Graduate Texts in Mathematics

GTM

Rabi Bhattacharya
Edward C. Waymire

Stationary Processes and Discrete Parameter Markov Processes

Springer

Graduate Texts in Mathematics 293

Graduate Texts in Mathematics

Graduate Texts in Mathematics bridge the gap between passive study and creative understanding, offering graduate-level introductions to advanced topics in mathematics. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Although these books are frequently used as textbooks in graduate courses, they are also suitable for individual study.

Rabi Bhattacharya • Edward C. Waymire

Stationary Processes and Discrete Parameter Markov Processes

Springer

Rabi Bhattacharya
Department of Mathematics
University of Arizona
Tucson, AZ, USA

Edward C. Waymire
Department of Mathematics
Oregon State Univeristy
Corvallis, OR, USA

ISSN 0072-5285 ISSN 2197-5612 (electronic)
Graduate Texts in Mathematics
ISBN 978-3-031-00941-9 ISBN 978-3-031-00943-3 (eBook)
https://doi.org/10.1007/978-3-031-00943-3

Mathematics Subject Classification: 60G10, 60G12, 60J05, 60J10, 60J20, 60G53, 60G42, 60G25, 60K35, 60F17, 60E15

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to Gouri Bhattacharya (in loving memory) and to Hermann Flaschka (in memory of a colleague, a brilliant mathematician, and a dear friend)

Preface

The author's recent book *Random Walk, Brownian Motion and Martingales* broadly provides an in-depth introduction to cornerstone elements of the theory of stochastic processes and their applications, and might be viewed as a first course. The present book, *Stationary Processes and Discrete Parameter Markov Processes*, singles out two particularly prominent areas of the general theory of stochastic processes for more focused investigations. Both treatments are stand alone, and otherwise depend on one's pedagogical goals and interests.

Two distinct theories of processes that evolve at random provide the dominant theme of this book. The first focuses on mean zero processes for which no distributional assumptions are made, except for the invariance under time shifts of the second order moments. That there exists for these weakly stationary processes an elegant and rather complete (stochastic) spectral representation theory, adequate for prediction and filtering, should come as a pleasant surprise ! The second theory, on the other hand, is based on the assumed Markov property which, given an initial state and one-step transition probabilities, completely identifies the distribution of the process. Still its breadth is enormous, with applications to most areas of physical, biological, and social sciences, as well as engineering.

The prerequisite is a one-semester/quarter of graduate level probability. Some familiarity with the standard models and methods introduced in Bhattacharya and Waymire (2021) will be helpful to have as background. However, efforts were made to make this book self-contained relative to the graduate level probability prerequisite. Throughout the book, the authors provide references to the second edition of their text Bhattacharya and Waymire (2016) *A Basic Course in Probability Theory* denoted BCPT, as an appendix for prerequisite material in analysis and probability as needed. However, there are many excellent texts and online resources that can be used for this purpose as well.

Much of the first part of this book is devoted to aspects of weakly stationary processes, or time series, and translation invariant random fields. The spectral theory for weakly stationary processes is introduced, including the necessary and sufficient condition of absolute continuity of the spectral measure for representing the process as a (linear) two-sided moving average. A further condition on integrability of the logarithm of the spectral density is then shown to be necessary and sufficient for a representation of a one-sided moving average, which leads to Kolmogorov's theory of prediction. This portion of the text is not used in the subsequent developments of the text and, accordingly, may be omitted on first reading.

An introduction to the ergodic theory of strictly stationary stochastic processes and dynamical systems, and their connections, is a major topic in this general framework. The central theme of the latter is Birkhoff's ergodic theorem.

The second part of this book concerns discrete parameter Markov processes. Discrete parameter stochastic processes are often viewed as models of temporal evolution for which there are definable past, present, and future periods of the evolution. When, at any given time, the future distribution of the process depends on the past and present only through the present state, the stochastic process is said to be a *Markov process*. Such processes comprise a manifestly important class of stochastic processes from points of view of both mathematical theory and application. The first half of this book is primarily devoted to the case of Markov processes taking values in a countable state space, referred to as Markov chains, while the second half of this book concerns processes with general state space. Much of the basic theory addresses questions pertaining to the time-asymptotic behavior of these processes. In particular, one seeks conditions for the existence of a unique invariant (steady state) distribution. Conditions for recurrence and ergodicity are developed in this connection. Also, under an invariant initial distribution, the process is stationary. Laws of large numbers and central limit theory are developed from the perspectives of renewal decompositions and, another, using martingale theory.

In the final chapters, rates of convergence to steady state are developed for possibly non-irreducible Markov processes on general state spaces by methods of a theory of i.i.d. iterated random maps. Non-irreducibility in the context of general state spaces provides exciting challenges and opportunities for the continued development of the theory.

An extended Perron–Frobenius theorem is presented for application to the Donsker–Varadhan theory of large deviations for Markov processes on a general state space, extending Cramér's large deviation theory for i.i.d. random variables, and Sanov's theorem on large deviations of empirical measures.

Special topics chapters include applications of the large deviation theory developed in Chapter 21. Others are a simple exposition of the Kalman filter, and another on the theory of "positive dependence" of the type found in areas ranging from reliability theory to statistical physics and interacting particle systems. This latter chapter culminates with Newman's central limit theorem for associated random fields and Pitt's theorem for positively correlated Gaussian random vectors. An application to a two-dimensional bond percolation model is included for illustration. Another special topics chapter on coupling methods includes proofs of Choquet–Deny theorem for harmonic functions, Strassen's theorem for stochastic ordering, and the role of log-convexity in Holley's inequalities and the FKG inequalities, together with a proof of the FKG inequalities for Ising ferromagnets. The latter has strong ties with the special topics chapter on associated random fields as well.

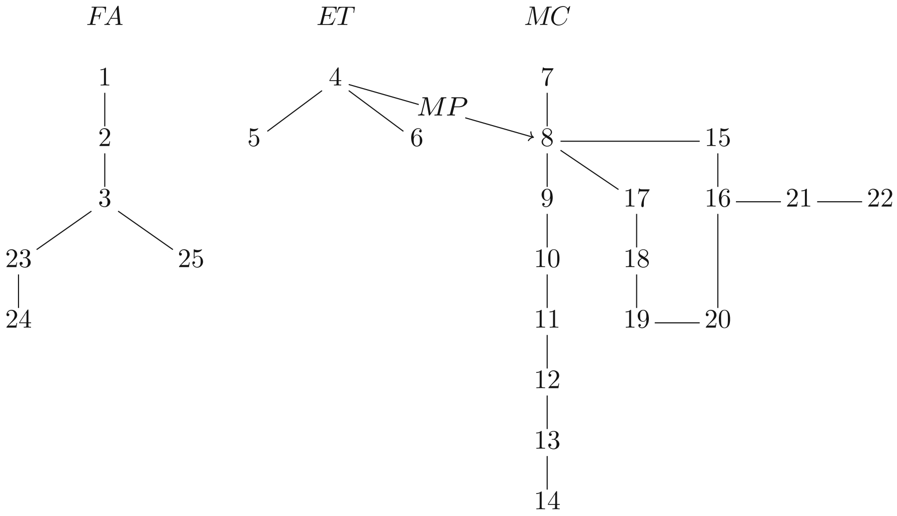
The authors are grateful to Springer Editor Loretta Bartolini for patience and guidance. We would like to thank our colleagues Sunder Sethuraman, William Faris. and Enrique Thomann for their continued advice and encouragement. We are grateful to the University of Arizona graduate students Duncan Bennettt and Eric Roon for technical assistance with some TikZ and LaTeX code. The authors gratefully acknowledge partial support (DMS1811317, DMS-1408947) of the National Science Foundation during the preparation of this book.

Tucson, AZ, USA Rabi Bhattacharya
Corvallis, OR, USA Edward C. Waymire

Ten-Week Course Suggestions

(A) Stationary Processes: 1-6, 23, 25
(B) Markov Chains: 4, 5, 7-11, 12-14
(C) Markov Processes: 4, 8-12, 15-21

$FA = \text{FourierAnalysis}$
$ET = \text{ErgodicTheory}$
$MC = \text{MarkovChains}$
$MP = \text{MarkovProcesses}$



Chapter Dependency Diagram

Contents

Symbol Definition List

Special Sets and Functions:
In the classic notation of G.H. Hardy, one writes $a(x) = O(b(x))$ to mean that there is a constant c (independent of x) such that $|a(x)| \leq c|b(x)|$ for all x. Also $a(x) = o(b(x))$ indicates that the ratio $a(x)/b(x) \to 0$ according to specified limit.

$\mathbb{Z}_+$, set of non-negative integers
$\mathbb{Z}_{++}$, set of positive integers
$\mathbb{R}_+$, set of non-negative real numbers
$\mathbb{R}_{++}$, set of positive real numbers
$\mathbb{Z}_2$, the group of integers modulo two
$\mathbb{Z}_2^m$, m-dimensional hypercube (product space)
$D(0 : r)$, disc of radius r centered at 0.
∂A, boundary of set A
A^o, interior of set A
A^-, closure of set A
A^c, complement of set A
$\mathbf{1}_B(x)$, indicator of the set B
$[X \in B]$, inverse image of the set B under X
$\# A$, $|A|$, cardinality for finite set A
δ_x, Dirac delta (point mass)
$\otimes$, σ-field product
τ_B, time of first arrival in B
$\tau_B^{(r)}, r \geq 1$, r-th return time to B
$\tau|n$, restriction of tree graph to first n generations
$v|n$, restriction of tree vertex to first n generations
$\overline{c, d}$, closed interval $[c, d]$
$\bigotimes$, product of σ-fields
$\mathcal{S}^{\bigotimes n}$, the n-fold product σ-field of $\mathcal{S}$
$\mathcal{B}$, Borel σ-field
$\mathcal{J}$, invariant σ-field
$\oplus$, orthogonal sum

$i \to j$, j is accessible from i

$i \leftrightarrow j$, i and j communicate, or graph connectivity in percolation

$\partial \Lambda$, boundary points of Λ

∂_e, edge boundary in percolation

$\rhd$, Sarkovskii order symbol for dynamical systems

$[0, n]_0 = \{0, 1, \ldots, n-1, n\}$

$\lceil x \rceil$, the ceiling function

$[x]$, greatest integer function

$\mathcal{R}(\gamma)$, the range of the map γ

$\preceq$, partial order

$\leq^s$, stochastically less than or equal

$p(x, dy)$, homogeneous (stationary) one-step discrete parameter transition probability

$f_{X|Y}(x|y)$, density of conditional distribution of X on $[Y = y]$.

$((p_{ij}))$, countable state (one step) Markov transition probability matrix

$((p_{ij}^{(n)}))$, countable state n-step Markov transition probability matrix

$p(x, y), q(x, y), q(y|x)$, (variously) a one-step transition probability density

$p(t; x, dy)$, homogeneous (stationary) continuous parameter transition probability

$p(s, t : x, dy)$, nonhomogeneous (nonstationary) continuous parameter transition probability

$p_A^{(n)}(x, B)$, transition probability from x to B in n steps before reaching A

${}_A p(x, B)$, transition probability from x to B eventually and before reaching A

$D(\nu||\mu)$, Kulback-Liebler divergence of ν with respect to μ

$H(\cdot)$, Shannon entropy

$H_\alpha(\cdot)$, Renyi entropy

ARE, asymptotic relative efficiency

$A \circ B$, disjoint occurrence of events A, B in percolation

$\wedge\ \vee$, lattice min and max operations

Function Spaces, Elements and Operations:

$C[0, 1]$, set of continuous, real-valued functions defined on $[0, 1]$

$\mathbb{R}^\infty$, infinite sequence space

$C([0, \infty) : \mathbb{R}^k)$, set of continuous functions on $[0, \infty)$ with values in $\mathbb{R}^k$

$C_b(S)$, set of continuous bounded, real-valued functions on a metric (or topological) space S

$B(S)$, set of bounded, measurable real-valued functions on a measurable space $(S, \mathcal{S})$

BL, the space of bounded Lipschitz functions

d_P, Prohorov metric on $\mathcal{P}(S)$

d_{BL}, bounded-Lipschitz metric on $\mathcal{P}(S)$

$C_b^0(S)$, continuous functions on a metric or topological space vanishing at infinity

$C(S : \mathbb{C})$, set of complex-valued functions on S

$\mathcal{P}(S)$, space of probability measures on S

$||\cdot||_{tv}$, total variation norm

$||\cdot||_{op}$, operator norm

$\otimes_{e_\xi}$, Navier-Stokes projected convolution

e_i, **e** i-th coordinate of unit vector

$i.o.$, infinitely often

$f * g$, convolution of functions

$Q_1 * Q_2$, convolution of measures Q_1, Q_2

Cov, covariance

Var, variance

$\Rightarrow$, weak convergence

A^t, v^t matrix transpose

Chapter 1
Fourier Analysis: A Brief Survey

A few of the basic concepts and definitions from Fourier analysis that will be used in the next few chapters are recalled.

As remarked in the **Preface**, *throughout the text the authors' footnote references to the second edition of their text Bhattacharya and Waymire (2016), A Basic Course in Probability Theory, denoted BCPT, are used as an Appendix for prerequisite material in analysis and probability as needed. However there are many excellent texts and online resources that can be used for this purpose.*

The idea that general functions, including those with discontinuities, may be expressed as superpositions, or linear combinations, of periodic functions with different periods and amplitudes is due to the legendary French mathematician/physicist Joseph Fourier (1768–1830). Although exceptions were pointed out by some other mathematicians, Fourier's brilliant idea brought forth a revolution in mathematics. Fourier analysis is a major tool used in this book, especially the first few chapters comprising Part I.

We begin with Fourier series, namely Fourier analysis on the unit circle $\mathbb{T}$ in the plane represented as $[-\pi, \pi]$, with $-\pi$ and π identified. One may also conveniently represent $\mathbb{T}$ as the unit circle in the complex plane, $\mathbb{T} = \{\exp\{i\theta\} : -\pi < \theta \leq \pi\}$. A (Borel measurable complex-valued) function on $\mathbb{T}$ may be thought of as a (Borel measurable complex-valued) periodic function on the real line $\mathbb{R}$, with period 2π. The simplest such functions are $\exp\{inx\}$ and their superpositions. One looks for representing, or approximating, a more general periodic function f by such a superposition $f \sim \sum_n d_n \exp\{inx\}$; a finite sum of this type is called a trigonometric polynomial. According to a result of Weierstrass (Exercise 2), such polynomials are dense in the set of all complex-valued continuous functions

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_1

$C(\mathbb{T} : \mathbb{C})$ endowed with the uniform norm: $||f||_u = \sup_{x\in\mathbb{T}} |f(x)|$. It follows that trigonometric polynomials are dense in the Hilbert space $L^2(\mathbb{T})$ of square integrable functions on $\mathbb{T}$ with squared norm $||f||^2 = \frac{1}{2\pi}\int_{\mathbb{T}} |f(x)|^2 dx$. It is simple to check that the functions $\exp\{inx\}$, $n \in \mathbb{Z}$, form an orthonormal sequence in $L^2(\mathbb{T})$, and, by the density in uniform norm of trigonometric polynomials in $C(\mathbb{T} : \mathbb{C})$, they are a complete orthonormal sequence in $L^2(\mathbb{T})$ (Exercise 3). Thus $f(\cdot) = \sum_{n\in\mathbb{Z}}\langle f, \exp\{in\cdot\}\rangle \exp\{in\cdot\}$ or

$$f(x) = \sum_{n\in\mathbb{Z}} c_n \exp\{inx\}, \tag{1.1}$$

where

$$c_n = \langle f, \exp\{in\cdot\}\rangle = \frac{1}{2\pi}\int_{[-\pi,\pi]} f(x)\exp\{-inx\}dx \tag{1.2}$$

is the n-th Fourier coefficient of f, and the expansion (1.1) is the Fourier series of f. The equality in (1.1) is in $L^2(\mathbb{T})$. It follows from this orthogonal expansion that

$$||f||^2 = \sum_{n\in\mathbb{Z}} |c_n|^2, \quad f \in L^2(\mathbb{T}). \tag{1.3}$$

Thus the Fourier series is an isometry between $L^2(\mathbb{T})$ and $L^2(\mathbb{Z})$, where $L^2(\mathbb{Z})$ is the space of sequences $\{a_n : n \in \mathbb{Z}\}$ endowed with the squared norm $\sum_n |a_n|^2$. This immediately implies that (Exercise 4)

$$\langle f, g\rangle_{\mathbb{T}} = \langle\{a_n\}, \{d_n\}\rangle_{\mathbb{Z}}, \tag{1.4}$$

where $\{a_n\}$ and $\{d_n\}$ are the Fourier coefficients of f and g, respectively. The subscripts $\mathbb{T}$ and $\mathbb{Z}$ in (1.4) are used to distinguish the inner products in the two Hilbert spaces: $\langle f, g\rangle_{\mathbb{T}} = \frac{1}{2\pi}\int_{[-\pi,\pi]} f(x)\overline{g}(x)dx$ and $\langle\{a_n\}, \{d_n\}\rangle_{\mathbb{Z}} = \sum_{n\in\mathbb{Z}} c_n\overline{d}_n$. We will drop these subscripts for the norms and inner products in the future when there is no possibility of confusion. One may extend the notion of Fourier coefficients to finite signed measures. The Fourier coefficients c_n of a finite signed measure μ are defined by

$$c_n = \frac{1}{2\pi}\int_{[-\pi,\pi]} \exp\{-inx\}\mu(dx) \quad (n \in \mathbb{Z}). \tag{1.5}$$

By usual approximation by functions (see BCPT , Proposition 6.3), one can show that the Fourier coefficients of μ determine it. One may think of this as a generalization of Fourier coefficients of integrable functions f by letting $\mu(dx) = f(x)dx$, noting that square integrable functions on $\mathbb{T}$ are integrable. An important question arises in the theory pursued in the following chapters: *Which sequences*

$\{c_n\}$ *are Fourier coefficients of a finite measure on* $\mathbb{T}$ *?* The answer to this question is provided by the important theorem by Herglotz[1]

Theorem 1.1 *(Herglotz).* A sequence $\{c_n\}$ is the sequence of Fourier coefficients of a finite measure μ on $\mathbb{T}$ if and only if the sequence is positive-definite, i.e., if and only if for every finite sequence $z_j (j = 1, \dots, n)$ of complex numbers, one has $\sum_{1 \le j,k \le n} c_{j-k} z_j \bar{z}_k \ge 0$.

Turning to the Fourier analysis on $\mathbb{R}$, one defines the Fourier transform of an integrable function (with respect to Lebesgue measure) f as

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(x) \exp\{i\xi x\} dx, \quad \xi \in \mathbb{R}. \tag{1.6}$$

If one used $\exp\{-i\xi x\}$ instead of $\exp\{i\xi x\}$, it would correspond to the definition (1.2) of the Fourier coefficient. But we will rather follow the usual convention in the probability literature here. If f and $\widehat{f}$ both are integrable with respect to Lebesgue measure, then one has the inversion formula (BCPT, Theorem 6.7 (a))

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\xi) \exp\{-i\xi x\} d\xi, \quad x \in \mathbb{R}. \tag{1.7}$$

This corresponds to the Fourier series representation (1.1), except for the sign in the exponent of exp. Because $L^1(\mathbb{R}, dx)) \cap L^2(\mathbb{R}, dx)$ is dense in $L^2(\mathbb{R}, dx)$, with respect to the L^2-distance, one defines the Fourier transform (1.6) for functions in $L^2(\mathbb{R}, dx)$ also. One then has an isometry analogous to (1.3) (BCPT, Theorem 6.7(b)):

$$(2\pi)||f||_2^2 = ||\widehat{f}||_2^2 \quad f \in L^2(\mathbb{R}, dx). \tag{1.8}$$

This is known as the Plancherel identity. One also defines the Fourier transform $\widehat{\mu}$ of a finite signed measure μ as

$$\widehat{\mu}(\xi) = \int_{\mathbb{R}} \exp\{i\xi x\} \mu(dx), \quad \xi \in \mathbb{R}. \tag{1.9}$$

In the case μ is a probability measure, $\widehat{\mu}$ is called the *characteristic function* of μ.

Corresponding to the theorem of Hergloz, Bochner's theorem says that a complex-valued function φ on $\mathbb{R}$ is the Fourier transform of a finite signed measure if and only if it is positive-definite and continuous. Here a function φ is positive-definite if and only if for any finite sequence of complex number $\{z_1, \dots, z_n\}$, one has $\sum_{1 \le j,k \le n} z_j \bar{z}_k \varphi(z_j - z_k) \ge 0$ (BCPT, Theorem 6.13).

[1] For a proof, see BCPT, p. 111.

Exercises

1. (*Complex Stone-Weierstrass Theorem*) Let S be a compact metric space and Γ a subalgebra of $C(S : \mathbb{C})$ which separates points, contains constant functions, and is closed under complex conjugation. Prove that Γ is dense in $C(S : \mathbb{C})$. [*Hint*: Let $\Gamma_{\mathbb{R}}$ denote the set of real and imaginary parts of f, for all $f \in \Gamma$. Since for all $f \in \Gamma$, $Ref = (f + \overline{f})/2$, $Imf = (f - \overline{f})/2i$, it follows that $\Gamma_{\mathbb{R}}$ is a subalgebra of $C(S : \mathbb{R})$ which satisfies the hypothesis of the Stone–Weierstrass theorem for $C(S : \mathbb{R})$ and is, therefore, dense in it. But $\Gamma = \{g + ih : g, h \in \Gamma_{\mathbb{R}}\}$ is then clearly dense in $C(S : \mathbb{C})$.]
2. (*Weierstrass Approximation Theorem for* $C(\mathbb{T} : \mathbb{C})$) Prove that the set Γ of trigonometric polynomials, i.e., finite linear combinations of functions $\exp\{inx\}$, $n \in \mathbb{Z}$, is dense in $C(\mathbb{T} : \mathbb{C})$ in the uniform norm. [*Hint*: Show that Γ satisfies the hypothesis of Exercise 1.]
3. Prove that the sequence $\{\exp\{inx\}, n \in \mathbb{Z}\}$ is an orthonormal and complete basis of $L^2(\mathbb{T})$. [*Hint*: Apply Exercise 2 to prove completeness.]
4. (*Polarization Identity*) Give a proof of (1.4). [*Hint*: Use the fact that the inner product $\langle f, g\rangle$ in a complex Hilbert space H satisfies the so-called polarization identity: $4\langle f, g\rangle = ||f + g||^2 - ||f - g||^2 - i||f + ig||^2 + i||f - ig||^2$.]

