ell+1}\big) \\
&\quad = \int_A \cdots \int \exp\left(-\sum_{j=1}^{n_\ell+1} \xi_j\right) d\xi_1 \cdots d\xi_{n_\ell+1} \\
&\quad = \int_B \cdots \int e^{-\eta_{n_\ell+1}}\, d\eta_1 \cdots d\eta_{n_\ell+1} \\
&\quad = e^{-t_\ell} \prod_{j=1}^{\ell} \mathrm{vol}(\Delta_j) = \prod_{j=1}^{\ell} e^{-(t_j - t_{j-1})} \frac{(t_j - t_{j-1})^{n_j - n_{j-1}}}{(n_j - n_{j-1})!},
\end{aligned}$$

where

$$\begin{aligned}
A &= \left\{(\xi_1, \ldots, \xi_{n_\ell+1}) \in (0,\infty)^{n_\ell+1} : \sum_1^{n_j} \xi_k \leq t_j < \sum_1^{n_j+1} \xi_k \text{ for } 1 \leq j \leq \ell\right\}, \\
B &= \big\{(\eta_1, \ldots, \eta_{n_\ell+1}) \in (0,\infty)^{n_\ell+1} : \\
&\qquad \eta_i < \eta_{i+1} \text{ for } 1 \leq i \leq n_\ell \;\&\; \eta_{n_j} \leq t_j < \eta_{n_j+1} \text{ for } 1 \leq j \leq \ell\big\}, \\
\Delta_j &= \big\{(u_1, \ldots, u_{n_j - n_{j-1}}) \in \mathbb{R}^{n_j - n_{j-1}} : t_{j-1} \leq u_1 < \cdots < u_{n_j - n_{j-1}} \leq t_j\big\},
\end{aligned}$$

with the obvious modifications when $n_j = 0$ for $1 \leq j \leq i$.

EXERCISE 4.5.2. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be commuting elements of $M_{\mathrm{u,v}}(\mathbb{S})$. After checking that

$$(\mathbf{M}_1 + \mathbf{M}_2)^m = \sum_{\ell=0}^{m} \binom{m}{\ell} \mathbf{M}_1^\ell \mathbf{M}_2^{m-\ell}$$

for all $m \in \mathbb{N}$, verify (4.2.12).

EXERCISE 4.5.3. Given a $\mathbf{Q}$-recurrent state i, set $C = \{j : i \overset{\mathbf{Q}}{\leftrightarrow} j\}$, and show that $R_i > 0 \implies R_j > 0$ for all $j \in C$.

EXERCISE 4.5.4. In this exercise we will give the continuous-time version of the ideas in Exercise 2.4.1. For this purpose, assume that $\mathbb{S}$ is irreducible and positive recurrent with respect to $\mathbf{Q}$, and use $\hat{\boldsymbol{\pi}}$ to denote the unique

probability vector which is stationary for each $\mathbf{P}(t)$, $t > 0$. Next, determine the *adjoint* semigroup $\{\mathbb{P}(t)^\top : t \geq 0\}$ so that

$$\big(\mathbf{P}(t)^\top\big)_{ij} = \frac{(\hat{\pi})_j \big(\mathbf{P}(t)\big)_{ji}}{(\hat{\pi})_i}.$$

(a) Show that $\mathbf{P}(t)^\top$ is a transition probability matrix and that $\hat{\pi}\mathbf{P}(t)^\top = \hat{\pi}$ for each $t \geq 0$. In addition, check that $\{\mathbf{P}(t)^\top : t \geq 0\}$ is the semigroup determined by the Q-matrix $\mathbf{Q}^\top$, where

$$\big(\mathbf{Q}^\top\big)_{ij} = \frac{(\hat{\pi})_j (\mathbf{Q})_{ji}}{(\hat{\pi})_i}.$$

(b) Let $\mathbb{P}$ and $\mathbb{P}^\top$ denote the probabilities computed for the Markov processes corresponding, respectively, to $\mathbf{Q}$ and $\mathbf{Q}^\top$ with initial distribution $\hat{\pi}$. Show that $\mathbb{P}^\top$ is the *reverse* of $\mathbb{P}$ in the sense that, for each $n \in \mathbb{N}$, $0 = t_0 < t_1 < \cdots < t_n$, and $(j_0, \ldots, j_n) \in \mathbb{S}^{n+1}$,

$$\mathbb{P}^\top\big(X(t_m) = j_m \text{ for } 0 \leq m \leq n\big) = \mathbb{P}\big(X(t_n - t_m) = j_m \text{ for } 0 \leq m \leq n\big).$$

(c) Define $\mathbf{P}^\top$ so that

$$(\mathbf{P}^\top)_{ij} = \frac{(\hat{\pi})_j (\mathbf{Q})_{ji}}{R_i (\hat{\pi})_i}.$$

Show that $\mathbf{P}^\top$ is again a transition probability matrix, and check that $\mathbf{Q}^\top = \mathbf{R}(\mathbf{P}^\top - \mathbf{I})$.

EXERCISE 4.5.5. Take $\mathbb{S} = \mathbb{N}$ and $(\mathbf{P})_{ij}$ equal to 1 or 0 according to whether $j = i + 1$ or not. Given a set of strictly positive rates $\mathfrak{R}$, show that, no matter what its initial distribution, the Markov process determined by $\mathfrak{R}$ and $\mathbf{P}$ explodes with probability 1 if $\sum_{i \in \mathbb{N}} R_i^{-1} < \infty$ and does not explode if $\sum_{i \in \mathbb{N}} R_i^{-1} = \infty$.

EXERCISE 4.5.6. Here is a more interesting example of explosion. Take $\mathbb{S} = \mathbb{Z}^3$, and let (cf. Exercise 3.3.5) $\mathbf{P}$ be the transition probability matrix for the symmetric, nearest neighbor random walk on $\mathbb{Z}^3$. Given a set of positive rates $\mathfrak{R}$ with the property that $\sum_{\mathbf{k} \in \mathbb{Z}^3} R_{\mathbf{k}}^{-1} < \infty$, show that, starting at every $\mathbf{k} \in \mathbb{Z}^3$, explosion occurs with probability 1 when $(\mathbf{Q})_{\mathbf{k}\boldsymbol{\ell}} = R_{\mathbf{k}}\big((\mathbf{P})_{\mathbf{k}\boldsymbol{\ell}} - \delta_{\mathbf{k},\boldsymbol{\ell}}\big)$.

Hint: Apply the criterion in the second half of Theorem 4.3.6 to a function of the form

$$\sum_{\boldsymbol{\ell} \in \mathbb{Z}^3} \frac{1}{R_{\boldsymbol{\ell}}} \left(\alpha^2 + \sum_{i=1}^{3} \big((\mathbf{k})_i - (\boldsymbol{\ell})_i\big)^2\right)^{-\frac{1}{2}},$$

and use the computation in Exercise 3.3.5.

EXERCISE 4.5.7. Even when $\mathbb{S}$ is finite, writing down a reasonably explicit expression for the solution to (4.2.7) is seldom easy and often impossible. Nonetheless, a little linear algebra often does quite a lot of good. Throughout this exercise, $\mathbf{Q}$ is a Q-matrix on the state space $\mathbb{S}$, and it is assumed that the associated Markov process exists (i.e., does not explode) starting from any point.

(**a**) If $\mathbf{u} \in \mathbb{C}^{\mathbb{S}}$ is a bounded, non-zero, right eigenvector for $\mathbf{Q}$ with eigenvalue $\alpha \in \mathbb{C}$, show that the real part of α must be less than or equal to 0.

(**b**) Assume that $N = \#\mathbb{S} < \infty$ and that $\mathbf{Q}$ admits a complete set of linearly independent, right eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_N \in \mathbb{C}^{\mathbb{S}}$ with associated eigenvalues $\alpha_1, \ldots, \alpha_N$. Let $\mathbf{U}$ be the matrix whose mth column is $\mathbf{u}_m$, and show that $e^{t\mathbf{Q}} = \mathbf{U}\Lambda(t)\mathbf{U}^{-1}$, where $\Lambda(t)$ is the diagonal matrix whose m diagonal entry is $e^{t\alpha_m}$.

EXERCISE 4.5.8. Here is the continuous analog of the result in Exercise 3.3.7. Namely, assume that i is Q-recurrent and $R_i > 0$, set

$$\hat{\mu}_j = \mathbb{E}\left[\int_0^{\sigma_i} \mathbf{1}_{\{j\}}\big(X(t)\big)\, dt \,\middle|\, X(0) = i\right] \in [0, \infty] \quad \text{for } j \in \mathbb{S},$$

and let $\hat{\boldsymbol{\mu}}$ be the row vector given by $(\hat{\boldsymbol{\mu}})_j = \hat{\mu}_j$ for each $j \in \mathbb{S}$.

(**a**) Show that $\hat{\mu}_i = \frac{1}{R_i}$, $\hat{\mu}_j < \infty$ for all $j \in \mathbb{S}$, and that $\hat{\mu}_j > 0$ if and only if $i \overset{\mathbf{Q}}{\leftrightarrow} j$.

(**b**) Show that $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}\mathbf{P}(t)$ for all $t > 0$.

Hint: Check that

$$\big(\hat{\boldsymbol{\mu}}\mathbf{P}(s)\big)_j = \mathbb{E}\left[\int_s^{s+\sigma_i} \mathbf{1}_{\{j\}}\big(X(t)\big)\, dt \,\middle|\, X(0) = i\right]$$

and that

$$\mathbb{E}\left[\int_{\sigma_i}^{s+\sigma_i} \mathbf{1}_{\{j\}}\big(X(t)\big)\, dt \,\middle|\, X(0) = i\right] = \mathbb{E}\left[\int_0^{s} \mathbf{1}_{\{j\}}\big(X(t)\big)\, dt \,\middle|\, X(0) = i\right].$$

(**c**) In particular, if i is $\mathbf{Q}$-positive recurrent and $C = \{j : i \overset{\mathbf{Q}}{\leftrightarrow} j\}$, show that $\hat{\boldsymbol{\mu}} = \left(\sum_j \hat{\mu}_j\right) \hat{\boldsymbol{\pi}}^C$. Equivalently,

$$(\hat{\boldsymbol{\pi}}^C)_j = \frac{\mathbb{E}\left[\int_0^{\sigma_i} \mathbf{1}_{\{j\}}\big(X(t)\big)\, dt \,\middle|\, X(0) = i\right]}{\mathbb{E}\big[\sigma_i \,\big|\, X(0) = i\big]}.$$

EXERCISE 4.5.9. Having given the continuous-time analog of Exercise 3.3.7, we now want to give the continuous-time analog of Exercise 3.3.8. For this purpose, again let i be a $\mathbf{Q}$-recurrent element of $\mathbb{S}$ with $R_i > 0$, and define the

measure $\boldsymbol{\mu}$ accordingly, as in Exercise 4.5.8. Next, assume that $\hat{\boldsymbol{\nu}} \in [0,\infty)^{\mathbb{S}}$ satisfies the conditions that: $(\hat{\boldsymbol{\nu}})_i > 0$, $(\hat{\boldsymbol{\nu}})_j = 0$ unless $\mathbb{P}(\sigma_j < \infty \,|\, X(0) = i) = 1$, and $\hat{\boldsymbol{\nu}}\mathbf{Q} = \mathbf{0}$ in the sense that $R_j(\hat{\boldsymbol{\nu}})_j = \sum_{k\neq j}(\hat{\boldsymbol{\nu}})_k \mathbf{Q}_{kj}$ for all $j \in \mathbb{S}$.[6] The goal is to show that $\hat{\boldsymbol{\nu}} = R_i(\hat{\boldsymbol{\nu}})_i \boldsymbol{\mu}$. Equivalently,

$$(\hat{\boldsymbol{\nu}})_j = R_i(\hat{\boldsymbol{\nu}})_i \mathbb{E}\left[\int_0^{\sigma_i} \mathbf{1}_{\{j\}}\big(X(t)\big)\,dt \,\middle|\, X(0) = i\right] \quad \text{for all } j \in \mathbb{S}.$$

In particular, by Exercise 4.5.8, this will mean that $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}(t)$ for all $t > 0$.

In the following, the transition probability $\mathbf{P}$ is chosen so that its diagonal entries vanish and $\mathbf{Q} = \mathbf{R}(\mathbf{P} - \mathbf{I})$, where, as usual, $\mathbf{R}$ is the diagonal matrix whose jth diagonal entry is R_j for each $j \in \mathbb{S}$.

(**a**) Begin by showing that i is $\mathbf{P}$-recurrent.

(**b**) Define the row vector $\boldsymbol{\nu}$ so that $(\boldsymbol{\nu})_j = \frac{R_j(\hat{\boldsymbol{\nu}})_j}{R_i(\hat{\boldsymbol{\nu}})_i}$, and show that $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}$.

(**c**) By combining (**b**) with Exercise 3.3.7, show that

$$(\boldsymbol{\nu})_j = \mathbb{E}\left[\sum_{m=0}^{\rho_i - 1} \mathbf{1}_{\{j\}}(X_m) \,\middle|\, X_0 = i\right],$$

where $\{X_n : n \geq 0\}$ is the Markov chain with transition probability matrix $\mathbf{P}$.

(**d**) Show that

$$R_j \mathbb{E}\left[\int_0^{\sigma_i} \mathbf{1}_{\{j\}}\big(X(t)\big)\,dt \,\middle|\, X(0) = i\right] = \mathbb{E}\left[\sum_{m=0}^{\rho_i - 1} \mathbf{1}_{\{j\}}(X_m) \,\middle|\, X_0 = i\right],$$

and, after combining this with (**c**), arrive at the desired conclusion.

EXERCISE 4.5.10. Following the strategy used in Exercise 3.3.9, show that, under the hypotheses in Corollary 4.4.10, one has the following continuous version of the individual ergodic theorem:

$$\mathbb{P}\left(\lim_{T\to\infty} \left\|\mathbf{L}_T - \hat{\boldsymbol{\pi}}^C\right\|_{\mathrm{v}} = 0\right) = 1,$$

where $C = \{i : j \overset{\mathbf{Q}}{\hookrightarrow} i\}$ and $\mathbf{L}_T$ is the *empirical measure* determined by

$$(\mathbf{L}_T)_i = \frac{1}{T}\int_0^T \mathbf{1}_{\{i\}}\big(X(t)\big)\,dt.$$

In addition, following the strategy in Exercise 3.3.11, show that, for any initial distribution,

$$\mathbb{P}\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{\{j\}}\big(X(t)\big)\,dt = 0\right) = 1$$

when j is not $\mathbf{Q}$-positive recurrent.

[6] The reason why the condition $\hat{\boldsymbol{\nu}}\mathbf{Q} = \mathbf{0}$ needs amplification is that, because $\mathbf{Q}$ has negative diagonal entries, possible infinities could cause ambiguities.

EXERCISE 4.5.11. Given a Markov process $\{X(t) : t \geq 0\}$ with Q-matrix $\mathbf{Q}$, one can produce a Markov process with Q-matrix $M\mathbf{Q}$ by speeding up the clock of $t \rightsquigarrow X(t)$ by a factor of M. That is, $\{X(Mt) : t \geq 0\}$ is a Markov process with Q-matrix $M\mathbf{Q}$. The purpose of this exercise is to show how to carry out the analogous procedure, known in the literature as *random time change*, for variable rates. To be precise, let $\mathbf{P}$ be a transition probability matrix on $\mathbb{S}$ with $(\mathbf{P})_{ii} = 0$ for all i. Choose and fix an $i \in \mathbb{S}$, let $\{X_n : n \geq 0\}$ be a Markov chain starting from i with transition probability matrix $\mathbf{P}$ and $\{N(t) : t \geq 0\}$ a simple Poisson process which is independent of $\{X_n : n \geq 0\}$, and set $X^0(t) = X^0_{N(t)}$ for $t \geq 0$. In particular, $\{X^0(t) : t \geq 0\}$ is a Markov process with Q-matrix $(\mathbf{P} - \mathbf{I})$ starting from i. Finally, let $\mathfrak{R}$ be a set of positive rates, and take $\mathbf{Q} = \mathbf{R}(\mathbf{P} - \mathbf{I})$ accordingly.

(**a**) Define

$$A(t) = \int_0^t \frac{1}{R_{X^0(\tau)}}\, d\tau \quad \text{for } t \in [0, \infty),$$

observe that $t \rightsquigarrow A(t)$ is strictly increasing, set $A(\infty) = \lim_{t \nearrow \infty} A(t) \in (0, \infty]$, and use $s \in \big[0, A(\infty)\big) \longmapsto A^{-1}(s) \in [0, \infty)$ to denote the inverse of $t \rightsquigarrow A(t)$. Show that

$$A^{-1}(s) = \int_0^s R_{X^0(A^{-1}(\sigma))}\, d\sigma, \quad s \in \big[0, A(\infty)\big).$$

(**b**) Set $X(s) = X^0\big(A^{-1}(s)\big)$ for $s \in \big[0, A(\infty)\big)$. Define $J^0_0 = 0 = J_0$ and, for $n \geq 1$, J^0_n and J_n to be the times of the nth jump of $t \rightsquigarrow X^0(t)$ and $s \rightsquigarrow X(s)$, respectively. After noting that $J_n = A(J^0_n)$, conclude that, for each $n \geq 1$, $s > 0$, and $j \in \mathbb{S}$,

$$\begin{aligned} &\mathbb{P}\big(J_n - J_{n-1} > s \;\&\; X(J_n) = j \,\big|\, X(\sigma),\, \sigma \in [0, J_n)\big) \\ &\quad = e^{-tR_{X(J_{n-1})}}(\mathbf{P})_{X(J_{n-1})j} \quad \text{on } \{J_{n-1} < \infty\}. \end{aligned}$$

(**c**) Show that the explosion time for the Markov process starting from i with Q-matrix $\mathbf{Q}$ has the same distribution as $A(\infty)$. In particular, if $A(\infty) = \infty$ with probability 1, use (**b**) to conclude that $\{X(s) : s \geq 0\}$ is a Markov process starting from i with Q-matrix $\mathbf{Q}$.

(**d**) As a consequence of these considerations, show that if the Markov process corresponding to $\mathbf{Q}$ does not explode, then neither does the one corresponding to $\mathbf{Q}'$, where $\mathbf{Q}'$ is related to $\mathbf{Q}$ by $(\mathbf{Q}')_{ij} = \alpha_i(\mathbf{Q})_{ij}$ and the $\{\alpha_i : i \in \mathbb{S}\}$ is a bounded subset of $(0, \infty)$.

CHAPTER 5

Reversible Markov Processes

This is devoted to the study of a class of Markov processes which admit an initial distribution with respect to which they are *reversible* in the sense that, on every time interval, the distribution of the process is the same when it is run backwards as when it is run forwards. That is, for any $n \geq 1$ and $(i_0, \ldots, i_n) \in E^{n+1}$, in the discrete time setting,

$$\mathbb{P}\big(X_m = i_m \text{ for } 0 \leq m \leq n\big) = \mathbb{P}\big(X_{n-m} = i_m \text{ for } 0 \leq m \leq n\big) \tag{5.0.1}$$

and in the continuous time setting,

$$\mathbb{P}\big(X(t_m) = i_m \text{ for } 0 \leq m \leq n\big) = \mathbb{P}\big(X(t_n - t_m) = i_m \text{ for } 0 \leq m \leq n\big) \tag{5.0.2}$$

whenever $0 = t_0 < \cdots < t_n$. Notice that the initial distribution of such a process is necessarily stationary. Indeed, depending on whether the setting is that of discrete or continuous time, we have

$$\begin{aligned} \mathbb{P}(X_0 = i \;\&\; X_n = j) &= \mathbb{P}(X_n = i \;\&\; X_0 = j) \\ \text{or } \mathbb{P}\big(X(0) = i \;\&\; X(t) = j\big) &= \mathbb{P}\big(X(t) = i \;\&\; X(0) = j\big), \end{aligned}$$

from which stationarity follows after one sums over j. In fact, what the preceding argument reveals is that reversibility says that the joint distribution of, depending on the setting, (X_0, X_n) or $\big(X(0), X(t)\big)$ is the same as that of (X_n, X_0) or $\big(X(t), X(0)\big)$. This should be contrasted with the stationarity which gives equality only for the marginal distribution of the first components of these.

In view of the preceding, one should suspect that reversible Markov processes have ergodic properties which are better than those of general stationary processes, and in this chapter we will examine some of these special properties.

5.1 Reversible Markov Chains

In this section we will discuss irreducible, reversible Markov chains. Because the initial distribution of such a chain is stationary, we know (cf. Theorem 3.2.10) that the chain must be positive recurrent and that the initial distribution must be the probability vector (cf. (3.2.9)) $\boldsymbol{\pi} = \boldsymbol{\pi}^{\mathbb{S}}$ whose ith component

is $(\pi)_i \equiv \mathbb{E}[\rho_i | X_0 = i]^{-1}$. Thus, if $\mathbf{P}$ is the transition probability matrix, then, by taking $n = 1$ in (5.0.1), we see that

$$(\pi)_i(\mathbf{P})_{ij} = \mathbb{P}\big(X_0 = i \;\&\; X_1 = j\big) = \mathbb{P}\big(X_0 = j \;\&\; X_1 = i\big) = (\pi)_j(\mathbf{P})_{ji}.$$

That is, $\mathbf{P}$ satisfies[1]

$$(\pi)_i(\mathbf{P})_{ij} = (\pi)_j(\mathbf{P})_{ji}, \quad \text{the condition of } \textit{detailed balance}. \tag{5.1.1}$$

Conversely, (5.1.1) implies reversibility. To see this, one works by induction on $n \geq 1$ to check that

$$\pi_{i_0}(\mathbf{P})_{i_0 i_1} \cdots (\mathbf{P})_{i_{n-1} i_n} = \pi_{i_n}(\mathbf{P})_{i_n i_{n-1}} \cdots (\mathbf{P})_{i_1 i_0},$$

which is equivalent to (5.0.1).

5.1.1. Reversibility from Invariance: As we have already seen, reversibility implies invariance, and it should be clear that the converse is false. On the other hand, there are two canonical ways in which one can pass from an irreducible transition probability $\mathbf{P}$ with stationary distribution π to a transition probability for which π is reversible. Namely, define the *adjoint* $\mathbf{P}^\top$ of $\mathbf{P}$ so that

$$(\mathbf{P}^\top)_{ij} = \frac{(\pi)_j(\mathbf{P})_{ji}}{(\pi)_i}. \tag{5.1.2}$$

Obviously, π is reversible for $\mathbf{P}$ if and only if $\mathbf{P} = \mathbf{P}^\top$. More generally, because $\pi\mathbf{P} = \pi$, $\mathbf{P}^\top$ is again a transition probability. In addition, one can easily verify that both

$$\frac{\mathbf{P} + \mathbf{P}^\top}{2} \quad \text{and} \quad \mathbf{P}^\top\mathbf{P} \tag{5.1.3}$$

are transition probabilities which are reversible with respect to π. As is explained in Exercise 5.6.9 below, each of these constructions has its own virtue.

5.1.2. Measurements in Quadratic Mean: For reasons which will become increasingly clear, it turns out that we will here want to measure the size of functions using a Euclidean norm rather than the uniform norm $\|f\|_u$. Namely, we will use the norm

$$\|f\|_{2,\pi} \equiv \sqrt{\langle |f|^2 \rangle_\pi} \quad \text{where } \langle g \rangle_\pi \equiv \sum_i g(i)(\pi)_i \tag{5.1.4}$$

[1] The reader who did Exercise 2.4.1 should recognize that the condition below is precisely the same as the statement that $\mathbf{P} = \mathbf{P}^\top$. In particular, if one knows the conclusion of that exercise, then one has no need for the discussion which follows.

is the expected value of g with respect to π. Because $(\pi)_i > 0$ for each $i \in \mathbb{S}$, it is clear that $\|f\|_{2,\pi} = 0 \iff f \equiv 0$. In addition, if we define the *inner product* $\langle f, g\rangle_\pi$ to be $\langle fg\rangle_\pi$, then, for any $t > 0$,

$$0 \le \|tf \pm t^{-1}f\|_{2,\pi}^2 = t^2\|f\|_{2,\pi}^2 \pm 2\langle f, g\rangle_\pi + t^{-2}\|g\|_{2,\pi}^2,$$

and so $|\langle f, g\rangle_\pi| \le t^2\|f\|_{2,\pi}^2 + t^{-2}\|g\|_{2,\pi}^2$ for all $t > 0$. To get the best estimate, we minimize the right hand side with respect to $t > 0$. When either f or g is identically 0, then we see that $\langle f, g\rangle_\pi = 0$ by letting $t \to \infty$ or $t \to 0$. If neither f nor g vanishes identically, we can do best by taking $t = \left(\frac{\|g\|_{2,\pi}}{\|f\|_{2,\pi}}\right)^{\frac{1}{4}}$. Hence, in any case, we arrive at *Schwarz's inequality*

$$|\langle f, g\rangle_\pi| \le \|f\|_{2,\pi}\|g\|_{2,\pi}. \tag{5.1.5}$$

Given Schwarz's inequality, we know

$$\|f + g\|_{2,\pi}^2 = \|f\|_{2,\pi}^2 + 2\langle f, g\rangle_\pi + \|g\|_{2,\pi}^2 \le \left(\|f\|_{2,\pi} + \|g\|_{2,\pi}\right)^2$$

That is, we have the *triangle inequality*:

$$\|f + g\|_{2,\pi} \le \|f\|_{2,\pi} + \|g\|_{2,\pi}. \tag{5.1.6}$$

Thus, if $L^2(\pi)$ denotes the space of f for which $\|f\|_{2,\pi} < \infty$, then $L^2(\pi)$ is a linear space for which $(f, g) \rightsquigarrow \|f - g\|_{2,\pi}$ is a metric. In fact, this metric space is complete since, if $\lim_{m\to\infty} \sup_{n>m} \|f_n - f_m\|_{2,\pi} = 0$, then $\{f_n(i) : n \ge 0\}$ is Cauchy convergent in $\mathbb{R}$ for each $i \in \mathbb{S}$, and therefore there exists a limit function f such that $f_n(i) \longrightarrow f(i)$ for each $i \in \mathbb{S}$. Moreover, by Fatou's Lemma,

$$\|f - f_m\|_{2,\pi} \le \varliminf_{n\to\infty} \|f_n - f_m\|_{2,\pi} \longrightarrow 0 \quad \text{as } m \to \infty.$$

In this chapter, we will use the notation

$$\mathbf{P}f(i) = \sum_{j\in\mathbb{S}} f(j)(\mathbf{P})_{ij} = (\mathbf{Pf})_i,$$

where $\mathbf{f}$ is the column vector determined by f. When f is bounded, it is clear that $\mathbf{P}f(i)$ is well-defined and that $\|\mathbf{P}f\|_{\mathrm{u}} \le \|f\|_{\mathrm{u}}$, where $\|g\|_{\mathrm{u}} \equiv \sup_{i\in\mathbb{S}} |g(i)| = \|\mathbf{g}\|_{\mathrm{u}}$ when $\mathbf{g}$ is the column vector determined by g. We next want show that, even when $f \in L^2(\pi)$, $\mathbf{P}f(i)$ is well-defined for each $i \in \mathbb{S}$ and that $\mathbf{P}$ is a contraction in $L^2(\pi)$: $\|\mathbf{P}f\|_{2,\pi} \le \|f\|_{2,\pi}$. To check that $\mathbf{P}f(i)$ is well-defined, we will show that the the series $\sum_{j\in\mathbb{S}} f(j)(\mathbf{P})_{ij}$ is absolutely convergent. But, by the form of Schwarz's inequality in Exercise 1.3.1,

$$\sum_{j\in\mathbb{S}} |f(j)|(\mathbf{P})_{ij} \le \|f\|_{2,\pi}\left(\sum_{j\in\mathbb{S}} \frac{(\mathbf{P})_{ij}^2}{(\pi)_j}\right)^{\frac{1}{2}},$$

and, by (5.1.1),

$$\sum_{j\in\mathbb{S}} \frac{(\mathbf{P})_{ij}^2}{(\boldsymbol{\pi})_j} = \frac{1}{(\boldsymbol{\pi})_i} \sum_{j\in\mathbb{S}} (\mathbf{P})_{ij}(\mathbf{P})_{ji} = \frac{(\mathbf{P}^2)_{ii}}{(\boldsymbol{\pi}_i)} < \infty.$$

As for the estimate $\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} \le \|f\|_{2,\boldsymbol{\pi}}$, we use Exercise 5.6.2 below together with $\sum_j (\mathbf{P})_{ij} = 1$ to see that $\big(\mathbf{P}f(i)\big)^2 \le \mathbf{P}f^2(i)$ for each i. Thus, since $\boldsymbol{\pi}$ is $\mathbf{P}$-stationary,

$$\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} \le \|f\|_{2,\boldsymbol{\pi}}. \tag{5.1.7}$$

An important consequence of (5.1.7) is the fact that, in general (cf. (2.2.4)),

$$\lim_{n\to\infty} \|\mathbf{A}_n f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} = 0 \quad \text{for all } f \in L^2(\boldsymbol{\pi}) \tag{5.1.8}$$

and

$$\mathbf{P} \text{ is aperiodic} \implies \lim_{n\to\infty} \|\mathbf{P}^n f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} = 0 \quad \text{for all } f \in L^2(\boldsymbol{\pi}). \tag{5.1.9}$$

To see these, first observe that, by (3.2.11), (3.2.15), and Lebesgue's Dominated Convergence Theorem, there is nothing to do when f vanishes off of a finite set. Thus, if $\{F_N : N \ge 1\}$ is an exhaustion of $\mathbb{S}$ by finite sets and if, for $f \in L^2(\boldsymbol{\pi})$, $f_N \equiv \mathbf{1}_{F_N} f$, then, for each $N \in \mathbb{Z}^+$,

$$\begin{aligned}\|\mathbf{A}_n f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} &\le \|\mathbf{A}_n(f - f_N)\|_{2,\boldsymbol{\pi}} + \|\mathbf{A}_n f_N - \langle f_N\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} + \langle |f_N - f|\rangle_{\boldsymbol{\pi}} \\ &\le 2\|f - f_N\|_{2,\boldsymbol{\pi}} + \|\mathbf{A}_n f_N - \langle f_N\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}},\end{aligned}$$

where, in the passage to the second line, we have used $\|\mathbf{A}_n g\|_{2,\boldsymbol{\pi}} \le \|g\|_{2,\boldsymbol{\pi}}$, which follows immediately from (5.1.7). Thus, for each N, $\overline{\lim}_{n\to\infty} \|\mathbf{A}_n f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} \le 2\|f - f_N\|_{2,\boldsymbol{\pi}}$, which gives (5.1.8) when $N \to \infty$. The argument for (5.1.9) is essentially the same, and, as is shown in Exercise 5.6.6 below, all these results hold even in the non-reversible setting.