Chapter 2
Weakly Stationary Processes and Their Spectral Measures

Stationary stochastic processes are analyzed at the level of their first and second order characteristics, mean and covariance, using Fourier methods.

Some historic considerations that naturally lead to spectral considerations of the type covered in the next few chapters often involved such phenomena as electrical currents in a vacuum tube or components of turbulent fluid velocities in a wind tunnel. In such arrangements, it is somewhat natural to at least formally view the physical process as a sum of a mean process (the “signal”) and stochastic fluctuations (the “noise”). The noise is often regarded to be stationary in the sense that its distribution is invariant under translations in time. Second order quantities such as the average electrical power dissipated by a current across a resistor (Joule’s law) or the mean-square kinetic energy of the fluid after the mean is removed are both proportional to the variance.

From a mathematical perspective, since the basic second order structure and properties embodied in variances and covariances (correlations) can be conveniently analyzed through the Fourier analysis, one may expect Fourier theory to provide the appropriate framework. The survey of the basic notions and results from Fourier analysis given in Chapter 1, and an occasional review of other concepts[1] as they occur, should be sufficient background orientation for the development of spectral theory to follow.

[1] See BCPT, Chap. VI.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_2

Definition 2.1. Let $T = \mathbb{Z}^+$, $\mathbb{Z}$, $\mathbb{R}^+$ or $\mathbb{R}$. A stochastic process $\{X_t : t \in T\}$ with values in a measurable space $(S, \mathcal{S})$ is said to be *stationary*, or *strictly stationary*, if for every finite set of indices $t_1 < t_2 < \cdots < t_k$ $(k \geq 1)$ and every $s > 0$ such that $t_i + s \in T$ for all i, the distribution of $(X_{t_1+s}, X_{t_2+s}, \dots, X_{t_k+s})$ is the same as that of $(X_{t_1}, X_{t_2}, \dots, X_{t_k})$.

The qualifying term "strictly" in Definition 2.1 is sometimes applied to a stationary process to distinguish it from processes that are stationary in a much weaker sense as follows.

Definition 2.2. Let T be as above, but take $S = \mathbb{C}$ with its Borel σ-field for $\mathcal{S}$. A complex-valued process $\{X_t = U_t + iV_t : t \in T\}$ is said to be *weakly stationary* if its first two moment sequences are each finite and translation invariant, that is,

$$\mathbb{E}X_t = \mathbb{E}X_0 = \mu, \quad \mathrm{Cov}(X_s, X_{t+s}) = \mathrm{Cov}(X_0, X_t) = r(t) \quad \text{for all } t, t+s \in T,\ s \geq 0.$$

Here $\mathbb{E}X_t = \mathbb{E}U_t + i\mathbb{E}V_t$, $\mathrm{Cov}(X_s, X_{t+s}) = \mathbb{E}(X_s - \mu)(\overline{X_{t+s} - \mu})$. In particular, if the X_t's are real-valued, then one can omit V_t and the conjugation sign in the covariance.

Example 1. Let $\{X_n : n \in \mathbb{Z}^+\}$ be a real- or complex-valued *(uncorrelated)* process with constant mean μ and $\mathrm{Cov}(X_m, X_n) = \delta_{m,n}\sigma^2$ for some $\sigma^2 \geq 0$. Then $\{X_n : n \in \mathbb{Z}^+\}$ is weakly stationary. In particular, a weakly stationary sequence $\{X_n : n \in \mathbb{Z}^+\}$ of independent random variables will also be strictly stationary if and only if it is an i.i.d. sequence.

Example 2 *(Ornstein–Uhlenbeck Process).* Let $\{B(t) : t \geq 0\}$ denote standard Brownian motion started at $B(0) = 0$. Since $\mathrm{Var}(B(t)) = t$ grows with $t \geq 0$, standard Brownian motion is clearly non-stationary on the linear time scale. However, consider the process $X_t = e^{-\frac{t}{2}}B(e^t), -\infty < t < \infty$. $\{X_t : t \in \mathbb{R}\}$ is Gaussian with $\mathbb{E}X_t = 0$ and, recalling $\mathrm{Cov}(B(s), B(t)) = \min(s,t)$, $\mathrm{Cov}(X_s, X_t) = e^{-\frac{|t-s|}{2}}$, $s, t \in \mathbb{R}$. In particular $\{X_t : t \in \mathbb{R}\}$ is a (strictly) stationary process, referred to as the *Ornstein–Uhlenbeck process.*

Proposition 2.1. Let S be a Polish space with Borel σ-field $\mathcal{S}$. (a) If $\{X_n : n \in \mathbb{Z}^+\}$ is a stationary process with values in S, one can construct a stationary process $\{Y_n : n \in \mathbb{Z}\}$ such that $\{Y_n : n \in \mathbb{Z}^+\}$ has the same distribution as that of $\{X_n : n \in \mathbb{Z}^+\}$. (b) If $\{X_t : t \in \mathbb{R}^+\}$ is a stationary stochastic process with values in S, one can construct a stationary process $\{Y_t : t \in \mathbb{R}\}$ such that the distributions of $(Y_{t_1}, Y_{t_2}, \dots)$ and $(X_{t_1}, X_{t_2}, \dots)$ are the same for every sequence $0 \leq t_1 < t_2 < \cdots$ in $\mathbb{R}^+$.

Proof. (a) The distribution of $\{X_n : n \in \mathbb{Z}^+\}$ is a probability measure on the measurable space $(S^{\mathbb{Z}^+}, \mathcal{S}^{\otimes\mathbb{Z}^+})$, where $\mathcal{S}^{\otimes\mathbb{Z}^+}$ is the product σ-field. Let $\mu_{n_1,n_2,\dots,n_k}$ be the distribution of $(X_{n_1}, \dots, X_{n_k})$, $0 \leq n_1 < n_2 < \cdots < n_k$. For every k-tuple of integers $n_1 < n_2 < \cdots < n_k$ in $\mathbb{Z}$ $(k \geq 1)$, define $\mu_{n_1,n_2,\dots,n_k} := \mu_{0,n_2-n_1,\cdots,n_k-n_1}$. This defines a consistent family of finite dimensional distributions on $(S^{\mathbb{Z}}, \mathcal{S}^{\otimes\mathbb{Z}})$.

By Kolmogorov's existence theorem,[2] the coordinate projections $\{Y_n : n \in \mathbb{Z}\}$ on this space have the desired property. (b) The proof of part (b) is entirely analogous. ■

Remark 2.1. A different proof of Proposition 2.1(a) may be given by defining for each $m \geq 1$ the *left-shifted process* $X^{(m)}$ as $X_n^{(m)} := X_{n+m}$ for $n \geq -m$ and extending it to indices $n < -m$ by setting $X_n^{(m)} = X_0$. Then $X^{(m)}$ converges in distribution as $m \to \infty$ to the desired stationary process. A similar argument applies to part (b) if X_t has continuous sample paths (a.s.) on $\mathbb{R}^+$.

Let $\{X_t : t \in T\}$ be a real- or complex-valued process with finite second moments on an index set T contained in $\mathbb{R}$ or $\mathbb{Z}$. The *covariance function* of the process is

$$r(s,t) = \mathrm{Cov}(X_s, X_t) = \mathbb{E}(X_s - \mathbb{E}X_s)\overline{(X_t - \mathbb{E}X_t)}. \tag{2.1}$$

It is clearly (i) Hermitian symmetric in the sense that $r(s,t) = \overline{r(t,s)}$, and it is (ii) *positive definite*. In the sense that for $n \geq 1$ and arbitrary $t_1, \ldots, t_n \in T$, and $(a_1, a_2, \ldots, a_n) \in \mathbb{C}^n$, one has

$$\sum_{1 \leq j,k \leq n} a_j \overline{a}_k r(t_j, t_k) = \mathbb{E}|\sum_{j=1}^{n} a_j (X_{t_j} - \mathbb{E}X_{t_j})|^2 \geq 0. \tag{2.2}$$

Strictly speaking the term *nonnegative definite* is more appropriate here. However this abuse of terminology is somewhat standard.

Remark 2.2. The term *autocovariance function* is also often used in reference to $r(s,t)$. Also, since variances are constant under (weak) stationarity, covariance and correlation are in constant proportion to each other.

In the case $T = \mathbb{Z}$ or $\mathbb{R}$ and the process is weakly stationary, we often write

$$c_t = \mathrm{Cov}(X_s, X_{s+t}) = \mathrm{Cov}(X_0, X_t), \quad t \in T.$$

One has for all $n \geq 1$

$$\sum_{1 \leq j,k \leq n} a_j \overline{a}_k c_{t_k - t_j} \geq 0 \quad ((a_1, \ldots, a_n) \in \mathbb{C}^n, \quad t_j \in T \quad \text{for } 1 \leq j \leq n. \tag{2.3}$$

One may check from (2.3) that (i) $c_0 \geq 0$ and (ii) $c_{-t} = \overline{c}_t$ (Exercise 3). The normalized quantity $\rho_t = c_t / c_0$, where by weak stationarity $\mathrm{Var}(X_t) = \mathrm{Var}(X_0) = c_0$, $t \in T$, defines the *correlation function*.

[2] BCPT, p. 168.

Since for any $t_j, t_k, t_k - t_j \in T$, $\mathrm{Cov}(X_{t_j}, X_{t_k}) = \mathrm{Cov}(X_0, X_{t_k - t_j})$, one sees that $\{c_t : t \in T\}$ satisfying (2.3) is also *positive definite* in the sense of Herglotz' and Bochner's theorems. Let us first consider this in the case of real-valued processes indexed by the discrete parameter set $T = \mathbb{Z}$.

Theorem 2.2. There is a one-to-one correspondence between the set of all covariance sequences $\{c_j : j \in \mathbb{Z}\}$ of real-valued weakly stationary processes $\{X_n : n \in \mathbb{Z}\}$ and the set of all finite symmetric measures F on $[-\pi, \pi]$, with $\{c_j : j \in \mathbb{Z}\}$ as the sequence of Fourier coefficients of F, i.e., $c_j = \frac{1}{2\pi}\int_{[-\pi,\pi]} e^{-ij\lambda} F(d\lambda)$, $j \in \mathbb{Z}$.

Proof. The existence of the measure F follows from (2.3) and Herglotz' theorem[3] symmetry of F is due to the fact that c_j's are real, so that $c_{-j} = c_j$ for all j which, since the Fourier coefficients determine the measure, holds if and only if F equals the measure induced by the change of variable $\lambda \to -\lambda$ on $[-\pi, \pi]$. ■

Remark 2.3. For purposes of integration on the unit circle $[-\pi, \pi)$ with respect to F, it is convenient to integrate on $[-\pi, \pi]$, treating π as distinct from $-\pi$, but always with $F(\{\pi\}) = F(\{-\pi\})$, by splitting the mass at the point on the circle into equal halves.

The finite measure F in Theorem 2.2 is called the *spectral measure*. Note that the variance may be viewed as the accumulated total spectral mass,

$$\mathrm{Var}(X_n) = \mathrm{Var}(X_0) = c_0 = \frac{1}{2\pi}\int_{[-\pi,\pi]} F(d\lambda) = \frac{F([-\pi,\pi])}{2\pi}.$$

In the next chapter we will derive a representation of the process $\{X_n : n \in \mathbb{Z}\}$ as a superposition of sinusoidal oscillations with random amplitudes for which $F(d\lambda)/2\pi$ may be interpreted as the contribution to the variance from the variance in amplitude of an oscillation in the frequency range λ to $\lambda + d\lambda$.

From Theorem 2.2, one has the simple corollary.

Corollary 2.2. Finite symmetric (spectral) measures are in a one-to-one correspondence with mean-zero real-valued stationary Gaussian processes. The covariances are given by

$$c_j = \frac{1}{2\pi}\int_{[-\pi,\pi]} e^{-ij\lambda} F(d\lambda) = \frac{1}{2\pi}\int_{[-\pi,\pi]} \cos(j\lambda) F(d\lambda) \quad (j \in \mathbb{Z}). \tag{2.4}$$

In particular, $c_0 = F([-\pi, \pi])/2\pi = \sigma^2 \equiv \mathrm{Var}(X_j)$.

Example 3 *(Dirac Point Mass Spectral Measure).* If the support of F is $\{0\}$, i.e., $F(d\lambda) = b\delta_{\{0\}}(d\lambda)$ for a positive constant b, then $c_n = F(\{0\})/2\pi = b/2\pi = \sigma^2$ for all $n \in \mathbb{Z}$. This implies $\mathrm{Var}(X_n) = c_0 = \sigma^2$ and $\mathrm{Var}(X_n - X_0) = 2\sigma^2 - 2c_n = 0$ for all n. Hence, with probability one, $X_n = X_0$ for all n. One may choose

[3] See BCPT, p.110.

X_0 arbitrarily to be any random variable with variance σ^2. This is a case of *extreme dependence* in the sense of no decay in the correlations with increasing n. In particular, the sequence is "deterministic" in the sense that its value at one n determines its values for all n. Note that the point mass of F at $\{0\}$ may be equivalently viewed as a jump discontinuity in the *spectral distribution function* $\lambda \to F[-\pi, \lambda]$, $-\pi \leq \lambda \leq \pi$, at $\lambda = 0$.

Example 4 *(Lebesgue Spectral Measure).* Let $F(d\lambda) = bd\lambda$ be a multiple b of Lebesgue measure for some $b > 0$. Then $\sigma^2 = c_0 = b$, $c_n = 0$ for all $n \neq 0$. Hence the sequence $\{X_n : n \in \mathbb{Z}\}$ is *uncorrelated.* In particular, all i.i.d. sequences with (common) variance $\sigma^2 = b$ have this form of the spectral measure.

The previous two examples are somewhat extreme illustrations of a connection between the smoothness of the spectral distribution function $\lambda \to F(-\pi, \lambda]$ and the rate of decay of correlations, see Exercise 15.

Example 5 *(Fractional Gaussian Noise Sequence).* Fix $0 < h < 1$ and consider the (one-dimensional) continuous parameter fractional Brownian motion process $\{B_t^{(h)} : t \geq 0\}$. The fractional Brownian motion is defined by a mean-zero Gaussian process starting at $B_0^{(h)} = 0$ with stationary increments such that

$$\mathrm{Cov}(B_s^{(h)}, B_t^{(h)}) = \Gamma_h(s,t) = \frac{\sigma_0^2}{2}\left\{|s|^{2h} + |t|^{2h} - |s-t|^{2h}\right\}. \tag{2.5}$$

Remark 2.4. The nomenclature, the reason for Gaussian, and the form of the correlation function are suggested by a representation of this process as a "fractional derivative of Brownian motion" to be developed in the next chapter. However, the covariance function (2.5) characterizes fractional Brownian motion among mean-zero Gaussian processes $\{X_t : t \geq 0\}$ starting at zero, having stationary increments and variance scaling as $\mathbb{E}X_t^2 = \sigma_0^2 t^{2h}$ for an exponent $h \in (0,1)$; see Exercise 7. This is a modification of the original definition of fractional Brownian motion given by Lévy (1953). Lévy's definition is valid for all $h > 0$.

For the present, we wish to consider the discrete parameter process

$$\{W_n^{(h)} = B_{n+1}^{(h)} - B_n^{(h)} : n = 0, 1, \ldots\}.$$

This defines a stationary Gaussian sequence with, for $n = 0, 1, 2, \ldots,$

$$c_n = \mathbb{E}\left\{(B_{n+1}^{(h)} - B_n^{(h)})B_1^{(h)}\right\} = \frac{\sigma_0^2}{2}\left(|n+1|^{2h} - 2|n|^{2h} + |n-1|^{2h}\right).$$

The computation of the spectral distribution $F(d\lambda)$ will be postponed to an application of the spectral representation theory to be developed in the next chapter. This stationary process defined by the increments of the fractional Brownian motion may be extended backward to a process again denoted $\{W_n^{(h)} : n \in \mathbb{Z}\}$ and referred to as *fractional Gaussian noise*. In the case $h = 1/2$, this is simply an i.i.d. mean-zero Gaussian sequence. In particular one has a trivial "central limit theorem" in the sense that the distribution of $\frac{1}{\sqrt{n}}\sum_{j=1}^{n} W_j^{(\frac{1}{2})}$ has a standard normal distribution for any n and hence in the limit as $n \to \infty$. However, for $h \neq \frac{1}{2}$, although the distribution of the sum is Gaussian, to obtain a non-degenerate limit distribution, the corresponding scaling is by n^{-h}, not $1/\sqrt{n}$. Since the fractional Gaussian noise is a Gaussian process, this is a reflection of the statistical dependence exhibited in the decay of the correlations. One may check that for $h \neq 1/2$ (Exercise 8)

$$c_n \sim \sigma_0^2 h(2h-1)n^{2h-2} \quad \text{as } n \to \infty.$$

In particular, while $c_n \to 0$ as $n \to \infty$ for any value of $h \in (0,1)$, the series $\sum_{n=0}^{\infty} |c_n|$ diverges for $1/2 < h < 1$. Note also that the correlations are *positive*[4] in the case $1/2 < h < 1$. For future reference, let us also note that the fractional Brownian motion has a natural extension backward in time to a process indexed by $t \in \mathbb{R}$, defined via the Kolmogorov extension theorem, as a mean-zero Gaussian process with covariance function defined by (2.5).

The absolute summability of the covariance function is more generally related to the absolute continuity of the spectral distribution $F(d\lambda)$ and associated continuous positive density as follows: the proof is left as Exercise 15. However it is not used in the remainder of this chapter.

Proposition 2.4. Let $F(d\lambda)$ denote the spectral measure of a covariance function $\{c_n : n \in \mathbb{Z}\}$ such that $\sum_{n\in\mathbb{Z}} |c_n| < \infty$. Then $F(d\lambda)$ is absolutely continuous with respect to Lebesgue measure with a continuous positive density given by

$$f(\lambda) = \sum_{n\in\mathbb{Z}} c_n e^{in\lambda}, \quad -\pi \le \lambda \le \pi.$$

In this context the following sample statistic provides a natural quantity associated with the spectral distribution.

Definition 2.3. For a positive integer N, the N-sample *periodogram* of a discrete parameter real-valued weakly stationary process $\{X_n : n \in \mathbb{Z}\}$ is defined for frequencies of the form $\lambda_j = \frac{2\pi j}{N} \in [-\pi, \pi]$, $j \in \mathbb{Z}$, by

$$S_N(\lambda_j) = \frac{1}{N}|\sum_{n=1}^{N} X_n e^{in\lambda_j}|^2.$$

For $\lambda \in [-\pi, \pi]$, consider the partition of $[-\pi, \pi]$ into subintervals of lengths $2\pi/N$

[4] According to a celebrated Theorem 23.9 in Chapter 23, this also provides an example of a special type of *associated* statistical dependence.

$$S_N(\lambda) = \begin{cases} S_N(\lambda_j) \text{ if } \frac{2\pi j}{N} - \frac{\pi}{N} < \lambda \le \frac{2\pi j}{N} + \frac{\pi}{N}, 0 \le \lambda \le \pi, \\ S_N(-\lambda) \text{ if } -\pi \le \lambda < 0. \end{cases}$$

Observe that

$$S_N(0) = N|\frac{1}{N}\sum_{n=1}^{N} X_n|^2.$$

Moreover, since for $\lambda_j \neq 0$ one has for the simple (geometric) sum

$$\sum_{n=1}^{N} e^{in\lambda_j} = \sum_{n=1}^{N} e^{-in\lambda_j} = 0,$$

it follows that the definition of $S_N(\lambda_j)$ is unchanged by centering each term X_n of the sum by $X_n - \frac{1}{N}\sum_{k=1}^{N} X_k$. For that matter, one may replace each term X_n by $X_n - m$ without changing the value of $S_N(\lambda_j)$, and therefore $S_N(\lambda), \lambda \neq 0$. In particular, the following form is convenient for computations:

$$S_N(\lambda_j) = \sum_{|n|<N} \frac{1}{N} \sum_{k=1}^{N-|n|} \mathbb{E}(X_k - m)(X_{k+|n|} - m)e^{in\lambda_j}.$$

Corollary 2.5. Suppose that $\{X_n : n \in \mathbb{Z}\}$ is a weakly stationary process with mean m and absolutely summable covariance function $\{c_n : n \in \mathbb{Z}\}$ and spectral density f. Then, in the limit as $N \to \infty$, one has

$$\mathbb{E}S_N(0) - Nm^2 \to f(0),$$

$$\mathbb{E}S_N(\lambda) \to f(\lambda) \quad \lambda \neq 0.$$

If $m = 0$, then the second limit holds uniformly for all $\lambda \in [-\pi, \pi]$.

Proof. Note that

$$\begin{aligned} \lim_{N\to\infty}\left(\mathbb{E}S_N(0) - Nm^2\right) &= \lim_{N\to\infty} N \operatorname{Var}\left(\frac{1}{N}\sum_{k=1}^{N} X_k\right) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{|n|<N} (N - |n|)c_n \quad (2.6) \\ &= \sum_{n\in\mathbb{Z}} c_n = f(0), \quad (2.7) \end{aligned}$$

where the last limit follows, for example, by the dominated convergence theorem since $\sum_{n\in\mathbb{Z}} |c_n| < \infty$. For $0 < \lambda \leq \pi$, choose the smallest positive integer $j_N(\lambda) = j$ to minimize $|\lambda - \lambda_j|$, and define $j_N(\lambda) = -j_N(-\lambda)$ for $-\pi \leq \lambda < 0$. Then

$$\begin{aligned}\mathbb{E}S_N(\lambda) &= \mathbb{E}S_N(\lambda_{j_N(\lambda)}) \\ &= \sum_{|n|<N} \frac{1}{N} \sum_{k=1}^{N-|n|} \mathbb{E}(X_k - m)(X_{k+|n|} - m)e^{in\lambda_{j_N(\lambda)}} \qquad (2.8) \\ &= \sum_{|n|<N} \left(1 - \frac{|n|}{N}\right) c_n e^{in\lambda_{j_N(\lambda)}}.\end{aligned}$$

Now, since $\sum_{n\in\mathbb{Z}} |c_n| < \infty$, it follows, for example, using dominated convergence, that $\sum_{|n|<N} \left(1 - \frac{|n|}{N}\right) c_n e^{in\lambda} \to f(\lambda)$ uniformly. Thus, since $\lambda_{j_N(\lambda)} \to \lambda$, it follows that $\mathbb{E}S_N(\lambda) \to f(\lambda)$. The uniform convergence holds for all $\lambda \in [-\pi, \pi]$ in the case $m = 0$ since the convergence $\lambda_{j_N(\lambda)} \to \lambda$ as $N \to \infty$ is uniform, and the continuous function f on the compact set $[-\pi, \pi]$ must be uniformly continuous. ■

Remark 2.5. Unfortunately, the variance of $S_N(\lambda)$ does not go to zero as $N \to \infty$ and, indeed, $S_N(\lambda)$ is not a consistent estimate of $f(\lambda)$, as shown in Grenander (1981, Theorem 1,p. 362). However, by smoothing it by convolution with a continuous symmetric density $\gamma_N(\lambda)$ which converges slowly to δ_0 as $N \to \infty$, a consistent estimate can be obtained (Grenander 1981, Theorem 2, p. 367, and Brockwell and Davis 1991, pp. 350–354). This useful result is known as the Wiener–Khinchin theorem.

The following related proposition provides an alternative to the definition for checking positive definiteness of an absolutely summable sequence of complex numbers (see Exercise 5(c) for an application).

Proposition 2.6. *A sequence* $\{c_n : n \in \mathbb{Z}\}$ *of complex numbers such that* $\sum_{n\in\mathbb{Z}} |c_n| < \infty$ *is positive definite if and only if* $\sum_{n\in\mathbb{Z}} c_n e^{in\lambda}$ *is a positive (real) number for each* $\lambda \in [-\pi, \pi]$.

Proof. If $f(\lambda) = \sum_{n\in\mathbb{Z}} c_n e^{in\lambda} > 0$ for each $\lambda \in [-\pi, \pi]$, then positive definiteness follows from Herglotz' theorem, since $\{c_n : n \in \mathbb{Z}\}$ are the Fourier coefficients of the continuous periodic function f. Conversely, if $\{c_n : n \in \mathbb{Z}\}$ is positive definite and summable, then $f(\lambda)$ is the density of a positive measure $F(d\lambda)$ on the circle. ■

The fractional Brownian motion process is *self-similar* with exponent h in the sense that for any $t_1, \ldots, t_m \in \mathbb{R}$ $(m \geq 1)$, and $\lambda > 0$, the distributions of $(X_{\lambda t_1}, \ldots, X_{\lambda t_m})$ and $(\lambda^h X_{t_1}, \ldots, \lambda^h X_{t_m})$ coincide (see Example 5). The non-stationarity of fractional Brownian motion is intimately tied to its self-similarity and sample path regularity[5] in a manner which is made clear as follows. First we introduce a very mild regularity condition, e.g., satisfied by the Poisson process for which the sample paths are (discontinuous) step functions (Exercise 12).

Definition 2.4. A complex-valued stochastic process $\{X_t : t \in \mathbb{R}\}$ is said to have *stochastically continuous* sample paths if for each $t \in \mathbb{R}$, $X_{t_n} \to X_t$ in probability whenever $t_n \to t$ ($t_n \in \mathbb{R}, n \geq 1$) as $n \to \infty$.

Definition 2.5. A complex-valued stochastic process $\{X_t : t \in \mathbb{R}\}$ is said to be *self-similar* with exponent h if for any $t_1, \ldots, t_m \in \mathbb{R}$ ($m \geq 1$), and $\lambda > 0$, the distributions of $(X_{\lambda t_1}, \ldots, X_{\lambda t_m})$ and $(\lambda^h X_{t_1}, \ldots, \lambda^h X_{t_m})$ coincide.

Proposition 2.7. If $X = \{X_t : t \in \mathbb{R}\}$ is a stationary and self-similar stochastic process with stochastically continuous sample paths, then X is a.s. constant.

Proof. By stationarity, $X_{nt} = X_{t+(n-1)t}$ and X_t have the same distribution for each $n = 1, 2, \ldots$. By self-similarity, therefore, $n^h X_t$ and X_t have the same distribution. Thus $h = 0$. Thus $X_{\frac{1}{n}t}$ and X_t have the same distribution for each n. But $X_{\frac{1}{n}t} \to X_0$ as $n \to \infty$ by stochastic continuity. Since limits in probability are a.s. unique, it follows that for each $t \in \mathbb{R}$, one has $X_t = X_0$ with probability one. ■

Remark 2.6. The definitions of stationarity (translation invariance), self-similarity, and stochastic continuity readily generalize to complex-valued *random fields* $\{X_t : t \in \mathbb{R}^k\}$ indexed by $T = R^k$. The proof of the above proposition readily extends to this setting as well.

In the full generality of $T = \mathbb{Z}$ and $S = \mathbb{C}$, Herglotz' theorem also provides a representation of covariance functions of complex-valued weakly stationary processes or of complex-valued (strictly) stationary Gaussian process, on $\mathbb{Z}$. However here one must adopt a convention to resolve *nonuniqueness* of the processes associated with a given spectral measure through the covariance function. To see this, let $\{X_n = U_n + iV_n : n \in \mathbb{Z}\}$ be a weakly stationary process with $\mathbb{E}X_n = \theta + i\eta$ ($\mathbb{E}U_n = \theta$, $\mathbb{E}V_n = \eta$), and $\alpha_j = \mathrm{Cov}(U_0, U_j)$, $\beta_j = \mathrm{Cov}(V_0, V_j)$, $\delta_j = \mathrm{Cov}(V_0, U_j)$, $\gamma_j = \mathrm{Cov}(U_0, V_j)$. Then

$$\begin{aligned} c_j &= \mathrm{Cov}(X_n, X_{n+j}) = \mathrm{Cov}(X_0, X_j) \\ &= \mathrm{Cov}(U_0, U_j) + \mathrm{Cov}(V_0, V_j) + i\{\mathrm{Cov}(U_j, V_0) - \mathrm{Cov}(U_0, V_j)\} \\ &= \alpha_j + \beta_j + i(\delta_j - \gamma_j). \end{aligned} \tag{2.9}$$

The property $c_{-j} = \overline{c}_j$ yields

$$\alpha_{-j} + \beta_{-j} = \alpha_j + \beta_j, \quad \delta_{-j} - \gamma_{-j} = -\delta_j + \gamma_j \quad (j \in \mathbb{Z}). \tag{2.10}$$

[5] Stationary (translation invariant) and self-similar phenomena are of rather widespread interest in the sciences, especially in connection with critical phenomena. Thus the stochastic models for such phenomena typically involve random measures and/or generalized functions.

Since $c_j = a_j + ib_j$, say, has only two parameters, while the right side in (2.9) has four, it is clear that the correspondence between the Fourier coefficients $\{c_j : j \in \mathbb{Z}\}$ of a spectral measure and the covariance function is unfortunately not one-to-one in this case (Exercises 10 and 11). To make the correspondence one-to-one, one generally restricts attention to those processes for which one has

$$\begin{aligned} \alpha_j = \beta_j = a_j/2 \quad &(\text{i.e., } \operatorname{Cov}(U_0, U_j) = \operatorname{Cov}(V_0, V_j)), \\ \delta_j = -\gamma_j = b_j/2 \quad &(\text{i.e., } \operatorname{Cov}(V_0, U_j) = -\operatorname{Cov}(U_0, V_j)). \end{aligned} \tag{2.11}$$

With these somewhat arbitrary restrictions, we have the following consequence of Herglotz' theorem.

Theorem 2.8.

(a) There is a one-to-one correspondence between the set of all finite measures F on the unit circle (or, F on $[-\pi, \pi]$ with $F(\{-\pi\}) = F(\{\pi\})$) and the set of all covariance sequences of complex-valued weakly stationary processes $\{X_n : n \in \mathbb{Z}\}$ satisfying (2.11), with the covariances given by Fourier coefficients of F:

$$c_j \equiv \operatorname{Cov}(X_n, X_{n+j}) = \frac{1}{2\pi} \int_{[-\pi,\pi]} e^{-ij\lambda} F(d\lambda) \quad (j \in \mathbb{Z}). \tag{2.12}$$

(b) There is a one-to-one correspondence between complex zero-mean stationary Gaussian sequences, satisfying ((2.11)), and finite (spectral) measures on $[-\pi, \pi]$.

Remark 2.7. It is important to note that even for a symmetric measure $F \neq 0$ on $[-\pi, \pi]$, the complex-valued (weakly) stationary process $\{X_n : n \in \mathbb{Z}\}$ satisfying (2.11) has a nonzero imaginary part, as well as a nonzero real part. For, by (2.11), $\operatorname{Var}(U_j) = \operatorname{Var}(V_j) = c_0/2 > 0$ (also see Exercise 11). Thus, except for trivial sequences $X_n = \theta$ for all n (for some $\theta \in \mathbb{R}$), corresponding to $F = 0$, no (weakly) stationary process considered in Theorem 2.8 is real-valued. Theorem 2.2, therefore, provides a different representation from Theorem 2.8 when the spectral measure is symmetric. However, the spectral representation of a process in the next chapter depends on the process itself and is unaffected by this apparent ambiguity.

Example 6 *(Dirac Point Mass Spectral Measure).* Fix $\lambda_0 \in [-\pi, \pi], b > 0$, and consider the spectral measure $F(d\lambda) = b\delta_{\lambda_0}$. Then $c_n = \frac{b}{2\pi} e^{-i\lambda_0 n}$. Just as in the case $\lambda_0 = 0$, one arrives at $\mathbb{E}|X_n - e^{-in\lambda_0} X_0|^2 = 0$, for all $n \in \mathbb{Z}$, where $X_j = U_j + iV_j$ such that, under the uniqueness convention, $\operatorname{Cov}(U_0, U_j) = \operatorname{Cov}(V_0, V_j) = \frac{b}{4\pi} \cos(\lambda_0 j)$, and $\operatorname{Cov}(U_0, V_j) = -\operatorname{Cov}(V_0, U_j) = \frac{b}{4\pi} \sin(\lambda_0 j)$. Notice that the process $X_n = e^{-in\lambda_0} X_0$, $n \in \mathbb{Z}$, is determined by its value at a single value of n.

Now let us consider the continuous parameter case. In view of Bochner's theorem,[6] Theorem 2.8 extends to continuous parameter weakly stationary processes. The continuity requirement in Bochner's theorem is met by the following rather mild condition.

Definition 2.6. Suppose T is an interval in $\mathbb{R}$. A complex-valued stochastic process $\{X_t : t \in T\}$ is *mean-square continuous* at $t \in T$ if

$$X_s \xrightarrow{L^2} X_t \quad \text{as} \quad s \longrightarrow t. \tag{2.13}$$

A process is said to be *mean-square continuous* if it is mean-square continuous at each $t \in T$.

Remark 2.8. One may easily check that since the increment $X_t - X_s, s < t$, of a Poisson process has mean and variance proportional to $t - s$, the process is mean-square continuous (Exercise 12). Notice that for any $\epsilon > 0$, by Chebyshev's inequality,

$$P(|X_t - X_s| > \epsilon) \leq \frac{\mathbb{E}|X_t - X_s|^2}{\epsilon^2} = \frac{\|X_t - X_s\|^2}{\epsilon^2},$$

so that mean-square continuity easily implies stochastic continuity.

Lemma 1. Let $\{X_t : t \in T\}$ be a mean-zero square-integrable complex-valued process on an interval $T \subset \mathbb{R}$, finite or infinite. Let $r(s, t) = \mathbb{E}X_x\overline{X_t}$ $(s, t \in T)$ be its covariance function. Then $(s, t) \to r(s, t)$ is continuous on $T \times T$ if and only if (2.13) holds.

Proof. (*Sufficiency*). Suppose (2.12) holds. Then, writing $< Y, Z >= \mathbb{E}Y\overline{Z}$, $\|Y\|^2 = \mathbb{E}Y\overline{Y} = \mathbb{E}|Y|^2$, by adding and subtracting terms and using the Cauchy–Schwarz inequality, one has

$$\begin{aligned}
&|r(s + h_1, t + h_2) - r(s, t)| \\
&\quad = | < X_{s+h_1}, X_{t+h_2} > - < X_s, X_t > | \\
&\quad \leq | < X_{s+h_1}, X_{t+h_2} > - < X_{s+h_1}, X_t > | + | < X_{s+h_1}, X_t > - < X_s, X_t > | \\
&\quad \leq \|X_{s+h_1}\| \cdot \|X_{t+h_2} - X_t\| + \|X_t\| \cdot \|X_{s+h_1} - X_s\| \to 0 \text{ as} h_1 \to 0,\ h_2 \to 0,
\end{aligned}$$

since $\|X_{s+h_1} - X_s\|^2 = \mathbb{E}|X_{s+h_1} - X_s|^2 \to 0$ and $\|X_{t+h_2} - X_t\|^2 \to 0$.

(Necessity). Suppose $(s, t) \to r(s, t)$ is continuous. Fix $t \in T$. Then

$$\begin{aligned}
\mathbb{E}|X_{t+h} - X_t|^2 &= \mathbb{E}|X_{t+h}|^2 + \mathbb{E}|X_t|^2 - \mathbb{E}X_{t+h}\overline{X}_t - \mathbb{E}\overline{X}_{t+h}X_t \\
&= r(t + h, t + h) + r(t, t) - r(t + h, t) - r(t, t + h) \to 0
\end{aligned}$$

[6] See BCPT, p.119.

as $h \to 0$. ■

As an immediate consequence of the lemma, it follows that the *covariance function* $t \to r(t) = \mathrm{Cov}(X_0, X_t)$ of a weakly stationary process $\{X_t : t \in \mathbb{R}\}$ is continuous if and only if (2.13) holds. Applying Bochner's theorem, we get the following analogues of Theorems 2.2 and 2.8. For complex-valued weakly stationary processes $X_t = U_t + iV_t, t \in \mathbb{R}$, we assume the analogue of (2.11):

$$\mathrm{Cov}(U_0, U_t) = \mathrm{Cov}(V_0, V_t) = \frac{1}{2}\mathcal{R}er(t); \mathrm{Cov}(V_0, U_t) = -\mathrm{Cov}(U_0, V_t) = \frac{1}{2}\mathcal{I}mr(t), \tag{2.14}$$

where $\mathcal{R}er(t)$ and $\mathcal{I}mr(t)$ are the real and imaginary parts of $r(t)$, respectively.

Theorem 2.9.

(a) There is a one-to-one correspondence between the set of all covariance functions $r(\cdot)$ of real-valued weakly stationary processes $\{X_t : t \in \mathbb{R}\}$ satisfying (2.13) and the set of all finite symmetric measures F on $\mathbb{R}$, called the *spectral measure* of $\{X_t : t \in \mathbb{R}\}$, with $r(\cdot)$ being the Fourier transform of F:

$$r(t) = \int_{-\infty}^{\infty} e^{it\lambda} F(d\lambda) = \int_{-\infty}^{\infty} \cos(t\lambda) F(d\lambda), \quad (t \in \mathbb{R}). \tag{2.15}$$

(b) There is a one-to-one correspondence, via (2.15), between the set of all real-valued mean-zero stationary Gaussian processes $\{X_t : t \in \mathbb{R}\}$ satisfying (2.13) and the set of all finite symmetric measures on $\mathbb{R}$.

For complex-valued processes, one has the following.

Theorem 2.10. There is a one-to-one correspondence between the set of all covariance functions $r(\cdot)$ of complex-valued weakly stationary processes $\{X_t = U_t + iV_t : t \in \mathbb{R}\}$ satisfying (2.13) and (2.14) and the set of all finite measures F on $\mathbb{R}$, called the *spectral measure* of $\{X_t : t \in \mathbb{R}\}$, with $r(\cdot)$ being the Fourier transform of F, as given by the first equality in (2.15).

Remark 2.9. Once again (see Remark 2.7), one should note that Theorem 2.9 is not a special case of Theorem 2.10, since the latter only yields complex-valued processes with both real and imaginary parts nonzero, except in the trivial case $F = 0$.

Example 7. Let the spectral measure of a real-valued weakly stationary process be $F(d\lambda) = b\exp\{-c|\lambda|\}d\lambda$, for some $b > 0, c > 0$. Then, using the Fourier inversion formula or Cauchy's residue theorem from complex variables, or simply recalling the characteristic function of the Cauchy distribution, one has $r(s) = 2b[c(1 + s^2/c^2)]^{-1}$

Example 8. Let $F(d\lambda) = b[c(1 + \lambda^2/c^2)]^{-1}d\lambda$ $(b > 0, c > 0)$. Then $r(s) = b\pi e^{-c|s|}$. This follows by a direct evaluation of the integral or from the previous

Example 7 through the Fourier inversion formula.[7] In the case this process is real Gaussian, then, recalling the covariance structure in Example 2, it is referred to as an *Ornstein–Uhlenbeck process*.

Example 9. Let $F(d\lambda) = b\exp\{-c\lambda^2\}d\lambda$ $(b > 0, c > 0)$. Then $r(s) = (b\sqrt{\pi/c}\,)\exp\{-s^2/4c\}$.

Since finite dimensional distributions of mean-zero real-valued stationary Gaussian processes are completely determined by their covariance functions $r(s)$, and since in Examples 7–9 one can derive versions of these processes with continuous sample paths, the distributions of these processes on the space $C(-\infty,\infty)$ of continuous functions (paths) are entirely determined (Exercise 8). It is instructive to determine the corresponding mean-zero complex-valued stationary Gaussian processes satisfying (2.11) (Exercises 13–14).

Exercises

1. (*Trend Removal by Differencing*) Suppose that $X_t = m(t) + Y_t, t = \dots, -2, -1, 0, 1, 2, \dots$, where $m(t) = \sum_{j=0}^{k} m_j t^j$ is a polynomial trend of degree k and Y is a mean-zero stationary process. Define a *differencing operator* $\Delta f(t) = f(t) - f(t-1)$, with iterates defined iteratively by $\Delta^m f = \Delta^{m-1}\Delta f$. Show that $\Delta^k X$ is a stationary process with mean $k! m_k$.
2. (*Seasonality Removal by Lag-Differencing*) Suppose that $X_t = m(t) + s(t) + Y_t, t = \dots - 2, -1, 0, 1, 2, \dots$, where $m(t)$ is a trend and $s(t)$ a periodic component with period $d \geq 2$ and Y is a stationary process. Define a *lag-d differencing operator* Δ_d by $\Delta_d f(t) = f(t) - f(t-d)$. Show that $\Delta_d X_t = m_d(t) + \Delta_d Y_t$, where $m_d(t) = m(t) - m(t-d)$, and $\Delta_d Y$ is a stationary process.
3. Let $T = \mathbb{Z}$ or $\mathbb{R}$, and suppose $\{c_t : t \in T\}$ is positive definite in the sense (2.3).
   (a) Show that c_0 is real and nonnegative. [*Hint*: Take $n = 1, a_1 = 1$, t_1 arbitrary.]
   (b) Show that $c_{-t} = \overline{c}_t$ for all t. [*Hint*: First take $n = 2$, $a_1 = a_2 = 1$, $t_1 = 0$, $t_2 = t$, to prove that $\mathcal{I}mc_{-t} = -\mathcal{I}mc_t$. Then take $n = 2$, $a_1 = i$, $a_2 = 1$, $t_1 = 0$, $t_2 = t$, to show that $\mathcal{R}ec_{-t} = \mathcal{R}ec_t$.]
4. Show that, for any real-valued X_0 with mean-zero and finite variance, the process $X_n = (-1)^n X_0$, $n \in \mathbb{Z}$, is weakly stationary, and compute its covariance function and spectral measure.
5. Determine which of the following is a covariance function. If so, compute the corresponding spectral measure.

[7] See BCPT, p. 112.

(a) $c_n = \sin(n), n \in \mathbb{Z}$.
(b) $c_n = cos(n), n \in \mathbb{Z}$.
(c) $c_n = ba^{|n|}, n \in \mathbb{Z}$, for fixed $b > 0$, and $a \in (-1, 1)$.