Finally, (5.1.1) leads to

$$\langle g, \mathbf{P}f\rangle_{\boldsymbol{\pi}} = \sum_{(i,j)} (\boldsymbol{\pi})_i g(i)(\mathbf{P})_{ij} f(j) = \sum_{(i,j)} (\boldsymbol{\pi})_j f(i)(\mathbf{P})_{ji} g(j) = \langle \mathbf{P}g, f\rangle_{\boldsymbol{\pi}}.$$

In other words, $\mathbf{P}$ is *symmetric* on $L^2(\boldsymbol{\pi})$ in the sense that

$$\langle g, \mathbf{P}f\rangle_{\boldsymbol{\pi}} = \langle \mathbf{P}g, f\rangle_{\boldsymbol{\pi}} \quad \text{for } (f,g) \in \big(L^2(\boldsymbol{\pi})\big)^2. \tag{5.1.10}$$

5.1.3. The Spectral Gap: The equation (5.1.7) combined with (5.1.10) say that $\mathbf{P}$ a *self-adjoint contraction* on the Hilbert space[2] is $L^2(\boldsymbol{\pi})$. For the

[2] A Hilbert space is a vector space equipped with an inner product which determines a norm for which the associated metric is complete.

reader who is unfamiliar with these concepts at this level of abstraction, think about the case when $\mathbb{S} = \{1, \dots, N\}$. Then the space of functions $f : \mathbb{S} \longrightarrow \mathbb{R}$ can be identified with $\mathbb{R}^N$. (Indeed, we already made this identification when we gave, in §2.1.3, the relation between functions and column vectors.) After making this identification, the inner product on $L^2(\boldsymbol{\pi})$ becomes the inner product $\sum_1^N (\boldsymbol{\pi})_i (\mathbf{v})_i (\mathbf{w})_i$ for column vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^N$. Hence (5.1.7) says that the matrix $\mathbf{P}$ acts as a symmetric, contraction on $\mathbb{R}^N$ with this inner product. Alternatively, if $\tilde{\mathbf{P}} \equiv \boldsymbol{\Pi}^{\frac{1}{2}} \mathbf{P} \boldsymbol{\Pi}^{-\frac{1}{2}}$, where $\boldsymbol{\Pi}$ is the diagonal matrix whose ith diagonal entry is $(\boldsymbol{\pi})_i$, then, by (5.1.1), $\tilde{\mathbf{P}}$ is symmetric with respect to the standard inner product $(\mathbf{v}, \mathbf{w})_{\mathbb{R}^N} \equiv \sum_1^N (\mathbf{v})_i (\mathbf{w})_i$ on $\mathbb{R}^N$. Moreover, because, by (5.1.7),

$$\|\tilde{\mathbf{P}}\mathbf{f}\|^2_{\mathbb{R}^N} \equiv (\tilde{\mathbf{P}}\mathbf{f}, \tilde{\mathbf{P}}\mathbf{f})_{\mathbb{R}^N} = \sum_i (\boldsymbol{\pi})_i \big(\mathbf{P}\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{f}\big)_i^2 = \|\mathbf{P}g\|^2_{2,\boldsymbol{\pi}} \leq \|g\|^2_{2,\boldsymbol{\pi}} = \|\mathbf{f}\|^2_{\mathbb{R}^N},$$

where g is the function determined by the column vector $\boldsymbol{\Pi}^{-\frac{1}{2}}\mathbf{f}$, we see that, as an operator on $\mathbb{R}^N$, $\tilde{\mathbf{P}}$ is length contracting. Now, by the standard theory of symmetric matrices on $\mathbb{R}^N$, we know that $\tilde{\mathbf{P}}$ admits eigenvalues $1 \geq \lambda_1 \geq \cdots \lambda_N \geq -1$ with associated eigenvectors $(\mathbf{e}_1, \dots, \mathbf{e}_N)$ which are orthonormal for $(\,\cdot\,,\,\cdot\,)_{\mathbb{R}^N}$: $(\mathbf{e}_k, \mathbf{e}_\ell)_{\mathbb{R}^N} = \delta_{k,\ell}$. Moreover, because $\sqrt{(\boldsymbol{\pi})_i} = \sum_{j=1}^N (\tilde{\mathbf{P}})_{ij} \sqrt{(\boldsymbol{\pi})_j}$, we know that $\lambda_1 = 1$ and can take $(\mathbf{e}_1)_i = \sqrt{(\boldsymbol{\pi})_i}$. Finally, by setting $\mathbf{g}_\ell = (\boldsymbol{\Pi})^{-\frac{1}{2}} \mathbf{e}_\ell$ and letting g_ℓ be the associated function on $\mathbb{S}$, we see that $\mathbf{P}g_\ell = \lambda_\ell g_\ell$, $g_1 \equiv 1$, and $\langle g_k, g_\ell \rangle_{\boldsymbol{\pi}} = \delta_{k,\ell}$. To summarize, when $\mathbb{S}$ has N elements, we have shown that $\mathbf{P}$ on $L^2(\boldsymbol{\pi})$ has eigenvalues $1 = \lambda_1 \geq \cdots \geq \lambda_N \geq -1$ with corresponding eigenfunctions $g_1, \dots, g_N$ which are orthonormal with respect to $\langle\,\cdot\,,\,\cdot\,\rangle_{\boldsymbol{\pi}}$. Of course, since $L^2(\boldsymbol{\pi})$ has dimension N and, by ortho-normality, the g_ℓ's are linearly independent, $(g_1, \dots, g_N)$ is an orthonormal basis in $L^2(\boldsymbol{\pi})$. In particular,

$$\mathbf{P}^n f - \langle f \rangle_{\boldsymbol{\pi}} = \sum_{\ell=2}^N \lambda_\ell^n \langle f, g_\ell \rangle_{\boldsymbol{\pi}} g_\ell \quad \text{for all } n \geq 0 \text{ and } f \in L^2(\boldsymbol{\pi}), \tag{5.1.11}$$

and so

$$\|\mathbf{P}^n f - \langle f \rangle_{\boldsymbol{\pi}}\|^2_{2,\boldsymbol{\pi}} = \sum_{\ell=2}^N \lambda_\ell^{2n} \langle f, g_\ell \rangle^2_{\boldsymbol{\pi}} \leq (1-\beta)^{2n} \|f - \langle f \rangle_{\boldsymbol{\pi}}\|^2_{2,\boldsymbol{\pi}},$$

where $\beta = (1 - \lambda_2) \wedge (1 + \lambda_N)$ is the *spectral gap* between $\{-1, 1\}$ and $\{\lambda_2, \dots, \lambda_N\}$. In other words,

$$\big\|\mathbf{P}^n f - \langle f \rangle_{\boldsymbol{\pi}}\big\|_{2,\boldsymbol{\pi}} \leq (1-\beta)^n \|f - \langle f \rangle_{\boldsymbol{\pi}}\|_{L^2(\boldsymbol{\pi})} \quad \text{for all } n \geq 0 \text{ and } f \in L^2(\boldsymbol{\pi}). \tag{5.1.12}$$

When $\mathbb{S}$ is not finite, it will not be true in general that one can find an orthonormal basis of eigenfunctions for $\mathbf{P}$. Instead, the closest approximation

to the preceding development requires a famous result, known as the Spectral Theorem (cf. §107 in [7]), about bounded, symmetric operators on a Hilbert space. Nonetheless, seeing as it is the estimate (5.1.12) in which we are most interested, we can get away without having to invoke the Spectral Theorem. To be more precise, observe that (5.1.12) holds when

$$\beta \equiv 1 - \sup\{\|\mathbf{P}f - \langle f\rangle_\pi\|_{2,\pi} : f \in L^2(\pi) \text{ with } \|f\|_{2,\pi} = 1\}. \tag{5.1.13}$$

To see this, first note that, when β is given by (5.1.13),

$$\begin{aligned}\|\mathbf{P}f - \langle f\rangle_\pi\|_{2,\pi} &= \|f\|_{2,\pi}\left\|\mathbf{P}\left(\frac{f}{\|f\|_{2,\pi}}\right) - \left\langle \frac{f}{\|f\|_{2,\pi}}\right\rangle_\pi\right\|_{2,\pi} \\ &\le (1-\beta)\|f\|_{2,\pi} \quad \text{when } f \in L^2(\pi)\setminus\{\mathbf{0}\},\end{aligned}$$

and that $\|\mathbf{P}f - \langle f\rangle_\pi\|_{2,\pi} \le \|f\|_{2,\pi}$ trivially when $f = \mathbf{0}$. Next, because $\langle \mathbf{P}^n f\rangle_\pi = \langle f\rangle_\pi$ for all $n \ge 0$,

$$\begin{aligned}\left\|\mathbf{P}^{n+1}f - \langle f\rangle_\pi\right\|_{2,\pi} &= \left\|\mathbf{P}\left(\mathbf{P}^n f - \langle \mathbf{P}^n f\rangle_\pi\right)\right\|_{2,\pi} \\ &\le (1-\beta)\left\|\mathbf{P}^n f - \langle \mathbf{P}^n f\rangle_\pi\right\|_{2,\pi} = (1-\beta)\left\|\mathbf{P}^n f - \langle f\rangle_\pi\right\|_{2,\pi}.\end{aligned}$$

Thus, by induction on n, $\|\mathbf{P}^n f - \langle f\rangle_\pi\|_{2,\pi} \le (1-\beta)^n\|f\|_{2,\pi}$ for all $n \ge 0$ and $f \in L^2(\pi)$. Hence, if $f \in L^2(\pi)$ and $\bar{f} = f - \langle f\rangle_\pi$, then $\|\mathbf{P}^n f - \langle f\rangle_\pi\|_{2,\pi} = \|\mathbf{P}^n \bar{f}\|_{2,\pi} \le (1-\beta)^n\|\bar{f}\|_{2,\pi} = (1-\beta)^n\|f - \langle f\rangle_\pi\|_{2,\pi}$. That is, (5.1.12) holds with the β in (5.1.13). Observe that when $\mathbb{S}$ is finite the β in (5.1.13) coincides with the one in the preceding paragraph. Hence, we have made a first step toward generalizing the contents of that paragraph to situations when (5.1.11) does not apply.

5.1.4. Reversibility and Periodicity: Clearly, the constant β in (5.1.13) can be as small as 0, in which case (5.1.12) tells us nothing. There are three ways in which this might happen. One way is that there exist an $f \in L^2(\pi)$ with the property that, $\|f\|_{2,\pi} = 1$, $\langle f\rangle_\pi = 0$, and $\mathbf{P}f = f$. However, irreducibility rules out the existence of such an f. Indeed, because of irreducibility, we would have (cf. (5.1.8)) the contradiction that $0 = \langle f\rangle_\pi = \lim_{n\to\infty}(\mathbf{A}_n\mathbf{f})_i = f(i)$ for all $i \in \mathbb{S}$ and that $f(i) \ne 0$ for some $i \in \mathbb{S}$. Thus, we can ignore this possibility because it never occurs. A second possibility is that there exists an $f \in L^2(\pi)$ with $\|f\|_{2,\pi} = 1$ such that $\mathbf{P}f = -f$. In fact, if f is such a function, then $\langle f\rangle_\pi = \langle \mathbf{P}f\rangle_\pi = -\langle f\rangle_\pi$, and so $\langle f\rangle_\pi = 0$. Hence, we would have that $\|\mathbf{P}f - \langle f\rangle_\pi\|_{2,\pi} = 1$ and therefore that $\beta = 0$. The third possibility is that there is no non-zero solution to $\mathbf{P}f = -f$ but that, nonetheless, there exists a sequence $\{f_n\}_1^\infty \subseteq L^2(\pi)$ with $\|f_n\|_{2,\pi} = 1$ and $\langle f_n\rangle_\pi = 0$ such that $\|\mathbf{P}f_n\|_{2,\pi}$ tends to 1.

Because the analysis of this last possibility requires the Spectral Theorem, we will not deal with it. However, as the next theorem shows, the second possibility has a pleasing and simple probabilistic interpretation. See Exercise 5.6.7 below for an extension of these considerations to non-reversible $\mathbf{P}$'s.

5.1.14 THEOREM. *If $\mathbf{P}$ is an irreducible transition probability for which there is a reversible initial distribution, which is necessarily $\boldsymbol{\pi}$, then the period of $\mathbf{P}$ is either 1 or 2. Moreover, the period is 2 if and only if there exists an $f \in L^2(\boldsymbol{\pi}) \setminus \{\mathbf{0}\}$ for which $f = -\mathbf{P}f$.*

PROOF: We begin by showing that the period d must be less than or equal to 2. To this end, remember that, because of irreducibility, $(\boldsymbol{\pi})_i > 0$ for all i's. Hence, the detailed balance condition, (5.1.1), implies that $(\mathbf{P})_{ij} > 0 \iff (\mathbf{P})_{ji} > 0$. In particular, since, for each i, $(\mathbf{P})_{ij} > 0$ for some j and therefore $(\mathbf{P}^2)_{ii} = \sum_j (\mathbf{P})_{ij}(\mathbf{P})_{ji} > 0$, we see that the period must divide 2.

To complete the proof at this point, first suppose that $d = 1$. If $f \in L^2(\boldsymbol{\pi})$ satisfies $f = -\mathbf{P}f$, then, as noted before, $\langle f \rangle_{\boldsymbol{\pi}} = 0$, and yet, because of aperiodicity and (5.1.9), $\lim_{n\to\infty} \mathbf{P}^n f(i) = \langle f \rangle_{\boldsymbol{\pi}} = 0$ for each $i \in \mathbb{S}$. Since $f = \mathbf{P}^{2n} f$ for all $n \geq 0$, this means that $f \equiv 0$. Conversely, if $d = 2$, take $\mathbb{S}_0$ and $\mathbb{S}_1$ accordingly, as in §3.2.7, and consider $f = \mathbf{1}_{\mathbb{S}_0} - \mathbf{1}_{\mathbb{S}_1}$. Because of (3.2.19), $\mathbf{P}f = -f$, and clearly $\|f\|_{2,\boldsymbol{\pi}} = 1$. □

As an immediate corollary of the preceding, we can give the following graph theoretic picture of aperiodicity for irreducible, reversible Markov chains. Namely, if we use $\mathbf{P}$ to define a graph structure in which the elements of $\mathbb{S}$ are the "vertices" and an "edge" between i and j exists if and only if $(\mathbf{P})_{ij} > 0$, then the first part of Theorem 5.1.14, in combination with the considerations in §3.2.7, says that the resulting graph is *bipartite* (i.e., splits into two parts in such a way that all edges run from one part to the other) if and only if the chain fails to be aperiodic, and the second part says that this is possible if and only if there exists an $f \in L^2(\boldsymbol{\pi}) \setminus \{\mathbf{0}\}$ satisfying $\mathbf{P}f = -f$.

5.1.5. Relation to Convergence in Variation: Before discussing methods for finding or estimating the β in (5.1.13), it might be helpful to compare the sort of convergence result contained in (5.1.12) to the sort of results we have been getting heretofore. To this end, first observe that

$$\left\|\mathbf{P}^n f - \langle f \rangle_{\boldsymbol{\pi}}\right\|_{2,\boldsymbol{\pi}} \leq \left\|\mathbf{P}^n f - \langle f \rangle_{\boldsymbol{\pi}}\right\|_{\mathrm{u}} \leq \sup_i \|\boldsymbol{\delta}_i \mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \|f\|_{\mathrm{u}}.$$

In particular, if one knows, as one does when Theorem 2.2.1 applies, that

$$(*) \qquad \sup_i \|\boldsymbol{\delta}_i \mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \leq C(1-\epsilon)^n$$

for some $C < \infty$ and $\epsilon \in (0, 1]$, then one has that

$$\|\mathbf{P}^n f - \langle f \rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} \leq C(1-\epsilon)^n \|f\|_{\mathrm{u}},$$

which looks a lot like (5.1.12). Indeed, the only difference is that on the right hand side of (5.1.12), $C = 1$ and the norm is $\|f\|_{2,\boldsymbol{\pi}}$ instead of $\|f\|_{\mathrm{u}}$. Thus, one should suspect that (*) implies that the β in (5.1.12) is at least as large as the ϵ in (*). In, as will always be the case when $\mathbb{S}$ is finite, there exists a $g \in L^2(\boldsymbol{\pi})$

with the properties that $\|g\|_{2,\pi} = 1$, $\langle g\rangle_\pi = 0$, and either $\mathbf{P}g = (1-\beta)g$ or $\mathbf{P}g = -(1-\beta)g$, this suspicion is easy to verify. Namely, set $f = g\mathbf{1}_{[-R,R]}(g)$ where $R > 0$ is chosen so that $a \equiv \langle f, g\rangle_\pi \geq \frac{1}{2}$, and set $\bar{f} = f - \langle f\rangle_\pi$. Then, after writing $\bar{f} = ag + (\bar{f} - ag)$ and noting that $\langle g, \bar{f} - ag\rangle_\pi = 0$, we see that, for any $n \geq 0$,

$$\begin{aligned}\|\mathbf{P}^n\bar{f}\|_{2,\pi}^2 &= a^2(1-\beta)^{2n} \pm 2a(1-\beta)^n\big\langle g, \mathbf{P}^n(\bar{f} - ag)\big\rangle_\pi + \big\|\mathbf{P}^n(\bar{f} - ag)\big\|_{2,\pi}^2 \\ &\geq \frac{1}{4}(1-\beta)^{2n},\end{aligned}$$

since

$$\pm\big\langle g, \mathbf{P}^n(\bar{f} - ag)\big\rangle_\pi = \big\langle \mathbf{P}^n g, \bar{f} - ag\big\rangle_\pi = (1-\beta)^n\langle g, \bar{f} - ag\rangle_\pi = 0.$$

On the other hand, $\|\mathbf{P}^n\bar{f}\|_{2,\pi}^2 \leq C^2(1-\epsilon)^{2n}\|\bar{f}\|_{\mathrm{u}}^2 \leq (CR)^2(1-\epsilon)^{2n}$. Thus, $\frac{1}{4}(1-\beta)^{2n} \leq (CR)^2(1-\epsilon)^{2n}$ for all $n \geq 0$, which is possible only if $\beta \geq \epsilon$. When no such g exists, the same conclusion holds, only one has to invoke the Spectral Theorem in order to arrive at it.

As the preceding shows, uniform estimates on the variation distance between $\boldsymbol{\mu}\mathbf{P}^n$ and $\boldsymbol{\pi}$ imply estimates like the one in (5.1.12). However, going in the opposite direction is not always possible. To examine what can be done, let a probability vector $\boldsymbol{\mu}$ be given, and define f so that $f(i) = \frac{(\boldsymbol{\mu})_i}{(\boldsymbol{\pi})_i}$. Then, since $\langle f\rangle_\pi = 1$ and $\langle g\rangle_\pi^2 \leq \langle g^2\rangle_\pi$ for all $g \in L^2(\boldsymbol{\pi})$,

$$\begin{aligned}\|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} &= \sum_j\big|(\boldsymbol{\mu}\mathbf{P}^n)_j - (\boldsymbol{\pi})_j\big| = \sum_j\big|\langle f, \mathbf{P}^n\mathbf{1}_{\{j\}}\rangle_\pi - (\boldsymbol{\pi})_j\big| \\ &= \sum_j\big|\langle \mathbf{P}^n f - \mathbf{1}, \mathbf{1}_{\{j\}}\rangle_\pi\big| = \sum_j\big|\big\langle \mathbf{P}^n f - \langle f\rangle_\pi, \mathbf{1}_{\{j\}}\big\rangle_\pi\big| \\ &= \sum_j(\boldsymbol{\pi})_j\big|\mathbf{P}^n f(j) - \langle f\rangle_\pi\big| \leq \left(\sum_j(\boldsymbol{\pi})_j\big|\mathbf{P}^n f(j) - \langle f\rangle_\pi\big|^2\right)^{\frac{1}{2}} \\ &= \|\mathbf{P}^n f - \langle f\rangle_\pi\|_{2,\pi}.\end{aligned}$$

Hence, (5.1.12) implies that

$$\|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \leq \left(\sum_i \frac{(\boldsymbol{\mu})_i^2}{(\boldsymbol{\pi})_i} - 1\right)^{\frac{1}{2}}(1-\beta)^n. \tag{5.1.15}$$

In the case when $\mathbb{S}$ is finite, and therefore there exists an $\lambda \in (0,1)$ for which $(\boldsymbol{\pi})_i \geq \lambda$, (5.1.15) yields the Doeblin type estimate

$$\|\boldsymbol{\mu}\mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \leq \left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{2}}(1-\beta)^n.$$

However, when $\mathbb{S}$ is infinite, (5.1.15), as distinguished from Doeblin, does not give a rate of convergence which is independent of $\boldsymbol{\mu}$. In fact, unless $\sum_i \frac{(\boldsymbol{\mu})_i^2}{(\boldsymbol{\pi})_i} < \infty$, it gives no information at all. Thus, one might ask why we are considering estimates like (5.1.15) when Doeblin does as well and sometimes does better. The answer is that, although Doeblin may do well when it works, it seldom works when $\mathbb{S}$ is infinite and, even when $\mathbb{S}$ is finite, it usually gives a far less than optimal rate of convergence. See, for example, Exercise 5.6.15 below.

5.2 Dirichlet Forms and Estimation of β

Our purpose in this section will be to find methods for estimating the optimal (i.e., largest) value of β for which (5.1.12) holds. Again, we will assume that the chain is reversible and irreducible. Later, we will add the assumption that it is aperiodic.

5.2.1. The Dirichlet Form and Poincaré's Inequality: Our first step requires us to find other expressions the right hand side of (5.1.13). To this end, we begin with the observation that

$$1-\beta = \sup\bigl\{\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} : f \in L_0^2(\boldsymbol{\pi}) \text{ with } \|f\|_{2,\boldsymbol{\pi}} = 1\bigr\} \quad \text{where } L_0^2(\boldsymbol{\pi}) \equiv \{f \in L^2(\boldsymbol{\pi}) : \langle f\rangle_{\boldsymbol{\pi}} = 0\}. \tag{5.2.1}$$

It is obvious that the supremum on the right hand side of (5.1.13) dominates the supremum on the right above. On the other hand, if $f \in L^2(\boldsymbol{\pi})$ with $\|f\|_{2,\boldsymbol{\pi}} = 1$, then either $\|f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} = 0$ and therefore $\|\mathbf{P}f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} = 0$, or $1 \geq \|f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}} > 0$, in which case

$$\bigl\|\mathbf{P}f - \langle f\rangle_{\boldsymbol{\pi}}\bigr\|_{2,\boldsymbol{\pi}} \leq \left\|\mathbf{P}\left(\frac{f - \langle f\rangle_{\boldsymbol{\pi}}}{\|f - \langle f\rangle_{\boldsymbol{\pi}}\|_{2,\boldsymbol{\pi}}}\right)\right\|_{2,\boldsymbol{\pi}}$$

is also dominated by the right hand side of (5.2.1).

To go further, we need to borrow the following simple result from the theory of symmetric operators on a Hilbert space. Namely,

$$\sup\bigl\{\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} : f \in L_0^2(\boldsymbol{\pi}) \;\&\; \|f\|_{2,\boldsymbol{\pi}} = 1\bigr\} = \sup\bigl\{|\langle f, \mathbf{P}f\rangle_{\boldsymbol{\pi}}| : f \in L_0^2(\boldsymbol{\pi}) \;\&\; \|f\|_{2,\boldsymbol{\pi}} = 1\bigr\}. \tag{5.2.2}$$

That the right hand side of (5.2.2) is dominated by the left is Schwarz's inequality: $|\langle f, \mathbf{P}f\rangle_{\boldsymbol{\pi}}| \leq \|f\|_{2,\boldsymbol{\pi}}\|\mathbf{P}f\|_{2,\boldsymbol{\pi}}$. To prove the opposite inequality, let $f \in L_0^2(\boldsymbol{\pi})$ with $\|f\|_{2,\boldsymbol{\pi}} = 1$ be given, assume that $\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} > 0$, and set $g = \frac{\mathbf{P}f}{\|\mathbf{P}f\|_{2,\boldsymbol{\pi}}}$. Then $g \in L_0^2(\boldsymbol{\pi})$ and $\|g\|_{2,\boldsymbol{\pi}} = 1$. Hence, if γ denotes the supremum on the right hand side of (5.2.2), then, from the symmetry of $\mathbf{P}$,

$$\begin{aligned} 4\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} &= 4\langle g, \mathbf{P}f\rangle_{\boldsymbol{\pi}} = \langle (f+g), \mathbf{P}(f+g)\rangle_{\boldsymbol{\pi}} - \langle (f-g), \mathbf{P}(f-g)\rangle_{\boldsymbol{\pi}} \\ &\leq \gamma\bigl(\|f+g\|_{2,\boldsymbol{\pi}}^2 + \|f-g\|_{2,\boldsymbol{\pi}}^2\bigr) = 2\gamma\bigl(\|f\|_{2,\boldsymbol{\pi}}^2 + \|g\|_{2,\boldsymbol{\pi}}^2\bigr) = 4\gamma. \end{aligned}$$

The advantage won for us by (5.2.2) is that, in conjunction with (5.2.1), it shows that

$$\begin{aligned} &\beta = \beta_+ \wedge \beta_- \\ &\text{where } \beta_\pm \equiv \inf\bigl\{\langle f, (\mathbf{I} \mp \mathbf{P})f\rangle_\pi : f \in L_0^2(\pi) \ \&\ \|f\|_{2,\pi} = 1\bigr\} \end{aligned} \tag{5.2.3}$$

Notice that, because $(\mathbf{I} - \mathbf{P})f = (\mathbf{I} - \mathbf{P})(f - \langle f\rangle_\pi)$,

$$\begin{aligned} \beta_+ = \inf\bigl\{\langle f, (\mathbf{I} - \mathbf{P})f\rangle_\pi : f \in L^2(\pi) \\ \& \ \mathrm{Var}_\pi(f) \equiv \|f - \langle f\rangle_\pi\|_\pi^2 = 1\bigr\}. \end{aligned} \tag{5.2.4}$$

At the same, because

$$\bigl\langle (f + c), (\mathbf{I} + \mathbf{P})(f + c)\bigr\rangle_\pi = \bigl\langle f_n, (\mathbf{I} + \mathbf{P})f_n\bigr\rangle_\pi + c^2 \quad \text{for } f \in L_0^2(\pi) \text{ and } c \in \mathbb{R},$$

it is clear that the infemum in the definition of β_- is the same whether we consider all $f \in L^2(\pi)$ or just those in $L_0^2(\pi)$. Hence, another expression for β_- is

$$\beta_- = \inf\bigl\{\langle f, (\mathbf{I} + \mathbf{P})f\rangle_\pi : f \in L^2(\pi) \ \&\ \mathrm{Var}_\pi(f) = 1\bigr\}. \tag{5.2.5}$$

Two comments are in order here. First, when $\mathbf{P}$ is *non-negative definite*, abbreviated by $\mathbf{P} \geq 0$, in the sense that $\langle f, \mathbf{P}f\rangle_\pi \geq 0$ for all $f \in L^2(\pi)$, then $\beta_+ \leq 1 \leq \beta_-$, and so $\beta = \beta_+$. Hence,

$$\mathbf{P} \geq 0 \implies \beta = \inf\bigl\{\langle f, (\mathbf{I} - \mathbf{P})f\rangle_\pi : f \in L^2(\pi) \ \&\ \mathrm{Var}_\pi(f) = 1\bigr\}. \tag{5.2.6}$$

Second, by Theorem 5.1.14, we know that, in general, $\beta_- = 0$ unless the chain is aperiodic, and (cf. (5.1.11) and the discussion at the beginning of §5.1.4), when $\mathbb{S}$ is finite, that $\beta > 0$ if and only if $\mathbb{S}$ is aperiodic.

The expressions for β_+ and β_- in (5.2.4) and (5.2.5) lead to important calculational tools. Namely, observe that, by (5.1.1),

$$\begin{aligned} \langle f, (\mathbf{I} - \mathbf{P})f\rangle_\pi &= \sum_{(i,j)} f(i)(\pi)_i(\mathbf{P})_{ij}\bigl(f(i) - f(j)\bigr) \\ &= \sum_{(i,j)} f(i)(\pi)_j(\mathbf{P})_{ji}\bigl(f(i) - f(j)\bigr) = \sum_{(i,j)} f(j)(\pi)_i(\mathbf{P})_{ij}\bigl(f(j) - f(i)\bigr). \end{aligned}$$

Hence, when we add the second expression to the last, we find that

$$\langle f, (\mathbf{I} - \mathbf{P})f\rangle_\pi = \mathcal{E}(f, f) \equiv \frac{1}{2}\sum_{i \neq j} (\pi)_i(\mathbf{P})_{ij}\bigl(f(j) - f(i)\bigr)^2. \tag{5.2.7}$$

Because the quadratic form $\mathcal{E}(f, f)$ is a discrete analog of the famous quadratic form $\frac{1}{2}\int |\nabla f|^2(x)\,dx$ introduced by Dirichlet, it is called a *Dirichlet form*. Extending this metaphor, one interprets β_+ as the *Poincaré constant*

$$\beta_+ = \inf\bigl\{\mathcal{E}(f, f) : f \in L^2(\pi) \ \&\ \mathrm{Var}_\pi(f) = 1\bigr\} \tag{5.2.8}$$

in the *Poincaré inequality*

$$\beta_+ \mathrm{Var}_\pi(f) \le \mathcal{E}(f,f), \quad f \in L^2(\pi). \tag{5.2.9}$$

To make an analogous application of (5.2.5), observe that

$$\sum_{(i,j)} \big(f(j)+f(i)\big)^2 (\pi)_i(\mathbf{P})_{ij} = \langle \mathbf{P}f^2\rangle_\pi + 2\langle f, \mathbf{P}f\rangle_\pi + \|f\|_{2,\pi}^2$$
$$= 2\|f\|_{2,\pi}^2 + 2\langle f, \mathbf{P}f\rangle_\pi,$$

and therefore

$$\langle f, (\mathbf{I}+\mathbf{P})f\rangle_\pi = \tilde{\mathcal{E}}(f,f) \equiv \frac{1}{2}\sum_{(i,j)} (\pi)_i(\mathbf{P})_{ij}\big(f(j)+f(i)\big)^2. \tag{5.2.10}$$

Hence

$$\begin{aligned} \beta_- &= \inf\big\{\tilde{\mathcal{E}}(f,f) : f \in L^2(\pi) \ \&\ \mathrm{Var}_\pi(f) = 1\big\} \\ &= \inf\big\{\tilde{\mathcal{E}}(f,f) : f \in L_0^2(\pi) \ \&\ \|f\|_{2,\pi} = 1\big\}. \end{aligned} \tag{5.2.11}$$

In order to give an immediate application of (5.2.9) and (5.2.11), suppose (cf. Exercise 2.4.3) that $(\mathbf{P})_{ij} \ge \epsilon_j$ for all (i,j), set $\epsilon = \sum_j \epsilon_j$, assume that $\epsilon > 0$, and define the probability vector $\boldsymbol{\mu}$ so that $(\boldsymbol{\mu})_i = \frac{\epsilon_i}{\epsilon}$. Then, by Schwarz's inequality for expectations with respect to $\boldsymbol{\mu}$ and the variational characterization of $\mathrm{Var}_\pi(f)$ as the minimum value of $a \rightsquigarrow \langle (f-a)^2\rangle_\pi$,

$$\begin{aligned} 2\mathcal{E}(f,f) &= \sum_{(i,j)} (\pi)_i(\mathbf{P})_{ij}\big(f(j)-f(i)\big)^2 \\ &\ge \epsilon \sum_{(i,j)} \big(f(j)-f(i)\big)^2 (\pi)_i(\boldsymbol{\mu})_j \ge \epsilon \sum_i \big(\langle f\rangle_\mu - f(i)\big)^2 (\pi)_i \ge \epsilon \mathrm{Var}_\pi(f), \end{aligned}$$

and, similarly, $\tilde{\mathcal{E}}(f,f) \ge \frac{\epsilon}{2}\mathrm{Var}_\pi(f)$. Hence, by (5.2.9), (5.2.11), and (5.2.3), $\beta \ge \frac{\epsilon}{2}$. Of course, this result is a significantly less strong than the one we get by combining Exercise 2.4.3 with the reasoning in §5.1.4. Namely, by that exercise we know that $\|\boldsymbol{\delta}_i \mathbf{P}^n - \boldsymbol{\pi}\|_{\mathrm{v}} \le 2(1-\epsilon)^n$, and so the reasoning in §5.1.4 tells us that $\beta \ge \epsilon$, which is twice as good as the estimate we are getting here. On the other hand, as we already advertised, there are circumstances when the considerations here yield results when a Doeblin-type approach does not.

5.2.2. Estimating β_+: The origin of many applications of (5.2.9) and (5.2.11) to estimate β_+ and β_- is the simple observation that

$$\mathrm{Var}_\pi(f) = \frac{1}{2}\sum_{ij} \big(f(i)-f(j)\big)^2 (\pi)_i(\pi)_j, \tag{5.2.12}$$

which is easily checked by expanding $\big(f(i) - f(j)\big)^2$ and seeing that the sum on the right equals $2\langle f^2\rangle_\pi - 2\langle f\rangle_\pi^2$.

The importance of (5.2.12) is that it expresses $\mathrm{Var}_\pi(f)$ in terms of difference between the values of f at different points in $\mathbb{S}$, and clearly $\mathcal{E}(f,f)$ is also given in terms of such differences. However, the differences which appear in $\mathcal{E}(f,f)$ are only between the values of f at pairs of points (i,j) for which $(\mathbf{P})_{ij} > 0$, whereas the right hand of side of (5.2.12) entails sampling all pairs (i,j). Thus, in order to estimate $\mathrm{Var}_\pi(f)$ in terms of $\mathcal{E}(f,f)$, it is necessary to choose, for each (i,j) with $i \neq j$, a *path* $p(i,j) = (k_0,\dots,k_n) \in \mathbb{S}^{n+1}$ with $k_0 = i$ and $k_n = j$, which is *allowable* in the sense $(\mathbf{P})_{k_{m-1}\,k_m} > 0$ for each $1 \le m \le n$, and to write

$$\big(f(i) - f(j)\big)^2 = \left(\sum_{e\in p(i,j)} \Delta_e f\right)^2$$

where the summation in e is taken over the oriented segments (k_{m-1}, k_m) in the path $p(i,j)$, and, for $e = (k,\ell)$, $\Delta_e f \equiv f(\ell) - f(k)$. At this point there are various ways in which one can proceed. For example, given any $\{\alpha(e) : e \in p(i,j)\} \subseteq (0,\infty)$, Schwarz's inequality (cf. Exercise 1.3.1) shows that the quantity on the right is dominated by

$$\left(\sum_{e\in p(i,j)} \frac{\alpha(e)}{\rho(e)}\right)\left(\sum_{e\in p(i,j)} \frac{\rho(e)}{\alpha(e)}\big(\Delta_e f\big)^2\right)$$
$$\le \max_{e\in p(i,j)} \frac{1}{\alpha(e)}\left(\sum_{e\in p(i,j)} \frac{\alpha(e)}{\rho(e)}\right)\sum_{e\in p(i,j)} \big(\Delta_e f\big)^2\rho(e),$$

where $\rho(e) \equiv (\boldsymbol{\pi})_k(\mathbf{P})_{k\ell}$ when $e = (k,\ell)$. Thus, for any selection of paths $\mathcal{P} = \{p(i,j) : (i,j) \in \mathbb{S}^2 \setminus D\}$ (D here denotes the diagonal $\{(i,j) \in \mathbb{S}^2 : i = j\}$) and coefficients $\mathcal{A} = \{\alpha(e,p) : e \in p \in \mathcal{P}\} \subseteq (0,\infty)$,

$$\mathrm{Var}_\pi(f) \le \frac{1}{2}\sum_{p\in\mathcal{P}} w_A(p)\sum_{e\in p}\big(\Delta_e f\big)^2\rho(e)$$
$$= \frac{1}{2}\sum_e \big(\Delta_e f\big)^2\rho(e)\left(\sum_{p\ni e} w_A(p)\right) \le W(\mathcal{P},\mathcal{A})\mathcal{E}(f,f),$$

where

$$w_{\mathcal{A}}(p) \equiv (\boldsymbol{\pi})_i(\boldsymbol{\pi})_j\left(\max_{e\in p}\frac{1}{\alpha(e,p)}\right)\sum_{e\in p}\frac{\alpha(e,p)}{\rho(e)} \quad \text{if } p = p(i,j),$$

and

$$W(\mathcal{P},\mathcal{A}) = \sup_e \sum_{p\ni e} w_{\mathcal{A}}(e,p).$$

Hence, we have now shown that

$$\beta_+ \geq \frac{1}{W(\mathcal{P},\mathcal{A})} \tag{5.2.13}$$

for every choice of allowable paths $\mathcal{P}$ and coefficients $\mathcal{A}$.

The most effective applications of (5.2.13) depend on making a choice of $\mathcal{P}$ and $\mathcal{A}$ which takes advantage of the particular situation under consideration. Given a selection $\mathcal{P}$ of paths, one of the most frequently made choices of $\mathcal{A}$ is to take $\alpha(e,p) \equiv 1$. In this case, (5.2.13) gives

$$\beta_+ \geq \frac{1}{W(\mathcal{P})} \quad \text{where } W(\mathcal{P}) \equiv \sup_e \sum_{p \ni e} \sum_{e' \in p} \frac{(\boldsymbol{\pi}(p))_- (\boldsymbol{\pi}(p))_+}{\rho(e')}, \tag{5.2.14}$$

where $(\boldsymbol{\pi}(p))_- = (\boldsymbol{\pi})_i$ and $(\boldsymbol{\pi}(p))_+ = (\boldsymbol{\pi})_j$ when p begins at i and ends at j.

Finally, it should be recognized that, in general, (5.2.13) gives no information. Indeed, although irreducibility guarantees that there is always at least one path connecting every pair of points, when $\mathbb{S}$ is infinite there is no guarantee that $\mathcal{P}$ and $\mathcal{A}$ can be chosen so that $W(\mathcal{P},\mathcal{A}) < \infty$. Moreover, even when $\mathbb{S}$ is finite, and therefore $W(\mathcal{P},\mathcal{A}) < \infty$ for every choice, only a judicious choice will make (5.2.13) yield a good estimate.