6. Calculate the covariance function c_n for the spectral distributions on the unit circle $U = [-\pi, \pi]$ given by $F(d\lambda) = \frac{1}{2}\delta_{\pi/m}(d\lambda) + \frac{1}{2}\delta_{-\pi/m}(d\lambda)$, for a fixed positive integer m.
7. Suppose that $X = \{X_t : t \geq 0\}$ is a mean-zero Gaussian process starting at $X_0 \equiv 0$ having stationary increments with variance scaling as $\mathbb{E}X_t^2 = \sigma_0^2 t^{2h}$ for some exponent $h \in (0, 1)$. Show that X is distributed as fractional Brownian motion.
8. Show that for the fractional Gaussian noise with $h \neq 1/2$, one has $c_n \sim \sigma_0^2 h(2h - 1)n^{2h-2}$ as $n \to \infty$. [*Hint*: Factor out n^{2h} and use Taylor expansions for $(1 \pm x)^{2h}$, respectively.]
9. Compute the covariance function $\{c_n : n \in \mathbb{Z}\}$ corresponding to the spectral measure F having the triangular density $f(\lambda) = \pi - |\lambda|$ $(|\lambda| \leq \pi)$. [*Hint*: The triangular distribution on $[-\pi, \pi]$ is the convolution of two uniform distributions on $[-\pi/2, \pi/2]$.]
10. (a) Let $\{U_n : n \in \mathbb{Z}\}$ and $\{V_n : n \in \mathbb{Z}\}$ be both i.i.d. standard normal sequences independent of each other. Show that the complex-valued Gaussian sequence $X_n = U_n + iV_n$ $(n \in \mathbb{Z})$ satisfies the restrictions (2.11), and compute its spectral measure.
    (b) Let $\{(U_n, V_n) : n \in \mathbb{Z}\}$ be an i.i.d. two-dimensional Gaussian sequence with $\mathbb{E}U_n = \mathbb{E}V_n = 0$, $\mathbb{E}U_n^2 = \mathbb{E}V_n^2 = 1$, $\rho = \text{Cov}(U_n, V_n)$. Show that $\{X_n + iV_n : n \in \mathbb{Z}\}$ has the same spectral measure as in (a), no matter what ρ is $(-1 \leq \rho \leq 1)$, and that for $\rho \neq 0$, this process does not satisfy (2.11).
    (c) Which real stationary Gaussian process has the same spectral measure as in (a) and (b)?
11. (a) Let $\{U_n : n \in \mathbb{Z}\}$ and $\{V_n : n \in \mathbb{Z}\}$ be both i.i.d. standard normal and independent of each other. Define the stationary Gaussian sequence $X_n = U_n + iV_n + U_{n+1} + iV_{n+1}$ $(n \in \mathbb{Z})$. Compute the corresponding spectral measure.
    (b) Consider now a two-dimensional i.i.d. Gaussian sequence $\{(U_n, V_n) : n \in \mathbb{Z}\}$ with $\mathbb{E}U_n = \mathbb{E}V_n = 0$, $\mathbb{E}U_n^2 = \mathbb{E}V_n^2 = 1$, but $\rho \equiv \text{Cov}(U_n, V_n) \neq 0$. Show that the stationary process $X_n = U_n + iV_n + U_{n+1} + iV_{n+1}$ has the same spectral measure as in (a), but that the two Gaussian processes have different distributions.
    (c) Construct a real-valued stationary Gaussian process with the same spectral measure as in (a) and (b).
    (d) Show that of all the Gaussian processes in (a)–(c), only the one in (a) satisfies the restrictions (2.11).
12. Show that a homogeneous Poisson process is both stochastically continuous and mean-square continuous.

13. Use the Kolmogorov–Chentsov criterion[8] to prove that there exists a continuous version of a mean-zero (stationary) Gaussian process with $r(s)$ as specified in each of Examples 7–9.
14. For each of Examples 7–9, specify the distribution of a mean-zero stationary complex-valued Gaussian process $\{X_t = U_t + iV_t : t \in \mathbb{R}\}$ with the covariance function as specified, obeying the restrictions (2.14). [*Hint*: For each $n \geq 0$, specify the joint distribution of $(U_0, V_0, U_1, V_1, \ldots, U_n, V_n)$.]
15. (a) Suppose the spectral measure F on $[-\pi, \pi]$ (with $F(\{-\pi\}) = F(\{\pi\})$ is such that $\sum_{n\geq 1} |c_n| < \infty$. Show that F is absolutely continuous, with a version of the density which is continuous on $[-\pi, \pi]$. [*Hint*: Consider the function defined by the Fourier series $f(\lambda) = \sum_{n\in\mathbb{Z}} c_n e^{in\lambda}$, and compute $\frac{1}{2\pi} \int_{[-\pi,\pi]} e^{-in\lambda} f(\lambda) d\lambda$.]
    (b) Suppose the spectral measure F on $(-\infty, \infty)$ is such that $r(s)$ is integrable with respect to Lebesgue measure Show that F is absolutely continuous and has a (version of the) density which is uniformly continuous. [*Hint*: Consider Fourier transform in place of Fourier series.]
16. A stationary sequence $\{X_n : n \in \mathbb{Z}\}$ is said to be *time-reversible*[9] if for any integers m and $n_1, \ldots, n_m$, the finite dimensional distributions of $(X_{n_1}, \ldots, X_{n_m})$ and $(X_{-n_1}, \ldots, X_{-n_m})$ coincide.

    (a) Show that any stationary Gaussian sequence is time-reversible.
    (b) Assume $\{X_n : n \in \mathbb{Z}\}$ is stationary and suppose that $Y_n = g(X_n), n \in \mathbb{Z}$, where g is a one-to-one function. Show that $\{Y_n : n \in \mathbb{Z}\}$ is time-reversible if and only if $\{X_n : n \in \mathbb{Z}\}$ is time-reversible.

[8] See Bhattacharya and Waymire (2021) or BCPT, p.180.

[9] A metaphor for a stationary sequence is a process viewed as a data stream movie for which statistically it does not matter when one arrives at the theater. Time-reversibility permits it to also be run backward without changing the stochastic structure.

Chapter 3
Spectral Representation of Stationary Processes

The spectral representation of a stationary process or random field broadly refers to a stochastic Fourier representation in a mean-square sense.

Our goal in this chapter is to exploit the second order spectral structure to represent weakly stationary processes, as well as more general processes with finite second moments, as stochastic integrals with respect to orthogonal processes. The basic idea for such a representation is formally as follows. Suppose one wishes to represent a centered (mean-zero) weakly stationary process $X = \{X_t : t \in T\}$, say real-valued for purposes of illustration, as a superposition of sinusoidal elements $e^{it\lambda}$ with random amplitudes $Z(d\lambda)$ in the frequency range λ to $\lambda + d\lambda$, i.e.,

$$X_t = \int_\Lambda e^{it\lambda} Z(d\lambda), \qquad t \in T.$$

Then, from the previous chapter, one has (formally)

$$\int_\Lambda e^{it\lambda} \mu(d\lambda) = \mathbb{E} X_0 X_t = \int_\Lambda \int_\Lambda e^{it\lambda} \mathbb{E}\{Z(d\lambda) Z(d\lambda')\},$$

with

$$\Lambda = [-\pi, \pi], \quad \mu(d\lambda) = F(d\lambda)/2\pi$$

when $T = \mathbb{Z}$, and, assuming continuity of the covariance function,

$$\Lambda = \mathbb{R}, \quad \mu(d\lambda) = F(d\lambda),$$

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_3

when $T = \mathbb{R}$. In particular, one needs to introduce a suitable notion of the "stochastic integral" with respect to an additive random field $Z(d\lambda)$ with covariance structure $\mathbb{E}Z(d\lambda)Z(d\lambda')$ concentrated on the diagonal $\lambda = \lambda'$ with value $\mu(d\lambda)$. With this as the goal, we proceed as follows.

Definition 3.1. Let $(\Lambda, \mathcal{L}, \mu)$ be a σ-finite measure space. An *orthogonal random field* for this space is a family of complex-valued random variables

$$Z = \{Z(B) : B \in \mathcal{L}, \mu(B) := \mathbb{E}|Z(B)|^2 < \infty\}$$

defined on some probability space $(\Omega, \mathcal{F}, P)$, satisfying

(i) $\mathbb{E}Z(B) = 0$.
(ii) $Z(B_1 \cup B_2) = Z(B_1) + Z(B_2)$ (a.s.), if $B_1 \cap B_2 = \emptyset$.
(iii) $(Z(B_1), Z(B_2)) := \mathbb{E}Z(B_1)\overline{Z}(B_2) = \mu(B_1 \cap B_2)$.

Let $L^2(Z)$ denote the closure (in the complex Hilbert space $L^2(\Omega, \mathcal{F}, P)$) of the space of all finite linear combinations (with complex coefficients) of the random variables $Z(B)$, $B \in \mathcal{L}$. So, in particular, elements of $L^2(Z)$ are (equivalence classes of) random variables on Ω obtained as mean-square limits of "integral sums" of the form $\sum_{j=1}^m a_j Z(B_j)$, equipped with the inner product $(\sum_1^m a_j Z(B_j), \sum_1^n b_k Z(C_k)) = \sum_{j,k} a_j \overline{b}_k \mu(B_j \cap C_k)$. Extended by taking limits (in $L^2(\Omega, \mathcal{F}, P)$), the resulting space $L^2(Z)$ is a complex Hilbert space.

We will use $L^2(\mu)$ to denote the Hilbert space of complex-valued square integrable (with respect to μ) functions on Λ with the inner product

$$\langle f, g\rangle := \int_\Lambda f\overline{g}d\mu.$$

By a *simple function* in $L^2(\mu)$, we mean a function of the form

$$f = \sum_1^m a_j \mathbf{1}_{B_j},$$

with $\mu(B_j) < \infty$ for all j, B_j's pairwise disjoint, and $a_1, \ldots, a_m$ complex numbers $(m \geq 1)$.

Theorem 3.1. The map φ defined on the class of all simple functions on Λ, by

$$\varphi\left(\sum_1^m a_j \mathbf{1}_{B_j}\right) = \sum_1^m a_j Z(B_j), \tag{3.1}$$

extends to a linear isometry of $L^2(\mu)$ onto $L^2(Z)$:

$$(\varphi(f), \varphi(g)) = \langle f, g\rangle, \qquad f, g \in L^2(\mu). \tag{3.2}$$

Proof. If $f = \sum_1^m a_j \mathbf{1}_{B_j}$ and $g = \sum_1^n b_k \mathbf{1}_{C_k}$ are two simple functions, then

$$(\varphi(f), \varphi(g)) = \sum_{j,k} a_j \bar{b}_k \mu(B_j \cap C_k) = \langle f, g \rangle. \tag{3.3}$$

Since simple functions are dense in $L^2(\mu)$, for any given f in $L^2(\mu)$, there exists a sequence f_n $(n \geq 1)$ of simple functions such that $f_n \to f$ in $L^2(\mu)$. Define $\varphi(f)$ to be the limit in $L^2(\Omega, \mathcal{F}, P)$ of $\varphi(f_n)$. This limit is well defined, and it is simple to check (3.2) for this extension. ■

We will write φ for the extension (of the map (3.1)) in Theorem 3.1. From here on, for an orthogonal random field $\{Z(B) : B \in \mathcal{L}, \mu(B) := \mathbb{E}|Z(B)|^2 < \infty\}$, we will also define

$$\int_\Lambda f dZ \equiv \int_\Lambda f(\lambda) Z(d\lambda) := \varphi(f) \qquad \text{for all } f \in L^2(\mu), \tag{3.4}$$

i.e., as an $L^2(\Omega, \mathcal{F}, P)$-limiting form of (3.1). Thus, from this point of view, one has the following:

Proposition 3.2. The stochastic integral defined by

$$f \to \int_\Lambda f(\lambda) Z(d\lambda),\, f \in L^2(\mu)$$

is a linear isometry between the Hilbert spaces $L^2(\mu)$ and $L^2(Z)$ for a given orthogonal random field Z.

It is sometimes also convenient to denote the orthogonal random field $Z = \{Z(B) : B \in \mathcal{L}\}$ simply by $Z(d\lambda)$ in this context.

The following general result will be used repeatedly in this chapter to obtain a variety of representations. The spectral representations are important special cases, where the essential idea for obtaining the orthogonal random field $Z(d\lambda)$, i.e., $\{Z(B) : B \in \mathcal{L}\}$, from the process $X = \{X_t : t \in t\}$ is to first use the spectral representation of the covariance function to define an isometry φ between $L^2(\mu)$ and the Hilbert space $L^2(X)$ which is the closure in $L^2(\Omega, \mathcal{F}, P)$ of the linear span of $\{X_t : t \in T\}$. In particular,

$$X_t = \varphi(f)$$

for a suitable $f \in L^2(\mu)$ (depending on t). On the other hand, owing to the spectral representation of the covariance function, by defining

$$Z(B) = \varphi(\mathbf{1}_B),\, B \in \mathcal{L},$$

such an isometry φ will induce an isometry between $L^2(\mu)$ and $L^2(Z)$. The isometry so obtained, therefore, both defines the integral

$$\int_{\Lambda} f(\lambda)Z(d\lambda) = \varphi(f),\ f \in L^2(\mu),$$

and links it to X_t for a suitably chosen $f \in L^2(\mu)$ (depending on t).

Theorem 3.3 *(General Orthogonal Representation of Processes).* Let $\{X_t : t \in T\}$ be a mean-zero real- or complex-valued process with finite second moments on $T = \mathbb{Z}$, or $\mathbb{Z}^+$, or a finite or infinite subinterval of $\mathbb{R}$. Assume there exists a family of functions $\psi(t,\cdot) \in L^2(\mu)$, for all $t \in T$, such that

$$r(s,t) \equiv \operatorname{Cov}(X_s, X_t) = \int_{\Lambda} \psi(s,\lambda)\overline{\psi(t,\lambda)}\mu(d\lambda) \quad \text{for all } s,t, \in T. \tag{3.5}$$

Assume further that $L^2(\mu)$ is separable. Then there exists an orthogonal random field $Z(\cdot)$ for $L^2(\mu)$ such that

$$X_t = \int_{\Lambda} \psi(t,\lambda)Z(d\lambda) \qquad \text{for all } t \in T. \tag{3.6}$$

Proof. Define $\varphi(\psi(t,\cdot)) = X_t, t \in T$. Let $L^2(X)$ be the closure (in $L^2(\Omega, \mathcal{F}, P)$) of the linear span of the random variables X_t $(t \in T)$. For functions of the form $f = \sum_1^m a_j\psi(s_j,\cdot)$, $s_j \in T$, *define* $\varphi(f) = \sum_1^m a_j X_{s_j}$. If $f = \sum_1^m a_j\psi(s_j,\cdot)$, $g = \sum_1^m b_k\psi(t_k,\cdot)$, then, denoting by $(\cdot,\cdot)$ the inner product in $L^2(X)$, one has, by (3.5),

$$\begin{aligned}(\varphi(f),\varphi(g)) &= \sum_{j,k} a_j\bar{b}_k r(s_j,t_k) = \sum_{j,k} a_j\bar{b}_k \int_{\Lambda} \psi(s_j,\lambda)\overline{\psi(t_k,\lambda)}\mu(d\lambda)\\ &= \int_{\Lambda} f(\lambda)\overline{g(\lambda)}\mu(d\lambda) = \langle f,g\rangle. \end{aligned}\tag{3.7}$$

Suppose the set G of functions f (or g) of the above form is dense in $L^2(\mu)$. Then one can extend the map φ to a unique linear isometry on $L^2(\mu)$ onto the Hilbert space $L^2(X)$. Now define

$$Z(B) = \varphi(\mathbf{1}_B) \qquad (\mu(B) < \infty). \tag{3.8}$$

Clearly,

(i) $\mathbb{E}Z(B) = 0$.
(ii) If $B_1 \cap B_2 = \emptyset$, then

$$Z(B_1 \cup B_2) \equiv \varphi(\mathbf{1}_{B_1 \cup B_2}) = \varphi(\mathbf{1}_{B_1} + \mathbf{1}_{B_2}) = \varphi(\mathbf{1}_{B_1}) + \varphi(\mathbf{1}_{B_2}) = Z(B_1) + Z(B_2).$$

(iii) $\mathbb{E}Z(B_1)\overline{Z(B_2)} = (\varphi(\mathbf{1}_{B_1}), \varphi(\mathbf{1}_{B_2})) = \langle \mathbf{1}_{B_1}, \mathbf{1}_{B_2} \rangle = \mu(B_1 \cap B_2)$ for all B_1, B_2 with $\mu(B_j) < \infty$ $(j = 1, 2)$.

Since $\varphi(\psi(t, \cdot)) = X_t$, by definition of φ, (3.6) follows (as a limiting form of (3.1)).

Consider now the case when the set G of functions of the form $\sum_1^m a_j \psi(s_j, \cdot)$ $(m \geq 1)$ is not dense in $L^2(\mu)$. Separability of $L^2(\mu)$ is equivalent to the existence of a countable orthonormal basis since any countable dense subset yields an orthonormal basis by the Gram–Schmidt process[1] and, conversely, any countable orthonormal basis yields a countable dense subset by taking finite linear combinations with coefficients from a countable dense subset of $\mathbb{C}$. Let $h_{t'_n}$ $(n = 1, 2, \ldots)$ be a complete orthonormal sequence (finite or infinite, as the case may be) for the Hilbert space $G^\perp$ of functions orthogonal to G. The index set $T' = \{t'_1, t'_2, \cdots\}$ is chosen to be disjoint from T. Let $Y_{t'_n}$ $(n = 1, 2, \ldots)$ be real-valued i.i.d. standard normal random variables independent of $\{X_t : t \in T\}$ (constructed, if necessary, by enlarging $(\Omega, \mathcal{F}, P)$). First let

$$\begin{aligned} \varphi(h_{t'_n}) &= Y_{t'_n} \qquad (n = 1, 2, \cdots), \\ \varphi\left(\sum_{n=1}^m a_n h_{t'_n}\right) &= \sum_{n=1}^m a_n Y_{t'_n}, \end{aligned} \tag{3.9}$$

noting that if $f_1 = \sum_{n=1}^m a_n h_{t'_n}$, $f_2 = \sum_{n=1}^k b_n h_{t'_n}$, then

$$(\varphi(f_1), \varphi(f_2)) = \mathbb{E}\left(\sum_{n=1}^m a_n Y_{t'_n} \cdot \sum_{n=1}^k \overline{b}_n Y_{t'_n}\right) = \sum_{n,n'} a_n \overline{b}_{n'} \delta_{nn'} = \langle f_1, f_2 \rangle. \tag{3.10}$$

Now extend φ to $G^\perp$ by taking limits in L^2, using the isometry between $L^2(G^\perp, \mu)$ and the subspace spanned by linear combinations of $Y_{t'_n}$ $(n = 1, 2, \ldots)$, which follows from (3.10). Finally, for an arbitrary $h \in L^2(\mu)$, let $h = f + g$ be the unique decomposition of h with $f \in \overline{G}$ and $g \in \overline{G}^c = G^c$, and define

$$\varphi(h) = \varphi(f) + \varphi(g). \tag{3.11}$$

It is simple to check, by taking limits and using (3.7) and (3.10), and using the orthogonality $(\varphi(f), \varphi(g)) = 0 = \langle f, g \rangle$ if $f \in \overline{G}$ and $g \in G^\perp$, that

$$(\varphi(f), \varphi(g)) = \langle f, g \rangle \qquad \text{for all } f, g \in \overline{G} \oplus G^\perp = L^2(\mu). \tag{3.12}$$

[1] See BCPT p. 249.

Now apply the first part of the proof to the family of random variables $\{X_t : t \in T \cup T'\}$ with $X_{t'_n} := Y_{t'_n}$ noting that

$$r(s,t) = \begin{cases} \int \psi(s,\lambda)\overline{\psi(t,\lambda)}\mu(d\lambda) & \text{if } s,t \in T \\ 0 & \text{if } s \in T, t \in T' \text{ or vice versa} \\ \delta_{st} = \int h_s(\lambda)\overline{h_t(\lambda)}\mu(d\lambda) & \text{if } s,t \in T'. \end{cases} \tag{3.13}$$

The hypothesis of the theorem then holds with an extended family given by $\widetilde{\psi}(s,\lambda) = \psi(s,\lambda)$ if $s \in T$ and $\widetilde{\varphi}(s,\lambda) = h_s(\lambda)$ if $s \in T'$, with the linear span of $\{\widetilde{\psi}(s,\lambda) : s \in T \cup T'\}$ dense in $L^2(\mu)$. So the orthogonal random field $Z(\cdot)$ is constructed by (3.8) with the extension of φ defined by (3.9) and (3.11). ■

Definition 3.2. For a given symmetric positive definite complex-valued function $r(s,t), s,t \in T \subset \mathbb{R}$, a Hilbert space $H(r)$ of functions on T such that for all $t \in T$, one has (a) $r(t,\cdot) \in H(r)$, and (b) $(f, r(t,\cdot))_{H(r)} = f(t), f \in H(r), t \in T$, is called a *reproducing kernel Hilbert space*.

Property (b) of the definition says that the pointwise evaluation of f at t is via a bounded linear functional, i.e., as an inner product with $r(t,\cdot)$. In this context, $r(t,\cdot)$ is referred to as the *representative evaluator* at t. This function is also called a *reproducing kernel* for the Hilbert space. Applying (b) to $r(s,\cdot)$, one sees that

$$r(s,t) = \langle r(s,\cdot), r(t,\cdot)\rangle, \quad s,t \in T.$$

This is the origin of the terminology "reproducing kernel."

The Moore–Aronszajn[2] theorem asserts the existence of a unique reproducing Hilbert space associated with a positive definite function $r(s,t)$ and conversely. In particular, it can be stated as follows.

Theorem 3.4 *(Moore–Aronszajn).* The function $r(s,t)$ is positive definite if and only if it is a reproducing kernel. In particular, the reproducing Hilbert space $H(r)$ is uniquely determined by (a) and (b), as the closure of the set of functions of the form $f(s) = \sum_{j=1}^{n} a_j r(t_j, s), s \in T$, in the norm $\|f\|_{H(r)}^2 := \sum_{k,j=1}^{n} a_j r(t_j, t_k)\bar{a}_k$ induced by the inner product $r(s,t) = \langle r(s,\cdot), r(t,\cdot)\rangle$.

Although explicit use is not being made of this terminology, the previous Theorem 3.3 may be viewed as an element of that framework with $\Lambda = T$, $\psi(s,t) = r(s,t)$.

Theorems 3.5 and 3.6 below are immediate consequences of Theorem 3.3.

Theorem 3.5 *(Spectral Representation of Weakly Stationary Processes on $\mathbb{Z}$).* Let $\{X_n : n \in \mathbb{Z}\}$ be a real- or complex-valued mean-zero weakly stationary process with spectral measure $F(d\lambda)$ on $\Lambda = U \equiv [-\pi, \pi]$ $(F(\{-\pi\}) = F(\{\pi\}))$. Then

[2] Aronszajn (1950).

there exists an orthogonal random field $Z(\cdot)$ for $L^2(\mu)$, with $\mu(d\lambda) := F(d\lambda)/2\pi$, such that

$$X_n = \int_{[-\pi,\pi]} e^{in\lambda} Z(d\lambda) \qquad \text{for all } n \in \mathbb{Z}. \tag{3.14}$$

Proof. The hypotheses of Theorem 3.3 are satisfied with $T = \mathbb{Z}$, $\psi(t,\lambda) = e^{it\lambda}$ $(t \in \mathbb{Z})$ and $\mu(d\lambda) = F(d\lambda)/2\pi$ on $[-\pi,\pi] = \Lambda$ (see (2.12)). Here $\varphi(e^{in\lambda}) = X_n$ $(n \in \mathbb{Z})$ extends to a linear isometry between $L^2(\mu)$. and $L^2(X)$. ■

Theorem 3.6 *(Cramér's Spectral Representation of Weakly Stationary Processes on* $\mathbb{R}$*).* Let $\{X_t : t \in \mathbb{R}\}$ be a real- or complex-valued weakly stationary process, with continuous covariance function and with spectral measure F on $\Lambda = \mathbb{R}$. Then there exists an orthogonal random field $Z(\cdot)$ for $L^2(F)$ such that

$$X_t = \int_{\mathbb{R}} e^{it\lambda} Z(d\lambda) \qquad \text{for all } t \in \mathbb{R}. \tag{3.15}$$

Proof. In Theorem 3.3, take $T = \mathbb{R}$, $\mu = F$, $\psi(t,\lambda) = e^{-it\lambda}$ $(t \in \mathbb{R})$. Also, $\varphi(e^{-it\lambda}) = X_t$ $(t \in \mathbb{R})$ extends to an isometry between $L^2(\mu)$ and $L^2(X)$. ■

Definition 3.3. A stochastic process $\{X_t : t \in T\}$, which admits a representation of the form in Theorem 3.5 in the case $T = \mathbb{Z}$ or Theorem 3.6 in the case $T = \mathbb{R}$, is said to be *(weakly) harmonizable*.

Example 1 *(Point Mass Spectral Distributions).* It is interesting to consider the case in which the spectral distribution has a point mass, i.e., jump discontinuity at λ_0, say $b = F(\{\lambda_0\}) > 0$. In the case $T = \mathbb{Z}$, one has

$$X_n = \int_{[-\pi,\pi]\backslash\{\lambda_0\}} e^{in\lambda} Z(d\lambda) + \left(Z(\lambda_0) - Z(\lambda_0^-)\right) e^{in\lambda_0}, \quad n \in \mathbb{Z}.$$

Moreover,

$$\operatorname{Var}\left(Z(\lambda_0) - Z(\lambda_0^-)\right) = F(\{\lambda_0\}) = b.$$

The process $Y_n = \left(Z(\lambda_0) - Z(\lambda_0^-)\right) e^{in\lambda_0}, \quad n \in \mathbb{Z}$, is a sinusoidal process of frequency λ_0. Moreover, it is a *deterministic* function of time since its values are known for all time if a value is known for one time point n. Similar considerations apply to the case $T = \mathbb{R}$ and when $F(d\lambda)$ consists of a finite number of point masses, say $\lambda_0, \ldots, \lambda_k$.

Theorem 3.7. For real-valued mean-zero weakly stationary sequences $\{X_n : n \in \mathbb{Z}\}$, (3.14) may be expressed in the following form:

$$X_n = U(\{0\}) + \int_{(0,\pi]} (2\cos n\lambda) U(d\lambda) - \int_{(0,\pi]} (2\sin n\lambda) V(d\lambda), \tag{3.16}$$

where $U(\cdot)$ and $V(\cdot)$ are mutually orthogonal real-valued orthogonal random fields on $[0, \pi]$ and

$$\mathbb{E}(U(d\lambda))^2 = \mathbb{E}(V(d\lambda))^2 = F(d\lambda)/4\pi.$$

Moreover,

$$V(\{0\}) = 0 \text{ a.s.}, \quad \mathbb{E}|U(\{0\})|^2 = F(\{0\})/2\pi.$$

Proof. Write $Z(\cdot)$ in Theorem 3.5 as

$$Z(B) = U(B) + iV(B),$$

where $U(\cdot)$ and $V(\cdot)$ are real-valued orthogonal random fields. The map $\varphi(e^{in\lambda}) = X_n$ $(n \in \mathbb{Z})$ extends to a linear isometry between $L^2(\mu)$ and $L^2(X)$. Hence if B is Borel, one may express

$$\mathbf{1}_B(\lambda) = \sum_{-\infty<n<\infty} a_n e^{in\lambda},$$

i.e., the series converging to $\mathbf{1}_B$ in $L^2(\mu)$. Hence

$$Z(B) = \sum_{-\infty<n<\infty} a_n X_n, \text{ a.s.}$$

Since

$$\mathbf{1}_{-B}(-\lambda) = \overline{\mathbf{1}_B(\lambda)} = \sum_{-\infty<n<\infty} \overline{a}_n e^{-in\lambda} = \sum_{-\infty<n<\infty} a_{-n} e^{in\lambda},$$

one has $Z(-B) = \sum_{-\infty<n<\infty} a_{-n} X_n = \overline{Z(B)}$. That is,

$$U(-B) + iV(-B) = U(B) - iV(B),$$

and hence

$$U(-B) = U(B) \quad \text{and} \quad V(-B) = -V(B) \text{ a.s.} \tag{3.17}$$

If in particular, $B \subset (0, \pi]$, then B and $-B$ are disjoint and $\mathbf{1}_B$ and $\mathbf{1}_{-B}$ are orthogonal, so that

$$0 = \mathbb{E}Z(B) \cdot \overline{Z(-B)} = \mathbb{E}Z^2(B) = \mathbb{E}(U^2(B) - V^2(B)) + 2i\mathbb{E}U(B)V(B),$$

implying (together with $F(B)/2\pi = \mathbb{E}|Z(B)|^2 = \mathbb{E}U^2(B) + \mathbb{E}V^2(B)$)

$$\mathbb{E}U^2(B) = \mathbb{E}V^2(B) = F(B)/4\pi, \quad \mathbb{E}U(B)V(B) = 0 \quad \text{for all Borel } B \subset (0, \pi]. \tag{3.18}$$

Furthermore, if B and C are disjoint Borel subsets of $(0, \pi]$, then from the orthogonality of $\mathbf{1}_B$, $\mathbf{1}_C$ and $\mathbf{1}_B$, $\mathbf{1}_{-C}$, it follows that

$$\mathbb{E}Z(B)\overline{Z(C)} = 0, \quad \mathbb{E}Z(B)Z(C) = 0,$$

so that $\mathbb{E}Z(B)U(C) = 0$, which implies $\mathbb{E}V(B)U(C) = 0$.

Now let B and C be arbitrary Borel subsets of $(0, \pi]$. Then

$$\begin{aligned}\mathbb{E}U(C)V(B) &= \mathbb{E}(U(B \cap C) + U(C \backslash B \cap C)) \cdot (V(B \cap C) + V(B \backslash B \cap C)) \\ &= \mathbb{E}U(B \cap C) \cdot V(B \cap C) = 0,\end{aligned} \tag{3.19}$$

by (3.18).

Finally, (3.14) leads to

$$\begin{aligned}X_n &= Z(\{0\}) + \int_{[-\pi,\pi]\backslash\{0\}} [(\cos n\lambda) + i(\sin x\lambda)](U(d\lambda) + iV(d\lambda)) \\ &= Z(\{0\}) + \int_{[-\pi,\pi]\backslash\{0\}} \{(\cos n\lambda)U(d\lambda) - (\sin n\lambda)V(d\lambda)\} \\ &\qquad + i\int_{[-\pi,\pi]\backslash\{0\}} \{(\cos n\lambda)V(d\lambda) + (\sin n\lambda)U(d\lambda)\} \\ &= Z(\{0\}) + \int_{(0,\pi]} (2\cos n\lambda)U(d\lambda) - \int_{(0,\pi]} (2\sin n\lambda)V(d\lambda).\end{aligned}$$

$Z(\{0\})$ must be real, $Z(\{0\}) = U(\{0\})$, and $V(\{0\}) = 0$. From this, the representation (3.16) follows. ■

An argument entirely analogous to the preceding yields (from Theorem 3.6) the following:

Theorem 3.8. If $\{X_t : t \in \mathbb{R}\}$ is a real-valued mean-zero weakly stationary process, then

$$X_t = U(\{0\}) + \int_{[0,\infty)} (2\cos t\lambda)U(d\lambda) - \int_{(0,\infty)} (2\sin t\lambda)V(d\lambda), \tag{3.20}$$

where $U(\cdot)$ and $V(\cdot)$ are two mutually orthogonal real-valued orthogonal random fields on $[0, \infty)$ satisfying

$$\mathbb{E}(U(d\lambda))^2 = \mathbb{E}(V(d\lambda))^2 = F(d\lambda)/2.$$

Moreover,

$$\mathbb{E}V(\{0\})^2 = V^2(\{0\}) = 0, \quad \mathbb{E}U(\{0\})^2 = \mathbb{E}|Z(\{0\})|^2 = F(\{0\}).$$

Example 2. Let $\{X_n : n \in \mathbb{Z}\}$ be a mean-zero weakly stationary real-valued process with spectral measure F with finite support $\{\lambda_j : -k \leq j \leq k\}$, $\lambda_{-j} = -\lambda_j$ $(1 \leq j \leq k)$, and

$$-\pi \leq \lambda_{-k} < \lambda_{-k+1} < \cdots < \lambda_0 = 0 < \lambda_1 < \cdots < \lambda_k \leq \pi,$$

with

$$F(\{\lambda_{-j}\}) = F(\{\lambda_j\}) = p_j > 0, (j = 0, 1, \ldots, k).$$

Then (3.16) yields

$$X_n = U(\{0\}) + \sum_{j=1}^{k}(2\cos n\lambda_j)U(\{\lambda_j\}) + \sum_{j=1}^{k}(2\sin n\lambda_j)V(\{\lambda_j\}), \quad (n \in \mathbb{Z}), \tag{3.21}$$

where $U(\{0\})$, $U(\{\lambda_j\})$ $(1 \leq j \leq k)$, and $V(\lambda_j)$ $(1 \leq j \leq k)$ are uncorrelated mean-zero random variables, and

$$\mathbb{E}U(\{\lambda_j\})^2 = p_j/4\pi = \mathbb{E}V(\{\lambda_j\})^2, (1 \leq j \leq k), \mathbb{E}U(\{0\})^2 = p_0/2\pi.$$

One may express (3.21) as

$$X_n = \sum_{-k \leq j \leq k} a_{n,j} Z_j, \qquad (n \in \mathbb{Z}), \tag{3.22}$$

where $\{Z_j : -k \leq j \leq k\}$ are *orthonormal,* i.e., mean-zero uncorrelated random variables each with variance one, defined by

$$Z_0 = U(\{0\})/\sqrt{p_0/2\pi}\,, \quad a_{n,0} := \sqrt{\frac{p_0}{2\pi}}\,;$$

$$Z_j = U(\{\lambda_j\})/\sqrt{p_j/4\pi}\,, \quad a_{n,j} := (2\cos n\lambda_j)\sqrt{\frac{p_j}{4\pi}} \quad (1 \leq j \leq k);$$

$$Z_{-j} = V(\{\lambda_j\})/\sqrt{p_j/4\pi}\,, \quad a_{n,-j} := (2\sin n\lambda_j)\sqrt{\frac{p_j}{4\pi}} \quad (1 \leq j \leq k).$$

From equations such as (3.22), one may compute Z_j in terms of X_n ($n \in \mathbb{Z}$). In general, one may expect an infinite series on the right in (3.22). Also, (weak) stationarity would suggest that X_n is a *moving average* as defined by the representation:

$$X_n = \sum_{-\infty<j<\infty} a_{n-j} Z_j \equiv \sum_{-\infty<j<\infty} a_j Z_{n-j} \qquad (n \in \mathbb{Z}). \tag{3.23}$$

It turns out, however, that if $\{X_n : n \in \mathbb{Z}\}$ has a discrete spectral measure F, then it cannot be represented by moving averages!

Theorem 3.9. A weakly stationary process $\{X_n : n \in \mathbb{Z}\}$ has a moving average representation if and only if its spectral measure is absolutely continuous.

Proof. (Sufficiency). Suppose (3.23) holds. Then $\sum |a_j|^2 < \infty$ and

$$c_m \equiv \mathrm{Cov}(X_n, X_{n+m}) = \sum_{-\infty<j<\infty} a_{n-j}\overline{a}_{n+m-j} = \sum_{-\infty<j<\infty} a_j \overline{a}_{m+j},$$

so that with

$$g(\lambda) = \sum a_j e^{-ij\lambda}, \quad |g(\lambda)|^2 = \sum_{j,j'} a_j \overline{a}_{j'} e^{i(j'-j)\lambda} = \sum_{j,m} a_j \overline{a}_{m+j} \overline{e}^{im\lambda}, \tag{3.24}$$

one has

$$c_m = \frac{1}{2\pi} \int_{[-\pi,\pi]} e^{-im\lambda} |g(\lambda)|^2 d\lambda \qquad (m \in \mathbb{Z}). \tag{3.25}$$

Hence the spectral measure F is absolutely continuous with density $|g(\lambda)|^2$.

(Necessity). Suppose, conversely, that F is absolutely continuous with density f. Then, with $g = \sqrt{f}$, one has $g \in L^2([-\pi, \pi), dx)$ with a Fourier expansion $g(\lambda) = \sum b_j e^{ij\lambda}$, with $\sum |b_j|^2 < \infty$. Hence

$$\begin{aligned}
\mathrm{Cov}(X_n, X_{n+m}) = c_m &= \frac{1}{2\pi} \int_{[-\pi,\pi]} e^{-im\lambda} |g(\lambda)|^2 d\lambda \\
&= \langle e^{-im\lambda} g(\lambda), g(\lambda) \rangle = \left\langle \sum b_j e^{-i(m-j)\lambda}, \sum b_{j'} e^{ij'\lambda} \right\rangle \\
&= \sum_{j,j'} b_j \overline{b}_{j'} \left(\frac{1}{2\pi}\right) \int_{[-\pi,\pi]} e^{-i(m-j)\lambda - ij'\lambda} d\lambda \\
&= \sum_j b_j \overline{b}_{j-m} = \sum \overline{b}_j b_{m+j} = \sum_j \overline{b}_{n+j} b_{n+m+j}.
\end{aligned}$$

Therefore, the general representation Theorem 3.3 applies with $T = \Lambda = \mathbb{Z}$, μ counting measure, $\psi(n, j) = \overline{b}_{n+j}$. One may then write (3.6) as

$$X_n = \sum_j \overline{b}_{n+j} Z_j = \sum_j \overline{b}_{n-j} Z_{-j} = \sum_j a_{n-j} W_j,$$

with $a_j = \overline{b}_j$, and $W_j = Z_{-j}$. ■

A linear transformation of $\{Z_j\}$ involved in the moving average representation (3.23) is referred to as a *filter*. Specific features of the filter depend on the choice of the coefficients a_j, $j \in \mathbb{Z}$. This and corresponding forms of covariance functions and spectral densities are aptly illustrated by the examples below.

Example 3 *(Autoregressive $AR(1)$ Models).* Consider first an autoregressive model AR(1) of order 1 defined by

$$Y_{n+1} = \beta Y_n + Z_{n+1} \qquad (n \geq 0), \tag{3.26}$$

where $\{Z_n : n \geq 1\}$ is a real- or complex-valued uncorrelated sequence, $\mathbb{E}Z_n = 0$, $\mathbb{E}Z_n\overline{Z_n} = \sigma^2 > 0$. If $|\beta| < 1$, then repeated iteration yields

$$Y_n = \beta^n Y_0 + \sum_{j=1}^{n} \beta^{n-j} Z_j = \beta^n Y_0 + \sum_{j=0}^{n-1} \beta^j Z_{n-j}, \tag{3.27}$$

which is easily checked to be a Cauchy sequence in $L^2(\Omega, \mathcal{F}, P)$ and, therefore, converges in L^2 to some Y_∞, say, where

$$\delta^2 := \operatorname{Var} Y_\infty = \sigma^2 \sum_{j=0}^{\infty} |\beta|^{2j} = \sigma^2/(1 - |\beta|^2). \tag{3.28}$$

Now let Y_0 be a mean-zero random variable with variance δ^2, uncorrelated with $\{Z_j : j \geq 1\}$. Then it is simple to check that $\{Y_n : n \geq 0\}$ is a weakly stationary process with the (summable) covariance function

$$c_m := \operatorname{Cov}(Y_n, Y_{n+m}) = \delta^2\overline{\beta}^m = \sigma^2\overline{\beta}^m/(1 - |\beta|^2) \quad (m \geq 0). \tag{3.29}$$

Motivated by (3.27), one may consider the process defined by

$$X_n := \sum_{j=0}^{\infty} \beta^j Z_{n-j} \qquad (n \in \mathbb{Z}), \tag{3.30}$$

where $\{Z_n : n \in \mathbb{Z}\}$ is a mean-zero uncorrelated sequence with variance σ^2 ($|\beta| < 1$). Then it is easy to check that $\{X_n : n \in \mathbb{Z}\}$ is a weakly stationary process on

$\mathbb{Z}$ with the same (summable) covariance function as that of $\{Y_n : n \in \mathbb{Z}\}$ in (3.29) (and $c_{-m} = \overline{c}_m$ for $m < 0$). Note that $\{X_n : n \in \mathbb{Z}\}$ is a *(one-sided) moving average process* in the sense of the representation (3.23), with coefficients $a_j = 0$ for all $j < 0$. Note also that X_n satisfies the same equation as Y_n extended to all $n \in \mathbb{Z}$. Namely,

$$X_{n+1} = \beta X_n + Z_{n+1} \qquad (n \in \mathbb{Z}). \tag{3.31}$$

Finally, the spectral density is readily computed using either Proposition 2.4 or Theorem 3.9, as a sum of a geometric series (also see Exercise 7)

$$f(\lambda) = \sum_{m\in\mathbb{Z}} c_m e^{im\lambda} = \frac{\sigma^2}{|1 - \beta e^{-i\lambda}|^2}, \quad -\pi \le \lambda \le \pi. \tag{3.32}$$

The spectral density changes qualitatively in the parameter regimes $-1 < \beta < 0$, $\beta = 0$, $0 < \beta < 1$, see Exercise 6.

In the case $|\beta| > 1$, one may rewrite (3.26) as

$$Y_n = (1/\beta)Y_{n+1} + W_{n+1} \qquad (W_n := -(1/\beta)Z_n), \tag{3.33}$$

where $\{W_n : n \ge 1\}$ is a sequence of uncorrelated mean-zero random variables, $\mathrm{Cov}(Y_n, W_{n+1}) = 0$ and $\mathrm{Var}(W_n) = \sigma_0^2 := |\beta|^{-2}\sigma^2$. Repeated iteration yields

$$Y_n = \beta^{-n} Y_{2n} + \sum_{j=0}^{n-1} \beta^{-j} W_{n+j+1}, \qquad n \ge 0.$$

This leads to the definition of the weakly stationary process given by

$$X_n := \sum_{j=0}^{\infty} \beta^{-j} W_{n+j+1} \qquad (n \in \mathbb{Z}), \tag{3.34}$$

which has the covariance function $c_m \equiv \mathrm{Cov}(X_n, X_{n+m}) = \sigma^2 \beta^{-m}(|\beta|^2 - 1)^{-1}$ $(m \ge 0)$, $c_{-m} = \overline{c}_m$ $(m > 0)$. The process (3.34) satisfies

$$X_{n+1} = \beta X_n + Z_{n+1} \qquad (n \in \mathbb{Z}), \tag{3.35}$$

and, by (3.34), is a one-sided moving average. However, unlike (3.30), the uncorrelated random variables $\{W_{n+1+j} : j \ge 0\}$ in (3.34) are from the *future,* as opposed to those in (3.30) which are from the *past* and the *present.* The representation (3.30), corresponding to the case $|\beta| < 1$, is called *causal*, while that in (3.34), corresponding to $|\beta| > 1$, is said to be *noncausal*. In particular, causality refers to the nature of the transformation (filter) of $\{Z_n : n \in \mathbb{Z}\}$ representing $\{X_n : n \in \mathbb{Z}\}$.