5.2.3. Estimating β_-: The estimation of β_- starting from (5.2.11) is a bit more contrived than the one of β_+ starting from (5.2.9). For one thing, we already know (cf. Theorem 5.1.14) that $\beta_- = 0$ unless the chain is aperiodic. Thus, we will now require that the chain be aperiodic. As a consequence of aperiodicity and irreducibility, we know (cf. (3.1.13)) that, for each $i \in \mathbb{S}$, there always is a path $p(i) = (k_0, \ldots, k_{2n+1})$ which is allowable (i.e. $(\mathbf{P})_{k_{m-1}k_m} > 0$ for each $1 \leq m \leq 2n+1$), is closed (i.e., $k_0 = k_{2n+1}$), and starts at i (i.e., $k_0 = i$). Note our insistence that this path have an odd number of steps. The reason for our doing so is that when the number of steps is odd, an elementary exercise in telescoping sums shows that

$$2f(i) = \sum_{m=0}^{2n} (-1)^m \big(f(k_m) + f(k_{m+1})\big).$$

Thus, if $\tilde{\mathcal{P}}$ be a selection of such paths, one path $p(i)$ for each $i \in \mathbb{S}$, and we make an associated choice of coefficients $\mathcal{A} = \{\alpha(e,p) : e \in p \in \tilde{\mathcal{P}}\} \subseteq (0,\infty)$. Then, just as in the preceding section,

$$\begin{aligned} 4\|f\|^2_{2,\boldsymbol{\pi}} &= \sum_i \left(\sum_{e \in p(i)} (-1)^{m(e)} \tilde{\Delta}_e f \right)^2 (\boldsymbol{\pi})_i \\ &\leq \sum_{p \in \tilde{\mathcal{P}}} \boldsymbol{\pi}(p) \left(\sum_{e \in p} \frac{\alpha(e,p)}{\rho(e)} \right) \left(\sum_{e \in p} (\tilde{\Delta}_e f)^2 \frac{\rho(e)}{\alpha(e,p)} \right), \end{aligned}$$

where, when $p = (k_0, \dots, k_{2n+1})$, $\boldsymbol{\pi}(p) = (\boldsymbol{\pi})_{k_0}$, and, for $0 \le m \le 2n$, $m(e) = m$ and $\tilde{\Delta}_e f \equiv f(k_m) + f(k_{m+1})$ if $e = (k_m, k_{m+1})$. Hence, if

$$\tilde{w}(p) = \boldsymbol{\pi}(p) \left(\max_{e \in p} \frac{1}{\alpha(e,p)} \right) \sum_{e \in p} \frac{\alpha(e,p)}{\rho(e)},$$

then

$$2\|f\|^2_{2,\boldsymbol{\pi}} \le \tilde{W}(\tilde{\mathcal{P}}, \mathcal{A}) \tilde{\mathcal{E}}(f,f) \quad \text{where } \tilde{W}(\tilde{\mathcal{P}}, \mathcal{A}) \equiv \sup_e \sum_{p \ni e} \tilde{w}(p),$$

and, by (5.2.11), this proves that

$$\beta_- \ge \frac{2}{\tilde{W}(\tilde{\mathcal{P}}, \mathcal{A})}$$

for any choice of paths $\tilde{\mathcal{P}}$ and coefficients $\mathcal{A}$ satisfying the stated requirements. When we take $\alpha(e,p) \equiv 1$, then this specializes to

$$\beta_- \ge \frac{2}{\tilde{W}_1(\tilde{\mathcal{P}})} \quad \text{where } \tilde{W}_1(\tilde{\mathcal{P}}) \equiv \sup_e \sum_{p \ni e} \sum_{e' \in p} \frac{\boldsymbol{\pi}(p)}{\rho(e')}, \tag{5.2.15}$$

It should be emphasized that the preceding method for getting estimates on β_- is inherently flawed in that it appears incapable of recognizing spectral properties of $\mathbf{P}$ like non-negative definiteness. That is, when $\mathbf{P}$ is non-negative definite, then $\beta_- \ge 1$, but it seems unlikely that one could get that conclusion out of the arguments being used here. On the other hand, because our real interest is in $\beta = \beta_+ \wedge \beta_-$ and the estimates given by (5.2.14) and (5.2.15) are likely to be comparable, the inadequacy of (5.2.15) causes less damage than one otherwise might fear.

5.3 Reversible Markov Processes in Continuous Time

Here we will see what the preceding theory looks like in the continuous time context and will learn that it is both easier and more æsthetically pleasing there.

We will be working with the notation and theory developed in Chapter 4. In particular, $\mathbf{Q}$ will denote a matrix of the form $\mathbf{R}(\mathbf{P} - \mathbf{I})$, where $\mathbf{R}$ is a diagonal matrix whose diagonal entries are the rates $\mathfrak{R} = \{R_i : i \in \mathbb{S}\}$ and $\mathbf{P}$ is a transition probability matrix. We will assume throughout that $\mathbf{Q}$ is irreducible in the sense that (cf. (4.4.2)), for each pair $(i,j) \in \mathbb{S}^2$, $(\mathbf{Q}^n)_{ij} > 0$ for some $n \ge 0$.

5.3.1. Criterion for Reversibility: Let $\mathbf{Q}$ be given, assume that the associated Markov process never explodes (cf. §4.3.1), and use $\{\mathbf{P}(t) : t \ge 0\}$ to denote the semigroup determined by $\mathbf{Q}$ (cf. Corollary 4.3.2). Our purpose

in this subsection is to show that if $\hat{\boldsymbol{\mu}}$ is a probability vector for which the *detailed balance condition*

$$(\hat{\boldsymbol{\mu}})_i(\mathbf{Q})_{ij} = (\hat{\boldsymbol{\mu}})_j(\mathbf{Q})_{ji} \quad \text{for all } (i,j) \in \mathbb{S}^2, \tag{5.3.1}$$

holds relative to $\mathbf{Q}$, then the detailed balance condition

$$(\hat{\boldsymbol{\mu}})_i\big(\mathbf{P}(t)\big)_{ij} = (\hat{\boldsymbol{\mu}})_j\big(\mathbf{P}(t)\big)_{ji} \quad \text{for all } t > 0 \text{ and } (i,j) \in \mathbb{S}^2 \tag{5.3.2}$$

also holds.

The proof that (5.3.1) implies (5.3.2) is trivial in the case when the rates $\mathfrak{R}$ are bounded. Indeed, all that we need to do in that case is first use a simple inductive argument to check that (5.3.1) implies $(\hat{\boldsymbol{\mu}})_i(\mathbf{Q}^n)_{ij} = (\hat{\boldsymbol{\mu}})_j(\mathbf{Q}^n)_{ji}$ for all $n \geq 0$ and then use the expression for $\mathbf{P}(t)$ given in (4.2.13). When the rates are unbounded, we will use the approximation procedure introduced in §4.3.1. Namely, refer to §4.3.1, and take $\mathbf{Q}^{(N)}$ corresponding to the choice rates $\mathfrak{R}^{(N)}$ described there. Equivalently, take $(\mathbf{Q}^{(N)})_{ij}$ to be $(\mathbf{Q})_{ij}$ if $i \in F_N$ and 0 when $i \notin F_N$. Using induction, one finds first that $((\mathbf{Q}^{(N)})^n)_{ij} = 0$ for all $n \geq 1$ and $i \notin F_N$ and second that

$$(\hat{\boldsymbol{\mu}})_i\big((\mathbf{Q}^{(N)})^n\big)_{ij} = (\hat{\boldsymbol{\mu}})_j\big((\mathbf{Q}^{(N)})^n\big)_{ji} \quad \text{for all } n \geq 0 \text{ and } (i,j) \in F_N^2.$$

Hence, if $\{\mathbf{P}^{(N)}(t) : t > 0\}$ is the semigroup determined by $\mathbf{Q}^{(N)}$, then, since the rates for $\mathbf{Q}^{(N)}$ are bounded, (4.2.13) shows that

$$\begin{aligned} \big(\mathbf{P}^{(N)}(t)\big)_{ij} &= \delta_{i,j} && \text{if } i \notin F_N \\ (\hat{\boldsymbol{\mu}})_i\big(\mathbf{P}^{(N)}(t)\big)_{ij} &= (\hat{\boldsymbol{\mu}})_j\big(\mathbf{P}^{(N)}(t)\big)_{ji} && \text{if } (i,j) \in F_N^2. \end{aligned} \tag{5.3.3}$$

In particular, because, by (4.3.3), $\big(\mathbf{P}^{(N)}(t)\big)_{ij} \longrightarrow \big(\mathbf{P}(t)\big)_{ij}$, we are done.

As a consequence of the preceding, we now know that (5.3.1) implies that $\hat{\boldsymbol{\mu}}$ is stationary for $\mathbf{P}(t)$. Hence, because we are assuming that $\mathbf{Q}$ is irreducible, the results in §4.4.2 and §4.4.3 allow us to identify $\hat{\boldsymbol{\mu}}$ as the probability vector $\hat{\boldsymbol{\pi}} \equiv \hat{\boldsymbol{\pi}}^{\mathbb{S}}$ introduced in Theorem 4.4.8 and discussed in §4.4.3, especially (4.4.11). To summarize, *if $\hat{\boldsymbol{\mu}}$ is a probability vector for which (5.3.1) holds, then $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\pi}}$.*

5.3.2. Convergence in $L^2(\hat{\boldsymbol{\pi}})$ for Bounded Rates: In view of the results just obtained, from now on we will be assuming that $\hat{\boldsymbol{\pi}}$ is a probability vector for which (5.3.1) holds when $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\pi}}$. In particular, this means that

$$(\hat{\boldsymbol{\pi}})_i\big(\mathbf{P}(t)\big)_{ij} = (\hat{\boldsymbol{\pi}})_j\big(\mathbf{P}(t)\big)_{ji} \quad \text{for all } t > 0 \text{ and } (i,j) \in \mathbb{S}^2. \tag{5.3.4}$$

Knowing (5.3.4), one is tempted to use the ideas in §5.2 to get an estimate on the rate, as measured by convergence in $L^2(\hat{\boldsymbol{\pi}})$, at which $\mathbf{P}(t)f$ tends to $\langle f \rangle_{\hat{\boldsymbol{\pi}}}$. To be precise, first note that, for each $h > 0$,

$$\langle f, \mathbf{P}(h)f \rangle_{\hat{\boldsymbol{\pi}}} = \big\|\mathbf{P}\big(\tfrac{h}{2}\big)f\big\|^2_{2,\hat{\boldsymbol{\pi}}} \geq 0,$$

and therefore (cf. (5.1.13),(5.2.6), and (5.2.7)) that

$$\begin{aligned}\beta(h) &\equiv 1 - \sup\{|\langle f, \mathbf{P}(h)f\rangle_{\hat{\boldsymbol{\pi}}}| : f \in L_0^2(\hat{\boldsymbol{\pi}}) \text{ with } \|f\|_{2,\hat{\boldsymbol{\pi}}} = 1\} \\ &= \inf\{\langle f, \mathbf{I} - \mathbf{P}(h)f\rangle_{\hat{\boldsymbol{\pi}}} : \mathrm{Var}_{\hat{\boldsymbol{\pi}}}(f) = 1\} = \inf\{\mathcal{E}_h(f,f) : \mathrm{Var}_{\hat{\boldsymbol{\pi}}}(f) = 1\},\end{aligned}$$

where

$$\mathcal{E}_h(f,f) \equiv \frac{1}{2}\sum_{i\neq j}(\hat{\boldsymbol{\pi}})_i\big(\mathbf{P}(h)\big)_{ij}\big(f(j) - f(i)\big)^2$$

is the Dirichlet form for $\mathbf{P}(h)$ on $L^2(\hat{\boldsymbol{\pi}})$. Hence (cf. (5.1.12)), for any $t > 0$ and $n \in \mathbb{Z}^+$,

$$\big\|\mathbf{P}(t)f - \langle f\rangle_{\hat{\boldsymbol{\pi}}}\big\|_{L^2(\hat{\boldsymbol{\pi}})} \leq \Big(1 - \beta\big(\tfrac{t}{n}\big)\Big)^n \|f - \langle f\rangle_{\hat{\boldsymbol{\pi}}}\|_{2,\hat{\boldsymbol{\pi}}}.$$

To take the next step, we add the assumption that the rates $\mathfrak{R}$ are bounded. One can then use (4.2.8) to see that, uniformly in f satisfying $\mathrm{Var}_{\hat{\boldsymbol{\pi}}}(f) = 1$,

$$\lim_{h\searrow 0}\frac{\mathcal{E}_h(f,f)}{h} = \mathcal{E}^{\mathbf{Q}}(f,f)$$

where

$$\mathcal{E}^{\mathbf{Q}}(f,f) \equiv \frac{1}{2}\sum_{i\neq j}(\hat{\boldsymbol{\pi}})_i(\mathbf{Q})_{ij}\big(f(j) - f(i)\big)^2; \tag{5.3.5}$$

and from this it follows that the limit $\lim_{h\searrow 0} h^{-1}\beta(h)$ exists and is equal to

$$\lambda \equiv \inf\{\mathcal{E}^{\mathbf{Q}}(f,f) : f \in L^2(\boldsymbol{\pi}) \;\&\; \mathrm{Var}_{\hat{\boldsymbol{\pi}}}(f) = 1\}. \tag{5.3.6}$$

Thus, at least when the rates are bounded, we know that

$$\big\|\mathbf{P}(t)f - \langle f\rangle_{\hat{\boldsymbol{\pi}}}\big\|_{2,\hat{\boldsymbol{\pi}}} \leq e^{-\lambda t}\|f - \langle f\rangle_{\hat{\boldsymbol{\pi}}}\|_{2,\hat{\boldsymbol{\pi}}}. \tag{5.3.7}$$

5.3.3. $L^2(\hat{\boldsymbol{\pi}})$-Convergence Rate in General: When the rates are unbounded, the preceding line of reasoning is too naïve. In order to treat the unbounded case, one needs to make some additional observations, all of which have their origins in the following lemma.[3]

5.3.8 LEMMA. *Given $f \in L^2(\hat{\boldsymbol{\pi}})$, the function $t \in [0,\infty) \longmapsto \|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2$ is continuous, non-increasing, non-negative, and convex. In particular, $t \in (0,\infty) \longmapsto \frac{\langle f,(\mathbf{I}-\mathbf{P}(t))f\rangle_{\hat{\boldsymbol{\pi}}}}{t}$ is non-increasing, and therefore*

$$\lim_{h\searrow 0}\frac{\|f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(h)f\|_{2,\hat{\boldsymbol{\pi}}}^2}{h} \quad \textit{exists in } [0,\infty].$$

[3] If one knows spectral theory, especially Stone's Theorem, the rather cumbersome argument which follows can be avoided.

In fact (cf. (5.3.5)),

$$\lim_{h \searrow 0} \frac{\|f\|_{2,\hat{\pi}}^2 - \|\mathbf{P}(h)f\|_{2,\hat{\pi}}^2}{h} \geq 2\mathcal{E}^{\mathbf{Q}}(f,f).$$

PROOF: Let $f \in L^2(\hat{\pi})$ be given, and (cf. the notation in §4.3.1) define $f_N = \mathbf{1}_{F_N} f$. Then, by Lebesgue's Dominated Convergence Theorem, $\|f - f_N\|_{2,\hat{\pi}} \longrightarrow 0$ as $N \to \infty$. Because $\|\mathbf{P}(t)g\|_{2,\hat{\pi}} \leq \|g\|_{2,\hat{\pi}}$ for all $t > 0$ and $g \in L^2(\hat{\pi})$, we know that

$$\big|\|\mathbf{P}(t)f\|_{2,\hat{\pi}} - \|\mathbf{P}(t)f_N\|_{2,\hat{\pi}}\big| \leq \big\|\mathbf{P}(t)(f - f_N)\big\|_{2,\hat{\pi}} \leq \|f - f_N\|_{2,\hat{\pi}} \longrightarrow 0$$

uniformly in $t \in (0,\infty)$ as $N \to \infty$. Hence, by part (**a**) of Exercise 5.6.1 below, in order to prove the initial statement, it suffices to do so when f vanishes off of F_M for some M. Now let f be a function which vanishes off of F_M, and set $\psi(t) = \|\mathbf{P}(t)f\|_{2,\hat{\pi}}^2$. At the same time, set $\psi_N(t) = \|\mathbf{P}^{(N)}(t)f\|_{2,\hat{\pi}}^2$ for $N \geq M$. Then, because by (4.3.3), $\psi_N \longrightarrow \psi$ uniformly on finite intervals, another application of part (**a**) in Exercise 5.6.1 allows us to restrict our attention to the ψ_N's. That is, we will have proved that ψ is a continuous, non-increasing, non-negative, convex function as soon as we show that each ψ_N is. The non-negativity requires no comment. To prove the other properties, we apply (4.2.7) to see that

$$\dot{\psi}_N(t) = \big\langle \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f, \mathbf{P}^{(N)}(t)f\big\rangle_{\hat{\pi}} + \big\langle \mathbf{P}^{(N)}(t)f, \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f\big\rangle_{\hat{\pi}}.$$

Next, by the first line of (5.3.3), we know that, because $N \geq M$, $\mathbf{P}^{(N)}(t)f$ vanishes off of F_N, and so, because $(\hat{\pi})_i \mathbf{Q}_{ij}^{(N)} = (\hat{\pi})_j \mathbf{Q}_{ji}^{(N)}$ for $(i,j) \in F_N^2$, the preceding becomes

$$\dot{\psi}_N(t) = 2\big\langle \mathbf{P}^{(N)}(t)f, \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f\big\rangle_{\hat{\pi}}.$$

Similarly, we see that

$$\begin{aligned}\ddot{\psi}_N(t) &= 2\big\langle \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f, \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f\big\rangle_{\hat{\pi}} \\ &\quad + 2\big\langle \mathbf{P}^{(N)}(t)f, (\mathbf{Q}^{(N)})^2\mathbf{P}^{(N)}(t)f\big\rangle_{\hat{\pi}} \\ &= 4\big\langle \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f, \mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f\big\rangle_{\hat{\pi}} = 4\big\|\mathbf{Q}^{(N)}\mathbf{P}^{(N)}(t)f\big\|_{2,\hat{\pi}}^2 \geq 0.\end{aligned}$$

Clearly, the second of these proves the convexity of ψ_N. In order to see that the first implies that ψ_N is non-increasing, we will show that

$$\text{(*)} \qquad \begin{aligned}\langle g, -\mathbf{Q}^{(N)}g\rangle_{\hat{\pi}} &= \frac{1}{2}\sum_{\substack{(i,j)\in F_N^2 \\ i\neq j}} (\hat{\pi})_i(\mathbf{Q})_{ij}\big(g(j) - g(i)\big)^2 + \sum_{i\in F_N} (\hat{\pi})_i V_i^{(N)} g(i)^2 \\ &\text{if } g = 0 \text{ off } F_N \quad \text{and } V^{(N)}(i) \equiv \sum_{j\notin F_N} Q_{ij} \quad \text{for } i \in F_N.\end{aligned}$$

To check (*), first observe that

$$\langle g, -\mathbf{Q}^{(N)} g\rangle_{\hat{\boldsymbol{\pi}}} = -\sum_{\substack{(i,j)\in F_N^2 \\ i\neq j}} (\hat{\boldsymbol{\pi}})_i (\mathbf{Q})_{ij} g(i) g(j)$$

$$= -\sum_{\substack{(i,j)\in F_N^2 \\ i\neq j}} (\hat{\boldsymbol{\pi}})_i (\mathbf{Q})_{ij} g(i)\big(g(j)-g(i)\big) + \sum_{i\in F_N} (\hat{\boldsymbol{\pi}})_i V^{(N)}(i) g(i)^2.$$

Next, use $(\hat{\boldsymbol{\pi}})_i (\mathbf{Q})_{ij} = (\hat{\boldsymbol{\pi}})_j (\mathbf{Q})_{ji}$ for $(i,j) \in F_N^2$ to see that

$$-\sum_{\substack{(i,j)\in F_N \\ i\neq j}} (\hat{\boldsymbol{\pi}})_i (\mathbf{Q})_{ij} g(i)\big(g(j)-g(i)\big) = \sum_{\substack{(i,j)\in F_N \\ i\neq j}} (\hat{\boldsymbol{\pi}})_i (\mathbf{Q})_{ij} g(j)\big(g(j)-g(i)\big),$$

and thereby arrive at (*). Finally, apply (*) with $g = \mathbf{P}^{(N)}(t) f$ to conclude that $\dot{\psi}_N \leq 0$.

Turning to the second and third assertions, let f be any element of $L^2(\hat{\boldsymbol{\pi}})$. Now that we know that the corresponding ψ is a continuous, non-increasing, non-negative, convex function, it is easy (cf. part (**d**) in Exercise 5.6.1) to check that $t \nearrow \frac{\psi(0)-\psi(t)}{t}$ is non-increasing and therefore that $\lim_{h\searrow 0} \frac{\psi(0)-\psi(h)}{h}$ exists in $[0,\infty]$. Next, remember (5.2.7), apply it when $\mathbf{P} = \mathbf{P}(2h)$, and conclude that

$$\psi(0) - \psi(h) = \big\langle f, \big(\mathbf{I} - \mathbf{P}(2h)\big) f\big\rangle_{\hat{\boldsymbol{\pi}}} = \frac{1}{2} \sum_{\substack{(i,j) \\ i\neq j}} (\hat{\boldsymbol{\pi}})_i \big(\mathbf{P}(2h)\big)_{ij} \big(f(j)-f(i)\big)^2.$$

Hence, since, by (4.3.4),

$$\lim_{h\searrow 0} \frac{\big(\mathbf{P}(2h)\big)_{ij}}{h} = 2(\mathbf{Q})_{ij} \quad \text{for } i \neq j,$$

the required inequality follows after an application of Fatou's Lemma. □

5.3.9 LEMMA. *If $0 < s < t$, then for any $f \in L^2(\hat{\boldsymbol{\pi}})$*

$$\frac{\|f\|_{2,\hat{\boldsymbol{\pi}}}^2}{s} \geq \frac{\|\mathbf{P}(s)f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2}{t-s} \geq 2\mathcal{E}^{\mathbf{Q}}\big(\mathbf{P}(t)f, \mathbf{P}(t)f\big).$$

PROOF: Set $\psi(t) = \|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2$. We know that ψ is a continuous, non-increasing, non-negative, convex function. Hence, by part (**a**) of Exercise 5.6.1,

$$\frac{\psi(0)}{s} \geq \frac{\psi(0)-\psi(s)}{s} \geq \frac{\psi(s)-\psi(t)}{t-s} \geq \frac{\psi(t)-\psi(t+h)}{h}$$

for any $h > 0$. Moreover, because,

$$\frac{\psi(t) - \psi(t+h)}{h} = \frac{\|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(h)\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2}{h}$$

the last part of Lemma 5.3.8 applied with $\mathbf{P}(t)f$ replacing f yields the second asserted estimate. □

With the preceding at hand, we can now complete our program. Namely, by writing

$$\|f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(t)\|_{2,\hat{\boldsymbol{\pi}}}^2 = \sum_{m=0}^{n-1} \Bigl(\bigl\|\mathbf{P}\bigl(\tfrac{mt}{n}\bigr)\bigr\|_{2,\hat{\boldsymbol{\pi}}}^2 - \bigl\|\mathbf{P}\bigl(\tfrac{(m+1)t}{n}\bigr)\bigr\|_{2,\hat{\boldsymbol{\pi}}}^2 \Bigr),$$

we can use the result in Lemma 5.3.9 to obtain the estimate

$$\|f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2 \geq \frac{2t}{n} \sum_{m=1}^{n} \mathcal{E}^{\mathbf{Q}}\Bigl(\mathbf{P}\bigl(\tfrac{mt}{n}\bigr)f, \mathbf{P}\bigl(\tfrac{mt}{n}\bigr)f\Bigr).$$

Hence, if λ is defined as in (5.3.6), then, for any $f \in L_0^2(\hat{\boldsymbol{\pi}})$,

$$\|f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2 \geq \frac{2\lambda t}{n} \sum_{m=1}^{n} \bigl\|\mathbf{P}\bigl(\tfrac{mt}{n}\bigr)f\bigr\|_{2,\hat{\boldsymbol{\pi}}}^2,$$

which, when $n \to \infty$, leads to

$$\|f\|_{2,\hat{\boldsymbol{\pi}}}^2 - \|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2 \geq 2\lambda \int_0^t \|\mathbf{P}(\tau)f\|_{2,\hat{\boldsymbol{\pi}}}^2 \, d\tau.$$

Finally, by *Gronwall's inequality* (cf. Exercise 5.6.4), the preceding yields the estimate $\|\mathbf{P}(t)f\|_{2,\hat{\boldsymbol{\pi}}}^2 \leq e^{-2\lambda t}\|f\|_{2,\hat{\boldsymbol{\pi}}}^2$. After replacing a general $f \in L^2(\hat{\boldsymbol{\pi}})$ by $f - \langle f \rangle_{\hat{\boldsymbol{\pi}}}$, we have now proved that (5.3.7) holds even when $\mathfrak{R}$ is unbounded.

5.3.4. Estimating λ: Proceeding in the exactly the same way that we did in §5.2.2, we can estimate the λ in (5.3.6) in the same way as we estimated β_+ there. Namely, we make a selection $\mathcal{P}$ consisting of paths $p(i,j)$, one for each from pair $(i,j) \in \mathbb{S} \setminus D$, with the properties that if $p(i,j) = (k_0, \ldots, k_n)$, then $k_0 = i$, $k_n = j$, and $p(i,j)$ is *allowable* in the sense that $(\mathbf{Q})_{k_{m-1}k_m} > 0$ for each $1 \leq m \leq n$. Then, just as in §5.2.2, we can say that

$$\lambda \geq \frac{1}{W(\mathcal{P})} \quad \text{where } W(\mathcal{P}) \equiv \sup_e \sum_{p \ni e} \sum_{e' \in p} \frac{(\hat{\boldsymbol{\pi}}(p))_- - (\hat{\boldsymbol{\pi}}(p))_+}{\rho(e')}, \tag{5.3.10}$$

where the supremum is over oriented edges $e = (k,\ell)$ with $(\mathbf{Q})_{k\ell} > 0$, the first sum is over $p \in \mathcal{P}$ in which the edge e appears, the second sum is over edges e' which appear in the path p, $(\hat{\boldsymbol{\pi}}(p))_- = (\hat{\boldsymbol{\pi}})_i$ if the path p starts at i, $(\hat{\boldsymbol{\pi}}(p))_+ = (\hat{\boldsymbol{\pi}})_j$ if the path p ends at j, and $\rho(e') = (\hat{\boldsymbol{\pi}})_k(\mathbf{Q})_{k\ell}$ if $e' = (k,\ell)$.

5.4 Gibbs States and Glauber Dynamics

Loosely speaking, the physical principle underlying statistical mechanics can be summarized in the statement that, when a system is in equilibrium, *states with lower energy are more likely than those with higher energy.* In fact, J.W. Gibbs sharpened this statement by saying that the probability of a state i will be proportional to $e^{-\frac{H(i)}{kT}}$, where k is the Boltzmann constant, T is temperature, and $H(i)$ is the energy of the system when it is in state i. For this reason, a distribution which assigns probabilities in this Gibbsian manner is called a *Gibbs state.*

Since a Gibbs state is to be a model of equilibrium, it is only reasonable to ask what is the dynamics for which it is the equilibrium. From our point of view, this means that we should seek a Markov process for which the Gibbs state is the stationary distribution. Further, because dynamics in physics should be reversible, we should be looking for Markov processes which are reversible with respect to the Gibbs state, and, because such processes were introduced in this context by R. Glauber, we will call a Markov process which is reversible with respect to a Gibbs state a *Glauber dynamics* for that Gibbs state.

In this section, we will give a rather simplistic treatment of Gibbs states and their associated Glauber dynamics.

5.4.1. Formulation: Throughout this section, we will be working in the following setting. As usual, $\mathbb{S}$ is either a finite or countably infinite space. On $\mathbb{S}$ there is given some "natural" background assignment $\boldsymbol{\nu} \in (0,\infty)^{\mathbb{S}}$ of (not necessarily summable) weights, which should be thought of as a row vector. In many applications, $\boldsymbol{\nu}$ is uniform: it assigns each i weight 1, but in other situations it is convenient to not have to assume that it is uniform. Next, there is a function $H : \mathbb{S} \longrightarrow [0,\infty)$ (alias, the energy function) with the property that

$$Z(\beta) \equiv \sum_{i\in\mathbb{S}} e^{-\beta H(i)}(\boldsymbol{\nu})_i < \infty \quad \text{for each } \beta \in (0,\infty). \tag{5.4.1}$$

In the physics metaphor, $\beta = \frac{1}{kT}$ is, apart from Boltzmann's constant k, the reciprocal temperature, and physicists would call $\beta \rightsquigarrow Z(\beta)$ the *partition function.* Finally, for each $\beta \in (0,\infty)$, the *Gibbs state* $\boldsymbol{\gamma}(\beta)$ is the probability vector given by

$$\bigl(\boldsymbol{\gamma}(\beta)\bigr)_i = \frac{1}{Z(\beta)} e^{-\beta H(i)}(\boldsymbol{\nu})_i \quad \text{for } i \in \mathbb{S}. \tag{5.4.2}$$

From a physical standpoint, everything of interest is encoded in the partition function. For example, it is elementary to compute both the average and variance of the energy by taking logarithmic derivatives:

$$\langle H\rangle_{\boldsymbol{\gamma}(\beta)} = -\frac{d}{d\beta}\log Z(\beta) \quad \text{and} \quad \mathrm{Var}_{\boldsymbol{\gamma}(\beta)}(H) = \frac{d^2}{d\beta^2}\log Z(\beta). \tag{5.4.3}$$

The final ingredient is the description of the Glauber dynamics. For this purpose, we start with a matrix $\mathbf{A}$ all of whose entries are non-negative and whose diagonal entries are 0. Further, we assume that $\mathbf{A}$ is irreducible in the sense that

$$\sup_{n\geq 0}(\mathbf{A}^n)_{ij} > 0 \quad \text{for all } (i,j) \in \mathbb{S}^2 \tag{5.4.4}$$

and that it is reversible in the sense that

$$(\boldsymbol{\nu})_i(\mathbf{A})_{ij} = (\boldsymbol{\nu})_j(\mathbf{A})_{ji} \quad \text{for all } (i,j) \in \mathbb{S}^2. \tag{5.4.5}$$

Finally, we insist that

$$\sum_{j\in\mathbb{S}} e^{-\beta H(j)}(\mathbf{A})_{ij} < \infty \quad \text{for each } i \in \mathbb{S} \text{ and } \beta > 0. \tag{5.4.6}$$

At this point there are many ways in which to construct a Glauber dynamics. However, for our purposes, the one which will serve us best is the one whose Q-matrix is given by

$$\begin{aligned} \big(\mathbf{Q}(\beta)\big)_{ij} &= e^{-\beta(H(j)-H(i))^+}(\mathbf{A})_{ij} \quad \text{when } j \neq i \\ \big(\mathbf{Q}(\beta)\big)_{ii} &= -\sum_{j\neq i}\big(\mathbf{Q}(\beta)\big)_{ij}, \end{aligned} \tag{5.4.7}$$

where $a^+ \equiv a \vee 0$ is the non-negative part of the number $a \in \mathbb{R}$. Because,

$$\big(\boldsymbol{\gamma}(\beta)\big)_i\big(\mathbf{Q}(\beta)\big)_{ij} = Z(\beta)^{-1}e^{-\beta H(i)\vee H(j)}(\boldsymbol{\nu})_i(\mathbf{A})_{ij} \quad \text{for } i \neq j, \tag{5.4.8}$$

$\boldsymbol{\gamma}(\beta)$ clearly is reversible for $\mathbf{Q}(\beta)$. There are many other possibilities, and the optimal choice is often dictated by special features of the situation under consideration. However, whatever choice is made, it should be made in such a way that, for each $\beta > 0$, $\mathbf{Q}(\beta)$ *determines a Markov process which never explodes.*

5.4.2. The Dirichlet Form: In this subsection we will modify the ideas developed in §5.2.3 to get a lower bound on

$$\begin{aligned} &\lambda_\beta \equiv \inf\big\{\mathcal{E}_\beta(f,f) : \operatorname{Var}_\beta(f) = 1\big\} \\ &\text{where } \mathcal{E}_\beta(f,f) \equiv \frac{1}{2}\sum_{j\neq i}\big(\boldsymbol{\gamma}(\beta)\big)_i\big(\mathbf{Q}(\beta)\big)_{ij}\big(f(j)-f(i)\big)^2 \end{aligned} \tag{5.4.9}$$

and $\operatorname{Var}_\beta(f)$ is shorthand for $\operatorname{Var}_{\boldsymbol{\gamma}(\beta)}(f)$, the variance of f with respect to $\boldsymbol{\gamma}(\beta)$. For this purpose, we introduce the notation

$$\operatorname{Elev}(p) = \max_{0\leq m\leq n} H(i_m) \quad \text{and} \quad e(p) = \operatorname{Elev}(p) - H(i_0) - H(i_n)$$

for a path $p = (i_0, \ldots, i_n)$. Then, when $\mathbf{Q}(\beta)$ is given by (5.4.7), one sees that, when $p = (i_0, \ldots, i_n)$

$$w_\beta(p) \equiv \sum_{m=1}^{n} \frac{(\boldsymbol{\gamma}(\beta))_{i_0}(\boldsymbol{\gamma}(\beta))_{i_n}}{(\boldsymbol{\gamma}(\beta))_{i_{m-1}}(\mathbf{Q}(\beta))_{i_{m-1}i_m}} \le Z(\beta)^{-1} e^{\beta e(p)} w(p),$$

$$\text{where } w(p) \equiv \sum_{m=1}^{n} \frac{(\boldsymbol{\nu})_{i_0}(\boldsymbol{\nu})_{i_n}}{(\boldsymbol{\nu})_{i_{m-1}}\mathbf{A}_{i_{m-1}i_m}}.$$

Hence, for any choice of paths $\mathcal{P}$, we know that (cf. (5.3.10))

$$W_\beta(\mathcal{P}) \equiv \sup_e \sum_{p \ni e} w_\beta(p) \le Z(\beta)^{-1} e^{\beta E(\mathcal{P})} W(\mathcal{P}),$$

$$\text{where } W(\mathcal{P}) \equiv \sup_e \sum_{p \ni e} w(p) \text{ and } E(\mathcal{P}) \equiv \sup_{p \in \mathcal{P}} e(p),$$

and therefore that

$$\lambda_\beta \ge \frac{Z(\beta) e^{-\beta E(\mathcal{P})}}{W(\mathcal{P})}. \tag{5.4.10}$$

On the one hand, it is clear that (5.4.10) gives information only when $W(\mathcal{P}) < \infty$. At the same time, it shows that, at least if ones interest is in large β's, then it is important to choose $\mathcal{P}$ so that $E(\mathcal{P})$ is as small as possible. When $\mathbb{S}$ is finite, reconciling these two creates no problem. Indeed, the finiteness of $\mathbb{S}$ guarantees that $W(\mathcal{P})$ will be finite for every choice of allowable paths. In addition, finiteness allows one to find for each (i, j) a path $p(i, j)$ which minimizes $\text{Elev}(p)$ among allowable paths from i to j, and clearly any $\mathcal{P}$ consisting of such paths will minimize $E(\mathcal{P})$. Of course, it is sensible to choose such a $\mathcal{P}$ so as to minimize $W(\mathcal{P})$ as well. In any case, whenever $\mathbb{S}$ is finite and $\mathcal{P}$ consists of paths $p(i, j)$ which minimize $\text{Elev}(p)$ among paths p between i and j, $E(\mathcal{P})$ has a nice interpretation. Namely, think of $\mathbb{S}$ as being sites on a map and of H as giving the altitude of the sites. That is, in this metaphor, $H(i)$ is the distance of i "above sea level." Without loss in generality, we will assume that at least one site k_0 is at sea level: $H(k_0) = 0$.[4] When such an k_0 exists, the metaphorical interpretation of $E(\mathcal{P})$ is as the least upper bound on the altitude a hiker must gain, no matter where he starts or what allowable path he chooses to follow, in order to reach the sea. To see this, first observe that if p and p' are a pair of allowable paths and if the end point of p is the initial point of p', then the path q is allowable and $\text{Elev}(q) \le \text{Elev}(p) \vee \text{Elev}(p')$ when q is obtained by concatenating p and

[4] If that is not already so, we can make it so by choosing k_0 to be a point at which H takes its minimum value and replacing H by $H - H(k_0)$. Such a replacement leaves both $\boldsymbol{\gamma}(\beta)$ and $\mathbf{Q}(\beta)$ as will as the quantity on the right hand side of (5.4.10) unchanged.

p': if $p = (i_0, \dots, i_n)$ and $p' = (i'_0, \dots, i'_{n'})$, then $q = (i_0, \dots, i_n, i'_1, \dots, i'_{n'})$. Hence, for any (i,j), $e\big(p(i,j)\big) = e\big(p(i,k_0)\big) \vee e\big(p(j,k_0)\big)$, from which it should be clear that $E(\mathcal{P}) = \max_i e\big(p(i,k_0)\big)$. Finally, since, for each i, $e\big(p(i,k_0)\big) = H(\ell) - H(i)$, where ℓ is a highest point along the path $p(i,k_0)$, the explanation is complete. When $\mathbb{S}$ is infinite, the same interpretation is valid in various circumstances. For example, it applies when H "tends to infinity at infinity" in the sense that $\{i : H(i) \le M\}$ is finite for each $M < \infty$.

When $\mathbb{S}$ is finite, we can show that, at least for large β, (5.4.10) is quite good. To be precise, we have the following result:

5.4.11 THEOREM. *Assume that $\mathbb{S}$ is finite and that $\mathbf{Q}(\beta)$ is given by (5.4.7). Set $\mathfrak{m} = \min_{i\in\mathbb{S}} H(i)$ and $\mathbb{S}_0 = \{i : H(i) = \mathfrak{m}\}$, and let $\mathfrak{e}$ be the minimum value $E(\mathcal{P})$ takes as $\mathcal{P}$ runs over all selections of allowable paths. Then $\mathfrak{e} \ge -\mathfrak{m}$, and $\mathfrak{e} = -\mathfrak{m}$ if and only if for each $(i,j) \in \mathbb{S} \times \mathbb{S}_0$ there is an allowable path p from i to j with $\mathrm{Elev}(p) = H(i)$. (See also Exercise 5.6.5 below.) More generally, whatever the value of $\mathfrak{e}$, there exist constants $0 < c_- \le c_+ < \infty$, which are independent of H, such that*

$$c_- e^{-\beta(\mathfrak{e}+\mathfrak{m})} \le \lambda_\beta \le c_+ e^{-\beta(\mathfrak{e}+\mathfrak{m})} \quad \text{for all } \beta \ge 0.$$

PROOF: Because neither $\gamma(\beta)$ nor $\mathbf{Q}(\beta)$ is changed if H is replaced by $H - \mathfrak{m}$ whereas $\mathfrak{e}$ changes to $\mathfrak{e} + \mathfrak{m}$, we may and will assume that $\mathfrak{m} = 0$.

Choose a collection $\mathcal{P} = \{p(i,j) : (i,j) \in \mathbb{S}^2\}$ of allowable paths so that, for each (i,j), $e\big(p(i,j)\big)$ minimizes $e(p)$ over allowable paths from i to j. Next choose and fix a $k_0 \in \mathbb{S}_0$. By the reasoning given above,

$$(*) \qquad \mathfrak{e} = \max_{i\in\mathbb{S}} e\big(p(i,k_0)\big).$$

In particular, since $e\big(p(i,k_0)\big) = \mathrm{Elev}\big(p(i,k_0)\big) - H(i) \ge 0$ for all $i \in \mathbb{S}$, this proves that $\mathfrak{e} \ge 0$ and makes it clear that $\mathfrak{e} = 0$ if and only if $H(i) = \mathrm{Elev}\big(p(i,k_0)\big)$ for all $i \in \mathbb{S}$.

Turning to the lower bound for λ_β, observe that, because $\mathfrak{m} = 0$, $Z(\beta) \ge (\boldsymbol{\nu})_{k_0} > 0$ and therefore, by (5.4.10), that we can take $c_- = \frac{(\boldsymbol{\nu})_{k_0}}{W(\mathcal{P})}$.

Finally, to prove the upper bound, choose $\ell_0 \in \mathbb{S} \setminus \{k_0\}$ so that $e(p_0) = \mathfrak{e}$ when $p_0 \equiv p(\ell_0, k_0)$, let Γ be the set of $i \in \mathbb{S}$ with the property that either $i = k_0$ or $\mathrm{Elev}\big(p(i)\big) < \mathrm{Elev}(p_0)$ for the path $p(i) \equiv p(i,k_0) \in \mathcal{P}$ from i to k_0, and set $f = \mathbf{1}_\Gamma$. Then, because $k_0 \in \Gamma$ and $\ell_0 \notin \Gamma$,

$$\begin{aligned}\mathrm{Var}_\beta(f) &= \left(\sum_{i\in\Gamma} \big(\gamma(\beta)\big)_i\right)\left(\sum_{j\notin\Gamma} \big(\gamma(\beta)\big)_j\right) \\ &\ge \big(\gamma(\beta)\big)_{k_0}\big(\gamma(\beta)\big)_{\ell_0} = \frac{(\boldsymbol{\nu})_{k_0}(\boldsymbol{\nu})_{\ell_0}}{Z(\beta)^2} e^{-\beta(H(k_0)+H(\ell_0))}.\end{aligned}$$

At the same time

$$\mathcal{E}_\beta(f,f) = \sum_{(i,j)\in\Gamma\times\Gamma\mathbb{C}} \big(\gamma(\beta)\big)_i \big(\mathbf{Q}(\beta)\big)_{ij} = \frac{1}{Z(\beta)} \sum_{(i,j)\in\Gamma\times\Gamma\mathbb{C}} (\boldsymbol{\nu})_i (\mathbf{A})_{ij} e^{-\beta H(i)\vee H(j)}.$$

But if $i \in \Gamma \setminus \{k_0\}$, $j \notin \Gamma$, and $(\mathbf{A})_{ij} > 0$, then $H(j) \geq \text{Elev}(p_0)$. To see this, consider the path q obtained by going in one step from j to i and then following $p(i)$ from i to k_0. Clearly, q is an allowable path from j to k_0, and therefore $\text{Elev}(q) \geq \text{Elev}\big(p(j)\big) \geq \text{Elev}(p_0)$. But this means that

$$\text{Elev}(p_0) \leq \text{Elev}\big(p(j)\big) \leq \text{Elev}(q) = \text{Elev}\big(p(i)\big) \vee H(j),$$

which, together with $\text{Elev}\big(p(i)\big) < \text{Elev}(p_0)$, forces the conclusion that $H(j) \geq \text{Elev}(p_0)$. Even easier is the observation that $H(j) \geq \text{Elev}(p_0)$ if $j \notin \Gamma$ and $(\mathbf{A})_{k_0 j} > 0$, since in that case the path (j, k_0) is allowable and

$$H(j) = \text{Elev}\big((j, k_0)\big) \geq \text{Elev}\big(p(j)\big) \geq \text{Elev}(p_0).$$

Hence, after plugging this into the preceding expression for $\mathcal{E}_\beta(f, f)$, we get

$$\mathcal{E}_\beta(f, f) \leq \frac{e^{-\beta \text{Elev}(p_0)}}{Z(\beta)} \sum_{(i,j) \in \Gamma \times \Gamma \complement} (\boldsymbol{\nu})_i (\mathbf{A})_{ij},$$

which, because $\mathfrak{e} = \text{Elev}(p_0) - H(k_0) - H(\ell_0)$, means that

$$\lambda_\beta \leq \frac{\mathcal{E}_\beta(f, f)}{\text{Var}_\beta(f)} \leq \frac{Z(\beta)}{(\boldsymbol{\nu})_{k_0} (\boldsymbol{\nu})_{\ell_0}} \left(\sum_{(i,j) \in \Gamma \times \Gamma \complement} (\boldsymbol{\nu})_i (\mathbf{A})_{ij} \right) e^{-\beta \mathfrak{e}}.$$

Finally, because $Z(\beta) \leq \|\boldsymbol{\nu}\|_{\text{v}}$, the upper bound follows. □

5.5 Simulated Annealing

This concluding section deals with an application of the ideas in the preceding section. Namely, given a function $H : \mathbb{S} \longrightarrow [0, \infty)$, we want to describe a procedure, variously known as the *simulated annealing* or the *Metropolis algorithm*, for locating a place where H achieves its minimum value.

In order to understand the intuition which underlies this procedure, let $\mathbf{A}$ be a matrix of the sort discussed in §5.4, assume that 0 is the minimum value of H, set $\mathbb{S}_0 = \{i : H(i) = 0\}$, and think about dynamic procedures which would lead you from any initial point to $\mathbb{S}_0$ via paths which are allowable according to $\mathbf{A}$ (i.e., $\mathbf{A}_{k\ell} > 0$ if k and ℓ are successive points along the path). One procedure is based on the steepest decent strategy. That is, if one is at k, one moves to any one of the points ℓ for which $\mathbf{A}_{k\ell} > 0$ and $H(\ell)$ is minimal if $H(\ell) \leq H(k)$ for at least one such point, and one stays put if $H(\ell) > H(k)$ for every ℓ with $(\mathbf{A})_{k\ell} > 0$. This procedure works beautifully as long as you avoid, in the metaphor suggested at the end of §5.4, getting trapped in some "mountain valley." The point is that the steepest decent procedure is the most efficient strategy for getting to some local minimum of H. However, if that minimum is not global, then, in general, you will get stuck! Thus, if you are going to avoid this fate, occasionally you will have to go "up hill" even when

you have the option to go "down hill." However, unless you have a detailed *a priori* knowledge of the whole terrain, there is no way to know when you should decide to do so. For this reason, it may be best to abandon rationality and let the decision be made randomly. Of course, after a while, you should hope that you will have worked your way out of the mountain valleys and that a steepest decent strategy should become increasingly reasonable.

5.5.1. The Algorithm: In order to eliminate as many technicalities as possible, we will assume throughout that $\mathbb{S}$ is finite and has at least 2 elements. Next, let $H : \mathbb{S} \longrightarrow [0, \infty)$ be the function for which we want to locate a place where it achieves its minimum, and, without loss in generality, we will assume that 0 is its minimum. Now take $\boldsymbol{\nu}$ so that $(\boldsymbol{\nu})_i = 1$ for all $i \in \mathbb{S}$, and choose a matrix $\mathbf{A}$ so that $(\mathbf{A})_{ii} = 0$ for all $i \in \mathbb{S}$, $(\mathbf{A})_{ij} = (\mathbf{A})_{ji} \geq 0$ if $j \neq i$, and $\mathbf{A}$ is irreducible (cf. (5.4.4)) on $\mathbb{S}$. In practice, the selection of $\mathbf{A}$ should be made so that the evaluation of $H(j) - H(i)$ when $(\mathbf{A})_{ij} > 0$ is as "easy as possible." For example, if $\mathbb{S}$ has some sort of natural neighborhood structure with respect to which $\mathbb{S}$ is connected and the computation of $H(j) - H(i)$ when j is a neighbor of i requires very little time, then it is reasonable to take $\mathbf{A}$ so that $(\mathbf{A})_{ij} = 0$ unless $j \neq i$ is a neighbor of i.

Now define $\boldsymbol{\gamma}(\beta)$ as in (5.4.2) and $\mathbf{Q}(\beta)$ as in (5.4.7). Clearly, $\boldsymbol{\gamma}(0)$ is just the normalized, uniform distribution on $\mathbb{S}$: $\big(\boldsymbol{\gamma}(0)\big)_i = L^{-1}$, where $L \equiv \#\mathbb{S} \geq 2$ is the number of elements in $\mathbb{S}$. On the one hand, as β gets larger, $\boldsymbol{\gamma}(\beta)$ becomes more concentrated on $\mathbb{S}_0$. More precisely, since $Z(\beta) \geq \#\mathbb{S}_0 \geq 1$

$$\big\langle 1_{\mathbb{S}_0\complement} \big\rangle_{\boldsymbol{\gamma}(\beta)} \leq Le^{-\beta\delta}, \quad \text{where } \delta \equiv \min\{H(j) : j \notin \mathbb{S}_0\}. \tag{5.5.1}$$

On the other hand, as β gets larger, Theorem 5.4.11 says that, at least when $\mathfrak{e} > 0$, $\lambda(\beta)$ will be getting smaller. Thus, we are confronted by a conflict.

In view of the introductory discussion, this conflict between the virtues of taking β large, which is tantamount to adopting an approximately steepest decent strategy, versus those of taking β small, which is tantamount to keeping things fluid and thereby diminishing the danger of getting trapped, should be expected. Moreover, a resolution is suggested at the end of that discussion. Namely, in order to maximize the advantages of each, one should start with $\beta = 0$ and allow β to increase with time.[5] That is, we will make β an increasing, continuous function $t \rightsquigarrow \beta(t)$ with $\beta(0) = 0$. In the interest of unencumbering our formulae, we will adopt the notation

$$Z(t) = Z\big(\beta(t)\big), \quad \boldsymbol{\gamma}_t = \boldsymbol{\gamma}\big(\beta(t)\big), \quad \langle\cdot\rangle_t = \langle\cdot\rangle_{\boldsymbol{\gamma}_t}, \quad \|\cdot\|_{2,t} = \|\cdot\|_{2,\boldsymbol{\gamma}_t},$$
$$\mathrm{Var}_t = \mathrm{Var}_{\boldsymbol{\gamma}_t}, \quad \mathbf{Q}(t) = \mathbf{Q}\big(\beta(t)\big), \quad \mathcal{E}_t = \mathcal{E}_{\beta(t)}, \quad \text{and } \lambda_t = \lambda_{\beta(t)}.$$

[5] Actually, there is good reason to doubt that monotonically increasing β is the best way to go. Indeed, the name "simulated annealing" derives from the idea that what one wants to do is simulate the annealing process familiar to chemists, material scientists, skilled carpenters, and followers of Metropolis. Namely, what these people do is alternately heat and cool to achieve their goal, and there is reason to believe we should be following their example. However, I have chosen not to follow them on the unforgivable, but understandable, grounds that my analysis is capable of handling only the monotone case.

Because, in the physical model, β is proportional to the reciprocal of temperature and β increases with time, $t \rightsquigarrow \beta(t)$ is called the *cooling schedule.*

5.5.2. Construction of the Transition Probabilities: Because the Q matrix here is time dependent, the associated transition probabilities will be *time-inhomogeneous.* Thus, instead of $t \rightsquigarrow \mathbf{Q}(t)$ determining a one parameter family of transition probability matrices, for each $s \in [0, \infty)$ it will determine a map $t \rightsquigarrow \mathbf{P}(s,t)$ from $[s, \infty)$ into transition probability matrices by the *time-inhomogeneous Kolmogorov forward equation*

$$\frac{d}{dt}\mathbf{P}(s,t) = \mathbf{P}(s,t)\mathbf{Q}(t) \quad \text{on } (s,\infty) \text{ with } \mathbf{P}(s,s) = \mathbf{I}. \tag{5.5.2}$$

Although (5.5.2) is not exactly covered by our earlier analysis of Kolmogorov equations, it nearly is. To see this, we solve (5.5.2) via an approximation procedure in which $\mathbf{Q}(t)$ is replaced on the right hand side by

$$\mathbf{Q}^{(N)}(t) \equiv \mathbf{Q}\big([t]_N\big) \quad \text{where } [t]_N = \frac{m}{N} \text{ for } t \in \big[\tfrac{m}{N}, \tfrac{(m+1)}{N}\big).$$

The solution $t \rightsquigarrow \mathbf{P}^{(N)}(s,t)$ to the resulting equation is then given by the prescription $\mathbf{P}^{(N)}(s,s) = \mathbf{I}$ and

$$\mathbf{P}^{(N)}(s,t) = \mathbf{P}^{(N)}\big(s, s \vee [t]_N\big) e^{(t-s)\mathbf{Q}([t]_N)} \quad \text{for } t > s.$$

As this construction makes obvious, $\mathbf{P}^{(N)}(s,t)$ is a transition probability matrix for each $N \geq 1$ and $t \geq s$, and $(s,t) \rightsquigarrow \mathbf{P}^{(N)}(s,t)$ is continuous. Moreover,

$$\begin{aligned}
\big\|\mathbf{P}^{(N)}(s,t) - \mathbf{P}^{(M)}(s,t)\big\|_{\mathrm{u,v}} \\
&\leq \int_s^t \big\|\mathbf{Q}\big([\tau]_N\big) - \mathbf{Q}\big([\tau]_M\big)\big\|_{\mathrm{u,v}} \big\|\mathbf{P}^{(N)}(s,\tau)\big\|_{\mathrm{u,v}}\, d\tau \\
&\quad + \int_s^t \big\|\mathbf{Q}([\tau]_M)\big\|_{\mathrm{u,v}} \big\|\mathbf{P}^{(N)}(s,\tau) - \mathbf{P}^{(M)}(s,\tau)\big\|_{\mathrm{u,v}}\, d\tau.
\end{aligned}$$

But

$$\big\|\mathbf{Q}(\tau)\big\|_{\mathrm{u,v}} \leq \|\mathbf{A}\|_{\mathrm{u,v}} \text{ and } \big\|\mathbf{Q}(\tau') - \mathbf{Q}(\tau)\big\|_{\mathrm{u,v}} \leq \|\mathbf{A}\|_{\mathrm{u,v}} \|H\|_{\mathrm{u}} |\beta(\tau') - \beta(\tau)|,$$

and so

$$\begin{aligned}
\big\|\mathbf{P}^{(N)}(s,t) - \mathbf{P}^{(M)}(s,t)\big\|_{\mathrm{u,v}} \leq & \|\mathbf{A}\|_{\mathrm{u,v}} \|H\|_{\mathrm{u}} \int_s^t \big|\beta\big([\tau]_N\big) - \beta\big([\tau]_M\big)\big|\, d\tau \\
& + \|\mathbf{A}\|_{\mathrm{u.v}} \int_s^t \big\|\mathbf{P}^{(N)}(\tau) - \mathbf{P}^{(M)}(\tau)\big\|_{\mathrm{u,v}}\, d\tau.
\end{aligned}$$

Hence, after an application of Gronwall's inequality, we find that

$$\sup_{0\le s\le t\le T}\left\|\mathbf{P}^{(N)}(s,t)-\mathbf{P}^{(M)}(s,t)\right\|_{\mathrm{u,v}}$$
$$\le \|\mathbf{A}\|_{\mathrm{u,v}}\|H\|_{\mathrm{u}}e^{\|\mathbf{A}\|_{\mathrm{u,v}}T}\int_0^T\left|\beta\big([\tau]_N\big)-\beta\big([\tau]_M\big)\right|d\tau.$$

Because $\tau\rightsquigarrow\beta(\tau)$ is continuous, this proves that the sequence $\{\mathbf{P}^{(N)}(s,t) : N\ge 1\}$ is Cauchy convergent in the sense that, for each $T>0$,

$$\lim_{M\to\infty}\sup_{N\ge M}\sup_{0\le s\le t\le T}\left\|\mathbf{P}^{(N)}(s,t)-\mathbf{P}^{(M)}(s,t)\right\|_{\mathrm{u,v}}=0.$$

As a consequence, we know that there exists a continuous $(s,t)\rightsquigarrow\mathbf{P}(s,t)$ to which the $\mathbf{P}^{(N)}(s,t)$'s converge with respect to $\|\cdot\|_{\mathrm{u,v}}$ uniformly on finite intervals. In particular, for each $t\ge s$, $\mathbf{P}(s,t)$ is a transition probability matrix and $t\in[s,\infty)\longmapsto\mathbf{P}(s,t)$ is a continuous solution to

$$\mathbf{P}(s,t)=\mathbf{I}+\int_s^t\mathbf{P}(s,\tau)\mathbf{Q}(\tau)\,d\tau,\quad t\in[s,\infty),$$

which is the equivalent, integrated form (5.5.2). Furthermore, if $t\in[s,\infty)\longmapsto\boldsymbol{\mu}_t\in M_1(\mathbb{S})$ is continuously differentiable, then

$$\text{(5.5.3)}\quad \dot{\boldsymbol{\mu}}_t\equiv\frac{d}{dt}\boldsymbol{\mu}_t=\boldsymbol{\mu}_t\mathbf{Q}(t)\text{ for }t\in[s,\infty)\iff\boldsymbol{\mu}_t=\boldsymbol{\mu}_s\mathbf{P}(s,t)\text{ for }t\in[s,\infty).$$

Since the "if" assertion is trivial, we turn to the "only if" statement. Thus, suppose that $t\in[s,\infty)\longmapsto\boldsymbol{\mu}_t\in M_1(\mathbb{S})$ satisfying $\dot{\boldsymbol{\mu}}_t=\boldsymbol{\mu}_t\mathbf{Q}(t)$ is given, and set $\boldsymbol{\omega}_t=\boldsymbol{\mu}_t-\boldsymbol{\mu}_s\mathbf{P}(s,t)$. Then

$$\boldsymbol{\omega}_t=\int_s^t\boldsymbol{\omega}_\tau\mathbf{Q}(\tau)\,d\tau,$$

and so,

$$\|\boldsymbol{\omega}_t\|_{\mathrm{v}}\le\|\mathbf{A}\|_{\mathrm{u,v}}\int_s^t\|\boldsymbol{\omega}_\tau\|_{\mathrm{v}}\,d\tau.$$

Hence, after another application of Gronwall's inequality, we see that $\boldsymbol{\omega}_t=\mathbf{0}$ for all $t\ge s$.

Of course, by applying this uniqueness result when $\boldsymbol{\mu}_t=\boldsymbol{\delta}_i\mathbf{P}(s,t)$ for each $i\in\mathbb{S}$, we learn that $(s,t)\rightsquigarrow\mathbf{P}(s,t)$ is the one and only solution to (5.5.2). In addition, it leads to the following time-inhomogeneous version of the Chapman–Kolmogorov equation:

$$\text{(5.5.4)}\qquad \mathbf{P}(r,t)=\mathbf{P}(r,s)\mathbf{P}(s,t)\quad\text{for }0\le r\le s\le t.$$

Indeed, set $\boldsymbol{\mu}_t=\boldsymbol{\delta}_i\mathbf{P}(r,t)$ for $t\ge s$, note that $t\rightsquigarrow\boldsymbol{\mu}_t$ satisfies (5.5.3) with $\boldsymbol{\mu}_s=\boldsymbol{\delta}_i\mathbf{P}(r,s)$, and conclude that $\boldsymbol{\mu}_t=\boldsymbol{\delta}_i\mathbf{P}(r,s)\mathbf{P}(s,t)$.

5.5.3. Description of the Markov Process: Given a probability vector $\boldsymbol{\mu}$, we now want to construct a Markov process $\{X(t) : t \geq 0\}$ which has $\boldsymbol{\mu}$ as its initial distribution and $(s,t) \rightsquigarrow \mathbf{P}(s,t)$ as its transition mechanism, in the sense that

$$\mathbb{P}\big(X(0) = i\big) = (\boldsymbol{\mu})_i \;\&\; \mathbb{P}\big(X(t) = j \,\big|\, X(\sigma),\ \sigma \in [0,s]\big) = \mathbf{P}(s,t)_{X(s)j}. \tag{5.5.5}$$

The idea which we will use is basically the same as the one which we used in §4.2.1 & 2.1.1. However, life here is made more complicated by the fact that the time inhomogeneity forces us to have an uncountable number of random variables at hand: a pair for each $(t,i) \in [0,\infty) \times \mathbb{S}$. To handle this situation, take, without loss in generality, $\mathbb{S} = \{1,\dots,L\}$ and, for $(t,i,j) \in [0,\infty) \times \mathbb{S}^2$, set $S(t,i,j) = \sum_{\ell=1}^{j} e^{-\beta(t)(H(\ell)-H(i))^+}(\mathbf{A})_{i\ell}$, take $S(t,i,0) = 0$, and define

$$\Psi(t,i,u) = \begin{cases} j & \text{if } \frac{S(t,i,j-1)}{S(t,i,L)} \leq u < \frac{S(t,i,j)}{S(t,i,L)} \\ i & \text{if } u > 1. \end{cases}$$

Also, determine $\mathcal{T} : [0,\infty) \times \mathbb{S} \times [0,\infty) \longrightarrow [0,\infty)$ by

$$\int_s^{s+\mathcal{T}(s,i,\xi)} S(\tau,i,L)^{-1}\, d\tau = \xi.$$

Next, let X_0 be an $\mathbb{S}$-valued random variable with distribution $\boldsymbol{\mu}$, let $\{E_n : n \geq 1\}$ be a sequence of unit exponential random variables which are independent of each other and of X_0, and let $\{U_n : n \geq 1\}$ be a sequence of random variables which are uniformly distributed on $[0,1)$ and independent of each other and of $\sigma\big(\{X_0\} \cup \{E_n : n \geq 1\}\big)$. Finally, set $J_0 = 0$ and $X(0) = X_0$, and, when $n \geq 1$, use induction to define

$$J_n - J_{n-1} = \mathcal{T}\big(J_{n-1}, X(J_{n-1}), E_n\big), \quad X(J_n) = \Psi\big(J_n, X(J_{n-1}), U_n\big),$$
$$\text{and } X(t) = X(J_{n-1}) \quad \text{for } J_{n-1} \leq t < J_n.$$

Without substantial change, the reasoning given in §2.1.1 combined with that in §4.2.2 allows one to show that (5.5.5) holds.

5.5.4. Choosing a Cooling Schedule: In this section, we will give a rational basis on which to choose the cooling schedule $t \rightsquigarrow \beta(t)$. For this purpose, it is essential to keep in mind what it is that we are attempting to do. Namely, we are trying to have the Markov process $\{X(t) : t \geq 0\}$ seek out the set $\mathbb{S}_0 = \{j : H(j) = 0\}$ in the sense that, as $t \to \infty$, $\mathbb{P}(X(t) \notin \mathbb{S}_0)$ should tend to 0 as fast as possible, and the way we hope to accomplish this is by making the distribution of $X(t)$ look as much like $\boldsymbol{\gamma}_t$ as possible. Thus, on the one hand, we need to give $\{X(t) : t \geq 0\}$ enough time to equilibrate, so that the distribution of $X(t)$ will look a lot like $\boldsymbol{\gamma}_t$. On the other hand, in spite of the fact that it may inhibit equilibration, unless we make $\beta(t)$ increase to infinity,

there is no reason for our wanting to make the distribution of $X(t)$ look like γ_t.

In order to understand how to deal with the concerns raised above, let $\boldsymbol{\mu}$ be a fixed initial distribution, and let $\boldsymbol{\mu}_t$ be the distribution at time $t \geq 0$ of the Markov process $\{X(t) : t \geq 0\}$ described in the §5.5.3 with initial distribution $\boldsymbol{\mu}$. Equivalently, $\boldsymbol{\mu}_t = \boldsymbol{\mu}\mathbf{P}(0,t)$, where $\{\mathbf{P}(s,t) : 0 \leq s \leq t < \infty\}$ is the family of transition probability matrices constructed in §5.5.2. Next, define $f_t : \mathbb{S} \longrightarrow [0,\infty)$ so that

$$f_t(i) = \frac{(\boldsymbol{\mu}_t)_i}{(\gamma_t)_i} \quad \text{for } t \geq 0 \text{ and } i \in \mathbb{S}.$$

It should be obvious that the size of f_t provides a good measure of the extent to which $\boldsymbol{\mu}_t$ resembles γ_t. For example, by Schwarz's inequality and (5.5.1),

$$\begin{aligned} \mathbb{P}\big(X(t) \notin \mathbb{S}_0\big) &= \langle \mathbf{1}_{\mathbb{S}_0\mathtt{C}} \rangle_{\boldsymbol{\mu}_t} = \langle f_t \mathbf{1}_{\mathbb{S}_0\mathtt{C}} \rangle_t \\ &\leq \|f_t\|_{2,t} \sqrt{\langle \mathbf{1}_{\mathbb{S}_0\mathtt{C}} \rangle_t} \leq L^{\frac{1}{2}} \|f_t\|_{2,t} e^{-\frac{\beta(t)\delta}{2}}, \end{aligned} \tag{5.5.6}$$

and so we will have made progress if we can keep $\|f_t\|_{2,t}$ under control.

With the preceding in mind, assume that $t \rightsquigarrow \beta(t)$ is continuously differentiable, note that this assumption makes $t \rightsquigarrow \|f_t\|_{2,t}^2$ also continuously differentiable, and, in fact, that

$$\begin{aligned} \frac{d}{dt}\|f_t\|_{2,t}^2 &= \frac{d}{dt}\left(\frac{1}{Z(t)} \sum_{i \in \mathbb{S}} \big(f_t(i)\big)^2 e^{-\beta(t)H(i)}\right) \\ &= 2\langle f_t, \dot{f}_t \rangle_t - \dot{\beta}(t)\big\langle H - \langle H \rangle_t, f_t^2 \big\rangle_t, \end{aligned}$$

since, by (5.4.3), $\dot{Z}(t) = -\dot{\beta}(t) Z(t) \langle H \rangle_t$. On the other hand, we can compute this same derivative another way. Namely, because $\langle \mathbf{P}(0,t) g \rangle_{\boldsymbol{\mu}} = \langle g \rangle_{\boldsymbol{\mu}_t} = \langle f_t, g \rangle_t$ for any function g,

$$\|f_t\|_{2,t}^2 = \langle f_t \rangle_{\boldsymbol{\mu}_t} = \langle \mathbf{P}(0,t) f_t \rangle_{\boldsymbol{\mu}},$$

and so we can use (5.5.2) to see that

$$\frac{d}{dt}\|f_t\|_{2,t}^2 = \big\langle \mathbf{P}(0,t)\mathbf{Q}(t) f_t \big\rangle_{\boldsymbol{\mu}} + \langle \mathbf{P}(0,t) \dot{f}_t \rangle_{\boldsymbol{\mu}} = -\mathcal{E}_t(f_t, f_t) + \langle f_t, \dot{f}_t \rangle_t.$$

Thus, after combining these to eliminate the term containing $\dot{f}_t$, we arrive at

$$\begin{aligned} \frac{d}{dt}\|f_t\|_{2,t}^2 &= -2\mathcal{E}_t(f_t, f_t) + \dot{\beta}(t)\big\langle H - \langle H \rangle_t, f_t^2 \big\rangle_t \\ &\leq -\big(2\lambda_t - \|H\|_{\mathrm{u}} \dot{\beta}(t)\big)\|f_t\|_{2,t}^2 + 2\lambda_t, \end{aligned}$$

where, in passing to the second line, we have used the fact that $\langle f_t \rangle_t = 1$ and therefore that $\mathrm{Var}_t(f) = \|f_t\|_{2,t}^2 - 1$. Putting all this together, we now know that

$$\|H\|_{\mathrm{u}}\dot{\beta}(t) \le \lambda_t \implies \frac{d}{dt}\|f_t\|_{2,t}^2 \le -\lambda_t \|f_t\|_{2,t}^2 + 2\lambda_t.$$

The preceding differential inequality for $\|f\|_{2,t}^2$ is easy to integrate. Namely, it says that

$$\frac{d}{dt}\left(e^{\Lambda(t)}\|f_t\|_{2,t}^2\right) \le 2\lambda_t e^{\Lambda(t)} \quad \text{where } \Lambda(t) = \int_0^t \lambda_\tau \, d\tau.$$

Hence,

$$\|f_t\|_{2,t}^2 \le e^{-\Lambda(t)}\|f_0\|_{2,0}^2 + 2\left(1 - e^{-\Lambda(t)}\right) \le \|f_0\|_{2,0}^2 \vee 2.$$

Moreover, since $(\gamma_0)_i = L^{-1}$, where $L = \#\mathbb{S} \ge 2$, $\|f_0\|_{2,0}^2 \le L$, and so

$$\|H\|_{\mathrm{u}}\dot{\beta}(t) \le \lambda_t \implies \|f_t\|_{2,t} \le L^{\frac{1}{2}}. \tag{5.5.7}$$

The final step is to find out how to choose $t \rightsquigarrow \beta(t)$ so that it satisfies the condition in (5.5.7); and, of course, we are only interested in the case when $\mathbb{S}_0 \neq \mathbb{S}$ or, equivalently, $\|H\|_{\mathrm{u}} > 0$. Next, by Theorem 5.4.11, we know that $\lambda_t \ge c_- e^{-\beta(t)\mathfrak{e}}$. Hence, we can take

$$\beta(t) = \begin{cases} \dfrac{1}{\mathfrak{e}} \log\left(1 + \dfrac{c_- \mathfrak{e} t}{\|H\|_{\mathrm{u}}}\right) & \text{when } \mathfrak{e} > 0 \\ \dfrac{c_- t}{\|H\|_{\mathrm{u}}} & \text{when } \mathfrak{e} = 0, \end{cases} \tag{5.5.8}$$

the case when $\mathfrak{e} = 0$ being obtained by an obvious limit procedure. After putting this together with (5.5.7) and (5.5.6), we have now proved that *when $\beta(t)$ is given by (5.5.8), then*

$$\mathbb{P}\big(X(t) \notin \mathbb{S}_0\big) \le L \begin{cases} \left(1 + \dfrac{c_- \mathfrak{e} t}{\|H\|_{\mathrm{u}}}\right)^{-\frac{\delta}{2\mathfrak{e}}} & \text{when } \mathfrak{e} > 0 \\ e^{-\frac{\delta c_- t}{2\|H\|_{\mathrm{u}}}} & \text{when } \mathfrak{e} = 0. \end{cases} \tag{5.5.9}$$

Remark: The result in (5.5.9) when $\mathfrak{e} = 0$ deserves some further comment. In particular, it should be observed that $\mathfrak{e} = 0$ does not guarantee success for a steepest decent strategy. Indeed, $\mathfrak{e} = 0$ only means that each i can be connected to $\mathbb{S}_0$ by an allowable path along which H is non-increasing (cf. Exercise 5.6.5), it does not rule out the possibility that, when using steepest decent, one will choose bad path and get stuck. Thus, even in this situation, one needs enough randomness to hunt around until one finds a good path.