Example 4 *(Autoregressive $AR(p)$ Models).* An autoregressive process $\{X_n : n \in \mathbb{Z}\}$ of order $p \geq 1$ is one in which X_n is a linear function of the past p values, plus noise:

$$X_n = \beta_0 X_{n-p} + \beta_1 X_{n-p+1} - \cdots - \beta_{p-1} X_{n-1} + Z_n \quad (n \in \mathbb{Z}), \tag{3.36}$$

where $\beta_0 \neq 0$, and $\{Z_n : n \in \mathbb{Z}\}$ is a mean-zero uncorrelated sequence, $\mathrm{Var}(Z_n) = \sigma^2 > 0$ for all n. In order to analyze conditions under which such a process may exist as a (weakly) stationary process, and to represent it as a moving average, it is useful to introduce the *backward shift operator* B:

$$BX_n = X_{n-1}, \quad BZ_n = Z_{n-1}, \tag{3.37}$$

and consider its linear extension to $L^2(\mathbf{X})$ and $L^2(\mathbf{Z})$, the closure of the linear span of X_n or Z_n, $(n \in \mathbb{Z})$ in $L^2(\Omega, \mathcal{F}, P)$. Then

$$B^j X_n = B^{j-1}(BX_n) = B^{j-1}X_{n-1} = \cdots = X_{n-j}(j \geq 0), \; B^j Z_n = Z_{n-j},$$

and one may express (3.36) as

$$(I - \psi(B))X_n = Z_n \qquad (n \geq 0), \tag{3.38}$$

where $\psi(z)$ is a polynomial, say

$$\psi(z) = \sum_0^{p-1} \beta_j z^{b-j} \quad (z \in \mathbb{C}), \tag{3.39}$$

and

$$\psi(B) = \sum_{j=0}^{p-1} \beta_j B^{p-j}. \tag{3.40}$$

Theorem 3.10. Let $\psi(z)$ be the polynomial (3.39), and suppose the zeros of the polynomial $1 - \psi(z)$ all lie outside the unit circle $\{z : z \in \mathbb{C}, |z| = 1\}$. Then there exists a weakly stationary process $\{X_n : n \in \mathbb{Z}\}$ satisfying (3.36), which has the one-sided moving average representation

$$X_n = \sum_{j=0}^{\infty} a_j Z_{n-j} \qquad (n \in \mathbb{Z}), \tag{3.41}$$

where $a_0 = 1$, and $\sum_{\cdot} |a_j| < \infty$.

Proof. The hypothesis means that the roots of the equation $1 - \psi(z) = 0$ all lie outside the unit circle. Then there exists $\delta > 0$ such that $1-\psi(z) = 0$ has no roots in $\{z \in \mathbb{C} : |z| < 1+\delta\}$. This implies $(1-\psi(z))^{-1}$ is analytic in $\{z \in \mathbb{C} : |z| < 1+\delta\}$, with an absolutely convergent power series $\sum_{j=0}^{\infty} a_j z^j$ expansion around $z = 0$, with radius convergence greater than one. This implies $\sum_{j=0}^{\infty} |a_j| < \infty$. Note that $a_0 = (1 - \psi(0))^{-1} = 1$. Using the identity $(1 - \psi(z))^{-1}(1 - \psi(z)) = 1$, one can now easily check that a solution to (3.38) is given for all $n \in \mathbb{Z}$ by

$$X_n = (I - \psi(B))^{-1} Z_n := \sum_{j=0}^{\infty} a_j B^j Z_n = \sum_{j=0}^{\infty} a_j Z_{n-j} \quad (n \in \mathbb{Z}).$$

Note that

$$\mathrm{Var}(X_n) = \sigma^2 \sum_{j=0}^{\infty} |a_j|^2, \quad c_m := \mathrm{Cov}(X_n, X_{n+m}) = \sum_{j=0}^{\infty} a_j \overline{a}_{j+m} \quad (m \geq 0),$$

$$c_{-m} = \overline{c}_m \qquad (m > 0).$$

■

It now follows that one may compute the spectral density as (see Exercise 7)

$$f(\lambda) = \frac{\sigma^2}{|1 - \psi(e^{-i\lambda})|^2}, \quad -\pi \leq \lambda \leq \pi. \tag{3.42}$$

Remark 3.1. The relation (3.36) implies that $L^2(\mathbf{Z}) \subset L^2(\mathbf{X})$, while (3.41) implies $L^2(\mathbf{X}) \subset L^2(\mathbf{Z})$. Hence, under the hypothesis of Theorem 3.10, $L^2(\mathbf{X}) = L^2(\mathbf{Z})$. The process $\{X_n : n \in \mathbb{Z}\}$ is *causal* since it depends only on the past and present values of Z_n. It is *invertible,* since Z_n can be expressed in terms of the past and present values of X_n. The backshift operator B is a *linear isometry* on

$$L^2(\mathbf{Z}) : \|BY\|^2 \equiv \mathbb{E}(BY \cdot \overline{BY}) = \|Y\|^2 = \mathbb{E}(Y \cdot \overline{Y}),$$

for all $Y = \sum_j b_j Z_j$, $(\sum_j |b_j|^2 < \infty)$. Hence

$$\|B\| := \sup\{\|BY\| : \|Y\| \leq 1\} = 1.$$

One may think of the expansion of $(I - \psi(B))^{-1}$ in powers of B to be convergent (with respect to this operator norm of B), noting that $\|\psi(B)\| < 1$, under the hypothesis of Theorem 3.10.

Example 5 *(Autoregressive-Moving Average Models $ARMA(p, q)$).* An autoregressive-moving average model ARMA (p, q) $(p \geq 1, q \geq 1)$ satisfies

$$X_n = \sum_{j=0}^{p-1} \beta_j X_{n-p+j} + Z_n + \sum_{j=0}^{q-1} \delta_j Z_{n-q+j} \qquad (n \in \mathbb{Z}), \tag{3.43}$$

where $\beta_0 \neq 0$, $\delta_0 \neq 0$, and $\{Z_n : n \in \mathbb{Z}\}$ are uncorrelated, with $\mathrm{Var}(Z_n) = \sigma^2 > 0$ for all n. In terms of the backshift operator B, one may express (3.43) as

$$(I - \psi(B))X_n = (I + \theta(B))Z_n, \qquad (n \in \mathbb{Z}), \tag{3.44}$$

where $\psi(B)$ is as in (3.40) and

$$\theta(B) = \sum_{j=0}^{q-1} \delta_j B^{q-j}. \tag{3.45}$$

Theorem 3.11. Suppose the equations $1 - \psi(z) = 0$ and $1 + \theta(z) = 0$ do not have any common root. (a) If all roots of the equation $1 - \psi(z) = 0$ lie outside the unit circle, then there exists a weakly stationary causal process $\{X_n : n \in \mathbb{Z}\}$ which satisfies (3.43) and has the one-sided moving average representation

$$X_n = \sum_{j=0}^{\infty} b_j Z_{n-j} \qquad (n \in \mathbb{Z}),$$

where $b_0 = 1$ and $\sum_j |b_j| < \infty$.

(b) If the roots of $1 + \theta(z) = 0$, as well as those of $1 - \psi(z) = 0$ all lie outside the unit circle, then the ARMA(p, q) model is invertible.

Proof.

(a) The proof follows the same argument as in the proof of Theorem 3.10 and expresses X_n as

$$\begin{aligned} X_n &= (I - \psi(B))^{-1}(I + \theta(B))Z_n \\ &= \left(\sum_{j=0}^{\infty} a_j B^j\right)\left(I + \sum_{j=0}^{q-1} \delta_j B^{q-j}\right) Z_n \\ &= \left(\sum_{j=0}^{\infty} b_j B^j\right) Z_n = \sum_{j=0}^{\infty} b_j Z_{n-j} \qquad (n \in \mathbb{Z}). \end{aligned}$$

For example, $b_0 = 1$, $b_1 = a_1 + a_0\delta_{q-1} = a_1 + \delta_{q-1}$, $b_2 = a_2 + a_1\delta_{q-1} + a_0\delta_{q-2} = a_2 + a_1\delta_{q-1} + \delta_{q-2}$, etc.

(b) Under the given hypothesis, $z \to (1+\theta(z))^{-1}$ is analytic in a circle $\{z \in \mathbb{C} : |z| < 1+\delta'\}$ of radius greater than one. Hence one may express (3.44) as

$$\begin{aligned} Z_n &= (I+\theta(B))^{-1}(I-\psi(B))X_n \\ &= \sum_{j=0}^{\infty} d_j X_{n-j}, \qquad\qquad (n \in \mathbb{Z}), \end{aligned}$$

say, with $\sum_{j=0}^{\infty} |d_j| < \infty$.

■

It follows that in case (a) the spectral density is given by (see Exercise 7)

$$f(\lambda) = |1+\theta(e^{-i\lambda})|^2 |1-\psi(e^{-i\lambda})|^{-2}, \quad -\pi \le \lambda \le \pi.$$

Remark 3.2. The $AR(p)$ and $ARMA(p,q)$ models are among the most widely used models in time series analysis, after some adjustments are made to remove trends and cyclical effects, if necessary. A different approach to these models may be based on iterated random maps, i.e., *random dynamical systems*. For example, the $AR(1)$ model is that of a first order linear difference equation driven by random (noise) forcing, and similarly $AR(p)$ is a p-th order linear difference equation driven by noise. In the analysis of these and other iterated random maps, one also considers the polynomial $\widetilde{\psi}(\lambda) := \beta_0 + \beta_1\lambda + \cdots + \beta_{p-1}\lambda^{p-1} - \lambda^p$ and the roots of the equation $\widetilde{\psi}(\lambda) = 0$. Since $\widetilde{\psi}(\lambda) = \lambda^p(\psi(1/\lambda) - 1)$, the zeros of $\widetilde{\psi}(\lambda)$ are the reciprocals of the zeros of $1 - \psi(z)$. There the convergence to a steady state for the random dynamical system involves the hypothesis that the zeros of $\widetilde{\psi}(\lambda)$ are all *inside* the unit circle. Finally, the assumption that there are no common roots of the equations $1-\psi(z) = 0$ and $1+\theta(z) = 0$ in Theorem 3.11 does not really restrict the result since common factors can be canceled out. Indeed, its purpose is to make sure that common zeroes (possibly lying on or inside the unit circle) are taken out of consideration, and the moving average is economically derived.

It will be shown in the forthcoming Theorem 3.15 that a necessary and sufficient condition for the one-sided moving average representation (3.30) is that, in addition to absolute continuity of the spectral measure $F(d\lambda)$, the logarithm of the spectral density $f(\lambda)$ be integrable. This is the case in most of the important examples, such as the $AR(p)$ and $ARMA(p,q)$ models described above (Exercise 1). At the moment, however, we present one of the most important results in the theory of weakly stationary processes, namely a general representation of weakly stationary processes due to H. Wold (1938). To state it involves a bit of nomenclature and notation as follows.

Let M_n denote the closure in $L^2(\Omega, \mathcal{F}, P)$ of the linear span of $\{X_m : -\infty < m \le n\}$, $M_{-\infty} = \cap_{-\infty<n<\infty} M_n$, and let $M = M_\infty$ be the closure of the linear space spanned by $\{X_m : -\infty < m < \infty\}$.

Definition 3.4. The process $\{X_n\}$ is said to be *deterministic* if $M_n = M_{-\infty}$ for all n, which implies (and is implied by) $\mathbb{P}_{M_n} X_{n+1} = X_n$ for all n, where $\mathbb{P}_L$ denotes the orthogonal projection onto a closed linear subspace L of M_∞.

Note that if the process is deterministic, then the expected squared error of prediction is

$$\sigma^2 = \mathbb{E}|X_{n+1} - \mathbb{P}_{M_n} X_{n+1}|^2 = 0.$$

A simple example of a deterministic process is $X_n = Y$ for all n, where Y is a mean-zero square-integrable random variable.

Definition 3.5. The process $\{X_n\}$ is purely nondeterministic if $M_{-\infty} = (0)$, the zero subspace of $L^2(\Omega, \mathcal{F}, P)$.

Theorem 3.12 *(The Wold Decomposition).* A weakly stationary process $\{X_n : n \in \mathbb{Z}\}$ has a decomposition into two orthogonal components given by

$$X_n = W_n + V_n, \tag{3.46}$$

where $\{W_n\}$ is purely nondeterministic and $\{V_n\}$ is deterministic and $\{W_n\}$ and $\{V_n\}$ are mutually orthogonal.

Proof. One can express X_n as $\mathbb{P}_{M_{n-1}} X_n + a_0 \xi_n$, where ξ_n is orthonormal to M (and $\xi_n \in M_n$), i.e., ξ_n has mean zero, is orthogonal to M_{n-1}, and is of norm one. Next, writing $\mathbb{P}_{M_{n-1}} X_n = Y_{n-1}$, one may express $Y_{n-1} = \mathbb{P}_{M_{n-2}} Y_{n-1} + a_1 \xi_{n-1} = Y_{n-2} + a_1 \xi_{n-1}$, $X_n = Y_{n-2} + a_1 \xi_{n-1} + a_0 \xi_n$. Suppose $a_m \neq 0$. Continuing this process indefinitely, one has the (one-sided) moving average representation

$$X_n = \sum_{m \geq 0} a_m \xi_{n-m} = W_n, \quad n \in \mathbb{Z}, \tag{3.47}$$

where $\{\xi_n : n \in \mathbb{Z}\}$ is an orthonormal sequence and a_m, $(m \geq 0)$ is a sequence of constants, as specified above, $\sum_{m \geq 0} |a_m|^2 = \mathbb{E}|X_0|^2 = \sigma^2$, say. In this case $V_n = 0$ for all n. If, on the other hand, the process of projections stops after n_0 steps, say, i.e., $\mathbb{P}_{M_{n-n_0}} Y_{n-n_0+1} = Y_{n-n_0+1}$, then $Y_n = Y_{n-n_0}$ for all $n \geq n_0$, and Y_{n-n_0+1} is deterministic, belonging to $M_{-\infty}$. In this case

$$X_n = \sum_{0 \leq m \leq n_0} a_m \xi_{n-m} + Y_{n-n_0+1} = W_n + V_n, n \in \mathbb{Z}, \tag{3.48}$$

where $W_n = \sum_{0 \leq m \leq n_0} a_m \xi_{n-m}$ and $V_n = Y_{n-n_0+1}$ are orthogonal to each other. ■

Corollary 3.3. The one-step prediction error defined as

$$\theta = \mathbb{E}|X_{n+1} - \mathbb{P}_{M_n} X_{n+1}|^2$$

is given by $|a_0|^2$, in the case $n_0 > 1$, i.e., the W component is nonzero. If $n_0 = 1$ so that $W_n = 0$ for all n, $\{X_n\}$ is deterministic, $X_n = V_n$ for all n, and $\theta = 0$.

Our next step is to determine conditions such that a one-sided moving average representation (3.47) is possible for a weakly stationary process.

Proposition 3.14. A necessary and sufficient condition for representing a weakly stationary process $X_n, -\infty < n < \infty$ as a (one-sided) moving average (3.47) is that its spectral measure is absolutely continuous with respect to Lebesgue measure on $[-\pi, \pi]$ with a density f satisfying $f(\lambda) = |g(\lambda)|^2$, where g is of the form

$$g(\lambda) = \sum_{m \geq 0} b_m e^{im\lambda}, \quad \sum_{m \geq 0} |b_m|^2 < \infty. \tag{3.49}$$

Proof. (Necessity). Let $\{X_n\}$ have the representation (3.47). Then

$$\begin{aligned} c_n &= \mathbb{E} X_0 \overline{X}_n \\ &= \sum_{m \geq 0} a_m \overline{a}_{m+n} \\ &= \frac{1}{2\pi} \int_{[-\pi,\pi]} \exp\{-in\lambda\} \left(\sum_{m \geq 0} a_m \exp\{-im\lambda\} \right) \left(\sum_{m \geq 0} \overline{a}_m \exp\{im\lambda\} \right) d\lambda, \end{aligned}$$

since the integral vanishes for all terms in the product of the two sums except the terms $a_m \overline{a}_{m+n}$. Letting $g(\lambda)$ be as in (3.49) with $b_m = \overline{a}_m$, one then has $c_n = \frac{1}{2\pi} \int_{[-\pi,\pi]} \exp\{-in\lambda\} |g(\lambda)|^2 d\lambda$. That is, the spectral measure is absolutely continuous with a density $f(\lambda) = |g(\lambda)|^2$ with $g(\lambda)$ of the form (3.49).

(Sufficiency). Assume that the spectral measure is absolutely continuous having a density $f = |g|^2$ with g of the form (3.49). Let the stochastic orthogonal measure in the spectral representation of X_n be $Z(d\lambda)$ (Theorem 3.5),

$$X_n = \int_{[-\pi,\pi]} \exp\{in\lambda\} Z(d\lambda), \quad \mathbb{E}|Z(d\lambda)|^2 = F(d\lambda) = f(\lambda) d\lambda. \tag{3.50}$$

Consider the orthogonal stochastic measure ζ on $[-\pi, \pi]$ defined by

$$\zeta(d\lambda) = \frac{1}{\overline{g}(\lambda)} Z(d\lambda). \tag{3.51}$$

Note that $\mathbb{E}|\zeta(d\lambda)|^2 = d\lambda$. In particular, $g(\lambda)$ can be zero on at most a set of Lebesgue measure zero since, otherwise, $\mathbb{E}|\zeta(d\lambda)|^2 = d\lambda$ would vanish on a set of positive Lebesgue measure. Thus,

$$\begin{aligned} X_n &= \int_{[-\pi,\pi]} \exp\{in\lambda\} Z(d\lambda) \\ &= \int_{[-\pi,\pi]} \exp\{in\lambda\} \overline{g}(\lambda) \zeta(d\lambda) \\ &= \sum_{m\geq 0} \overline{b}_m \int_{[-\pi,\pi]} \exp\{i(n-m)\lambda\} \zeta(d\lambda) \\ &= \sum_{m\geq 0} a_m \gamma_{n-m}, \end{aligned} \tag{3.52}$$

where $a_m = \overline{b}_m$, and $\gamma_k = \int_{[-\pi,\pi]} \exp\{ik\lambda\}\zeta(d\lambda)$ is a sequence of random variables satisfying

$$\mathbb{E}\gamma_k \gamma_r = \delta_{k,r}.$$

That is, $\{\gamma_k : -\infty < k < \infty\}$ is an orthonormal sequence, with

$$X_n = \sum_{m\geq 0} a_m \gamma_{n-m}.$$

■

Our next task is to find a simple condition on the spectral density f such that (3.49) holds. Its proof uses some results from complex variables, most of them standard.[3]

Theorem 3.15. The (one-sided) moving average representation (3.47) holds for a weakly stationary process $\{X_n : -\infty < n < \infty\}$ if and only if the spectral measure is absolutely continuous with density f satisfying

$$\int_{[-\pi,\pi]} |\ln f(\lambda)| d\lambda < \infty. \tag{3.53}$$

Proof. (Sufficiency). First, for each $\lambda \in [-\pi, \pi]$, consider the analytic function $z \to h(z;\lambda) = (e^{i\lambda} + z)/(e^{i\lambda} - z)$ in the unit circle $D(0:1) = \{z \in \mathbb{C} : |z| < 1\}$. A little algebra shows that the real part of $h(z;\lambda)$ is the so-called Poisson kernel (Exercise 12)

$$Reh(z;\lambda) = (1-r^2)/[1 - 2r\cos(\theta-\lambda) + r^2] := P(r, \theta - \lambda) \tag{3.54}$$

[3] An excellent source of all the results used is the complex variables part of the graduate text (Rudin, 1974).

for $z = re^{i\theta}(0 \leq r < 1, \theta \in [-\pi, \pi])$. Also, $w(z) = \int_{[-\pi,\pi]} \ln f(\lambda)h(z;\lambda)d\lambda$ is analytic, as is easily shown by the convergent expansion of $z \to h(z;\lambda)$ in $D(0:1)$ under the integral sign. The analytic function $h(z;\lambda)$ is a harmonic function in $D(0:1)$ for each λ, i.e., thought of as a function on the (x, y)-plane, with $z = x + iy$, its Laplacian $\Delta w(z)$ vanishes, where the Laplacian is defined as $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. For the real and imaginary parts of an analytic function are harmonic. Hence the real part of $\frac{1}{2\pi} w(z)$, namely,

$$\frac{1}{2\pi}\int_{[-\pi,\pi]} \ln f(\lambda)P(r,\theta-\lambda)d\lambda = u(r,\theta), \tag{3.55}$$

say, is harmonic since $\ln f(\lambda)$ is integrable, by hypothesis. Here, once again one[4] expresses the real part of $w(z)$ as a function of (x, y), via the relation $z = x+iy$, $r = |z| = (x^2 + y^2)^{\frac{1}{2}}$. By Jensen's inequality,

$$u(r,\theta) \leq \frac{1}{2\pi} \ln \int_{[-\pi,\pi]} f(\lambda)P(r,\theta-\lambda)d\lambda. \tag{3.56}$$

Define $g(z) = \exp\{\frac{w(z)}{4\pi}\}$. Then, by (3.56) and Jensen's inequality,

$$\begin{aligned}
|g(re^{i\theta})|^2 &= |g(z)|^2 = |\exp\{w(z)/2\pi\}| \\
&= |\exp\{Rew(z)/2\pi + iImw(z)/2\pi\}| \\
&= \exp\{Rew(z)/2\pi\} \\
&= \exp\{u(r,\theta\} \leq \frac{1}{2\pi}\int_{[-\pi,\pi]} f(\lambda)P(r,\theta-\lambda)d\lambda.
\end{aligned} \tag{3.57}$$

Since the Poisson kernel is a probability density function on $[-\pi, \pi]$ (see Remark 3.3), i.e., $\frac{1}{2\pi}\int_{[-\pi,\pi]} P(r,\theta-\lambda)d\lambda = 1$ for all θ, one has

$$\frac{1}{2\pi}\int_{[-\pi,\pi]} |g(rei\theta)|^2 d\theta \leq \frac{1}{2\pi}\int_{[-\pi,\pi]} f(\lambda)d\lambda < \infty, \quad \text{for all } 0 \leq r < 1. \tag{3.58}$$

Moreover, $\lim_{r\uparrow 1} |g(re^{i\theta})|^2 = \lim_{r\uparrow 1} \exp u(r,\theta) = f(\theta)$ for almost all θ, with respect to Lebesgue measure on $[-\pi, \pi]$. To understand this result, note that $P(r, \theta-\lambda)$ is the density of a distribution which converges weakly to the point mass $\delta_\theta(d\lambda)$ on $[-\pi, \pi]$, as $r \uparrow 1$ (see Remark 3.3 below). Hence if $\ln f$ is continuous, $u(r,\theta)$ in (3.55) converges to $\ln f(\theta)$, so that $|g(re^{i\theta})|^2$ converges to $f(\theta) = |g(e^{i\theta})|^2$. If $\ln f$ is just integrable, then the convergence is almost everywhere (Rudin 1974, Section 11.12). The function $g(z)$ is analytic in $D(0:1)$, so it has

[4] Rudin (1974), Sections 11.3–11.5.

an expansion $g(z) = \sum_{n\geq 0} b_n z^n$, the sum being absolutely convergent in $D(0:1)$. Therefore, $g(re^{i\theta}) = \sum_{n\geq 0} b_n r^n e^{in\theta}$ converges to $g(e^{i\theta}) = \sum_{n\geq 0} b_n e^{in\theta}$, as $r \uparrow 1$. The relation (3.58) now implies $\sum_{n\geq 0} |b_n|^2 < \infty$, and the hypothesis of Proposition 3.14 is satisfied.