5.5.5. Small Improvements: An observation, which is really only of interest when $\mathfrak{e} > 0$, is that one carry out the same sort of analysis to control the $\|f_t\|_{q,t} \equiv \bigl(\langle |f_t|^q \rangle_t\bigr)^{\frac{1}{q}}$ for each $q \in [2, \infty)$. As a result, one can show that there is, for each $\theta \in (0,1)$, a cooling schedule which makes $\mathbb{P}(X(t) \notin \mathbb{S}_0)$ go to 0 at least as fast as $t^{-\frac{\theta \delta t}{\mathfrak{e}}}$. The relationship between θ and q is given by $\theta = 1 - \frac{1}{q}$.

To carry this out, one begins by computing $\frac{d}{dt}\|f_t\|_{q,t}^q$ twice, once for each each of the following expressions:

$$\|f_t\|_{q,t}^q = \langle f_t^q \rangle_t \quad \text{and} \quad \|f_t\|_{q,t}^q = \langle \mathbf{P}(0,t) f_t^{q-1} \rangle_{\boldsymbol{\mu}}.$$

One then eliminates $\dot{f}_t$ from the resulting equations and thereby arrives at

$$\frac{d}{dt}\|f_t\|_{q,t}^q = -q\mathcal{E}_t\bigl(f_t, f_t^{q-1}\bigr) - \frac{\dot{\beta}(t)}{q'-1}\Bigl\langle f_t^q, H - \langle H \rangle_t \Bigr\rangle_t,$$

where $q' = \frac{q}{q-1}$ and

$$\mathcal{E}_t(\varphi, \psi) \equiv \frac{1}{2}\sum_{i \neq j} (\gamma_t)_i \bigl(\mathbf{Q}(t)\bigr)_{ij} \bigl(\varphi(j) - \varphi(i)\bigr)\bigl(\psi(j) - \psi(i)\bigr) = -\langle \varphi, \mathbf{Q}(t)\psi \rangle_t.$$

At this point one has to show that

$$\mathcal{E}_t(f_t, f_t^{q-1}) \geq \frac{4(q-1)}{q^2}\mathcal{E}_t\bigl(f_t^{\frac{q}{2}}, f_t^{\frac{q}{2}}\bigr),$$

and a little thought makes it clear that this inequality comes down to checking that, for any pair $(a,b) \in [0,\infty)^2$,

$$\bigl(b^{\frac{q}{2}} - a^{\frac{q}{2}}\bigr)^2 \leq \frac{q^2}{4(q-1)}(b-a)\bigl(b^{q-1} - a^{q-1}\bigr),$$

which, when looked at correctly, follows from the Fundamental Theorem of Calculus plus Schwarz's inequality. Hence, in conjunction with the preceding and (5.3.6), we find that

$$\frac{d}{dt}\|f_t\|_{q,t}^q \leq -\frac{1}{q'}\Bigl(4\lambda_t - q\dot{\beta}(t)\|H\|_{\mathrm{u}}\Bigr)\|f_t\|_{q,t}^q + \frac{4\lambda_t}{q'}\langle f_t^{\frac{q}{2}} \rangle_t^2.$$

In order to proceed further, we must learn how to control $\langle f_t^{\frac{q}{2}} \rangle_t^2$ in terms of $\|f_t\|_{q,t}^q$. In the case when $q = 2$, this quantity caused no problem because we knew it was equal to 1. When $q > 2$, we no longer have so much control over it. Nonetheless, by first writing $\langle f_t^{\frac{q}{2}} \rangle_t^2 = \langle f_t^{\frac{q}{2}-1} \rangle_{\boldsymbol{\mu}_t}^2$ and then using part (**c**) of Exercise 5.6.2 to see that

$$\langle f_t^{\frac{q}{2}-1} \rangle_{\boldsymbol{\mu}_t}^2 \leq \langle f_t^{q-1} \rangle_{\boldsymbol{\mu}_t}^{\frac{q-2}{q-1}} = \langle f_t^q \rangle_t^{\frac{q-2}{q-1}},$$

we arrive at $\langle f_t^{\frac{q}{2}}\rangle_t^2 \leq \langle f_t^q\rangle_t^{\frac{q-2}{q-1}}$. Armed with this estimate, we obtain the differential inequality

$$\frac{d}{dt}\|f_t\|_{q,t}^q \leq -\frac{1}{q'}\Big(4\lambda_t - q\|H\|_{\mathrm{u}}\dot{\beta}(t)\Big)\|f_t\|_{q,t}^q + \frac{4\lambda_t}{q'}\big(\|f_t\|_{q,t}^q\big)^{1-\frac{q'}{q}}.$$

Finally, by taking $\beta(t) = \frac{1}{\mathfrak{e}}\log\big(1 + \frac{3c_-\mathfrak{e}t}{q\|H\|_{\mathrm{u}}}\big)$, the preceding inequality can be replaced by

$$\frac{d}{dt}\|f_t\|_{q,t}^q \leq -\frac{\lambda_t}{q'}\|f_t\|_{q,t}^q + \frac{4\lambda_t}{q'}\big(\|f_t\|_{q,t}^q\big)^{1-\frac{q'}{q}},$$

which, after integration, can be made to yield

$$\|f_t\|_{q,t} \leq 2^{\frac{1}{q'}} \vee \|f_0\|_{q,0}.$$

Remark: Actually, it is possible to do even better if one is prepared to combine the preceding line of reasoning, which is basically a consequence of Poincaré's inequality, with analytic ideas which come under the general heading of Sobolev inequalities. The interested reader might want to consult [3], which is the source from which the contents of this whole section are derived.

5.6 Exercises

EXERCISE 5.6.1. A function $\psi : [0,\infty) \longrightarrow \mathbb{R}$ is said to be convex if the graph of ψ lies below the secant connecting any pair of points on its graph. That is, if it satisfies

(*) $\psi\big((1-\theta)s + \theta t\big) \leq (1-\theta)\psi(s) + \theta\psi(t)$ for all $0 \leq s < t$ and $\theta \in [0,1]$.

This exercise deals with various properties of convex functions, all of which turn on the property that the slope of a convex function is non-decreasing.

(**a**) If $\{\psi_n\}_1^\infty \cup \{\psi\}$ are functions on $[0,\infty)$ and $\psi_n(t) \longrightarrow \psi(t)$ for each $t \in [0,\infty)$, show that ψ is non-increasing if each of the ψ_n's is and that ψ is convex if each of the ψ_n's is.

(**b**) If $\psi : [0,\infty) \longrightarrow \mathbb{R}$ is continuous and twice continuously differentiable on $(0,\infty)$, show that ψ is convex on $[0,\infty)$ if and only if $\ddot{\psi} \geq 0$ on $(0,\infty)$.

Hint: The "only if" part is an easy consequence of

$$\ddot{\psi}(t) = \lim_{h \searrow 0} \frac{\psi(t+h) + \psi(t-h) - 2\psi(s)}{h^2} \quad \text{for } t \in (0,\infty).$$

To prove the "if" statement, let $0 < s < t$ be given, and for $\epsilon > 0$ set

$$\varphi_\epsilon(\theta) = \psi\big((1-\theta)s + \theta t\big) - (1-\theta)\psi(s) - \theta\psi(s) - \epsilon\theta(1-\theta), \quad \theta \in [0,1].$$

Note that $\varphi_\epsilon(0) = 0 = \varphi_\epsilon(1)$ and that $\ddot{\varphi}_\epsilon > 0$ on $(0,1)$. Hence, by the second derivative test, φ_ϵ cannot achieve a maximum value in $(0,1)$. Now let $\epsilon \searrow 0$.

(**c**) If ψ is convex on $[0,\infty)$, show that, for each $s \in [0,\infty)$,

$$t \in (s,\infty) \longmapsto \frac{\psi(s)-\psi(t)}{t-s} \quad \text{is non-increasing.}$$

(**d**) If ψ is convex on $[0,\infty)$ and $0 \le s < t \le u < w$, show that

$$\frac{\psi(s)-\psi(t)}{t-s} \ge \frac{\psi(u)-\psi(w)}{w-u}.$$

Hint: Reduce to the case when $u = t$.

EXERCISE 5.6.2. Given a probability vector $\boldsymbol{\mu} \in [0,1]^{\mathbb{S}}$, there are many ways to prove that $\langle f\rangle_{\boldsymbol{\mu}}^2 \le \langle f^2\rangle_{\boldsymbol{\mu}}$ for any $f \in L^2(\boldsymbol{\mu})$. For example, one can get this inequality as an application of Schwarz's inequality $|\langle f,g\rangle_{\boldsymbol{\mu}}| \le \|f\|_{2,\boldsymbol{\mu}}\|g\|_{2,\boldsymbol{\mu}}$ by taking $g = 1$. Alternatively, one can use $0 \le \mathrm{Var}_{\boldsymbol{\mu}}(f) = \langle f^2\rangle_{\boldsymbol{\mu}} - \langle f\rangle_{\boldsymbol{\mu}}^2$. However, neither of these approaches reveals the essential role that convexity plays here. Namely, the purpose of this exercise is to show that for any non-decreasing, continuous, convex function $\psi : [0,\infty) \longrightarrow [0,\infty)$ and any $f : \mathbb{S} \longrightarrow [0,\infty)$,

$$\psi\big(\langle f\rangle_{\boldsymbol{\mu}}\big) \le \langle \psi \circ f\rangle_{\boldsymbol{\mu}}, \tag{5.6.3}$$

where the meaning of the left hand side when $\langle f\rangle_{\boldsymbol{\mu}} = \infty$ is given by taking $\psi(\infty) \equiv \lim_{t\nearrow\infty}\psi(t)$. The inequality (5.6.3) is an example of more general statement known as *Jensen's inequality* (cf. Theorem 6.1.1 in [8]).

(**a**) Use induction on $n \ge 2$ to show that

$$\psi\left(\sum_{m=1}^n \theta_k x_k\right) \le \sum_{m=1}^n \theta_k \psi(x_k) \quad \text{for all}$$
$$(\theta_1,\dots,\theta_n) \in [0,1]^n \text{ with } \sum_{m=1}^n \theta_k = 1 \text{ and } (x_1,\dots,x_n) \in [0,\infty)^n.$$

(**b**) Let $\{F_N\}_1^\infty$ be a non-decreasing exhaustion of $\mathbb{S}$ by finite sets satisfying $\mu(F_N) \equiv \sum_{i\in F_N}(\boldsymbol{\mu})_i > 0$, apply part (**a**) to see that

$$\psi\left(\sum_{i\in F_N} f(i)(\boldsymbol{\mu})_i\right) \le \frac{\sum_{i\in F_N}\psi\big(f(i)\big)(\boldsymbol{\mu})_i}{\mu(F_N)} \le \frac{\langle \psi\circ f\rangle_{\boldsymbol{\mu}}}{\mu(F_N)}$$

for each N, and get the asserted result after letting $N \to \infty$.

(**c**) As an application of (5.6.3), show that, for any $0 < p \le q < \infty$ and $f : \mathbb{S} \longrightarrow [0,\infty)$, $\langle f^p\rangle_{\boldsymbol{\mu}}^{\frac{1}{p}} \le \langle f^q\rangle_{\boldsymbol{\mu}}^{\frac{1}{q}}$.

EXERCISE 5.6.4. Gronwall's is an inequality which has many forms, the most elementary of which states that if $u : [0,T] \longrightarrow [0,\infty)$ is a continuous function which satisfies

$$u(t) \leq A + B\int_0^t u(\tau)\,d\tau \quad \text{for } t \in [0,T],$$

then $u(t) \leq Ae^{Bt}$ for $t \in [0,T]$. Prove this form of *Gronwall's inequality.*

Hint: Set $U(t) = \int_0^t u(\tau)\,d\tau$, show that $\dot{U}(t) \leq A + BU(t)$, and conclude that $U(t) \leq \frac{A}{B}\left(e^{Bt} - 1\right)$.

EXERCISE 5.6.5. Define $\mathfrak{e}$ as in Theorem 5.4.11, and show that $\mathfrak{e} = -\mathfrak{m}$ if and only if for each $(i,j) \in \mathbb{S} \times \mathbb{S}_0$ there is an allowable path $(i_0, \ldots, i_n)$ starting at i and ending at j along which H is non-increasing. That is, $i_0 = i$, $i_n = j$, and, for each $1 \leq m \leq n$, $\mathbf{A}_{i_{m-1}i_m} > 0$ and $H(i_m) \leq H(i_{m-1})$.

EXERCISE 5.6.6. This exercise deals with the material in §5.1.3 and demonstrates that, with the exception of (5.1.10), more or less everything in that section extends to general irreducible, positive recurrent $\mathbf{P}$'s, whether or not they are reversible. Again let $\boldsymbol{\pi} = \boldsymbol{\pi}^{\mathbb{S}}$ be the unique $\mathbf{P}$-stationary probability vector. In addition, for the exercise which follows, it will be important to consider the space $L^2(\boldsymbol{\pi};\mathbb{C})$ consisting of those $f : \mathbb{S} \longrightarrow \mathbb{C}$ for which $|f| \in L^2(\boldsymbol{\pi})$.

(**a**) Define $\mathbf{P}^\top$ as in (5.1.2), show that $1 \geq (\mathbf{P}^\top\mathbf{P})_{ii} = (\boldsymbol{\pi})_i \sum_{j\in\mathbb{S}} \frac{(\mathbf{P})_{ij}^2}{(\boldsymbol{\pi})_j}$, and conclude that the series in the definition $\mathbf{P}f(i) \equiv \sum_{j\in\mathbb{S}} f(j)(\mathbf{P})_{ij}$ is absolutely convergent for each $i \in \mathbb{S}$ and $f \in L^2(\boldsymbol{\pi};\mathbb{C})$.

(**b**) Show that $\|\mathbf{P}f\|_{2,\boldsymbol{\pi}} \leq \|f\|_{2,\boldsymbol{\pi}}$ for all $f \in L^2(\boldsymbol{\pi};\mathbb{C})$, and conclude that (5.1.8) and (5.1.9) extend to the present setting for all $f \in L^2(\boldsymbol{\pi};\mathbb{C})$.

EXERCISE 5.6.7. Continuing with the program initiated in Exercise 5.6.6, we will now see that reversibility plays only a minor role in § 5.1.4. Thus, let $\mathbf{P}$ be any irreducible transition probability on $\mathbb{S}$ which is positive recurrent, let d be its period, and set $\theta_d = e^{\sqrt{-1}\,2\pi d^{-1}}$.

(**a**) Show that for each $0 \leq m < d$ there is a function $f_m : \mathbb{S} \longrightarrow \mathbb{C}$ with the properties that $|f_m| \equiv 1$ and $\mathbf{P}f_m = \theta_d^m f_m$.

Hint: Choose a cyclic decomposition $(\mathbb{S}_0, \ldots, \mathbb{S}_{d-1})$ as in § 3.2.7, and use (3.2.19).

(**b**) Given $\alpha \in \mathbb{R} \setminus \{0\}$, set $\theta_\alpha = e^{\sqrt{-1}\,2\pi\alpha^{-1}}$, and show that there exists an $f \in L^2(\boldsymbol{\pi};\mathbb{C}) \setminus \{\mathbf{0}\}$ satisfying $\mathbf{P}f = \theta_\alpha f$ if and only if $d = m\alpha$ for some $m \in \mathbb{Z} \setminus \{0\}$.

Hint: By part (**a**), it suffices to show that no f exists unless $d = m\alpha$ for some non-zero $m \in \mathbb{Z}$. Thus, suppose that f exists for some α which is not a rational number of the form $\frac{d}{m}$, choose $i \in \mathbb{S}$ so that $f(i) \neq 0$, and get a contradiction with the fact that $\lim_{n\to\infty} \mathbf{P}^{nd} f(i)$ exists.

(**c**) Suppose that $f \in L^2(\mathbb{S};\mathbb{C})$ is a non-trivial, bounded solution to $\mathbf{P}f = \theta_d^m f$ for some $m \in \mathbb{Z}$, and let $(\mathbb{S}_0, \dots, \mathbb{S}_{d-1})$ be a cyclic decomposition of $\mathbb{S}$. Show that, for each $0 \le r < d$, $f \restriction \mathbb{S}_r \equiv \theta_d^{rm} c_0$, where $c_0 \in \mathbb{C} \setminus \{0\}$. In particular, up to a multiplicative constant, for each integer $0 \le m < d$ there is exactly one non-trivial, bounded $f \in L^2(\mathbb{S};\mathbb{C})$ satisfying $\mathbf{P}f = \theta_d^m f$.

(**d**) Let H denote the subspace of $f \in L^2(\boldsymbol{\pi};\mathbb{C})$ satisfying $\mathbf{P}f = \theta f$ for some $\theta \in \mathbb{C}$ with $|\theta| = 1$. By combining parts (**b**) and (**c**), show that d is the dimension of H as a vector space over $\mathbb{C}$.

(**e**) Assume that $(\mathbf{P})_{ij} > 0 \implies (\mathbf{P})_{ji} > 0$ for all $(i,j) \in \mathbb{S}^2$. Show that $d \le 2$ and that $d = 2$ if and only if there is a non-trivial, bounded $f : \mathbb{S} \longrightarrow \mathbb{R}$ satisfying $\mathbf{P}f = -f$. Thus, this is the only property of reversible transition probability matrices of which we made essential use in Theorem 5.1.14.

EXERCISE 5.6.8. This exercise provides another way to think about the relationship between non-negative definiteness and aperiodicity. Namely, let $\mathbf{P}$ be a not necessarily irreducible transition probability matrix on $\mathbb{S}$, and assume that $\boldsymbol{\mu}$ is a probability vector for which the detailed balance condition $(\boldsymbol{\mu})_i(\mathbf{P})_{ij} = (\boldsymbol{\mu})_j(\mathbf{P})_{ji}$, $(i,j) \in \mathbb{S}^2$ holds. Further, assume that $\mathbf{P}$ is non-negative definite in $L^2(\boldsymbol{\mu})$: $\langle f, \mathbf{P}f\rangle_{\boldsymbol{\mu}} \ge 0$ for all bounded $f : \mathbb{S} \longrightarrow \mathbb{R}$. Show that $(\boldsymbol{\mu})_i > 0 \implies (\mathbf{P})_{ii} > 0$ and therefore that i is aperiodic if $(\boldsymbol{\mu})_i > 0$. What follows are steps which lead to this conclusion.

(**a**) Define the matrix $\mathbf{A}$ so that $(\mathbf{A})_{ij} = \langle \mathbf{1}_{\{i\}}, \mathbf{P}\mathbf{1}_{\{j\}}\rangle_{\boldsymbol{\mu}}$, and show that $\mathbf{A}$ is symmetric (i.e., $(\mathbf{A})_{ij} = (\mathbf{A})_{ji}$) and non-negative definite in the sense that $\sum_{ij}(\mathbf{A})_{ij}(\mathbf{x})_i(\mathbf{x})_j \ge 0$ for any $\mathbf{x} \in \mathbb{R}^{\mathbb{S}}$ with only a finite number of non-vanishing entries.

(**b**) Given $i \ne j$, consider the plane $\{\alpha\mathbf{1}_i + \beta\mathbf{1}_j : \alpha, \beta \in \mathbb{R}\}$ in $L^2(\boldsymbol{\pi})$, and, using the argument with which we derived (5.1.5), show that $(\mathbf{A})_{ij}^2 \le (\mathbf{A})_{ii}(\mathbf{A})_{jj}$. In particular, if $(\mathbf{A})_{ii} = 0$, then $(\mathbf{A})_{ij} = 0$ for all $j \in \mathbb{S}$.

(**c**) Complete the proof by noting that $\sum_{j\in\mathbb{S}}(\mathbf{A})_{ij} = (\boldsymbol{\mu})_i$.

(**d**) After examining the argument, show that we did not need $\mathbf{P}$ to be non-negative definite but only that, for a given $i \in \mathbb{S}$, each of the 2×2 submatrices

$$\begin{pmatrix} \langle \mathbf{1}_{\{i\}}, \mathbf{P}\mathbf{1}_{\{i\}}\rangle_{\boldsymbol{\mu}} & \langle \mathbf{1}_{\{i\}}, \mathbf{P}\mathbf{1}_{\{j\}}\rangle_{\boldsymbol{\mu}} \\ \langle \mathbf{1}_{\{j\}}, \mathbf{P}\mathbf{1}_{\{i\}}\rangle_{\boldsymbol{\mu}} & \langle \mathbf{1}_{\{j\}}, \mathbf{P}\mathbf{1}_{\{j\}}\rangle_{\boldsymbol{\mu}} \end{pmatrix}$$

be.

EXERCISE 5.6.9. Let $\mathbf{P}$ be an irreducible, positive recurrent transition probability matrix with stationary distribution $\boldsymbol{\pi}$. Refer to (5.1.2), and show that the first construction in (5.1.3) is again irreducible whereas the second one need not be. Also, show that the period of the first construction is never greater than that of $\mathbf{P}$ and that the second construction is always non-negative definite. Thus, if $\mathbf{P}^\top\mathbf{P}$ is irreducible, it is necessarily aperiodic.

EXERCISE 5.6.10. In many applications, the structure of a Markov chain is determined by a finite, simple graph $(\mathbb{S}, E)$. To be precise, the state space $\mathbb{S}$ is set of vertices of the graph and E is a symmetric subset of $\mathbb{S}^2$, the set of edges; and simplicity means that there are no loops or double edges. Next, we say that j is a nearest neighbor of i, denoted by $j \in \mathcal{N}(i)$, if and only if $(i,j) \in E$. Throughout, we assume that the degree $d(i) \equiv \#\mathcal{N}(i)$ is positive for each $i \in \mathbb{S}$. The transition probability $\mathbf{P}$ associated with $(\mathbb{S}, E)$ is determined so that $(\mathbf{P})_{ij} = \frac{1}{d(i)}$ if $j \in \mathcal{N}(i)$ and 0 otherwise. The interested reader will find more examples in [1].

(**a**) Show that $\mathbf{P}$ is irreducible if and only the graph is connected. That is, if and only if to each pair $(i,j) \in \mathbb{S}^2$ there corresponds a finite sequence $(i_0, \ldots, i_n) \in \mathbb{S}^{n+1}$ such that $i = i_0$, $j = i_n$, and $(i_{m-1}, i_m) \in E$ for $1 \leq m \leq n$. In addition, assuming irreducibility, show that the chain is aperiodic if and only if the graph is not bipartite. That is, if and only if there is no non-empty $\mathbb{S}' \subsetneq \mathbb{S}$ with the property that every edge connects a point in $\mathbb{S}'$ to one in $\mathbb{S} \setminus \mathbb{S}'$.

(**b**) Determine the probability vector $\boldsymbol{\pi}$ by $(\boldsymbol{\pi})_i = \frac{d(i)}{2\#E}$, and show that $\boldsymbol{\pi}$ is a reversible for $\mathbf{P}$.

(**c**) Assuming that the graph is connected, choose a set $\mathcal{P} = \{p(i,j) : (i,j) \in \mathbb{S}^2 \setminus D\}$ of allowable paths, as in §5.2.2, and show that

$$\beta_+ \leq 1 - \frac{2\#E}{D^2 L(\mathcal{P}) B(\mathcal{P})}, \tag{5.6.11}$$

where $D \equiv \max_{i \in \mathbb{S}} d(i)$, $L(\mathcal{P})$ is the maximal length (i.e., number of edges) of the paths in $\mathcal{P}$, and $B(\mathcal{P})$, the *bottleneck coefficient*, is the maximal number of paths $p \in \mathcal{P}$ which cross over an edge $e \in E$. Obviously, if one chooses $\mathcal{P}$ to consist of geodesics (i.e., paths of minimal length connecting their end points), then $L(\mathcal{P})$ is just the diameter of $(\mathbb{S}, E)$, and, as such, is as small as possible. On the other hand, because it may force there to be bad bottlenecks (i.e., many paths traversing a given edge), choosing geodesics may not be the optimal.

(**d**) Assuming that the graph is connected and not bipartite, choose $\mathcal{P} = \{p(i) : i \in \mathbb{S}\}$ to be of set of allowable closed paths of odd length, and show that

$$\beta_- \geq -1 + \frac{2}{DL(\mathcal{P})B(\mathcal{P})}. \tag{5.6.12}$$

EXERCISE 5.6.13. Here is an example to which the considerations in Exercise 5.6.10 apply and give close to optimal results. Namely, let $N \geq 2$, and consider the set $\mathbb{S}$ in the complex plain $\mathbb{C}$ consisting of the N roots of unity. Next, take E be the collection of pairs of adjacent roots of unity. That is, pairs of the form $\left(e^{\frac{\sqrt{-1}2\pi(m-1)}{N}}, e^{\frac{\sqrt{-1}2\pi m}{N}}\right)$. Finally, take $\mathbf{P}$ and $\boldsymbol{\pi}$ accordingly, as in the preceding.

(a) As an application of (c) in the preceding, show that $\beta_+ \leq 1 - \frac{8N}{(N-1)^2(N+1)}$.

(b) Assume that N is odd, and use (d) above to show that $\beta_- \geq -1 + \frac{1}{N^2}$.

EXERCISE 5.6.14. In this exercise we will give a very cursory introduction to a class of reversible Markov processes which provide somewhat naïve mathematical models of certain physical systems. In the literature, these are often called, for reasons which will be clear shortly, *spin-flip systems*, and they are among the earliest examples of Glauber dynamics. Here the state space $\mathbb{S} = \{-1,1\}^N$ is to be thought of as the configuration space for a system of N particles, each of which has "spin" $+1$ or -1. Because it is more conventional, we will use $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ or $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$ to denote generic elements of $\mathbb{S}$. Given $\boldsymbol{\omega} \in \mathbb{S}$ and $1 \leq k \leq N$, $\hat{\omega}^k$ will be the configuration obtained from $\boldsymbol{\omega}$ by "flipping" its kth spin. That is, the $\hat{\omega}^k = (\omega_1, \ldots, \omega_{k-1}, -\omega_k, \omega_{k+1}, \ldots, \omega_N)$. Next, given $\mathbb{R}$-valued functions f and g on $\mathbb{S}$, define

$$\Gamma(f,g)(\boldsymbol{\omega}) = \sum_{k=1}^{N} \bigl(f(\hat{\omega}^k) - f(\boldsymbol{\omega})\bigr)\bigl(g(\hat{\omega}^k) - g(\boldsymbol{\omega})\bigr),$$

which is a discrete analog of the dot product of the gradient of f with the gradient of g. Finally, given a probability vector $\boldsymbol{\mu}$ with $(\boldsymbol{\mu})_\omega > 0$ for all $\boldsymbol{\omega} \in \mathbb{S}$, define

$$\mathcal{E}^{\boldsymbol{\mu}}(f,g) = \frac{1}{2}\bigl\langle \Gamma(f,g) \bigr\rangle_{\boldsymbol{\mu}}, \quad \rho_k^{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \frac{(\boldsymbol{\mu})_\omega + (\boldsymbol{\mu})_{\hat{\omega}^k}}{2(\boldsymbol{\mu})_\omega}$$

$$\text{and } (\mathbf{Q}^{\boldsymbol{\mu}})_{\omega\eta} = \begin{cases} \rho^{\boldsymbol{\mu}}(\boldsymbol{\omega}) & \text{if } \boldsymbol{\eta} = \hat{\boldsymbol{\mu}}^k \\ 0 & \text{if } \boldsymbol{\eta} \notin \{\boldsymbol{\omega}, \hat{\omega}^1, \ldots, \hat{\omega}^N\} \\ -\sum_{k=1}^{N} \rho_k^{\boldsymbol{\mu}}(\boldsymbol{\omega}) & \text{if } \boldsymbol{\eta} = \boldsymbol{\omega}. \end{cases}$$

(a) Check that $\mathbf{Q}^{\boldsymbol{\mu}}$ is an irreducible Q-matrix on $\mathbb{S}$ and that the detailed balance condition $(\boldsymbol{\mu})_\omega \mathbf{Q}^{\boldsymbol{\mu}}_{\omega\eta} = (\boldsymbol{\mu})_\eta \mathbf{Q}^{\boldsymbol{\mu}}_{\eta\omega}$ holds. In addition, show that

$$-\bigl\langle g, \mathbf{Q}^{\boldsymbol{\mu}} f \bigr\rangle = \mathcal{E}^{\boldsymbol{\mu}}(f,g).$$

(b) Let $\boldsymbol{\lambda}$ be the uniform probability vector on $\mathbb{S}$. That is, $(\boldsymbol{\lambda})_\omega = 2^{-N}$ for each $\boldsymbol{\omega} \in \mathbb{S}$. Show that

$$\mathrm{Var}_{\boldsymbol{\mu}}(f) \leq M_{\boldsymbol{\mu}} \mathrm{Var}_{\boldsymbol{\lambda}}(f), \quad \text{where } M_{\boldsymbol{\mu}} \equiv 2^N \max_{\boldsymbol{\omega} \in \mathbb{S}} \boldsymbol{\mu}_\omega$$

and

$$\mathcal{E}^{\boldsymbol{\lambda}}(f,f) \leq \frac{1}{m_{\boldsymbol{\mu}}} \mathcal{E}^{\boldsymbol{\mu}}(f,f) \quad \text{where } m_{\boldsymbol{\mu}} = 2^N \min_{\boldsymbol{\omega} \in \mathbb{S}} \boldsymbol{\mu}_\omega.$$

(c) For each $S \subseteq \{1, \ldots, N\}$, define $\chi_S : \mathbb{S} \longrightarrow \{-1,1\}$ so that $\chi_S(\boldsymbol{\omega}) = \prod_{k \in S} \omega_k$. In particular, $\chi_\emptyset = \mathbf{1}$. Show that $\{\chi_S : S \subseteq \{1, \ldots, N\}\}$ is an orthonormal basis in $L^2(\boldsymbol{\lambda})$ and that $\mathbf{Q}^{\boldsymbol{\lambda}} \chi_S = -2(\# S)\chi_S$, and conclude from these that $\mathrm{Var}_{\boldsymbol{\lambda}}(f) \leq \frac{1}{2}\mathcal{E}^{\boldsymbol{\lambda}}(f,f)$.

(**d**) By combining (**b**) with (**c**), show that $\beta_{\boldsymbol{\mu}}\mathrm{Var}_{\boldsymbol{\mu}} \leq \mathcal{E}^{\boldsymbol{\mu}}(f,f)$ where $\beta_{\boldsymbol{\mu}} \equiv \frac{2m_{\boldsymbol{\mu}}}{M_{\boldsymbol{\mu}}}$. In particular, if $\{\mathbf{P}_t^{\boldsymbol{\mu}} : t \geq 0\}$ is the semigroup of transition probability matrices determined by $\mathbf{Q}^{\boldsymbol{\mu}}$, conclude that $\|\mathbf{P}_t^{\boldsymbol{\mu}} f - \langle f\rangle_{\boldsymbol{\mu}}\|_{2,\boldsymbol{\mu}} \leq e^{-\beta_{\boldsymbol{\mu}}}\|f - \langle f\rangle_{\boldsymbol{\mu}}\|_{2,\boldsymbol{\mu}}$.

EXERCISE 5.6.15. Refer to parts (**b**) and (**c**) in Exercise 5.6.14. It is somewhat surprising that the spectral gap for the uniform probability $\boldsymbol{\lambda}$ is 2, independent of N. In particular, this means that if $\{\mathbf{P}^{(N)}(t) : t \geq 0\}$ is the semigroup determined by the Q-matrix

$$(\mathbf{Q}^{(N)})_{\omega\eta} = \begin{cases} 1 & \text{if } \eta = \hat{\omega}^k \\ -N & \text{if } \eta = \omega \\ 0 & \text{otherwise} \end{cases}$$

on $\{-1,1\}^N$, then $\|\mathbf{P}^{(N)}(t)f - \langle f\rangle_{\boldsymbol{\lambda}^{(N)}}\|_{2,\boldsymbol{\lambda}^{(N)}} \leq e^{-2t}\|f\|_{2,\boldsymbol{\lambda}^{(N)}}$ for $t \geq 0$ and $f \in L^2(\boldsymbol{\lambda}^{(N)})$, where $\boldsymbol{\lambda}^{(N)}$ is the uniform probability measure on $\{-1,1\}^N$. In fact, the situation here provides convincing evidence of that the theory developed in this chapter works in situations where Doeblin's theory is doomed to failure. Indeed, the purpose of this exercise is to prove that, for any $t > 0$ and $\boldsymbol{\omega} \in \{-1,1\}^N$, $\lim_{N\to\infty}\|\boldsymbol{\mu}^{(N)}(t,\boldsymbol{\omega}) - \boldsymbol{\lambda}^{(N)}\|_{\mathrm{v}} = 2$ when $\big(\boldsymbol{\mu}^{(N)}(t,\boldsymbol{\omega})\big)_{\boldsymbol{\eta}} = \big(\mathbf{P}^{(N)}(t)\big)_{\boldsymbol{\omega}\boldsymbol{\eta}}$.

(**a**) Begin by showing that $\|\boldsymbol{\mu}^{(N)}(t,\boldsymbol{\omega}) - \boldsymbol{\lambda}^{(N)}\|_{\mathrm{v}}$ is independent of $\boldsymbol{\omega}$.

(**b**) Show that for any two probability vectors $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ on a countable space $\mathbb{S}$, $2 \geq \|\boldsymbol{\nu} - \boldsymbol{\nu}'\|_{\mathrm{v}} \geq 2|\boldsymbol{\nu}(A) - \boldsymbol{\nu}'(A)|$ for all $A \subseteq \mathbb{S}$.

(**c**) For $1 \leq k \leq N$, define the random variable X_k on $\{-1,1\}^N$ so that $X_k(\boldsymbol{\eta}) = \eta_k$ if $\boldsymbol{\eta} = (\eta_1,\dots,\eta_N)$. Show that, under $\boldsymbol{\lambda}^{(N)}$, the X_k's are mutually independent, $\{-1,1\}$-valued Bernoulli random variables with expectation 0. Next, let $t > 0$ be given, and set $\boldsymbol{\mu}^{(N)} = \boldsymbol{\mu}^{(N)}(t,\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ is the element of $\{-1,1\}$ whose coordinates are all -1. Show that, under $\boldsymbol{\mu}^{(N)}$, the X_k's are mutually independent, $\{-1,1\}$-valued Bernoulli random variables with expectation value $-e^{-2t}$.

(**d**) Continuing in the setting of (**c**), let $A^{(N)}$ be the set of $\boldsymbol{\eta}$ for which $\frac{1}{N}\sum_1^N X_k(\boldsymbol{\eta}) \leq -\frac{1}{2}e^{-2t}$, and show that $\boldsymbol{\mu}^{(N)}(A^{(N)}) - \boldsymbol{\lambda}^{(N)}(A^{(N)}) \geq 1 - \frac{8e^{4t}}{N}$. In fact, by using the sort of estimate developed at the end of §1.2.4, especially (1.2.16), one can sharpen this and get $\boldsymbol{\mu}^{(N)}(A^{(N)}) - \boldsymbol{\lambda}^{(N)}(A^{(N)}) \geq 1 - 2\exp\left(-\frac{Ne^{-4t}}{8}\right)$.

Hint: Use the usual Chebychev estimate with which the Weak Law is proved.

(**e**) By combining the preceding, conclude that $\|\boldsymbol{\mu}^{(N)}(t,\boldsymbol{\omega}) - \boldsymbol{\lambda}^{(N)}\|_{\mathrm{v}} \geq 2\big(1 - \frac{8e^{4t}}{N}\big)$ for all $t > 0$, $N \in \mathbb{Z}^+$, and $\boldsymbol{\omega} \in \{-1,1\}^N$.

CHAPTER 6

Some Mild Measure Theory

On Easter 1933, A.N. Kolmogorov published *Foundations of Probability*, a book which set out the foundations on which most of probability theory has rested ever since. Because Kolmogorov's model is given in terms of Lebesgue's theory of measures and integration, its full appreciation requires a thorough understanding that theory. Thus, although it is far too sketchy to provide anything approaching a thorough understanding, this chapter is an attempt to provide an introduction to Lebesgue's ideas and Kolmogorov's application of them in his model of probability theory.

6.1 A Description of Lebesgue's Measure Theory

In this section, we will introduce the terminology used in Lebesgue's theory. However, we will systematically avoid giving rigorous proofs of any hard results. There are many places in which these proofs can be found, one of them being [8].

6.1.1. Measure Spaces: The essential components in measure theory are a set Ω, the *space*, a collection $\mathcal{F}$ of subsets of Ω, the collection of *measurable, subsets*, and a function μ from $\mathcal{F}$ into $[0, \infty]$, called the *measure*. Being a space on which a measure might exist, the pair $(\Omega, \mathcal{F})$ is called a *measurable, space*, and when a measurable space $(\Omega, \mathcal{F})$ comes equipped with a measure μ, the triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*.

In order to avoid stupid trivialities, we will always assume that the space Ω is non-empty. Also, we will assume that the collection $\mathcal{F}$ of measurable sets forms a *σ-algebra* over Ω:

$$\Omega \in \mathcal{F}, \quad A \in \mathcal{F} \implies A\complement \equiv \Omega \setminus A \in \mathcal{F},$$
$$\text{and } \{A_n\}_1^\infty \subseteq \mathcal{F} \implies \bigcup_1^\infty A_n \in \mathcal{F}.$$

It is important to emphasize that, as distinguished from point-set topology (i.e., the description of open and closed sets) only finite or countable set theoretic operations are permitted in measure theory.

Using elementary set theoretic manipulations, it is easy to check that

$$A, B \in \mathcal{F} \implies A \cap B \in \mathcal{F} \text{ and } B \setminus A \in \mathcal{F}$$
$$\{A_n\}_1^\infty \subseteq \mathcal{F} \implies \bigcap_1^\infty A_n \in \mathcal{F}.$$

Finally, the measure μ will be function which assigns[1] 0 to $\emptyset$ and is *countably additive* in the sense that

$$\begin{aligned} &\{A_n\}_1^\infty \subseteq \mathcal{F} \text{ and } A_m \cap A_n = \emptyset \text{ when } m \neq n \\ &\implies \mu\left(\bigcup_1^\infty A_n\right) = \sum_1^\infty \mu(A_n). \end{aligned} \tag{6.1.1}$$

In particular, for $A, B \in \mathcal{F}$,

$$\begin{aligned} A \subseteq B &\implies \mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A) \\ \mu(A \cap B) < \infty &\implies \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B). \end{aligned} \tag{6.1.2}$$

The first line comes from writing B as the union of the disjoint sets A and $B \setminus A$, and the second line comes from writing $A \cup B$ as the union of the disjoint sets A and $B \setminus (A \cap B)$ and then applying the first line to B and $A \cap B$. The finiteness condition is needed when one moves the term $\mu(A \cap B)$ to the left hand side in $\mu(B) = \mu(A \cap B) + \mu\big(B \setminus (A \cap B)\big)$. That is, one wants to avoid having to subtract ∞ from ∞.

When the set Ω is finite or countable, there is no problem constructing measure spaces. Namely, one can take $\mathcal{F} = \{A : A \subseteq \Omega\}$, the set of all subsets of Ω, make any assignment of $\omega \in \Omega \longmapsto \mu(\{\omega\}) \in [0, \infty]$, at which point countable additivity demands that we take

$$\mu(A) = \sum_{\omega \in A} \mu(\{\omega\}) \quad \text{for } A \subseteq \Omega.$$

However, when Ω is uncountable, it is far from obvious that interesting measures can be constructed on a non-trivial collection of measurable sets. Indeed, it is reasonable to think that Lebesgue's most significant achievement was his construction of a measure space in which $\Omega = \mathbb{R}$, $\mathcal{F}$ is a σ-algebra of which every interval (open, closed, or semi-closed) is an element, and μ assigns each interval its length (i.e., $\mu(I) = b - a$ if I is an interval whose right and left end points are b and a).

Although the terminology is misleading, a measure space $(\omega, \mathcal{F}, \mu)$ is said to be *finite* if $\mu(\Omega) < \infty$. That is, the "finiteness" here is not determined by

[1] In view of additivity, it is clear that either $\mu(\emptyset) = 0$ or $\mu(A) = \infty$ for all $A \in \mathcal{F}$. Indeed, by additivity, $\mu(\emptyset) = \mu(\emptyset \cup \emptyset) = 2\mu(\emptyset)$, and therefore $\mu(\emptyset)$ is either 0 or ∞. Moreover, if $\mu(\emptyset) = \infty$, then $\mu(A) = \mu(A \cup \emptyset) = \mu(A) + \mu(\emptyset) = \infty$ for all $A \in \mathcal{F}$.

the size of Ω but instead by how large Ω looks to μ. Even if a measure space is not finite, it may be decomposable into a countable number of finite pieces, in which case it is said to be *σ-finite*. Equivalently, $(\Omega, \mathcal{F}, \mu)$ is σ-finite if there exists $\{\Omega_n\}_1^\infty \subseteq \mathcal{F}$ such that[2] $\Omega_n \nearrow \Omega$ and $\mu(\Omega_n) < \infty$ for each $n \geq 1$. Thus, for example, both Lebesgue's measure on $\mathbb{R}$ and counting measure on $\mathbb{Z}$ are σ-finite but not finite.

6.1.2. Some Consequences of Countable Additivity: Countable additivity is the *sine qua non* in this subject. In particular, it leads to the following continuity properties of measures:

$$\begin{aligned} &\{A_n\}_1^\infty \subseteq \mathcal{F} \text{ and } A_n \nearrow A \implies \mu(A_n) \nearrow \mu(A) \\ &\{A_n\}_1^\infty \subseteq \mathcal{F},\ \mu(A_1) < \infty, \text{ and } A_n \searrow A \implies \mu(A_n) \searrow \mu(A). \end{aligned} \tag{6.1.3}$$

Although we do not intend to prove many of the results discussed in this chapter, the proofs of those in (6.1.3) are too basic and easy to omit. Indeed, to prove the first line, simply take $B_1 = A_1$ and $B_{n+1} = A_{n+1} \setminus A_n$. Then $B_m \cap B_n = \emptyset$ when $m \neq n$, $\bigcup_1^n B_m = A_n$ for all $n \geq 1$, and $\bigcup_1^\infty B_m = A$. Hence, by (6.1.1),

$$\mu(A_n) = \sum_1^n \mu(B_m) \nearrow \sum_1^\infty \mu(B_m) = \mu\left(\bigcup_1^\infty B_m\right) = \mu(A).$$

To prove the second line, begin by noting that $\mu(A_1) = \mu(A_n) + \mu(A_1 \setminus A_n)$ and $\mu(A_1) = \mu(A) + \mu(A_1 \setminus A)$. Hence, since $A_1 \setminus A_n \nearrow A_1 \setminus A$ and $\mu(A_1) < \infty$, we have $\mu(A_1) - \mu(A_n) \nearrow \mu(A_1) - \mu(A)$ and therefore that $\mu(A_n) \searrow \mu(A)$. Just as in the proof of second line in (6.1.2), we need the finiteness condition here to avoid being forced to subtract ∞ from ∞.

Another important consequence of countable additivity is *countable subadditivity*:

$$\{A_n\}_1^\infty \subseteq \mathcal{F} \implies \mu\left(\bigcup_1^\infty A_n\right) \leq \sum_1^\infty \mu(A_n). \tag{6.1.4}$$

Like the preceding, this is easy. Namely, if $B_1 = A_1$ and $B_{n+1} = A_{n+1} \setminus \bigcup_1^n A_m$, then $\mu(B_n) \leq \mu(A_n)$ and

$$\mu\left(\bigcup_1^\infty A_n\right) = \mu\left(\bigcup_1^\infty B_n\right) = \sum_1^\infty \mu(B_n) \leq \sum_1^\infty \mu(A_n).$$

A particularly important consequence of (6.1.4) is the fact that

$$\mu\left(\bigcup_1^\infty A_n\right) = 0 \quad \text{if } \mu(A_n) = 0 \text{ for each } n \geq 1. \tag{6.1.5}$$

[2] We write $A_n \nearrow A$ when $A_n \subseteq A_{n+1}$ for all $n \geq 1$ and $A = \bigcup_1^\infty A_n$. Similarly, $A_n \searrow A$ means that $A_n \supseteq A_{n+1}$ for all $n \geq 1$ and $A = \bigcap_1^\infty A_n$. Obviously, $A_n \nearrow A$ if and only if $A_n{}^{\mathbf{C}} \searrow A^{\mathbf{C}}$.

That is, *the countable union of sets each of which has measure* 0 *is again a set having measure* 0. Here one begins to see the reason for restricting oneself to countable operations in measure theory. Namely, it is certainly not true that the uncountable union of sets having measure 0 will necessarily have measure 0. For example, in the case of Lebesgue's measure on $\mathbb{R}$, the second line of (6.1.3) implies

$$\mu(\{x\}) = \lim_{\delta \searrow 0} \mu\big((x-\delta, x+\delta)\big) = \lim_{\delta \searrow 0} 2\delta = 0$$

for each point $x \in \mathbb{R}$, and yet $(0,1) = \bigcup_{x\in(0,1)}\{x\}$ has measure 1.

6.1.3. Generating σ-Algebras: Very often one wants to make sure that a certain collection of subsets will be among the measurable subsets, and for this reason it is important to know the following constructions. First, suppose that $\mathcal{C}$ is a collection of subsets of Ω. Then there is a smallest σ-algebra $\sigma(\mathcal{C})$, called the σ-algebra *generated by* $\mathcal{C}$, over Ω which contains $\mathcal{C}$. Namely, consider the collection of all the σ-algebras over Ω which contain $\mathcal{C}$. This collection is non-empty because $\{A : A \subseteq \Omega\}$ is an element. In addition, as is easily verified, the intersection of any collection of σ-algebras is again a σ-algebra. Hence, $\sigma(\mathcal{C})$ is the intersection of all the σ-algebras which contain $\mathcal{C}$. When Ω is a topological space and $\mathcal{C}$ is the collection of all open subsets of Ω, then $\sigma(\mathcal{C})$ is called the *Borel σ-algebra* and is denoted by $\mathcal{B}_\Omega$.

One the most important reasons for knowing how a σ-algebra is generated is that one can often check properties of measures on $\sigma(\mathcal{C})$ by making sure that the property holds on $\mathcal{C}$. An important example of such a result is the one in the following uniqueness theorem.

6.1.6 THEOREM. *Suppose that $(\Omega, \mathcal{F})$ is a measurable space and that $\mathcal{C} \subseteq \mathcal{F}$ includes Ω and is closed under intersection (i.e., $A \cap B \in \mathcal{C}$ whenever $A, B \in \mathcal{C}$). If μ and ν are a pair of finite measures on $(\Omega, \mathcal{F})$ and if $\mu(A) = \nu(A)$ for each $A \in \mathcal{C}$, then $\mu(A) = \nu(A)$ for all $A \in \sigma(\mathcal{C})$.*

PROOF: We will say that $\mathcal{S} \subseteq \mathcal{F}$ is good if

(i) $\quad A, B \in \mathcal{S}$ and $A \subseteq B \implies B \setminus A \in \mathcal{S}$.

(ii) $\quad A, B \in \mathcal{S}$ and $A \cap B = \emptyset \implies A \cup B \in \mathcal{S}$.

(iii) $\quad \{A_n\}_1^\infty \subseteq \mathcal{S}$ and $A_n \nearrow A \implies A \in \mathcal{S}$.

Notice that if $\mathcal{S}$ is good, $\Omega \in \mathcal{S}$, and $A, B \in \mathcal{S} \implies A \cap B \in \mathcal{S}$, then $\mathcal{S}$ is a σ-algebra. Indeed, because of (i) and (iii), all that one has to do is check that $A \cup B \in \mathcal{S}$ whenever $A, B \in \mathcal{S}$. But, because $A \cup B = \big(A \setminus (A \cap B)\big) \cup B$, this is clear from (i), (ii), and the fact that $\mathcal{S}$ is closed under intersections. In addition, observe that if $\mathcal{S}$ is good, then, for any $\mathcal{D} \subseteq \mathcal{F}$,

$$\mathcal{S}' \equiv \{A \in \mathcal{S} : A \cap B \in \mathcal{S} \text{ for all } B \in \mathcal{D}\}$$

is again good.

Now set $\mathcal{B} = \{A \in \mathcal{F} : \mu(A) = \nu(A)\}$. From the properties of finite measures, in particular, (6.1.2) and (6.1.3), it is easy to check that $\mathcal{B}$ is good. Moreover, by assumption, $\mathcal{C} \subseteq \mathcal{B}$. Thus, if $\mathcal{B}' = \{A \in \mathcal{B} : A \cap C \in \mathcal{B} \text{ for all } C \in \mathcal{C}\}$, then, by the preceding observation, $\mathcal{B}'$ is again good. In addition, because $\mathcal{C}$ is closed under intersection, $\mathcal{C} \subseteq \mathcal{B}'$. Similarly, $\mathcal{B}'' \equiv \{A \in \mathcal{B}' : A \cap B \in \mathcal{B}' \text{ for all } B \in \mathcal{B}'\}$ is also good, and, by the definition of $\mathcal{B}'$, $\mathcal{C} \subseteq \mathcal{B}''$. Finally, if $A, A' \in \mathcal{B}''$ and $B \in \mathcal{B}'$, then $(A \cap A') \cap B = A \cap (A' \cap B) \in \mathcal{B}'$, and so $A \cap A' \in \mathcal{B}''$. Hence, $\mathcal{B}''$ is a σ-algebra which contains $\mathcal{C}$, $\mathcal{B}'' \subseteq \mathcal{B}$, and therefore μ equals ν on $\sigma(\mathcal{C})$. □

6.1.4. Measurable Functions: Given a pair $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ of measurable spaces, we will say that the map $F : \Omega_1 \longrightarrow \Omega_2$ is *measurable* if the inverse image of sets in $\mathcal{F}_2$ are elements of $\mathcal{F}_1$: $F^{-1}(\Gamma) \in \mathcal{F}_1$ for every $\Gamma \in \mathcal{F}_2$.[3] Notice that if $\mathcal{F}_2 = \sigma(\mathcal{C})$, F is measurable if $F^{-1}(C) \in \mathcal{F}_1$ for each $C \in \mathcal{C}$. In particular, if Ω_1 and Ω_2 are topological spaces and $\mathcal{F}_i = \mathcal{B}_{\Omega_i}$, then every continuous map from Ω_1 to Ω_2 is measurable.

It is important to know that when $\Omega_2 = \mathbb{R}$ and $\mathcal{F}_2 = \mathcal{B}_{\mathbb{R}}$, *measurability is preserved under sequential limit operations.* To be precise, if $\{f_n\}_1^\infty$ is a sequence of $\mathbb{R}$-valued measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, then

$$\begin{aligned} &\omega \rightsquigarrow \sup_n f_n(\omega), \quad \omega \rightsquigarrow \inf_n f_n(\omega), \\ &\omega \rightsquigarrow \overline{\lim_{n\to\infty}} f_n(\omega), \quad \text{and } \omega \rightsquigarrow \underline{\lim_{n\to\infty}} f_n(\omega) \end{aligned} \quad \text{are measurable.} \tag{6.1.7}$$

For example, the first of these can be proved by the following line of reasoning. Begin with the observation that $\mathcal{B}_{\mathbb{R}} = \sigma(\mathcal{C})$ when $\mathcal{C} = \{(a, \infty) : a \in \mathbb{R}\}$. Hence $f : \Omega \longrightarrow \mathbb{R}$ will be measurable if and only if[4] $\{f > a\}$ for each $a \in \mathbb{R}$, and so the first function in (6.1.7) is measurable because

$$\left\{\sup_n f_n > a\right\} = \bigcup_n \{f_n > a\} \quad \text{for every } a \in \mathbb{R}.$$

As a consequence the second line in (6.1.7), we know that the set Δ of points ω at which the limit $\lim_{n\to\infty} f_n(\omega)$ exists is measurable, and $\omega \rightsquigarrow \lim_{n\to\infty} f_n(\omega)$ is measurable if $\Delta = \Omega$.

Finally, measurable functions give rise to important instances of the construction in the preceding subsection. Namely, suppose that the space Ω_1 and the measurable space $(\Omega_2, \mathcal{F}_2)$ are given, and let $\mathfrak{F}$ be some collection of maps from Ω_1 into Ω_2. Then the *σ-algebra $\sigma(\mathfrak{F})$ over Ω_1 generated by $\mathfrak{F}$* is the smallest σ-algebra over Ω_1 with respect to which every element of $\mathfrak{F}$ is measurable. Equivalently, $\sigma(\mathfrak{F}) = \sigma(\mathcal{C})$ when $\mathcal{C} = \{F^{-1}(\Gamma) : F \in \mathfrak{F} \ \& \ \Gamma \in \mathcal{F}_2\}$.

[3] The reader should notice the striking similarity between this definition and the one for continuity in terms of inverse images of open sets.

[4] When there is no ambiguity caused by doing so, we use $\{F \in \Gamma\}$ to stand for $\{\omega : F(\omega) \in \Gamma\}$.

Finally, suppose that $\mathcal{F}_2 = \sigma(\mathcal{C}_2)$, where $\mathcal{C}_2 \subseteq \mathcal{F}_2$ contains Ω_2 and is closed under intersection (i.e., $A \cap B \in \mathcal{C}_2$ if $A, B \in \mathcal{C}_2$), and let $\mathcal{C}_1$ be the collection of sets of the form $F_1^{-1}(A_1) \cap \cdots \cap F_n^{-1}(A_n)$ for $n \in \mathbb{Z}^+$, $\{F_1, \ldots, F_n\} \subseteq \mathfrak{F}$, and $\{A_1, \ldots, A_n\} \subseteq \mathcal{C}_2$. Then $\sigma(\mathfrak{F}) = \sigma(\mathcal{C}_1)$, $\Omega_1 \in \mathcal{C}_1$, and $\mathcal{C}_1$ is closed under intersection. In particular, by Theorem 6.1.6, *if μ and ν are a pair of finite measures on $(\Omega_1, \mathcal{F}_1)$ and*

$$\begin{aligned}&\mu\big(\{\omega_1 : F_1(\omega_1) \in A_1, \ldots, F_n(\omega_1) \in A_n\}\big)\\ &\quad = \nu\big(\{\omega_1 : F_1(\omega_1) \in A_1, \ldots, F_n(\omega_1) \in A_n\}\big)\end{aligned}$$

for all $n \in \mathbb{Z}^+$, $\{F_1, \ldots, F_n\} \subseteq \mathfrak{F}$, and $\{A_1, \ldots, A_n\} \subseteq \mathcal{C}_2$, then μ equals ν on $\sigma(\mathfrak{F})$.

6.1.5. Lebesgue Integration: Given a measure space $(\Omega, \mathcal{F}, \mu)$, Lebesgue's theory of integration begins by giving a prescription for defining the integral with respect to μ of all non-negative, measurable functions (i.e., all measurable maps f from $(\Omega, \mathcal{F})$ into[5] $\big([0,\infty]; \mathcal{B}_{[0,\infty]}\big)$. Namely, his theory says that when $\mathbf{1}_A$ is the indicator function of a set $A \in \mathcal{F}$ and $a \in [0,\infty)$, then the integral of the function[6] $a\mathbf{1}_A$ should be equal to $a\mu(A)$. He then insists that the integral should be additive in the sense that the integral of $f_1 + f_2$ should be sum of the integrals of f_1 and f_2. In particular, this means that if f is a non-negative, measurable function which is *simple*, in the sense that it takes on only a finite number of values, then the integral $\int f\, d\mu$ of f must be

$$\sum_{x \in [0,\infty)} x\mu(f^{-1}(\{x\})),$$

where, because $f^{-1}(\{x\}) = \emptyset$ for all but a finite number of x, the sum involves only a finite number of non-zero terms. Of course, before one can insist on this additivity of the integral, one is obliged to check that additivity is consistent. Specifically, it is necessary to show that $\sum_1^n a_m \mu(A_m) = \sum_1^{n'} a'_{m'} \mu(A'_{m'})$ when $\sum_1^n a_m \mathbf{1}_{A_m} = f = \sum_1^{n'} a'_m \mathbf{1}_{A'_{m'}}$. However, once this consistency has been established, one knows how to define the integral of any non-negative, measurable, simple function in such a way that the integral is additive and gives the obvious answer for indicator functions. In particular, additivity implies *monotonicity*: $\int f\, d\nu \leq \int g\, d\mu$ when $f \leq g$.

To complete Lebesgue's program for non-negative functions, one has to first observe that if $f : E \longrightarrow [0,\infty]$ is measurable, then there exists a sequence $\{\varphi_n\}_1^\infty$ of non-negative, measurable, simple functions with the property that $\varphi_n(\omega) \nearrow f(\omega)$ for each $\omega \in \Omega$. For example, one can take

$$\varphi_n = \sum_{m=1}^{4^n} \frac{m}{2^n} \mathbf{1}_{A_{m,n}} \quad \text{where } A_{m,n} = \{\omega : m2^{-n} \leq f(\omega) < (m+1)2^{-n}\}.$$

[5] In this context, we are thinking of $[0,\infty]$ as the compact metric space obtained by mapping $[0,1]$ onto $[0,\infty]$ via the map $t \in [0,1] \longmapsto \tan(\frac{\pi}{2}t)$.

[6] In measure theory, the convention which works best is to take $0\infty = 0$.

Given $\{\varphi_n\}_1^\infty$, what Lebesgue says is that $\int f\,d\mu = \lim_{n\to\infty} \int \varphi_n\,d\mu$. Indeed, by monotonicity, $\int \varphi_n\,d\mu$ is non-decreasing in n, and therefore the indicated limit necessarily exists. On the other hand, just as there was earlier, there is a consistency problem which we must resolve before we can adopt Lebesgue's definition. This time the problem comes from the fact that there are myriad ways in which to construct the approximating simple functions φ_n, and one must make sure that the limit does not depend on which approximation scheme one chooses. That is, it is necessary to check that if $\{\varphi_n\}_1^\infty$ and $\{\psi_n\}_1^\infty$ are two non-decreasing sequences of non-negative, simple, measurable functions such that $\lim_{n\to\infty} \varphi_n(\omega) = \lim_{n\to\infty} \psi_n(\omega)$ for each $\omega \in \Omega$, then $\lim_{n\to\infty} \int \varphi_n\,d\mu = \lim_{n\to\infty} \int \psi_n\,d\mu$; and it is at this step that the full power of countable additivity must be brought to bear.

Having defined $\int f\,d\mu$ for all non-negative, measurable f's, one must check that the resulting integral is homogeneous and additive: $\int af\,d\mu = a\int f\,d\mu$ for $a \in [0,\infty]$ and $\int(f_1 + f_2)\,d\mu = \int f_1\,d\mu + \int f_2\,d\mu$. However, both these properties are easily seen to be inherited from the case of simple f's. Thus, the only remaining challenge in Lebesgue's construction is to get away from the restriction to non-negative functions and extend the integral to measurable functions which can take both signs. On the other hand, if one wants the resulting theory to be linear, then there is no doubt about how this extension must be made. Namely, given a measurable function $f : E \longrightarrow [-\infty,\infty]$, it is not hard to show that $f^+ \equiv f\vee 0$ and $f^- \equiv -(f\wedge 0) = (-f)^+$ are non-negative measurable functions. Hence, because $f = f^+ - f^-$, linearity demands that $\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu$; and this time there are two problems which have to be confronted. In the first place, at the very least, it is necessary to restrict ones attention to functions f for which at least one of the numbers $\int f^+\,d\mu$ or $\int f^-\,d\mu$ is finite, otherwise one ends up having to deal with $\infty-\infty$. Secondly, one must, once again, check consistency. That is, if $f = f_1 - f_2$, where f_1 and f_2 are non-negative and measurable, then one has to show that $\int f_1\,d\mu - \int f_2\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu$.

In most applications, the measurable functions which one integrates are either non-negative or have the property that $\int |f|\,d\mu < \infty$, in which case f is said to be an *integrable function*. Because $|a_1f_1+a_2f_2| \le |a_1||f_1|+|a_2||f_2|$, the set of integrable functions forms a vector space over $\mathbb{R}$ on which integration with respect to μ acts as a linear function. Finally, if f is a measurable function which is either non-negative or integrable and $A \in \mathcal{F}$, then the product $\mathbf{1}_A f$ is again a measurable function which is either non-negative or integrable and so

$$\int_A f\,d\mu \equiv \int \mathbf{1}_A f\,d\mu \tag{6.1.8}$$

is well defined.

6.1.6. Stability Properties of Lebesgue Integration: The power of Lebesgue's theory derives from the stability of the integral it defines, and its

stability is made manifest in the following three famous theorems. Throughout, $(\Omega, \mathcal{F}, \mu)$ is a measure space to which all references about measurability and integration refer. Also, the functions here are assumed to take their values in $(-\infty, \infty]$.

6.1.9 THEOREM. (**Monotone Convergence**) *Suppose that $\{f_n\}_1^\infty$ is a non-decreasing sequence of measurable functions, and, for each $\omega \in \Omega$, set $f(\omega) = \lim_{n\to\infty} f_n(\omega)$. If there exists a fixed integrable function g which is dominated by each f_n, then $\int f_n\, d\mu \nearrow \int f\, d\mu$. If, instead, $\{f_n\}_1^\infty$ is non-increasing and if there exists a fixed integrable function g which dominates each f_n, then $\int f_n\, d\mu \searrow \int f\, d\mu$.*

6.1.10 THEOREM. (**Fatou's Lemmas**) *Given any sequence $\{f_n\}_1^\infty$ of measurable functions, all of which dominate some fixed integrable function g,*

$$\varliminf_{n\to\infty} \int f_n\, d\mu \geq \int \varliminf_{n\to\infty} f_n\, d\mu.$$

If, instead, there is some fixed integrable function g which dominates all of the f_n's, then

$$\varlimsup_{n\to\infty} \int f_n\, d\mu \leq \int \varlimsup_{n\to\infty} f_n\, d\mu.$$

6.1.11 THEOREM. (**Lebesgue's Dominated Convergence**) *Suppose that $\{f_n\}_1^\infty$ is a sequence of measurable functions, and assume that there is a fixed integrable function g with the property that $\mu(\{\omega : |f_n(\omega)| > g(\omega)\}) = 0$ for each $n \geq 1$. Further, assume that there exists a measurable function f to which $\{f_n\}$ converges in either one of the following two senses:*

(**a**) $$\mu\left(\left\{\omega : f(\omega) \neq \lim_{n\to\infty} f_n(\omega)\right\}\right) = 0$$

(**b**) $$\lim_{n\to\infty} \mu(\{\omega : |f_n(\omega) - f(\omega)| \geq \epsilon\}) = 0 \quad \textit{for all } \epsilon > 0.$$

Then

$$\left|\int f_n\, d\mu - \int f\, d\mu\right| \leq \int |f_n - f|\, d\mu \longrightarrow 0 \quad \text{as } n \to \infty.$$

Before proceeding, a word should be said about the conditions in Lebesgue's Dominated Convergence Theorem. Because integrals cannot "see" what is happening on a *set of measure* 0 (i.e., an $A \in \mathcal{F}$ for which $\mu(A) = 0$) it is natural that conditions which guarantee certain behavior of integrals should be conditions which need only hold on the complement of a set of measure 0. Thus, condition (**a**) says that $f_n(\omega) \longrightarrow f(\omega)$ for all ω outside of a set of measure 0. In the jargon, conditions which hold off of a set of measure 0 are said to hold *almost everywhere*. In this terminology, the first hypothesis is that $|f_n| \leq g$ almost everywhere and (**a**) is saying that $\{f_n\}_1^\infty$ converges to f almost everywhere, and these statements would be abbreviated

in the literature by something like $|f_n| \le g$ a.e. and $f_n \longrightarrow f$ a.e.. The condition in (**b**) is related to, but significantly different from, the one in (**a**). In particular, it does not guarantee that $\{f_n(\omega)\}_1^\infty$ converges for any $\omega \in \Omega$. For example, take μ to be the measure of Lebesgue on $\mathbb{R}$ described above, and take $f_{m+2^n}(\omega) = \mathbf{1}_{[0,2^{-n})}(\omega - m2^{-n})$ for $n \ge 0$ and $0 \le m < 2^n$. Then $\mu\big(\{\omega : f_{m+2^n}(\omega) \neq 0\}\big) = 2^{-n}$, and so $\lim_{n\to\infty} \mu\big(\{\omega : |f_n(\omega)| \ge \epsilon\}\big) = 0$ for all $\epsilon > 0$. On the other hand, for each $\omega \in [0,1)$, $\overline{\lim}_{n\to\infty} f_n(\omega) = 1$ but $\underline{\lim}_{n\to\infty} f_n(\omega) = 0$. Thus, (**b**) most definitely does not imply (**a**). Conversely, although (**a**) implies (**b**) when $\mu(\Omega) < \infty$, (**a**) does not imply (**b**) when $\mu(\Omega) = \infty$. To wit, again when μ is Lebesgue's measure, and consider $f_n = \mathbf{1}_{\mathbb{R}\setminus[-n,n]}$.

In connection with the preceding discussion, there is a basic estimate, known as *Markov's inequality*, which plays a central role in all measure theoretic analysis. Namely, because, for any $\lambda > 0$, $\lambda \mathbf{1}_{[\lambda,\infty]} \circ f \le f \mathbf{1}_{[\lambda,\infty]} \circ f \le |f|$,

$$\mu\big(\{\omega : f(\omega) \ge \lambda\}\big) \le \frac{1}{\lambda} \int_{\{\omega : f(\omega) \ge \lambda\}} f \, d\mu \le \frac{1}{\lambda} \int |f| \, d\mu. \tag{6.1.12}$$

In particular, this leads to the conclusion that

$$\lim_{n\to\infty} \int |f_n - f| \, d\mu = 0 \implies \mu\big(\{\omega : |f_n(\omega) - f(\omega)| \ge \epsilon\}\big) = 0$$

for all $\epsilon > 0$. That is, the condition (**b**) is necessary for the conclusion in Lebesgue's theorem. In addition, (6.1.12) proves that

$$\int |f| \, d\mu = 0 \implies \mu\big(\{\omega : |f(\omega)| \ge \epsilon\}\big) = 0 \text{ for all } \epsilon > 0,$$

and therefore, by (6.1.3),

$$\int |f| \, d\mu = 0 \implies \mu\big(\{\omega : f(\omega) \neq 0\}\big) = 0. \tag{6.1.13}$$

Finally, the role of the Lebesgue dominant g is made clear by considering $\{f_n\}_1^\infty$ when either $f_n = n\mathbf{1}_{[0,\frac{1}{n})}$ or $f_n = \mathbf{1}_{[n-1,n)}$ and μ is Lebesgue's measure.

6.1.7. Lebesgue Integration in Countable Spaces: In this subsection we will see what Lebesgue's theory looks like in the relatively trivial case when Ω is countable, $\mathcal{F}$ is the collection of all subsets of Ω, and the measure μ is σ-finite. As we pointed out in §6.1.1, specifying a measure on $(\Omega, \mathcal{F})$ is tantamount to assigning a non-negative number to each element of Ω, and, because we want our measures to be σ-finite, no element is assigned ∞.

Because the elements of Ω can be counted, there is no reason to not count them. Thus, in the case when Ω is finite, there is no loss in generality if we write $\Omega = \{1, \dots, N\}$, where $N = \#\Omega$ is the number of elements in Ω, and, similarly, when Ω is countably infinite, we might as well, at least for abstract purposes, think of its being the set $\mathbb{Z}^+$ of positive integers. In fact, in order to

avoid having to worry about the finite and countably infinite cases separately, we will embed the finite case into the infinite one by simply noting that the theory for $\{1, \dots, N\}$ is exactly the same as the theory for $\mathbb{Z}^+$ restricted to measures μ for which $\mu(\{\omega\}) = 0$ when $\omega > N$. Finally, in order to make the notation here conform with the notation in the rest of the book, we will use $\mathbb{S}$ in place of Ω, i, j, or k to denote generic elements of $\mathbb{S}$, and we will identify a measure μ with the row vector $\boldsymbol{\mu} \in [0, \infty)^{\mathbb{S}}$ given by $(\boldsymbol{\mu})_i = \mu(\{i\})$.