(Necessity). By Proposition 3.14, the spectral measure is absolutely continuous with a density f of the form $f(\theta) = |g(\exp\{i\theta\})|^2$, where $g(\exp\{i\theta\}) = \sum_{n\geq 0} b_n e^{in\theta}$, with $\sum_{n\geq 0} |b_n|^2 < \infty$. Then

$$g(z) := \sum_{n\geq 0} b_n z^n = \sum_{n\geq 0} b_n r^n e^{in\theta} = g(r\exp\{i\theta\})$$

is analytic. Let $A = \{\theta \in [-\pi,\pi] : |g(r\exp i\theta)| \leq 1\}$, and $B = \{\theta \in [-\pi,\pi] : |g(r\exp\{i\theta\}| > 1\}$. Assume for the moment $g(0) = 1$. Write

$$\int_{[-\pi,\pi]} |\ln|g(r\exp\{i\theta\})||d\theta = \int_A + \int_B = 2\int_B - \int_{[-\pi,\pi]}. \tag{3.59}$$

By Jensen's formula (Rudin 1974, Theorem 15.18) (see Exercise 16),

$$\int_{[-\pi,\pi]} |\ln|g(r\exp i\theta)||d\theta = \ln \prod_{1\leq m\leq s} \frac{r}{|y_m|} \geq 0, \quad 0 < r < 1, \tag{3.60}$$

where y_m are the zeros of $g(z)$ in $\overline{D(0:r)}$. Hence, from (3.59) and the fact[5] that $\ln x \leq x$ for all $x \geq 0$,

$$\begin{aligned}\int_{[-\pi,\pi]} |\ln|g(r\exp i\theta)||d\theta &\leq 2\int_B \ln|g(r\exp\{i\theta\})|d\theta \\ &= \int_B \ln|g(r\exp\{i\theta\})|^2 d\theta \\ &\leq \int_B |g(r\exp\{i\theta\})|^2 d\theta \\ &\leq \int_{[-\pi,\pi]} |g(r\exp\{i\theta\})|^2 d\theta \\ &= \int_{[-\pi,\pi]} |\sum_n b_n r^n e^{in\theta}|^2 d\theta \\ &\leq (2\pi)\sum_{n\geq 0} |b_n|^2 < \infty. \end{aligned} \tag{3.61}$$

[5] Rudin (1974), Sections 11.12, 15.18.

As argued in the "sufficiency" proof, one then has $\lim_{r\uparrow 1}|g(re^{i\theta})|^2 = f(\theta)$ almost everywhere and, by Fatou's lemma,

$$\begin{aligned}\int_{[-\pi,\pi]} |\ln f(\theta)|d\theta &= \int_{[-\pi,\pi]} \lim_{r\uparrow 1} |\ln|g(r\exp\{i\theta\})||^2 d\theta \\ &\leq \limsup_{r\uparrow 1} \int_{[-\pi,\pi]} |\ln|g(r\exp\{i\theta\})||d\theta \\ &= 2\limsup_{r\uparrow 1} \int_{[-\pi,\pi]} |\ln|g(r\exp\{i\theta\})||d\theta \\ &\leq 2\pi \sum_{n\geq 0} |b_n|^2 < \infty. \end{aligned} \tag{3.62}$$

It only remains to consider the case $g(0) \neq 1$. In this case if the first nonzero term in the expansion of $g(z)$ is $a_k z^k$, then apply Jensen's formula to the function $\tilde{g}(z) = g(z)/a_k z^k$, so that $\tilde{g}(0) = 1$. Then, instead of the lower bound of 0, as the last inequality of (3.60), one gets a lower bound of a finite number c, perhaps negative, and the proof of finiteness of $\int_{[-\pi,\pi]} |\ln|g(r\exp\{i\theta\})||d\theta$ in (3.61) goes through (Exercise 16). ■

Remark 3.3. The function $\lambda \to P(r, \theta - \lambda)$ in (3.54) may be shown to be the density of B_τ, where τ is the first time a two-dimensional standard Brownian motion $\{B_t : t \geq 0\}$ hits the unit circle at a point $(\cos\theta, \sin\theta)$, represented as $\exp\{i\theta\}$, starting from a point (r, θ) in polar coordinates or the point $r\exp\{i\theta\}$ in complex coordinates, $0 < r < 1$ (Exercise 14).

It follows from the proof of Proposition 3.14 that the coefficients a_n in the moving average expansion of X_n and the coefficients of b_n in the expression for $g(\lambda)$ determine each other, namely, $a_n = \sqrt{2\pi} \cdot \overline{b}_n$. Especially, the prediction error $|a_0|^2$ can be expressed in terms of the spectral density f using this relationship (see Corollary 3.3).

Theorem 3.16 *(Szegö–Kolmogorov Formula).* Under the hypothesis of Theorem 3.15, the prediction error is given by

$$\mathbb{E}|X_{n+1} - \mathbb{P}_{M_n} X_{n+1}|^2 = |a_0|^2 = \exp\left\{\frac{1}{2\pi}\int_{[-\pi,\pi]} \ln f(\lambda) d\lambda\right\},$$

where f is the spectral density.

Proof. Recall from the proof of Theorem 3.15 the functions

$$g(z) = \exp\{w(z)/4\pi\},$$

and

$$w(z) = \int_{[-\pi,\pi]} \ln f(\lambda)[(e^{i\lambda} + z)/(e^{i\lambda} - z)]d\lambda. \tag{3.63}$$

Now,

$$\begin{aligned}(e^{i\lambda} + z)/(e^{i\lambda} - z) &= 1 + 2e^{i\lambda} z/(1 - e^{-i\lambda} z)\\ &= 1 + 2e^{i\lambda} z \sum_{n\geq 0} e^{-in\lambda} z^n,\end{aligned} \tag{3.64}$$

and

$$w(z) = \int_{[-\pi,\pi]} \ln f(\lambda)d\lambda + 2\int_{[-\pi,\pi]} \ln f(\lambda) \sum_{n\geq 1} e^{-in\lambda} z^n d\lambda, \tag{3.65}$$

so that

$$g(z) = \exp\left\{\frac{1}{4\pi}\int_{[-\pi,\pi]} \ln f(\lambda)d\lambda\right\} \cdot \exp\left\{\frac{1}{2\pi}\int_{[-\pi,\pi]} \ln f(\lambda) \sum_{n\geq 1} e^{-in\lambda} z^n d\lambda\right\}. \tag{3.66}$$

Equating this to $g(z) = \sum_{n\geq 0} b_n z^n$, one can express the coefficients b_n in terms of the spectral density f. As shown in the proof of Proposition 3.12, $a_n = \overline{b_n}$, where a_n are the coefficients a_n of the time series (3.47). In particular, $a_0 = \overline{b_0} = \exp\{\frac{1}{4\pi}\int_{[-\pi,\pi]} \ln f(\lambda)d\lambda\}$, and the prediction error is

$$|a_0|^2 = \exp\left\{\frac{1}{2\pi}\int_{[-\pi,\pi]} \ln f(\lambda)d\lambda\right\}.$$

■

Remark 3.4. If a moving average is expressed as $X_n = \sum_{0\leq m\leq\infty} a_m \zeta_{n-m}$, where, instead of being orthonormal, $\{\zeta_n\}$ is an orthogonal sequence with common variance σ^2, then the prediction error is $|a_0|^2\sigma^2$, given by the Szegö–Kolmogorov formula.

Remark 3.5. If one wishes to compute the h-step prediction error, then, provided that the representation is in terms of an orthonormal error sequence, one has

$$\theta(h) = \mathbb{E}|X_n - \mathbb{P}_{M_{n-h}} X_n| = |a_0 + a_1 + \cdots + a_{h-1}| = |\overline{b}_0 + \overline{b}_1 + \cdots + \overline{b}_{h-1}|.$$

These constants may be obtained by computing the coefficients of $z^k (k = 0, \ldots, h-1)$ in (3.66).

Remark 3.6. Prediction theory, presented above in its most general form, is due to Kolmogorov (1939, 1941a,b), with earlier contributions due to Wold (1938) and Szegö (1920). In particular, Kolmogorov proved that the formula in Theorem 3.16) holds for arbitrary weakly stationary processes in the Wold decomposition, with f being the absolutely continuous component of the spectral measure. Prediction theory in the continuous parameter case was obtained by Wiener in 1941, independently of Kolmogorov. We refer to Doob (1953), Chapter XII, for a comprehensive presentation of prediction theory. Among other valuable references are Gihman and Skokohod (1974), Chapter IV, Grenander (1981), Chapters 3–5, and Brockwell and Davis (1991), Chapters 4,5.

Next we will obtain representations based on covariance structure for processes which need not be weakly stationary. In this regard, the next result is the widely used Karhunen–Loève expansion[6] of a mean-zero real- or complex-valued process $\{X_t : t \in T\}$ with a continuous covariance function $(s,t) \to r(s,t)$ on a finite interval $T = [c,d]$, $c < d$ or more generally random field (see remarks below). Consider the integral operator $K : L^2([c,d]) \to L^2([c,d])$ defined by

$$(Kf)(s) = \int_{[a,b]} f(t)r(s,t)dt. \tag{3.67}$$

Then, by Mercer's theorem (see Appendix A), K has a sequence of positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$, with corresponding orthonormal eigenfunctions $\varphi_1, \varphi_2, \cdots$ such that $\sum_1^\infty \lambda_n^2 < \infty$ and

$$r(s,t) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(s)\overline{\varphi_n(t)} \qquad (s,t \in [c,d]), \tag{3.68}$$

where the convergence is absolute and uniform. Observe also that K is a *Hilbert–Schmidt operator* in the sense that $\sum_n \lambda_n^2 < \infty$, since

$$\infty > \int_{[c,d]} \int_{[c,d]} |r(s,t)|^2 dsdt = \sum_n \lambda_n^2.$$

Theorem 3.17 *(Karhunen–Loève Expansion).* Assume the covariance function $r(s,t)$ is continuous on $[c,d] \times [c,d]$. Define

$$Z_n = \lambda_n^{-\frac{1}{2}} \int_{[c,d]} X_t \overline{\varphi_n(t)} dt \qquad (n = 1, 2, \ldots). \tag{3.69}$$

Then $\{Z_n\}_{n=1}^\infty$ is an orthonormal sequence of random variables (in $L^2(\Omega, \mathcal{F}, P)$), and

[6] This theorem is due to Karhunen (1946) and Loève (1948).

$$X_t = \sum_{n=1}^{\infty} \sqrt{\lambda_n}\, \varphi_n(t) Z_n \qquad (t \in T = [c, d]), \tag{3.70}$$

with the sum on the right converging in $L^2(\Omega, \mathcal{F}, P)$.

Proof. In view of (3.68), the general orthogonal representation Theorem 3.3 applies with $\mu(\{n\}) = \lambda_n$ on $\Lambda = \{1, 2, \dots\}$, and $\psi(s, n) = \varphi_n(s)$ $(n = 1, 2, \dots)$. Hence there exists an uncorrelated sequence $Z(\{n\}), n = 1, 2, \dots,$ such that $\mathbb{E}|Z(\{n\})|^2 = \lambda_n$ for all $n \geq 1$ and

$$X_t = \sum_{n=1}^{\infty} \varphi_n(t) Z(\{n\}). \tag{3.71}$$

Now let $Z_n = \lambda_n^{-1/2} Z(\{n\})$. ■

Remark 3.7. The representation is also variously referred to as a *principal component decomposition*, *empirical orthogonal decomposition*, or *singular value decomposition*. The coefficients $\sqrt{\lambda_n}$ are referred to as the *singular values*. The above choice of $T = [c, d]$ with Lebesgue measure provides a *standard version* of the Karhunen–Loève expansion. However the above proof carries over without change to (centered) random fields $\{X_t : t \in T\}$ indexed by a compact metric space T and a finite measure $\pi(dt)$ on the Borel σ-field of T and *fully supported*[7] on T, for the Hilbert space $L^2(\pi)$ with the inner product defined by

$$\langle f, g \rangle := \int_T f(t)\overline{g(t)}\pi(dt), \quad f, g \in L^2(\pi).$$

Such generality has applications to random fields indexed by compact manifolds representing biological organs, for example. Also, viewed this way, one obtains the classical version of the singular value decomposition from linear algebra corresponding to a *finite* set $T = \{1, 2, \dots, M\}$ and $\pi(\{j\}) = 1$, for all $j \in T$, i.e., counting measure. In this case the operator K has the matrix representation by $((r(i, j)))_{1 \leq i, j \leq M}$.

Remark 3.8. If one has many independent realizations (samples) of $\{X_t : t \in T\}$, then, using the law of large numbers, one can estimate the mean $m_t = \mathbb{E}X_t$ for purposes of centering and the covariance kernel $r(s, t)$. If one has only a single realization, say for $t \in T = [0, S]$, with S *large*, then one needs to assume stationarity and ergodicity (to be treated in the next chapter), in order to properly estimate $r(s, t)$.

[7] The full support is needed in order to obtain Mercer's theorem in this generality; see the Appendix.

Remark 3.9. Pattern (or feature) extraction and *dimension reduction* are among the most popular uses of Karhunen–Loève decompositions. These naturally involve analysis of the eigenfunctions (i.e., principal components) and truncations corresponding to the largest eigenvalues.[8] Dimension reduction roughly occurs when only a few of the eigenvalues are large and capture most of the variance.

Example 6 *(Spectral Representation of Brownian Bridge and Brownian Motion).* Let $\{B_t^* : 0 \le t \le 1\}$ be the Brownian bridge defined by

$$B_t^* = B_t - tB_1, \quad 0 \le t \le 1,$$

where $\{B_t : 0 \le t \le 1\}$ is standard Brownian motion started at zero. The covariance function of $\{B_t^* : 0 \le t \le 1\}$ is readily computed as

$$r(s,t) = \begin{cases} s(1-t) \text{ if } s \le t, \\ t(1-s) \text{ if } s > t. \end{cases} \tag{3.72}$$

Consider the integral operator K on the real Hilbert space $L^2 = L^2([0,1], dt)$ having the kernel function $r(s,t)$. In view of Mercer's theorem and Theorem 3.17, one then has the Karhunen–Loève expansion of the Brownian bridge B_t^* in terms of an i.i.d. standard Gaussian sequence $\{Z_n\}_{n=1}^\infty$ (noting that the right side of (3.69) is Gaussian),

$$B_t^* = 2\sum_{n=1}^{\infty} \frac{\sin(n\pi t)}{n\pi} Z_n, \qquad 0 \le t \le 1. \tag{3.73}$$

Using the Karhunen–Loève expansion of the Brownian bridge, one may also represent standard Brownian motion B_t, $0 \le t \le 1$, as

$$B_t = B_t^* + tB_1 = 2\sum_{n=1}^{\infty} \frac{\sin n\pi t}{n\pi} Z_n + tZ_0, \quad 0 \le t \le 1, \tag{3.74}$$

where $\{Z_n : n = 0, 1, \dots\}$ is an i.i.d. standard Gaussian sequence. We have used the fact that $\{X_t = B_t - tB_1 : 0 \le t \le 1\}$ is independent of B_1,[9] to derive the representations (3.73) and (3.74). However, the expansion (3.74) is not really the Karhunen–Loève expansion for Brownian motion, whose covariance function is $r(s,t) = \min\{s,t\}$. If K is the operator on $L^2([0,1], dt)$ having this kernel, then $g(t) := Kf(t)$ ($f \in L^2([0,1], dt)$) is easily shown to satisfy

$$g''(t) = -g(t), \quad g(0) = 0, \quad g'(1) = 0. \tag{3.75}$$

[8] For applications in this connection, see Glavaski et al. (1998).

[9] Wiener (1923).

The functions $\sin(\frac{2n+1}{2}\pi t)$ $(n = 0, 1, \cdots)$ are eigenfunctions of $\mathbf{K}$ (but $\sin(n\pi t)$, $n \geq 1$, are not). In particular, it is straightforward to check that the Karhunen–Loève representation for standard Brownian motion starting at zero is given by

$$B_t = \sqrt{2}\sum_{n=0}^{\infty} Z_n \frac{2\sin(\frac{2n+1}{2}\pi t)}{(2n+1)\pi}, \quad 0 \leq t \leq 1. \tag{3.76}$$

The corresponding calculations readily extend to the Karhunen–Loève representation of Brownian sheet[10] as well (Exercise 3).

To conclude this chapter, we will consider further representations for the special case of Gaussian processes. Again, we do not assume weak stationarity for this development. In the end this will provide a way to compute the spectral density for fractional Gaussian noise.

Proposition 3.15. Let $(S, \mathcal{S}, m)$ be a σ-finite measure space such that $L^2(S, \mathcal{S}, m)$ is separable. Then there is a family of random variables $\{Z(h) : h \in L^2(S, \mathcal{S}, m)\}$ such that (i) $Z(ah_1 + bh_2) = aZ(h_1) + bZ(h_2)$ a.s. for $h_1, h_2 \in L^2(S, \mathcal{S}, m)$, $a, b \in \mathbb{R}$, and (ii) $Z(h)$ is Gaussian with mean zero and variance $\mathbb{E}Z^2(h) = \|h\|^2 \equiv \int_S |h(s)|^2 m(ds)$, for each $h \in L^2(S, \mathcal{S}, m)$.

Proof. Let $\{\varphi_n : n \geq 1\}$ be an orthonormal basis for $L^2(S, \mathcal{S}, m)$. One may apply the Kolmogorov construction[11] to obtain an i.i.d. sequence $Z_1, Z_2, \ldots$ of standard normal random variables on a probability space $(\Omega, \mathcal{F}, P)$. Define

$$Z(h) = \sum_{n=1}^{\infty} \langle h, \varphi_n \rangle Z_n, \quad h \in L^2(S, \mathcal{S}, m),$$

noting the L^2-convergence of each such series to an element of $L^2(S, \mathcal{S}, m)$. ■

In the case that $h = \mathbf{1}_B$, for $B \in \mathcal{S}$, with $m(B) < \infty$, one writes $Z(B) \equiv Z(\mathbf{1}_B)$. This defines an orthogonal random field $\{Z(B) : B \in \mathcal{S}\}$, sometimes referred to as a *Gaussian measure* with intensity m. Although for disjoint $B_1, B_2, \ldots$ in $\mathcal{S}$ such that $m(B_j) < \infty$ for each $j \geq 1$, one has $Z(\cup_{j\geq 1} B_j) = \sum_{j\geq 1} Z(B_j)$ a.s., and in $L^2(S, \mathcal{S}, m)$, the null set on which equality does not hold can depend on B. The term "measure" is a standard abuse of terminology which must be interpreted with care in this context. One may also denote it by $Z(ds)$ with the meaning defined by Proposition 3.15.

Example 7 *(Spectral Representation of Fractional Brownian Motion).* Recall that for fixed $0 < h < 1$ the covariance function for fractional Brownian motion

[10] See Adler and Taylor (2007). Also see Noda (1987), for the more difficult case of Lévy Brownian motion indexed by $\mathbb{R}^k$ and/or the k-dimensional sphere.

[11] See BCPT, p.168.

$\{B_t^{(h)} : t \in \mathbb{R}\}$ is given by

$$r(s,t) = \frac{1}{2}\left\{|s|^{2h} + |t|^{2h} - |s-t|^{2h}\right\}, \quad 0 \le s, t \in \mathbb{R}.$$

Although fractional Brownian motion is not a stationary process, it is a Gaussian process with stationary increments; see Example 5 in Chapter 2.