The first thing to observe is that

$$\int f \, d\mu = \sum_{i \in \mathbb{S}} f(i)(\boldsymbol{\mu})_i \tag{6.1.14}$$

whenever either $\int f^+ \, d\mu$ or $\int f^- \, d\mu < \infty$ is finite. Indeed, if $\varphi \geq 0$ is simple, $a_1, \dots, a_L$ are the distinct values f takes, and $A_\ell = \{i : \varphi(i) = a_\ell\}$, then

$$\begin{aligned} \int \varphi \, d\mu &= \sum_{\ell=1}^{L} a_\ell \mu(A_\ell) = \sum_{\ell=1}^{L} a_\ell \sum_{i \in A_\ell} (\boldsymbol{\mu})_i \\ &= \sum_{\ell=1}^{L} \sum_{i \in A_\ell} \varphi(i)(\boldsymbol{\mu})_i = \sum_{i \in \mathbb{S}} \varphi(i)(\boldsymbol{\mu})_i. \end{aligned}$$

Second, given any $f \geq 0$ on $\mathbb{S}$, set $\varphi_n(i) = f(i)$ if $1 \leq i \leq n$ and $\varphi_n(i) = 0$ if $i > n$. Then,

$$\int f \, d\mu = \lim_{n \to \infty} \int \varphi_n \, d\mu = \lim_{n \to \infty} \sum_{1 \leq i \leq n} f(i)(\boldsymbol{\mu})_i = \sum_{i \in \mathbb{S}} f(i)(\boldsymbol{\mu})_i.$$

Finally, if either $\int f^+ \, d\mu$ or $\int f^- \, d\mu$ is finite, then it is clear that

$$\begin{aligned} \int f \, d\mu &= \int f^+ \, d\mu - \int f^- \, d\mu \\ &= \sum_{\{i : f(i) \geq 0\}} f(i)(\boldsymbol{\mu})_i - \sum_{\{i : f(i) \leq 0\}} f(i)(\boldsymbol{\mu})_i = \sum_{i \in \mathbb{S}} f(i)(\boldsymbol{\mu})_i. \end{aligned}$$

We next want to see what the "big three" look like in this context.

The Monotone Convergence Theorem: First observe that it suffices to treat the case when $0 \leq f_n \nearrow f$. Indeed, one can reduce each of these statements to that case by replacing f_n with $f_n - g$ or $g - f_n$. When $0 \leq f_n \nearrow f$, it is obvious that

$$0 \leq \sum_{i \in \mathbb{S}} f_n(i)(\boldsymbol{\mu})_i \leq \sum_{i \in \mathbb{S}} f_{n+1}(i)(\boldsymbol{\mu})_i \leq \sum_{i \in \mathbb{S}} f(i)(\boldsymbol{\mu})_i.$$

Thus, all that remains is to note that

$$\lim_{n \to \infty} \sum_{i \in \mathbb{S}} f_n(i)(\boldsymbol{\mu})_i \geq \lim_{n \to \infty} \sum_{i=1}^{L} f_n(i)(\boldsymbol{\mu})_i = \sum_{i=1}^{L} f(i)(\boldsymbol{\mu})_i$$

for each $L \in \mathbb{Z}^+$, and therefore that the desired result follows after one lets $L \nearrow \infty$.

Fatou's Lemma: Again one can reduce to the case when $f_n \geq 0$ and the limit being taken is the limit inferior. But in this case,

$$\begin{aligned}\varliminf_{n\to\infty} \sum_{i\in\mathbb{S}} f_n(i)(\boldsymbol{\mu})_i &= \lim_{m\to\infty} \inf_{n\geq m} \sum_{i\in\mathbb{S}} f_n(i)(\boldsymbol{\mu})_i \\ &\geq \lim_{m\to\infty} \sum_{i\in\mathbb{S}} \inf_{n\geq m} f_n(i)(\boldsymbol{\mu})_i = \sum_{i\in\mathbb{S}} \varliminf_{n\to\infty} f_n(i)(\boldsymbol{\mu})_i,\end{aligned}$$

where the last equality follows from the Monotone Convergence Theorem applied to $0 \leq \inf_{n\geq m} f_n \nearrow \varliminf_{n\to\infty} f_n$.

Lebesgue's Dominated Convergence Theorem: First note that, if we eliminate those $i \in \mathbb{S}$ for which $(\boldsymbol{\mu})_i = 0$, none of the conclusions change. Thus, we will, from now on assume that $(\boldsymbol{\mu})_i > 0$ for all $i \in \mathbb{S}$. Next observe that, under this assumption, the hypotheses become $\sup_n |f_n(i)| \leq g(i)$ and $f_n(i) \longrightarrow f(i)$ for each $i \in \mathbb{S}$. In particular, $|f| \leq g$. Hence, by considering $f_n - f$ instead of $\{f_n\}_1^\infty$ and replacing g by $2g$, we may and will assume that $f \equiv 0$. Now let $\epsilon > 0$ be given, and choose L so that $\sum_{i>L} g(i)(\boldsymbol{\mu})_i < \epsilon$. Then

$$\left|\sum_{i\in\mathbb{S}} f_n(i)(\boldsymbol{\mu})_i\right| \leq \sum_{i\in\mathbb{S}} |f_n(i)|(\boldsymbol{\mu})_i \leq \sum_{i=1}^{L} |f_n(i)|(\boldsymbol{\mu})_i + \epsilon \longrightarrow \epsilon \quad \text{as } n \to \infty.$$

6.1.8. Fubini's Theorem: Fubini's Theorem deals with products of measure spaces. Namely, given measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, the product of $\mathcal{F}_1$ and $\mathcal{F}_2$ is the σ-algebra $\mathcal{F}_1 \times \mathcal{F}_2$ over $\Omega_1 \times \Omega_2$ which is generated by the set $\{A_1 \times A_2 : A_1 \in \mathcal{F}_1 \;\&\; A_2 \in \mathcal{F}_2\}$ of *measurable rectangles*. An important technical fact about this construction is that, if f is a measurable function on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$, then, for each $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, $\omega_2 \rightsquigarrow f(\omega_1, \omega_2)$ and $\omega_1 \rightsquigarrow f(\omega_1, \omega_2)$ are measurable functions on, respectively, $(\Omega_2, \mathcal{F}_2)$ and $(\Omega_1, \mathcal{F}_1)$.

In the following statement, it is important to emphasize exactly which variable is being integrated. For this reason, we use the more detailed notation $\int_\Omega f(\omega)\,\mu(d\omega)$ instead of the more abbreviated $\int f\,d\mu$.

6.1.15 THEOREM. **(Fubini)**[7] *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be a pair of σ-finite measure spaces, and set $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$. Then there is a unique measure $\mu = \mu_1 \times \mu_2$ on $(\Omega, \mathcal{F})$ with the property that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. Moreover, if f is a non-negative, measurable function on $(\Omega, \mathcal{F})$, then both*

$$\omega_1 \rightsquigarrow \int_{\Omega_2} f(\omega_1, \omega_2)\,\mu_2(d\omega_2) \quad \text{and} \quad \omega_2 \rightsquigarrow \int_{\Omega_1} f(\omega_1, \omega_2)\,\mu_1(d\omega_1)$$

[7] Although this theorem is usually attributed Fubini, it seems that Tonelli deserves, but seldom receives, a good deal of credit for it.

are measurable functions, and

$$\int_{\Omega_1}\left(\int_{\Omega_2} f(\omega_1,\omega_2)\,\mu_2(d\omega_2)\right)\mu_1(d\omega_1)$$
$$=\int_{\Omega} f\,d\mu=\int_{\Omega_2}\left(\int_{\Omega_1} f(\omega_1,\omega_2)\,\mu_1(d\omega_1)\right)\mu_2(d\omega_2).$$

Finally, for any integrable function f on $(\Omega,\mathcal{F},\mu)$,

$$A_1=\left\{\omega_1:\int_{\Omega_2}|f(\omega_1,\omega_2)|\,\mu_2(d\omega_2)<\infty\right\}\in\mathcal{F}_1,$$

$$A_2=\left\{\omega_2:\int_{\Omega_1}|f(\omega_1,\omega_2)|\,\mu_1(d\omega_1)<\infty\right\}\in\mathcal{F}_2,$$

$$\omega_1\rightsquigarrow f_1(\omega_1)\equiv 1_{A_1}(\omega_1)\int_{\Omega_2} f(\omega_1,\omega_2)\,\mu_2(d\omega_2)$$

and

$$\omega_2\rightsquigarrow f_2(\omega_2)\equiv 1_{A_2}(\omega_2)\int_{\Omega_1} f(\omega_1,\omega_2)\,\mu_1(d\omega_1)$$

are integrable, and

$$\int_{\Omega_1} f_1(\omega_1)\,\mu_1(d\omega_1)=\int_{\Omega} f(\omega)\,\mu(d\omega)=\int_{\Omega_2} f_2(\omega_2)\,\mu_2(d\omega_2).$$

In the case when Ω_1 and Ω_2 are countable, all this becomes very easy. Namely, in the notation which we used in §6.1.6, the measure $\mu_1\times\mu_2$ corresponds to row vector $\boldsymbol{\mu}_1\times\boldsymbol{\mu}_2\in[0,\infty)^{\mathbb{S}_1\times\mathbb{S}_2}$ given by $(\boldsymbol{\mu}_1\times\boldsymbol{\mu}_2)_{(i_1,i_2)}=(\boldsymbol{\mu})_{i_1}(\boldsymbol{\mu})_{i_2}$, and so, by (6.1.14), Fubini's Theorem reduces to the statement that

$$\sum_{i_1\in\mathbb{S}_1}\left(\sum_{i_2\in\mathbb{S}_2} a_{i_1i_2}\right)=\sum_{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2} a_{i_1i_2}=\sum_{i_2\in\mathbb{S}_2}\left(\sum_{i_i\in\mathbb{S}_1} a_{i_1i_2}\right)$$

when $\{a_{i_1i_2}:(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2\}\subseteq(-\infty,\infty]$ satisfies either $a_{i_1i_2}\geq 0$ for all (i_1,i_2) or $\sum_{(i_1,i_2)}|a_{i_1i_2}|<\infty$. In proving this, we may and will assume that $\mathbb{S}_1=\mathbb{Z}^+=\mathbb{S}_2$ throughout and will start with the case when $a_{i_1i_2}\geq 0$. Given any pair $(n_1,n_2)\in(\mathbb{Z}^+)^2$,

$$\sum_{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2} a_{i_1i_2}\geq\sum_{\substack{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2\\ i_1\leq n_1\ \&\ i_2\leq n_2}} a_{i_1i_2}=\sum_{i_2=1}^{n_2}\left(\sum_{i_1=1}^{n_1} a_{i_1i_2}\right).$$

Hence, by first letting $n_1\to\infty$ and then letting $n_2\to\infty$, we arrive at

$$\sum_{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2} a_{i_1i_2}\geq\sum_{i_2\in\mathbb{S}_2}\left(\sum_{i_1\in\mathbb{S}_1} a_{i_1i_2}\right).$$

Similarly, for any $n \in \mathbb{Z}^+$,

$$\sum_{\substack{(i_1,i_2)\mathbb{S}_1\times\mathbb{S}_2\\ i_1\vee i_2\le n}} a_{i_1i_2} = \sum_{i_2=1}^{n}\left(\sum_{i_1=1}^{n} a_{i_1i_2}\right) \le \sum_{i_2\in\mathbb{S}_1}\left(\sum_{i_2\in\mathbb{S}_2} a_{i_1i_2}\right),$$

and so, after letting $n \to \infty$, we get the opposite inequality. Next, when $\sum_{(i_1,i_2)} |a_{i_1i_2}| < \infty$,

$$\sum_{i_1\in\mathbb{S}_1} |a_{i_1i_2}| < \infty \quad \text{for all } i_2 \in \mathbb{S}_2,$$

and so

$$\begin{aligned}\sum_{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2} a_{i_1i_2} &= \sum_{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2} a^+_{i_1i_2} - \sum_{(i_1,i_2)\in\mathbb{S}_1\times\mathbb{S}_2} a^-_{i_1i_2}\\ &= \sum_{i_2\in\mathbb{S}_2}\left(\sum_{i_1\in\mathbb{S}_1} a^+_{i_1i_2}\right) - \sum_{i_2\in\mathbb{S}_2}\left(\sum_{i_1\in\mathbb{S}_1} a^-_{i_1i_2}\right)\\ &= \lim_{n\to\infty} \sum_{\substack{i_2\in\mathbb{S}_2\\ i_2\le n}}\left(\sum_{i_1\in\mathbb{S}_1} a_{i_1i_2}\right) = \sum_{i_2\in\mathbb{S}_2}\left(\sum_{i_1\in\mathbb{S}_1} a_{i_1i_2}\right).\end{aligned}$$

Finally, after reversing the roles 1 and 2, we get the relation with the order of summation reversed.

6.2 Modeling Probability

To understand how these considerations relate to probability theory, think about the problem of modeling the tosses of a fair coin. When the game ends after the nth toss, a Kolmogorov model is provided by taking $\Omega = \{0,1\}^n$, $\mathcal{F}$ the set of all subsets of Ω, and setting $\mu(\{\omega\}) = 2^{-n}$ for each $\omega \in \Omega$. More generally, any measure space in which Ω has total measure 1 can be thought of as a model of probability, for which reason, such a measure space is called a *probability space*, and the measure μ is called a *probability measure* and is often denoted by $\mathbb{P}$. In this connection, when dealing with probability spaces, ones intuition is aided by extending the metaphor to other objects. Namely, one calls Ω the *sample space*, its elements are called *sample points*, the elements of $\mathcal{F}$ are called *events*, and the number that $\mathbb{P}$ assigns an event is called the *probability* of that event. In addition, a measurable map is called a *random variable*, it tends to be denoted by X instead of F, and, when it is $\mathbb{R}$-valued, its integral is called its *expected value*. Moreover, the latter convention is reflected in the use of $\mathbb{E}[X]$, or, when more precision is required, $\mathbb{E}^{\mathbb{P}}[X]$ to denote $\int X\, d\mathbb{P}$. Also, $\mathbb{E}[X, A]$ or $\mathbb{E}^{\mathbb{P}}[X, A]$ is used to denote $\int_A X\, d\mathbb{P}$, the expected value of X on the event A. Finally, the *distribution* of a random variable whose values lie in the measurable space $(E, \mathcal{B})$ is the probability measure $X_*\mathbb{P}$ on

$(E, \mathcal{B})$ given by $\mu(B) = \mathbb{P}(X \in B)$. In particular, when X is $\mathbb{R}$-valued, its *distribution function* F_X is defined so that $F_X(x) = (X_*\mathbb{P})\big((-\infty, x]\big) = \mathbb{P}(X \leq x)$. Obviously, F_X is non-increasing. Moreover, as an application of (6.1.3), one can show that F_X is continuous from the right, in the sense that $F_X(x) = \lim_{y \searrow x} F_X(y)$ for each $x \in \mathbb{R}$, $\lim_{x \searrow -\infty} F_X(x) = 0$, and $\lim_{x \nearrow \infty} F_X(x) = 1$. At the same time, one sees that $\mathbb{P}(X < x) = F_X(x-) \equiv \lim_{y \nearrow x} F_X(y)$, and, as a consequence, that $\mathbb{P}(X = x) = F_X(x) - F_X(x-)$ is the jump in F_X at x.

6.2.1. Modeling Infinitely Many Tosses of a Fair Coin: Just as in the general theory of measure spaces, the construction of probability spaces presents no analytic (as opposed to combinatorial) problems as long as the sample space is finite or countable. The only change from the general theory is that the assignment of probabilities to the sample points must satisfy the condition that $\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$. The technical analytic problems arise when Ω is uncountable. For example, suppose that, instead of stopping the game after n tosses, one thinks about a coin tossing of indefinite duration. Clearly, the sample space will now be $\Omega = \{0,1\}^{\mathbb{Z}^+}$. In addition, when $A \subseteq \Omega$ depends only on tosses 1 through n, then A should be a measurable event and the probability assigned to A should be the same as the probability that would have been assigned had the game stopped after the nth toss. That is, if $\Gamma \subseteq \{0,1\}^n$ and[8] $A = \{\omega \in \Omega : (\omega(1), \ldots, \omega(n)) \in \Gamma\}$, then $\mathbb{P}(A)$ should equal $2^{-n}\#\Gamma$.

Continuing with the example of an infinite coin tossing game, one sees (cf. (6.1.3)) that, for any fixed $\eta \in \Omega$, $\mathbb{P}(\{\eta\})$ is equal to

$$\lim_{n \to \infty} \mathbb{P}\big(\{\omega : \big(\omega(1), \ldots, \omega(n)\big) = \big(\eta(1), \ldots, \eta(n)\big)\}\big) = \lim_{n \to \infty} 2^{-n} = 0.$$

Hence, in this case, nothing is learned from the way in which probability is assigned to points: every sample points has probability 0. In fact, it is far from obvious that there exists a probability measure on Ω with the required properties. Nonetheless, as we will now prove such a measure does exist.

To get started, for each $n \geq 1$, let $\Pi_n : \Omega \longrightarrow \Omega_n \equiv \{0,1\}^n$ be the projection map given by $\omega \rightsquigarrow \big(\omega(1), \ldots, \omega(n)\big)$, and set

$$\mathcal{A}_n = \big\{A \subseteq \Omega : A = \Pi_n^{-1}\big(\Pi_n(A)\big)\big\}.$$

Equivalently, $A \in \mathcal{A}_n$ if and only if A depends on the first n coordinates, in the sense that: $\omega \in A$ and $\Pi_n(\omega') = \Pi_n(\omega) \implies \omega' \in A$. Next, define $\mathbb{P}_n$ on $\mathcal{A}_n$ so that $\mathbb{P}_n(A) = 2^{-n}\#\Pi_n(A)$. Clearly $\mathcal{A}_n \subseteq \mathcal{A}_{n+1}$. In fact, if $A \in \mathcal{A}_n$, then $\Pi_{n+1}(A) = \big(\Pi_n(A) \times \{0,1\}\big)$, and so $\mathbb{P}_{n+1}(A) = \mathbb{P}_n(A)$. Hence, we can unambiguously define $\mathbb{P}$ on $\mathcal{A} \equiv \bigcup_{n=1}^{\infty} \mathcal{A}_n$ so that $\mathbb{P}(A) = \mathbb{P}_n(A)$ when $A \in \mathcal{A}_n$. Moreover, if $A, A' \in \mathcal{A}$ are disjoint, then, by choosing n so that $A, A' \in \mathcal{A}_n$, we see that $\mathbb{P}(A \cup A') = \mathbb{P}_n(A \cup A') = \mathbb{P}_n(A) + \mathbb{P}_n(A') = \mathbb{P}(A) + \mathbb{P}(A')$.

[8] It is convenient here to identify Ω with the set a mappings ω from $\mathbb{Z}^+$ into $\{0,1\}$. Thus, we will use $\omega(n)$ to denote the "nth coordinate" of ω.

Before taking the next step, it will be convenient to introduce a metric on Ω for which all elements of $\mathcal{A}$ are open sets. Namely, we take

$$\rho(\omega, \omega') \equiv \sum_{n=1}^{\infty} 2^{-n} |\omega(n) - \omega'(n)|.$$

It is an easy matter to check that ρ is a metric Ω. In fact, it is a metric for which the notion of convergence corresponds to convergence of each coordinate. In addition, if $\omega \in A \in \mathcal{A}_n$, then $\rho(\omega', \omega) < 2^{-n} \implies \omega' \in A$. Hence, each $A \in \mathcal{A}$ is open relative to topology induced by ρ. At the same time, each $A \in \mathcal{A}$ is also closed in ρ-topology. Indeed, if $A \in \mathcal{A}_n$, $\{\omega_k\}_1^\infty \subseteq A$, and $\rho(\omega_k, \omega) \longrightarrow 0$, then there is a $k \in \mathbb{Z}^+$ for which $\rho(\omega, \omega_k) < 2^{-n}$, and so, for this k, $\Pi_n(\omega) = \Pi_n(\omega_k)$. Another fact which we will need is that Ω is compact in the ρ-topology. To see this, suppose that $\{\omega_k\}_1^\infty$ is a sequence in Ω. To produce a convergent subsequence, we employ a diagonalization procedure. That is, because, for each m, either $\omega_k(m) = 0$ for infinitely many $k \in \mathbb{Z}^+$ or $\omega_k(m) = 1$ for infinitely many $k \in \mathbb{Z}^+$, we can use induction to construct $\{k_{m,\ell} : (m, \ell) \in (\mathbb{Z}^+)^2\}$ so that $\{k_{1,\ell} : \ell \in \mathbb{Z}^+\}$ is an increasing enumeration of $\{k : \omega_k(1) = 0\}$ or of $\{k : \omega_k(1) = 1\}$ depending on whether $\omega_k(1) = 0$ for infinitely or finitely many k's, and, when $m > 1$, $\{k_{m,\ell} : \ell \in \mathbb{Z}^+\}$ is an increasing enumeration of $\{k_{m-1,\ell} : \omega_{k_{m-1,\ell}}(m) = 0\}$ or of $\{k_{m-1,\ell} : \omega_{k_{m-1,\ell}}(m) = 1\}$ depending on whether $\omega_{k_{m-1,\ell}}(m) = 0$ for infinitely or finitely many ℓ's. Now set $k_\ell = k_{\ell,\ell}$, and check that $\{\omega_{k_\ell} : \ell \in \mathbb{Z}^+\}$ a convergent subsequence of $\{\omega_k\}_1^\infty$. Finally, there is another property which we will need to know about. Namely, for each open set $G \subseteq \Omega$ there is a sequence $\{A_m\}_1^\infty \subseteq \mathcal{A}$ of mutually disjoint sets such that $A_m \in \mathcal{A}_m$ and $G = \bigcup_1^\infty A_m$. To produce $\{A_m\}_1^\infty$ one can proceed as follows. Choose A_1 to be the largest element of $A \in \mathcal{A}_1$ with the property that $A \subseteq G$. Equivalently, A_1 is the union of all the $A \in \mathcal{A}_1$ contained in G. Next, given A_ℓ for $1 \leq \ell \leq m$, choose A_{m+1} to the largest $A \in \mathcal{A}_{m+1}$ contained in $G \setminus \bigcup_1^m A_m$. Obviously, these A_m's are mutually disjoint. To see that they cover G, suppose that $\omega \in G$, and choose $n \geq 2$ so that $\omega' \in G$ whenever $\rho(\omega', \omega) < 2^{-n+1}$. Then $A \equiv \{\omega' : \Pi_n(\omega') = \Pi_n(\omega)\} \subseteq G$, and either $\omega \in \bigcup_1^{n-1} A_m$ or $A \cap \bigcup_1^{n-1} A_m = \emptyset$, in which case $\omega \in A \subseteq A_n$.

Having made these preparations, we can continue our construction. First, define $\bar{\mathbb{P}}(\Gamma)$ for $\Gamma \subseteq \Omega$ to the infemum of $\sum_1^\infty \mathbb{P}(A_m)$ over all countable covers $\{A_m\}_1^\infty \subseteq \mathcal{A}$ of Γ. Equivalently, since all elements of $\mathcal{A}$ are open and all open sets can be written as the union of countably many elements of $\mathcal{A}$,

$$\bar{\mathbb{P}}(\Gamma) = \inf\{\bar{\mathbb{P}}(G) : G \supseteq \Gamma \text{ and } G \text{ open}\}. \tag{6.2.1}$$

The following lemma contains several elementary facts about $\bar{\mathbb{P}}$.

6.2.2 LEMMA. *For each $A \in \mathcal{A}$, $\bar{\mathbb{P}}(A) = \mathbb{P}(A)$. More generally, for all $\Gamma_1 \subseteq \Gamma_2 \subseteq \Omega$, $\bar{\mathbb{P}}(\Gamma_1 \leq \bar{\mathbb{P}}(\Gamma_2)$. Moreover, for any open $G \subseteq \Omega$ and any countable, exact cover $\{A_m\}_1^\infty$ of G by mutually disjoint elements of $\mathcal{A}$, $\bar{\mathbb{P}}(G) = \sum_1^\infty \mathbb{P}(A_m)$,*

and, in general, for any sequence $\{\Gamma_k\}_1^\infty$ *of subsets of* Ω,

$$\bar{\mathbb{P}}\left(\bigcup_1^\infty \Gamma_k\right) \leq \sum_1^\infty \bar{\mathbb{P}}(\Gamma_k). \tag{6.2.3}$$

Finally, if F *and* F' *are disjoint closed subsets of* Ω, *then* $\bar{\mathbb{P}}(F \cup F') = \bar{\mathbb{P}}(F) + \bar{\mathbb{P}}(F')$.

PROOF: The second assertion is obvious, since any cover of Γ_2 is also a cover of Γ_1. Next, suppose that $A \in \mathcal{A}$. Obviously, $\bar{\mathbb{P}}(A) \leq \mathbb{P}(A)$. On the other hand, if $\{A_m\}_1^\infty$ is a cover of A by elements of $\mathcal{A}$, then, by the Heine–Borel property of compact sets, we can find (remember that A is closed and therefore compact) an n such that $A \subseteq \bigcup_1^n A_m$. Now choose N so that $\{A\} \cup \{A_m\}_1^n \subseteq \mathcal{A}_N$. Then

$$\mathbb{P}(A) = \mathbb{P}_N(A) \leq \sum_1^n \mathbb{P}_N(A_m) = \sum_1^n \mathbb{P}(A_m) \leq \sum_1^\infty \mathbb{P}(A_m),$$

and so we can conclude that $\mathbb{P}(A) \leq \bar{\mathbb{P}}(A)$.

Next let G be open, and suppose that $\{A_m\}_1^\infty$ is an exact cover of G by mutually disjoint elements of $\mathcal{A}$. By definition, $\bar{\mathbb{P}}(G) \leq \sum_1^\infty \mathbb{P}(A_m)$. On the other hand, it is clear that, because $G \supseteq \bigcup_1^n A_m \in \mathcal{A}$, we know that

$$\begin{aligned}\bar{\mathbb{P}}(G) &\geq \bar{\mathbb{P}}\left(\bigcup_1^n A_m\right) = \mathbb{P}\left(\bigcup_1^n A_m\right) \\ &= \mathbb{P}_N\left(\bigcup_1^n A_m\right) = \sum_1^n \mathbb{P}_N(A_m) = \sum_1^n \mathbb{P}(A_m),\end{aligned}$$

where N is chosen so that $\{A_m\}_1^n \subseteq \mathcal{A}_N$. Hence, the desired conclusion follows after one lets $n \to \infty$.

Now suppose that $\{\Gamma_k\}_1^\infty$ is a sequence of subsets of Ω. Given $\epsilon > 0$, choose for each $k \in \mathbb{Z}^+$ a countable cover $\{A_{k,\ell} : \ell \in \mathbb{Z}^+\} \subseteq \mathcal{A}$ of Γ_k so that $\sum_\ell \mathbb{P}(A_{k,\ell}) \leq \bar{P}(\Gamma_k) + 2^{-k}\epsilon$. Then $\{A_{k,\ell} : (k,\ell) \in (\mathbb{Z}^+)^2\} \subseteq \mathcal{A}$ is a countable cover of $\bigcup_k \Gamma_k$, and so (by Fubini's Theorem for countable measure spaces)

$$\bar{\mathbb{P}}\left(\bigcup_k \Gamma_k\right) \leq \sum_{(k,\ell)\in(\mathbb{Z}^+)^2} \mathbb{P}(A_{k,\ell}) = \sum_{k\in\mathbb{Z}^+} \left(\sum_{\ell\in\mathbb{Z}^+} \mathbb{P}(A_{k,\ell})\right) \leq \sum_{k\in\mathbb{Z}^+} \bar{\mathbb{P}}(\Gamma_k) + \epsilon.$$

Thus, $\bar{\mathbb{P}}\left(\bigcup_k \Gamma_k\right) \leq \sum_k \bar{\mathbb{P}}(\Gamma_k)$.

Finally, given disjoint, closed subsets F and F' of Ω, the preceding implies that $\bar{\mathbb{P}}(F \cup F') \leq \bar{\mathbb{P}}(F) + \bar{\mathbb{P}}(F')$. To get the opposite inequality, first note that, because they are compact, there is an $N \geq 2$ such that $\rho(\omega, \omega') > 2^{-N+1}$ for all $\omega \in F$ and $\omega' \in F'$. Hence, if $B \equiv \Pi_N^{-1}\big(\Pi_N(F)\big)$ and $B' \equiv \Pi_N^{-1}\big(\Pi_N(F')\big)$, then $F \subseteq B \in \mathcal{A}$, $F' \subseteq B' \in \mathcal{A}$, and $B \cap B' = \emptyset$, since $\eta \in B \cap B'$ would imply there

exist $\omega \in F$ and $\omega' \in F'$ such that $\rho(\omega, \omega') \leq \rho(\omega, \eta) + \rho(\eta, \omega') \leq 2^{-N+1}$. Now suppose that $\{A_m\}_1^\infty \subseteq \mathcal{A}$ is a countable cover of $F \cup F'$, and set $B_m = B \cap A_m$ and $B'_m = B' \cap A'_m$. Then $\{B_m\}_1^\infty \subseteq \mathcal{A}$ and $\{B'_m\}_1^\infty \subseteq \mathcal{A}$ are countable covers of F and F'. In addition, because B_m and B'_m are disjoint elements of $\mathcal{A}$, $\mathbb{P}(A_m) \geq \mathbb{P}(B_m \cup B'_m) = \mathbb{P}(B_m) + \mathbb{P}(B'_m)$. Hence,

$$\sum_m \mathbb{P}(A_m) \geq \sum_m \mathbb{P}(B_m) + \sum_m \mathbb{P}(B'_m) \geq \bar{\mathbb{P}}(F) + \bar{\mathbb{P}}(F'). \quad \square$$

The final ingredient in our construction is the specification of which subsets of Ω are to be measurable. Although it may not be immediately apparent why, we will say that $\Gamma \subseteq \Omega$ is measurable and will write $\Gamma \in \bar{\mathcal{B}}$ if, for each $\epsilon > 0$, there exists an open $G \supseteq \Gamma$ such that $\bar{\mathbb{P}}(G \setminus \Gamma) \leq \epsilon$. Obviously, every open set is measurable. At the same time, it should be clear that $\Gamma \in \bar{\mathcal{B}}$ if $\bar{\mathbb{P}}(\Gamma) = 0$. Indeed, by (6.2.1), if $\bar{\mathbb{P}}(\Gamma) = 0$, then, for each $\epsilon > 0$ we can find a open $G \supseteq \Gamma$ such that $\bar{\mathbb{P}}(G \setminus \Gamma) \leq \bar{\mathbb{P}}(G) \leq \epsilon$. However, it is less obvious that every closed set $F \subseteq \Omega$ is measurable. To see this, let $\epsilon > 0$ be given, and choose an open set $G \supseteq F$ so that $\bar{\mathbb{P}}(G) \leq \bar{\mathbb{P}}(F) + \epsilon$. Then $G \setminus F$ is open, and so we can write $G \setminus F = \bigcup_m A_m$, where the A_m's are mutually disjoint elements of $\mathcal{A}$. Hence, by Lemma 6.2.2, $\bar{\mathbb{P}}(G \setminus F) = \sum_1^\infty \mathbb{P}(A_m)$. On the other hand, for each $n \geq 1$, $B_n \equiv \bigcup_1^n A_m$ is a closed subset of $G \setminus F$, and so, by the first and last parts of Lemma 6.2.2,

$$\sum_1^n \mathbb{P}(A_m) = \mathbb{P}(B_n) = \bar{\mathbb{P}}(B_n) = \bar{\mathbb{P}}(F \cup B_n) - \bar{\mathbb{P}}(F) \leq \bar{\mathbb{P}}(G) - \bar{\mathbb{P}}(F) \leq \epsilon.$$

Thus, $\bar{\mathbb{P}}(G \setminus F) = \sum_1^\infty \mathbb{P}(A_m) \leq \epsilon$.

Now that we know open sets, closed sets, and sets of $\bar{\mathbb{P}}$-measure 0 are all measurable, we can show that $\bar{\mathcal{B}}$ is a σ-algebra. To this end, first suppose that $\{\Gamma_k\}_1^\infty \subseteq \bar{\mathcal{B}}$, and, given $\epsilon > 0$, choose open sets $G_k \supseteq \Gamma_k$ so that $\bar{\mathbb{P}}(G_k \setminus \Gamma_k) \leq 2^{-k}\epsilon$. Then $G \equiv \bigcup_1^\infty G_k$ is open, $G \supseteq \Gamma \equiv \bigcup_1^\infty \Gamma_k$, and

$$\bar{\mathbb{P}}(G \setminus \Gamma) \leq \bar{\mathbb{P}}\left(\bigcup_1^\infty (G_k \setminus \Gamma_k)\right) \leq \sum_1^\infty \mathbb{P}(G_k \setminus \Gamma_k) \leq \epsilon.$$

Hence, $\Gamma \in \bar{\mathcal{B}}$, and so $\mathcal{B}$ is closed under countable unions. To see that it is also closed under complementation, let $\Gamma \in \bar{\mathcal{B}}$ be given, and, for each $n \geq 1$, choose an open set $G_n \supseteq \Gamma$ so that $\bar{\mathbb{P}}(G_n \setminus \Gamma) \leq \frac{1}{n}$. Then $D \equiv \bigcap_1^\infty G_n \supseteq \Gamma$ and $\bar{\mathbb{P}}(D \setminus \Gamma) \leq \bar{\mathbb{P}}(G_n \setminus \Gamma) \leq \frac{1}{n}$ for all $n \geq 1$. Thus, $\bar{\mathbb{P}}(D \setminus \Gamma) = 0$, and so $D \setminus \Gamma \in \bar{\mathcal{B}}$. Now set $F_n = G_n{}^{\complement}$ and $C = \bigcup_1^\infty F_n$. Because each F_n is closed, and therefore measurable, $C \in \bar{\mathcal{B}}$. Hence, since $\Gamma^{\complement} = C \cup (D \setminus \Gamma) \in \bar{\mathcal{B}}$, we are done.

6.2.4 THEOREM. *Referring to the preceding, $\mathcal{B}_\Omega = \sigma(\mathcal{A})$ and $\Gamma \in \bar{\mathcal{B}}$ if and only if there exist $C, D \in \mathcal{B}_\Omega$ such that $C \subseteq \Gamma \subseteq D$ and $\bar{\mathbb{P}}(D \setminus C) = 0$. Next, again use $\mathbb{P}$ to denote restriction $\bar{\mathbb{P}} \restriction \bar{\mathcal{B}}$ of $\bar{\mathbb{P}}$ to $\bar{\mathcal{B}}$. Then, $(\Omega, \bar{\mathcal{B}}, \mathbb{P})$ is a probability space with the property that $\mathbb{P} \restriction \mathcal{A}_n = \mathbb{P}_n$ for every $n \geq 1$.*

PROOF: We have already established that $\bar{\mathcal{B}}$ is a σ-algebra. Since all elements of $\mathcal{A}$ are open, the equality $\mathcal{B}_\Omega = \sigma(\mathcal{A})$ follows from the fact that every open set can be written as the countable union of elements of $\mathcal{A}$. To prove the characterization of $\bar{\mathcal{B}}$ in terms of $\mathcal{B}_\Omega$, first observe that, since open sets are in $\bar{\mathcal{B}}$, $\mathcal{B}_\Omega \subseteq \bar{\mathcal{B}}$. Moreover, since sets of $\bar{\mathbb{P}}$-measure 0 are also in $\bar{\mathcal{B}}$, $C \subseteq \Gamma \subseteq D$ for $C, D \in \mathcal{B}_\Omega$ with $\bar{\mathbb{P}}(D \setminus C) = 0$ implies that $\Gamma = C \cup (\Gamma \setminus C) \in \bar{\mathcal{B}}$, since $\bar{\mathbb{P}}(\Gamma \setminus C) \leq \bar{\mathbb{P}}(D \setminus C) = 0$. Conversely, if $\Gamma \in \bar{\mathcal{B}}$, then we can find sequences $\{G_n\}_1^\infty$ and $\{H_n\}_1^\infty$ of open sets such that $G_n \supseteq \Gamma$, $H_n \supseteq \Gamma\complement$, and $\bar{\mathbb{P}}(G_n \setminus \Gamma) \vee \bar{\mathbb{P}}(H_n \setminus \Gamma\complement) \leq \frac{1}{n}$. Thus, if $D = \bigcap_1^\infty G_n$ and $C = \bigcup_1^\infty H_n\complement$, then $C, D \in \mathcal{B}_\Omega$, $C \subseteq \Gamma \subseteq D$, and

$$\bar{\mathbb{P}}(D \setminus C) \leq \bar{\mathbb{P}}(G_n \setminus H_n\complement) \leq \bar{\mathbb{P}}(G_n \setminus \Gamma) + \bar{\mathbb{P}}(H_n \setminus \Gamma\complement) \leq \frac{2}{n}$$

for all $n \geq 1$. Hence, $\bar{\mathbb{P}}(D \setminus C) = 0$.

All that remains is to check that $\bar{\mathbb{P}}$ is countably additive on $\bar{\mathcal{B}}$. To this end, let $\{\Gamma_k\}_1^\infty \subseteq \bar{\mathcal{B}}$ be a sequence of mutually disjoint sets, and set $\Gamma = \bigcup_1^\infty \Gamma_k$. By the second part of Lemma 6.2.2, we know that $\bar{\mathbb{P}}(\Gamma) \leq \sum_1^\infty \bar{\mathbb{P}}(\Gamma_k)$. To get the opposite inequality, let $\epsilon > 0$ be given, and, for each k, choose an open $G_k \supseteq \Gamma_k\complement$ so that $\bar{\mathbb{P}}(G_k \setminus \Gamma_k\complement) \leq 2^{-k}\epsilon$. Then each $F_k \equiv G_k\complement$ is closed, and

$$\bar{\mathbb{P}}(\Gamma_k) \leq \bar{\mathbb{P}}(F_k) + \bar{\mathbb{P}}(G_k \setminus \Gamma_k\complement) \leq \mathbb{P}(F_k) + 2^{-k}\epsilon.$$

In addition, because the Γ_k's are, the F_k's are mutually disjoint. Hence, by the last part of Lemma 6.2.2, we know first that

$$\bar{\mathbb{P}}(\Gamma) \geq \bar{\mathbb{P}}\left(\bigcup_1^n F_k\right) = \sum_1^n \bar{\mathbb{P}}(F_k) \quad \text{for all } n \geq 1,$$

and then that

$$\sum_1^\infty \bar{\mathbb{P}}(\Gamma_k) \leq \sum_1^\infty \bar{\mathbb{P}}(F_k) + \epsilon \leq \bar{\mathbb{P}}(\Gamma) + \epsilon. \quad \square$$

6.3 Independent Random Variables

In Kolmogorov's model, independence is best described in terms of σ-algebras. Namely, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{F}_1$ and $\mathcal{F}_2$ are σ-subalgebras (i.e., are σ-algebras which are subsets) of $\mathcal{F}$, then we say that $\mathcal{F}_1$ and $\mathcal{F}_2$ are *independent* if $\mathbb{P}(\Gamma_1 \cap \Gamma_2) = \mathbb{P}(\Gamma_1)\mathbb{P}(\Gamma_2)$ for all $\Gamma_1 \in \mathcal{F}_1$ and $\Gamma_2 \in \mathcal{F}_2$. It should be comforting to recognize that, when $A_1, A_2 \in \mathcal{F}$ and, for $i \in \{1, 2\}$, $\mathcal{F}_i = \sigma(\{A_i\}) = \{\emptyset, A_i, A_i\complement, \Omega\}$, then, as is easily checked, $\mathcal{F}_1$ is independent of $\mathcal{F}_2$ precisely when, in the terminology of elementary probability theory, "A_1 is independent of A_2": $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$.

The notion of independence gets inherited by random variables. Namely, the members of a collection $\{X_\alpha : \alpha \in \mathcal{I}\}$ of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$

are said to be mutually independent if, for each pair of disjoint subsets $\mathcal{J}_1$ and $\mathcal{J}_2$ of $\mathcal{I}$, the σ-algebras $\sigma\big(\{X_\alpha : \alpha \in \mathcal{J}_1\}\big)$ and $\sigma\big(\{X_\alpha : \alpha \in \mathcal{J}_2\}\big)$ are independent. One can use Theorem 6.1.6 to show that this definition is equivalent to saying that if X_α takes its values in the measurable space $(E_\alpha, \mathcal{B}_\alpha)$, then, for every finite subset $\{\alpha_m\}_1^n$ of distinct elements of $\mathcal{I}$ and choice of $B_{\alpha_m} \in \mathcal{B}_{\alpha_m}$, $1 \le m \le n$,

$$\mathbb{P}\big(X_{\alpha_m} \in B_{\alpha_m} \text{ for } 1 \le m \le n\big) = \prod_1^n \mathbb{P}\big(X_{\alpha_m} \in B_{\alpha_m}\big).$$

As a dividend of this definition, it is essentially obvious that if $\{X_\alpha : \alpha \in \mathcal{I}\}$ are mutually independent and if, for each $\alpha \in \mathcal{I}$, F_α is a measurable map on the range of X_α, then $\{F_\alpha(X_\alpha) : \alpha \in \mathcal{I}\}$ are again mutually independent. Finally, by starting with simple functions, one can show that if $\{X_m\}_1^n$ are mutually independent and, for each $1 \le m \le n$, f_m is a measurable $\mathbb{R}$-valued function on the range of X_m, then

$$\mathbb{E}\big[f_1(X_1)\cdots f_n(X_n)\big] = \prod_1^n \mathbb{E}\big[f_m(X_m)\big]$$

whenever the f_m are all bounded or are all non-negative.

6.3.1. Existence of Lots of Independent Random Variables: In the preceding section, we constructed a countable family of mutually independent random variables. Namely, if $(\Omega, \bar{\mathcal{B}}, \mathbb{P})$ is the probability space discussed in Theorem 6.2.4 and $X_m(\omega) = \omega(m)$ is the mth coordinate of ω, then, for any choice of $n \ge 1$ and $(\eta_1, \dots, \eta_n) \in \{0,1\}$, $\mathbb{P}(X_m = \eta_m,\ 1 \le m \le n) = 2^{-n} = \prod_1^n \mathbb{P}(X_m = \eta_m)$. Thus, the random variables $\{X_m\}_1^\infty$ are mutually independent. Mutually independent random variables, like these, which take on only two values are called *Bernoulli random variables.*

As we are about to see, Bernoulli random variables can be used as building blocks to construct many other families of mutually independent random variables. The key to such constructions is contained in the following lemma.

6.3.1 LEMMA. *Given any family $\{B_m\}_1^\infty$ of mutually independent, $\{0,1\}$-valued Bernoulli random variables satisfying $\mathbb{P}(B_m = 0) = \frac{1}{2} = \mathbb{P}(B_m = 1)$ for all $m \in \mathbb{Z}^+$, set $U = \sum_1^\infty 2^{-m} B_m$. Then U is uniformly distributed on $[0,1)$. That is, $\mathbb{P}(U \le u)$ is 0 if $u < 0$, u if $u \in [0,1)$, and 1 if $u \ge 1$.*

PROOF: Given $N \ge 1$ and $0 \le n < 2^N$, we want to show that

$$(*) \qquad \mathbb{P}\big(n2^{-N} < U \le (n+1)2^{-N}\big) = 2^{-N}.$$

To this end, note that $n2^{-N} < U \le (n+1)2^{-N}$ if and only if $\sum_1^N 2^{-m} B_m = (n+1)2^{-N}$ and $B_m = 0$ for $m > N$, or $\sum_1^N 2^{-m} B_m = n2^{-N}$ and $B_m = 1$ for some $m > N$. Hence, since $\mathbb{P}(B_m = 0 \text{ for all } m > N) = 0$, the left hand

side of (*) is equal to the probability that $\sum_1^N 2^{-m}B_m = n2^{-N}$. However, elementary considerations show that, for any $0 \le n < 2^N$ there is exactly one choice of $(\eta_1, \dots, \eta_N) \in \{0,1\}^N$ for which $\sum_1^N 2^{-m}\eta_m = n2^{-N}$. Hence,

$$\mathbb{P}\left(\sum_1^N 2^{-m}B_m = n2^{-N}\right) = \mathbb{P}\big(B_m = \eta_m \text{ for } 1 \le m \le N\big) = 2^{-N}.$$

Having proved (*), the rest is easy. Namely, since $\mathbb{P}(U = 0) = \mathbb{P}(B_m = 0 \text{ for all } m \ge 1) = 0$, (*) tells us that, for any $1 \le k \le 2^N$

$$\mathbb{P}(U \le k2^{-N}) = \sum_{m=0}^{k-1} \mathbb{P}\big(m2^{-N} < U \le (m+1)2^{-N}\big) = k2^{-N}.$$

Hence, because $u \rightsquigarrow \mathbb{P}(U \le u)$ is continuous from the right, it is now clear that $F_U(u) = u$ for all $u \in [0,1)$. Finally, since $\mathbb{P}(U \in [0,1]) = 1$, this completes the proof. □

Now let $\mathcal{I}$ a non-empty, finite or countably infinite set. Then $\mathcal{I} \times \mathbb{Z}^+$ is countable, and so we can construct a 1-to-1 map $(\alpha, n) \rightsquigarrow N(\alpha, n)$ from $\mathcal{I} \times \mathbb{Z}^+$ onto $\mathbb{Z}^+$. Next, for each $\alpha \in \mathcal{I}$, define $\omega \in \Omega = \{0,1\}^{\mathbb{Z}^+} \longmapsto \mathbf{X}_\alpha(\omega) \in \Omega$ so that the nth coordinate of $\mathbf{X}_\alpha(\omega)$ is $\omega\big(N(\alpha, n)\big)$. Then, as random variables on the probability space $(\Omega, \bar{\mathcal{B}}, \mathbb{P})$ in Theorem 6.2.4, $\{\mathbf{X}_\alpha : \alpha \in \mathcal{I}\}$ are mutually independent and each has distribution $\mathbb{P}$. Hence, if $\Phi : \Omega \longrightarrow [0,1]$ is the continuous map given by

$$\Phi(\eta) \equiv \sum_{m=1}^{\infty} 2^{-m}\eta(m) \quad \text{for } \eta \in \Omega$$

and if $U_\alpha \equiv \Phi(\mathbf{X}_\alpha)$, then the random variables $\{U_\alpha \equiv \Phi(\mathbf{X}_\alpha) : \alpha \in \mathcal{I}\}$ are mutually independent and, by Lemma 6.3.1, each is uniformly distributed on $[0,1]$.

The final step in this section combines the preceding construction with the well-known fact that any $\mathbb{R}$-valued random variable can be represented in terms of a uniform random variable. More precisely, a map $F : \mathbb{R} \longrightarrow [0,1]$ is called a *distribution function* if F is non-decreasing, continuous from the right, and tends to 0 at $-\infty$ and 1 at $+\infty$. Given such an F, define

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \ge u\} \quad \text{for } u \in [0,1].$$

Notice that, by right continuity, $F(x) \ge u \iff F^{-1}(u) \le x$. Hence, if U is uniformly distributed on $[0,1]$, then $F^{-1}(U)$ is a random variable whose distribution function is F.

6.3.2 THEOREM. *Let $(\Omega, \bar{\mathcal{B}}, \mathbb{P})$ be the probability space in Theorem 6.2.4. Given any finite or countably infinite index set $\mathcal{I}$ and a collection $\{F_\alpha : \alpha \in \mathcal{I}\}$ of distribution functions, there exist mutually independent random variables $\{X_\alpha : \alpha \in \mathcal{I}\}$ on $(\Omega, \bar{\mathcal{B}}, \mathbb{P})$ with the property that, for each $\alpha \in \mathcal{I}$, F_α is the distribution of X_α.*

6.4 Conditional Probabilities and Expectations

Just as they are to independence, σ-algebras are central to Kolmogorov's definition of conditioning. Namely, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a sub-$\sigma$-algebra Σ, and a random variable X which is non-negative or integrable, Kolmogorov says that the random variable X_Σ is a *conditional expectation* of X given Σ if X_Σ is a non-negative random variable which is measurable with respect to Σ (i.e., $\sigma(\{X\}) \subseteq \Sigma$) and satisfies

$$\mathbb{E}[X_\Sigma, \Gamma] = \mathbb{E}[X, \Gamma] \quad \text{for all } \Gamma \in \Sigma. \tag{6.4.1}$$

When X is the indicator function of a set $B \in \mathcal{F}$, then the term conditional expectation is replaced to *conditional probability.*

To understand that this definition is an extension of the one given in elementary probability courses, begin by considering the case when Σ is the trivial σ-algebra $\{\emptyset, \Omega\}$. Because only constant random variables are measurable with respect to $\{\emptyset, \Omega\}$, it is clear that the one an only conditional expectation of X will be $\mathbb{E}[X]$. Next, suppose that $\Sigma = \sigma(\{A\}) = \{\emptyset, A, A^\complement, \Omega\}$ for some $A \in \mathcal{F}$ with $\mathbb{P}(A) \in (0,1)$. In this situation, it is an easy matter to check that, for any $B \in \mathcal{F}$,

$$\omega \rightsquigarrow \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \mathbf{1}_A(\omega) + \frac{\mathbb{P}(B \cap A^\complement)}{\mathbb{P}(A^\complement)} \mathbf{1}_{A^\complement}(\omega)$$

is a conditional probability of B given Σ. That is, the quantity $\mathbb{P}(B|A) \equiv \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$, which in elementary probability theory would be called "the conditional probability of B given A," appears here as the value on A of the map $\omega \rightsquigarrow \mathbb{P}(B|\Sigma)(\omega)$. More generally, if Σ is generated by a finite or countable partition $\mathcal{P} \subseteq \mathcal{F}$ of Ω, then, for any non-negative or integrable random variable X,

$$\omega \rightsquigarrow \sum_{\{A \in \mathcal{P} : \mathbb{P}(A) > 0\}} \frac{\mathbb{E}[X, A]}{\mathbb{P}(A)} \mathbf{1}_A(\omega)$$

will be a conditional expectation of X given Σ.

Of course, Kolmogorov's definition brings up two essential questions: existence and uniqueness. A proof of existence in general can be done in any one of many ways. For instance, when $\mathbb{E}[X^2] < \infty$, one can easily see that, just as $\mathbb{E}[X]\mathbf{1}$ is the minimum value of $\mathbb{E}[(X - X')^2]$ among all constant random variables X', so X_Σ will have to be the minimum of $\mathbb{E}[(X - X')^2]$ among all Σ-measurable random variables X'. In this way, the problem of existence can be related to a problem of orthogonal projection in the space of all square-integrable random variables, and, although they are outside the scope of this book, such projection results are familiar to anyone who has studied the theory Hilbert spaces.

Uniqueness, on the other hand, is both easier and more subtle than existence. Namely, there is no naïve uniqueness statement here, because, in

general, there will be uncountably many ways to take X_Σ.[9] On the other hand, every choice differs from any other choice on a set of measure at most 0. To see this, suppose that X'_Σ is a second non-negative random variable which satisfies (6.4.1). Then $A = \{X'_\Sigma > X_\Sigma\} \in \Sigma$, and so the only way that (6.4.1) can hold is if $\mathbb{P}(A) = 0$. Similarly, $\mathbb{P}(X_\Sigma > X'_\Sigma) = 0$, and so $\mathbb{P}(X_\Sigma \neq X'_\Sigma) = 0$.

In spite of the ambiguity caused by the sort of uniqueness problems just discussed, it is common to ignore, in so far as possible, this ambiguity and proceed as if a random variable possesses only one conditional expectation with respect to a given σ-algebra. In this connection, the standard notation for a conditional expectation of X given Σ is $\mathbb{E}[X|\Sigma]$ or, when $X = \mathbf{1}_B$, $\mathbb{P}(B|\Sigma)$, which is the notation which we adopted in the earlier chapters of this book.

6.4.1. Conditioning with Respect to Random Variables: In this book, essentially all conditioning is done when $\Sigma = \sigma(\mathfrak{F})$ (cf. § 6.1.4) for some family $\mathfrak{F}$ of measurable functions on $(\Omega, \mathcal{F}, \mathbb{P})$. When Σ has this form, the conditional expectation of a random variable X will be a measurable functions of the functions in $\mathfrak{F}$. For example, if $\mathfrak{F} = \{F_1, \dots, F_n\}$ and the functions F_m all take their values in a countable space $\mathbb{S}$, a conditional expectation value $\mathbb{E}\big[X \,\big|\, \sigma(\mathfrak{F})\big]$ of X given $\sigma(\mathfrak{F})$ has the form $\Phi(F_1, \dots, F_n)$, where $\Phi(i_1, \dots, i_n)$ is equal to

$$\frac{\mathbb{E}[X,\, F_1 = i_1, \dots, F_n = i_n]}{\mathbb{P}(F_1 = i_1, \dots, F_n = i_n)} \quad \text{or} \quad 0$$

according to whether $\mathbb{P}(F_1 = i_1, \dots, F_n = i_n)$ is positive or 0. In order to emphasize that conditioning with respect to $\sigma(\mathfrak{F})$ results in a function of $\mathfrak{F}$, we use the notation $\mathbb{E}[X|\mathfrak{F}]$ or $\mathbb{P}(B|\mathfrak{F})$ instead of $\mathbb{E}[X|\sigma(\mathfrak{F})]$ or $\mathbb{P}\big(B|\sigma(\mathfrak{F})\big)$.

To give more concrete examples of what we are talking about, first suppose that X and Y are independent random variables with values in some countable space $\mathbb{S}$, and set $Z = F(X, Y)$, where $F : \mathbb{S}^2 \longrightarrow \mathbb{R}$ is bounded. Then

$$\mathbb{E}[Z|X] = v(X) \quad \text{where } v(i) = \mathbb{E}[F(i, Y)] \text{ for } i \in \mathbb{S}. \tag{6.4.2}$$

A less trivial example is provided by our discussion of Markov chains. In Chapter 2, we encoded the Markov property in equations like

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

which displays this conditioning as a function, namely $(i_0, \dots, i_n) \rightsquigarrow (\mathbf{P})_{i_n j}$, of the random variables in terms of which the condition is made. (Of course, the distinguishing feature of the Markov property is that the function depends only on i_n and not $(i_0, \dots, i_{n-1})$.) Similarly, when we discussed Markov processes with a continuous time parameter, we wrote

$$\mathbb{P}\big(X(t) = j \,\big|\, X(\sigma),\, \sigma \in [0, s]\big) = \big(\mathbf{P}(t-s)\big)_{X(s)j},$$

which again makes it explicit that the conditioned quantity is a function random variables on which the condition is imposed.

[9] This non-uniqueness is the reason for our use of the article "a" instead of "the" in front of "conditional expectation."

Notation

Notation	Description	See
Z & Z^+	set of all integers and the subset of positive integers	
$\mathbb{N}$	set of non-negative integers	
$\#S$	number of elements in the set S	
$A\complement$	complement of the set A	
$\mathbf{1}_A$	indicator function of the set A: $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ if $x \notin A$	
$F \restriction S$	restriction of the function F to the set S	
$a \wedge b$ & $a \vee b$	minimum and the maximum of a, $b \in \mathbb{R}$	
a^+ & a^-	positive part $a \vee 0$ & negative part $(-a) \vee 0$ of $a \in \mathbb{R}$	
$i \to j$ & $i \leftrightarrow j$	state j is accessible from i & communicates with i	§3.1
δ_{ij}	Kronecker delta: δ_{ij} is 1 or 0 depending on whether i is equal or unequal to j	
$\boldsymbol{\delta}_i$	point measure at $i \in \mathbb{S}$: $(\boldsymbol{\delta}_i)_j = \delta_{ij}$ for $j \in \mathbb{S}$	
$\mathbb{E}[X,\, A]$	expected value of X on the event A	§ 6.2
$\mathbb{E}[X \mid A]$ & $\mathbb{E}[X \mid \Sigma]$	conditional expectatation value of X given the event A & the σ-algebra Σ	§6.4
$\langle \varphi \rangle_{\boldsymbol{\pi}}$	alternative notations for $\sum_{i \in \mathbb{S}} \varphi(i)(\boldsymbol{\pi})_i$	(5.1.4)
$\langle \varphi, \psi \rangle_{\boldsymbol{\pi}}$ & $\|\varphi, \psi\|_{2,\boldsymbol{\pi}}$	inner product and norm in $L^2(\boldsymbol{\pi})$	§5.1.2
$\mathrm{Stat}(\mathbf{P})$	set of stationary distribution for the transition probability matrix $\mathbf{P}$	§3.2.3
$\|\boldsymbol{\mu}\|_{\mathrm{v}}$	variation norm of the row vector $\boldsymbol{\mu}$	(2.1.5)
$\|f\|_{\mathrm{u}}$	uniform norm of the function f	(2.1.10)
$\|\mathbf{M}\|_{\mathrm{u,v}}$	uniform-variation norm of the matrix $\mathbf{M}$	(3.2.1)
$\mathrm{Var}_{\boldsymbol{\mu}}(f)$	variance of f relative to the probability vector $\boldsymbol{\mu}$	

References

1. Diaconis, P. & Stroock, D., *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab. **1 #1** (1991), 36–61.
2. Dunford, N. & Schwartz, J., *Linear Operators, part I*, Wiley Classics Lib., Wiley–Interscience, NY, 1988.
3. Holley, R. & Stroock, D., *Simulated annealing via Sobolev inequalities*, Comm. Math. Phys. **115 #4** (1988), 553–569.
4. Karlin, S. & Taylor, H, *A First Course in Stochastic Processes, 2nd ed.*, Academic Press, NY, 1975.
5. Norris, J.R., *Markov Chains*, Cambridge Series in Statistical & Probabilistic Mathematics, Cambridge Univ. Press, Cambridge, U.K., 1997.
6. Revuz, D., *Markov Chains*, Mathematical Library, vol. 11, North Holland, Amsterdam & New York, 1984.
7. Riesz, F. & Sz.-Nagy, B., *Functional Analysis*, translated from the French edition by F. Boron, reprint of 1955 original, Dover Books, NY, 1990.
8. Stroock, D., *A Concise Introduction to the Theory of Integration, 3rd ed.*, Birkhäuser–Boston, Cambridge, MA, USA, 1998.
9. Stroock, D., *Probability Theory, An Analytic View, 2nd ed.*, Cambridge Univ. Press, NY, 2000.

Index

(continued from page ii)

64 EDWARDS. Fourier Series. Vol. I. 2nd ed.
65 WELLS. Differential Analysis on Complex Manifolds. 2nd ed.
66 WATERHOUSE. Introduction to Affine Group Schemes.
67 SERRE. Local Fields.
68 WEIDMANN. Linear Operators in Hilbert Spaces.
69 LANG. Cyclotomic Fields II.
70 MASSEY. Singular Homology Theory.
71 FARKAS/KRA. Riemann Surfaces. 2nd ed.
72 STILLWELL. Classical Topology and Combinatorial Group Theory. 2nd ed.
73 HUNGERFORD. Algebra.
74 DAVENPORT. Multiplicative Number Theory. 3rd ed.
75 HOCHSCHILD. Basic Theory of Algebraic Groups and Lie Algebras.
76 IITAKA. Algebraic Geometry.
77 HECKE. Lectures on the Theory of Algebraic Numbers.
78 BURRIS/SANKAPPANAVAR. A Course in Universal Algebra.
79 WALTERS. An Introduction to Ergodic Theory.
80 ROBINSON. A Course in the Theory of Groups. 2nd ed.
81 FORSTER. Lectures on Riemann Surfaces.
82 BOTT/TU. Differential Forms in Algebraic Topology.
83 WASHINGTON. Introduction to Cyclotomic Fields. 2nd ed.
84 IRELAND/ROSEN. A Classical Introduction to Modern Number Theory. 2nd ed.
85 EDWARDS. Fourier Series. Vol. II. 2nd ed.
86 VAN LINT. Introduction to Coding Theory. 2nd ed.
87 BROWN. Cohomology of Groups.
88 PIERCE. Associative Algebras.
89 LANG. Introduction to Algebraic and Abelian Functions. 2nd ed.
90 BRØNDSTED. An Introduction to Convex Polytopes.
91 BEARDON. On the Geometry of Discrete Groups.
92 DIESTEL. Sequences and Series in Banach Spaces.
93 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part I. 2nd ed.
94 WARNER. Foundations of Differentiable Manifolds and Lie Groups.
95 SHIRYAEV. Probability. 2nd ed.
96 CONWAY. A Course in Functional Analysis. 2nd ed.
97 KOBLITZ. Introduction to Elliptic Curves and Modular Forms. 2nd ed.
98 BRÖCKER/TOM DIECK. Representations of Compact Lie Groups.
99 GROVE/BENSON. Finite Reflection Groups. 2nd ed.
100 BERG/CHRISTENSEN/RESSEL. Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.
101 EDWARDS. Galois Theory.
102 VARADARAJAN. Lie Groups, Lie Algebras and Their Representations.
103 LANG. Complex Analysis. 3rd ed.
104 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part II.
105 LANG. $SL_2(\mathbf{R})$.
106 SILVERMAN. The Arithmetic of Elliptic Curves.
107 OLVER. Applications of Lie Groups to Differential Equations. 2nd ed.
108 RANGE. Holomorphic Functions and Integral Representations in Several Complex Variables.
109 LEHTO. Univalent Functions and Teichmüller Spaces.
110 LANG. Algebraic Number Theory.
111 HUSEMÖLLER. Elliptic Curves. 2nd ed.
112 LANG. Elliptic Functions.
113 KARATZAS/SHREVE. Brownian Motion and Stochastic Calculus. 2nd ed.
114 KOBLITZ. A Course in Number Theory and Cryptography. 2nd ed.
115 BERGER/GOSTIAUX. Differential Geometry: Manifolds, Curves, and Surfaces.
116 KELLEY/SRINIVASAN. Measure and Integral. Vol. I.
117 J.-P. SERRE. Algebraic Groups and Class Fields.
118 PEDERSEN. Analysis Now.
119 ROTMAN. An Introduction to Algebraic Topology.
120 ZIEMER. Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.
121 LANG. Cyclotomic Fields I and II. Combined 2nd ed.
122 REMMERT. Theory of Complex Functions. *Readings in Mathematics*
123 EBBINGHAUS/HERMES et al. Numbers. *Readings in Mathematics*

124 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part III
125 BERENSTEIN/GAY. Complex Variables: An Introduction.
126 BOREL. Linear Algebraic Groups. 2nd ed.
127 MASSEY. A Basic Course in Algebraic Topology.
128 RAUCH. Partial Differential Equations.
129 FULTON/HARRIS. Representation Theory: A First Course.
Readings in Mathematics
130 DODSON/POSTON. Tensor Geometry.
131 LAM. A First Course in Noncommutative Rings. 2nd ed.
132 BEARDON. Iteration of Rational Functions.
133 HARRIS. Algebraic Geometry: A First Course.
134 ROMAN. Coding and Information Theory.
135 ROMAN. Advanced Linear Algebra.
136 ADKINS/WEINTRAUB. Algebra: An Approach via Module Theory.
137 AXLER/BOURDON/RAMEY. Harmonic Function Theory. 2nd ed.
138 COHEN. A Course in Computational Algebraic Number Theory.
139 BREDON. Topology and Geometry.
140 AUBIN. Optima and Equilibria. An Introduction to Nonlinear Analysis.
141 BECKER/WEISPFENNING/KREDEL. Gröbner Bases. A Computational Approach to Commutative Algebra.
142 LANG. Real and Functional Analysis. 3rd ed.
143 DOOB. Measure Theory.
144 DENNIS/FARB. Noncommutative Algebra.
145 VICK. Homology Theory. An Introduction to Algebraic Topology. 2nd ed.
146 BRIDGES. Computability: A Mathematical Sketchbook.
147 ROSENBERG. Algebraic *K*-Theory and Its Applications.
148 ROTMAN. An Introduction to the Theory of Groups. 4th ed.
149 RATCLIFFE. Foundations of Hyperbolic Manifolds.
150 EISENBUD. Commutative Algebra with a View Toward Algebraic Geometry.
151 SILVERMAN. Advanced Topics in the Arithmetic of Elliptic Curves.
152 ZIEGLER. Lectures on Polytopes.
153 FULTON. Algebraic Topology: A First Course.
154 BROWN/PEARCY. An Introduction to Analysis.
155 KASSEL. Quantum Groups.
156 KECHRIS. Classical Descriptive Set Theory.
157 MALLIAVIN. Integration and Probability.
158 ROMAN. Field Theory.
159 CONWAY. Functions of One Complex Variable II.
160 LANG. Differential and Riemannian Manifolds.
161 BORWEIN/ERDÉLYI. Polynomials and Polynomial Inequalities.
162 ALPERIN/BELL. Groups and Representations.
163 DIXON/MORTIMER. Permutation Groups.
164 NATHANSON. Additive Number Theory: The Classical Bases.
165 NATHANSON. Additive Number Theory: Inverse Problems and the Geometry of Sumsets.
166 SHARPE. Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.
167 MORANDI. Field and Galois Theory.
168 EWALD. Combinatorial Convexity and Algebraic Geometry.
169 BHATIA. Matrix Analysis.
170 BREDON. Sheaf Theory. 2nd ed.
171 PETERSEN. Riemannian Geometry.
172 REMMERT. Classical Topics in Complex Function Theory.
173 DIESTEL. Graph Theory. 2nd ed.
174 BRIDGES. Foundations of Real and Abstract Analysis.
175 LICKORISH. An Introduction to Knot Theory.
176 LEE. Riemannian Manifolds.
177 NEWMAN. Analytic Number Theory.
178 CLARKE/LEDYAEV/STERN/WOLENSKI. Nonsmooth Analysis and Control Theory.
179 DOUGLAS. Banach Algebra Techniques in Operator Theory. 2nd ed.
180 SRIVASTAVA. A Course on Borel Sets.
181 KRESS. Numerical Analysis.
182 WALTER. Ordinary Differential Equations.

183 MEGGINSON. An Introduction to Banach Space Theory.
184 BOLLOBAS. Modern Graph Theory.
185 COX/LITTLE/O'SHEA. Using Algebraic Geometry.
186 RAMAKRISHNAN/VALENZA. Fourier Analysis on Number Fields.
187 HARRIS/MORRISON. Moduli of Curves.
188 GOLDBLATT. Lectures on the Hyperreals: An Introduction to Nonstandard Analysis.
189 LAM. Lectures on Modules and Rings.
190 ESMONDE/MURTY. Problems in Algebraic Number Theory. 2nd ed.
191 LANG. Fundamentals of Differential Geometry.
192 HIRSCH/LACOMBE. Elements of Functional Analysis.
193 COHEN. Advanced Topics in Computational Number Theory.
194 ENGEL/NAGEL. One-Parameter Semigroups for Linear Evolution Equations.
195 NATHANSON. Elementary Methods in Number Theory.
196 OSBORNE. Basic Homological Algebra.
197 EISENBUD/HARRIS. The Geometry of Schemes.
198 ROBERT. A Course in p-adic Analysis.
199 HEDENMALM/KORENBLUM/ZHU. Theory of Bergman Spaces.
200 BAO/CHERN/SHEN. An Introduction to Riemann–Finsler Geometry.
201 HINDRY/SILVERMAN. Diophantine Geometry: An Introduction.
202 LEE. Introduction to Topological Manifolds.
203 SAGAN. The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions.
204 ESCOFIER. Galois Theory.
205 FÉLIX/HALPERIN/THOMAS. Rational Homotopy Theory. 2nd ed.
206 MURTY. Problems in Analytic Number Theory.
Readings in Mathematics
207 GODSIL/ROYLE. Algebraic Graph Theory.
208 CHENEY. Analysis for Applied Mathematics.
209 ARVESON. A Short Course on Spectral Theory.
210 ROSEN. Number Theory in Function Fields.
211 LANG. Algebra. Revised 3rd ed.
212 MATOUŠEK. Lectures on Discrete Geometry.
213 FRITZSCHE/GRAUERT. From Holomorphic Functions to Complex Manifolds.
214 JOST. Partial Differential Equations.
215 GOLDSCHMIDT. Algebraic Functions and Projective Curves.
216 D. SERRE. Matrices: Theory and Applications.
217 MARKER. Model Theory: An Introduction.
218 LEE. Introduction to Smooth Manifolds.
219 MACLACHLAN/REID. The Arithmetic of Hyperbolic 3-Manifolds.
220 NESTRUEV. Smooth Manifolds and Observables.
221 GRÜNBAUM. Convex Polytopes. 2nd ed.
222 HALL. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction.
223 VRETBLAD. Fourier Analysis and Its Applications.
224 WALSCHAP. Metric Structures in Differential Geometry.
225 BUMP: Lie Groups
226 ZHU. Spaces of Holomorphic Functions in the Unit Ball.
227 MILLER/STURMFELS. Combinatorial Communicative Algebra.
228 DIAMOND/SHURMAN. A First Course in Modular Forms.
229 EISENBUD. The Geometry of Syzygies.
230 STROOCK. An Introduction to Markov Processes.