Letting $Z(ds)$ denote a mean-zero Gaussian orthogonal random field with $S = \mathbb{R}$ and m Lebesgue measure on the Borel σ-field $\mathcal{S} = \mathcal{B}$, one has the following "moving average" type representation for fractional Brownian motion. Positive and negative parts of $a \in \mathbb{R}$ are denoted by $a_+ = a \vee 0$ and $a_- = -(a \wedge 0) = (-a)_+$, respectively.

Proposition 3.19. For $0 < h < 1, h \neq 1/2$, define

$$g(s,t) = \left((t-s)_+^{\frac{2h-1}{2}} - (-s)_+^{\frac{2h-1}{2}}\right), \quad s, t \in \mathbb{R}.$$

Then[12]

$$B_t^{(h)} = \frac{1}{c(h)} \int_{\mathbb{R}} g(s,t) Z(ds), \quad t \in \mathbb{R},$$

where

$$c(h) = \left\{\int_0^\infty \left((1+s)^{\frac{2h-1}{2}} - s^{\frac{2h-1}{2}}\right)^2 ds + \frac{1}{2h}\right\}^{\frac{1}{2}}.$$

In the case $h = \frac{1}{2}$, this representation may be extended to include Brownian motion with the convention that

$$B_t^{(\frac{1}{2})} = \int_0^t \mathbf{1}_{[0,\infty)}(t) Z(ds) - \int_t^0 \mathbf{1}_{(-\infty,0)}(t) Z(ds), \quad t \in \mathbb{R}.$$

Proof. Denoting the right side of the asserted representation by X_t, one may easily calculate $\mathbb{E}X_t^2 = |t|^{2h}$ by explicit integration considering each of the cases $t \ge 0$, $t < 0$, separately, and checking square-integrability in neighborhoods of singularities. For $t \ge 0$, one has

[12] This representation is sometimes loosely expressed as $B_t^{(h)} = \frac{1}{c(h)}\{\int_{-\infty}^t (t-s)^{\frac{2h-1}{2}} Z(ds) - \int_{-\infty}^0 (-s)^{\frac{2h-1}{2}} Z(ds)\}$, $t \in \mathbb{R}$; however such integrals do not exist separately. We prefer to avoid this particular abuse of notation, and however it does motivate the definition since the first integral may be viewed as a "fractional derivative" of the Brownian paths $Z(ds)$, and the second integral suggests a centering adjustment to get L^2-convergence of the integrand.

$$g^2(s,t) = \begin{cases} (t-s)^{2h-1} & \text{if } 0 \le s \le t \\ \left((t-s)^{\frac{2h-1}{2}} - (-s)^{\frac{2h-1}{2}}\right)^2 & \text{if } s < 0, \\ 0 & \text{if } s > t. \end{cases} \tag{3.77}$$

Making a change of variable $r = t/s$, one arrives at

$$\begin{aligned} c^2(h)\mathbb{E}|X_t|^2 &= \int_{\mathbb{R}} g^2(s,t)ds \\ &= t^{2h}\left\{\int_0^1 (1-r)^{2h-1}dr + \int_0^\infty \left((1+r)^{\frac{2h-1}{2}} - r^{\frac{2h-1}{2}}\right)^2 dr\right\}. \\ &= c^2(h)t^{2h}. \end{aligned} \tag{3.78}$$

The case $t < 0$ is similar. Also, for $0 < t_1 < t_2$,

$$(g(s,t_2) - g(s,t_1))^2 = \begin{cases} \left((t_2-s)^{\frac{2h-1}{2}} - (t_1-s)^{\frac{2h-1}{2}}\right)^2 & \text{if } s \le t_1 \\ (t_2-s)^{2h-1} & \text{if } t_1 < s < t_2, \\ 0 & \text{if } s > t_2. \end{cases} \tag{3.79}$$

This, together with a similar change of variables as above, first $u = s - t_1$ and then $r = u/(t_2 - t_1)$, yields

$$c^2(h)\mathbb{E}\left(X_{t_2} - X_{t_1}\right)^2 = c^2(h)(t_2 - t_1)^{2h}.$$

Thus, the desired calculation results from the identity

$$\mathbb{E}X_tX_s = \frac{1}{2}\left\{\mathbb{E}X_t^2 + \mathbb{E}X_s^2 - \mathbb{E}(X_t - X_s)^2\right\}.$$

In particular, therefore, $\{X_t : t \in \mathbb{R}\}$ is a version of $\{B_t^{(h)} : t \in \mathbb{R}\}$. ■

Remark 3.10. Moving average representations of the general form

$$X_t = \int_{\mathbb{R}} g(s,t)Z(ds), \quad t \in \mathbb{R}, \tag{3.80}$$

with

$$g(s,t) = \varphi(t-s) - \psi(0-s)$$

for square-integrable functions φ, ψ also include, for example, an Ornstein–Uhlenback process by taking $\varphi(t) = e^{-t}, t \ge 0$, and $\varphi(t) = 0, t < 0$, and $\psi \equiv 0$. In particular, the representation defines a stationary (Markov) process in this special case.

A "Fourier dual" to this can be obtained (guessed then proven) by considering a representation of the form $B_t^{(h)} = \int_R \hat{g}(\lambda, t)\hat{Z}(d\lambda)$, where $\hat{Z}(d\lambda) = \hat{Z}_1(d\lambda) + i\hat{Z}_2(d\lambda)$, for a pair of real-valued independent mean-zero Gaussian orthogonal random fields $\hat{Z}_1(ds), \hat{Z}_2(ds)$ for $L^2(m)$, where $m(ds) = \frac{1}{4}ds$, and where $\hat{Z}_1(B) = \overline{\hat{Z}_1(-B)}, \hat{Z}_2(B) = \overline{\hat{Z}_2(-B)}$, $B \in \mathcal{B}$ such that $m(B) < \infty$. For a complex square-integrable integrand $\hat{g} = \hat{g}_1 + i\hat{g}_2$ such that $\hat{g}_1$ is an even and $\hat{g}_2$ an odd function, i.e., $\overline{\hat{g}(-\lambda)} = \hat{g}(\lambda)$, the integral $\int_{\mathbb{R}} \hat{g}(\lambda)\hat{Z}(d\lambda)$ is defined by

$$\int_R \hat{g}(\lambda)\hat{Z}(d\lambda) := \int_R \hat{g}_1(\lambda)\hat{Z}_1(d\lambda) - \int_R \hat{g}_2(\lambda)\hat{Z}_2(d\lambda). \tag{3.81}$$

Remark 3.11. These definitions are suggested by considering a formal Parseval-like relation for which $\int_{\mathbb{R}} \hat{g}(\lambda)\hat{Z}(d\lambda) = \int_{\mathbb{R}} g(s)Z(ds)$. In addition, a further heuristic argument leading to the next representation using these definitions will be given following its proof.

Let us now see that such heuristics lead to a correct guess for the representation. Namely,

Proposition 3.20. Let $0 < h < 1, h \neq \frac{1}{2}$. Then

$$B_t^{(h)} = \frac{1}{\hat{c}(h)} \int_{\mathbb{R}} \frac{e^{i\lambda t} - 1}{i\lambda} |\lambda|^{-\frac{2h-1}{2}} \hat{Z}(d\lambda), \quad t \in \mathbb{R},$$

where

$$\hat{c}(h) = \left\{ \int_0^\infty \frac{\sin^2 r}{r^2} dr + \int_0^\infty \frac{(1-\cos(r))^2}{r^2} dr \right\}^{\frac{1}{2}}.$$

Proof. Since $|\frac{e^{it\lambda}-1}{i\lambda}|$ is bounded in neighborhood of $\lambda = 0$ and decays as $|\lambda|^{-1}$ as $\lambda \to \pm\infty$, the integrand is clearly square-integrable. Thus the representation is well defined. Denote it by X_t. Then, noting that $|e^{i\lambda} - 1|^2 = 4\sin^2(\frac{\lambda}{2})$ and using (3.81), one obtains by a change of variable that

$$\mathbb{E}X_t^2 = t^{2h} \frac{1}{\hat{c}(h)^2} \left\{ \int_0^\infty \frac{\sin^2 r}{r^2} dr + \int_0^\infty \frac{(1-\cos(r))^2}{r^2} dr \right\} = t^{2h}. \tag{3.82}$$

The computation of $\mathbb{E}|X_{t_2} - X_{t_1}|^2$ is similar. One has

$$\begin{aligned} \hat{c}^2(h)\mathbb{E}|X_{t_2} - X_{t_1}|^2 &= \int_{\mathbb{R}} \frac{|e^{i\lambda(t_2-t_1)} - 1|^2}{\lambda^2} |\lambda|^{-2h+1} d\lambda \\ &= |t_2 - t_1|^{2h} 4 \int_{\mathbb{R}} (1-\cos(r))^2 r^{-2h-1} dr. \end{aligned}$$

So the desired covariance structure again follows from the identity

$$\mathbb{E}X_tX_s = \frac{1}{2}\left\{\mathbb{E}X_t^2 + \mathbb{E}X_s^2 - \mathbb{E}(X_t - X_s)^2\right\}. \qquad \blacksquare$$

Remark 3.12. A heuristic basis for this representation as a "coloring" of Gaussian (white) noise can also be obtained by noticing that $\frac{e^{i\lambda t}-1}{i\lambda}$ is the Fourier transform of $\mathbf{1}_{[0,t]}(s)$. The purely formal form of the (nonexistent) Fourier transform of $s_+^{\frac{2h-3}{2}}$ would have the (divergent) form $\int_{\mathbb{R}} e^{i\lambda s} s_+^{\frac{2h-3}{2}} ds = |\lambda|^{-\frac{2h-1}{2}} \int_0^\infty e^{iy} y^{\frac{2h-3}{2}} dy$. Expressing the Fourier transform of convolution as a product of Fourier transforms, together with the above Parseval-like stipulations defining $\hat{Z}(d\lambda)$ and the integration formula, formally leads to the form of Proposition 3.20 that was rigorously *proven* to be correct.

As an application of this representation, it can be used to compute the spectral distribution for the fractional Gaussian noise $W_j^{(h)} = B_{j+1}^{(h)} - B_j^{(h)}$, $j \geq 0$.

$$\begin{aligned}
c_j = \mathbb{E}W_0^{(h)}\overline{W^{(h)}}_j &= \mathbb{E}B_1^{(h)}\overline{B^{(h)}}_{j+1} - \mathbb{E}B_1^{(h)}\overline{B^{(h)}}_j \\
&= \frac{1}{\hat{c}(h)^2}\int_{\mathbb{R}} e^{i\lambda j}\left|\frac{e^{i\lambda}-1}{i\lambda}\right|^2 |\lambda|^{-2h+1} d\lambda \\
&= \frac{1}{\hat{c}(h)^2}\sum_{n=-\infty}^{\infty}\int_{-\pi+2n\pi}^{\pi+2n\pi} e^{i\lambda j}|e^{i\lambda}-1|^2|\lambda|^{-2h-1}d\lambda \\
&= \frac{1}{\hat{c}(h)^2}\sum_{n=-\infty}^{\infty}\int_{-\pi}^{\pi} e^{i\lambda j}|e^{i\lambda}-1|^2|\lambda+2n\pi|^{-2h-1}d\lambda \\
&= \frac{1}{\hat{c}(h)^2}\int_{-\pi}^{\pi} e^{i\lambda j}|e^{i\lambda}-1|^2\sum_{n=-\infty}^{\infty}|\lambda+2n\pi|^{-2h-1}d\lambda. \qquad (3.83)
\end{aligned}$$

Thus[13]

$$F(d\lambda) = \frac{2\pi}{\hat{c}^2(h)}\left|e^{i\lambda}-1\right|^2\sum_{n=-\infty}^{\infty}|\lambda+2\pi n|^{-2h-1}\,d\lambda, \quad \lambda \in [-\pi, \pi].$$

[13] For this and a much more extensive treatment of fractional Brownian motion representations, see Samorodnitsky and Taqqu (1994). An application of the Karhunen–Loéve expansion for fractional Brownian motion to problems in fluid flow, which includes formal asymptotic estimates of eigenvalues and eigenfunctions for the fractional Brownian covariance kernel, is given in Bronski (2003). An insightful analysis in the contact of financial mathematics is given by Rogers (1997).

Writing $F(d\lambda) = f(\lambda)d\lambda$, then, since $|e^{i\lambda} - 1|^2 \sim |\lambda|^2/2$ as $\lambda \to 0$, splitting off the $n = 0$ term yields

$$f(\lambda) \sim \frac{2\pi}{\hat{c}^2(h)}|\lambda|^{1-2h} \qquad \text{as } \lambda \to 0.$$

A stationary sequence with a spectral density of the form $\frac{1}{|\lambda|^\theta}$ for an exponent $0 < \theta < 1$ is often referred to as a $\frac{1}{f}$*-noise*. Thus one obtains a $\frac{1}{f}$-noise in the case $1/2 < h < 1$. Notice that the convergence/divergence of $\sum_{n\in\mathbb{Z}} |c_n|$ is also reflected in the behavior of the spectral density as $\lambda \to 0$. Recall also that the correlations are *positive*, i.e., $c_n > 0$, in the case $1/2 \le h < 1$. The fractional Gaussian noise is sometimes said to exhibit *long-range dependence* when $1/2 < h < 1$. It is sometimes said to be *chaotic* in the case $0 < h \le 1/2$. The case $h = 1/2$ of uncorrelated increments is referred to as a *white noise*.

Finally one may note that when $h \neq 1/2$ the fractional Gaussian noise sequence provides an example of non-diffusive scaling for stationary processes in the sense that one must scale the partial sums $\sum_{j=1}^n Y_j$ by n^h to obtain a (Gaussian) limit distribution as $n \to \infty$, see Exercise 11.

Exercises

1. Compute the Wold decomposition for an $AR(1)$ model. [*Hint*: See Example 3.]
2. (*Poisson Random Field*) Let $(\Lambda, \mathcal{L}, \mu)$ be an arbitrary measure space with μ a σ-finite measure.

   (a) Show that there exists a random field $\{N(B) : B \in \mathcal{L}, \mu(B) < \infty\}$ such that (i) $N(B)$ is Poisson with mean $\mu(B)$, (ii) if $B_1 \cap B_2 = \emptyset$, then $N(B_1 \cup B_2) = N(B_1) + N(B_2)$ a.s., and (iii) if $B_1 \cap B_2 = \emptyset$, then $N(B_1)$ and $N(B_2)$ are independent. [*Hint*: First assume μ is a finite measure, and let η_i $(i = 1, 2, \cdots)$ be i.i.d. random variables with distribution $\mu(d\mu)/\mu(\Lambda)$, and let Y be a Poisson random variable with mean $\mu(S)$, independent of $\{\eta_i : i \ge 0\}$. Check that $N(B) := \sum_{i=1}^Y \mathbf{1}_B(\eta_i)$ $(\mathbf{1}_{[Y\ge1]}$ has the desired properties by using the moment generating function (mgf) of $N(B)$ and the joint mgf of $N(B_1)$ and $N(B_2)$ for disjoint B_1 and B_2. For the general case, let D_n be pairwise disjoint sets in $\mathcal{L}$ such that $0 < \mu(D_n) < \infty$ and $\cup_{n=1}^\infty D_n = \Lambda$. Then define an independent family of random variables $\{Y_n, \eta_i^{(n)} : n = 1, 2, \ldots ; i = 1, 2, \ldots\}$ with distributions given by $P(\eta_i^{(n)} \in B) = \mu(B \cap D_n)/\mu(D_n)$, $B \in \mathcal{S}$, and $P(Y_n = k) = e^{-\mu(D_n)}\mu^k(D_n)/k!$ $(k = 0, 1, \cdots)$. Let $N(B) = \sum_{n=1}^\infty \sum_{i=1}^{Y_n} \mathbf{1}_B(\eta_i^{(n)})\ \mathbf{1}_{[Y_n\ge1]}$.]

(b) Let $\{N(B) : B \in \mathcal{L}, \mu(B) < \infty\}$ be as in (a). Show that $\{\widetilde{N}(B) = N(B) - \mu(B) : B \in \mathcal{S}, \mu(B) < \infty\}$ is a real-valued orthogonal random field for $(\Lambda, \mathcal{L}, \mu)$. [*Hint*: Check $\mathbb{E}\widetilde{N}(B_1)\widetilde{N}(B_2) = \mu(B_1 \cap B_2)$.]

(c) Let $\Lambda = [-\pi, \pi]$, and let μ be a finite measure. Let $U(d\lambda)$ and $V(d\lambda)$ be two independent real-valued orthogonal random fields as in (b). Define X_n by (3.16) (or by (3.14) with $Z(\cdot) = U(\cdot) + iV(\cdot)$). Show that $\{X_n : n \in \mathbb{Z}\}$ is weakly stationary with spectral measure $F = 2\pi\mu$.

(d) For the case F has the two-point support: $F(\{-\frac{\pi}{2}\}) = F(\{\frac{\pi}{2}\}) = 2\pi$, compute $U(\cdot)$, $V(\cdot)$ as in (c), and show that $\{X_n : n \in \mathbb{Z}\}$ is not (strictly) stationary. [*Hint*: $U(\{\frac{\pi}{2}\}) = N_1 - \frac{1}{2}$, $V(\{\frac{\pi}{2}\}) = N_2 - \frac{1}{2}$ and $U([0,\pi]\backslash\{\frac{\pi}{2}\}) = 0 = V([0,\pi]\backslash\{\pi/2\})$ a.s., where N_1 and N_2 are independent Poisson random variables each with mean $\frac{1}{2}$. Show that, for n even, $X_n = 2(-1)^{n/2}(N_1 - \frac{1}{2})$ and, for n odd, $X_n = 2(-1)^{\frac{n+1}{2}}(N_2 - \frac{1}{2})$.]

(e) Check that, if $\mu(\Lambda) < \infty$, then $N(\cdot)$ and $\widetilde{N}(\cdot)$, as defined in (a) and (b) are *random measures:* $N(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} N(B_n)$ for every pairwise disjoint sequence $\{B_n\} \subset \mathcal{S}$. The same holds for $\widetilde{N}$. If $\mu(\Lambda) = \infty$, let $N(B) = \infty$ a.s. if $\mu(B) = \infty$. Then $N(\cdot)$ is a random measure on $(\Lambda, \mathcal{L})$.

3. (*Brownian motion and Brownian sheet*)

(a) Verify the Karhunen–Loève expansion for Brownian motion by computing the eigenfunctions and eigenvalues in the case of the kernel $K(s,t) = s \wedge t, 0 \le s, t \le 1$.

(b) The k-dimensional *Brownian sheet* is the mean-zero Gaussian random field $\{B(\mathbf{t}) : \mathbf{t} \in [0,1]^k\}$ having covariance kernel given by $K(\mathbf{s}, \mathbf{t}) = \prod_{j=1}^{k} s_j \wedge t_j, \mathbf{s} = (s_1, \ldots, s_k), \mathbf{t} = (t_1, \ldots, t_k) \in [0,1]^k$. Compute the Karhunen–Loève expansion for the Brownian sheet. [*Hint*: The eigenvalues and eigenfunctions are products of those obtained for the one-dimensional problem.]

4. (*Gaussian Orthogonal Random Fields, White Noise*)

(a) Let $\{X_n : n \in \mathbb{Z}\}$ be a stationary mean-zero complex-valued Gaussian process. Show that the associated orthogonal random field $Z(\cdot)$ in the representation (3.14) is Gaussian (i.e., $Z(B)$ is a complex-valued Gaussian random variable for all Borel $B \subset [-\pi, \pi]$).

(b) If $\{X_n : n \in \mathbb{Z}\}$ is a real-valued stationary Gaussian process with (symmetric) spectral measure F on $[-\pi, \pi]$, show that $U(\cdot)$ and $V(\cdot)$ in (3.16) are two independent real-valued orthogonal Gaussian random fields on $[0, \pi]$ and that for Borel $B_1, B_2, \cdots, B_n \subset (0, \pi]$ the k random variables $U(B_j)$, $1 \le j \le k$, are jointly (mean-zero) Gaussian with covariances $\mathrm{Cov}(U(B_j), U(B_k)) = F(B_j \cap B_k)/4\pi$. The same is true for $V(\cdot)$. However, $U(\{0\})$ is mean-zero Gaussian with variance $F(\{0\})/2\pi$, while $V(\{0\}) = 0$ a.s. In the particular case $F/4\pi$ is Lebesgue measure on $[0, \pi]$, $U(\cdot)$ (as well as $V(\cdot)$) is called white noise on $[0, \pi]$.

(c) Construct (Gaussian) white noise $W(\cdot)$ as an orthogonal real-valued random field on $\mathbb{R}$, and show that its restriction $B_t := W([0,t])$, $t \geq 0$, has the same finite dimensional distribution as Brownian motion.

5. Identify the random fields $U(\cdot)$ and $V(\cdot)$ in the representation (3.20) for the Ornstein–Uhlenbeck process.
6. (*Red, White, Blue AR(1) Spectra*) Physicists assign colors to spectral distributions by considering colors displayed by equivalent frequency bands of light. Red is displayed by decreasing power $f(\lambda)d\lambda$ in bands of increasing frequency $|\lambda|$, blue is displayed by increasing power $f(\lambda)d\lambda$ in bands with increasing $|\lambda|$, and white has constant power across equal frequency bands $\lambda \pm d\lambda$. Explain the classification of the spectral distributions for $AR(1)$ accordingly in each of the cases of $0 < \beta < 1$, $\beta = 0$, and $-1 < \beta < 0$, as red, white, and blue.
7. (a) Show that the spectral density of the $AR(p)$ process, under the hypothesis of Theorem 3.10, is $|1-\psi(e^{-i\lambda})|^{-2}$. [*Hint*: Spectral density of $\{Z_n : n \in \mathbb{Z}\}$ is $1_{[-\pi,\pi)}$, and $\sum_{j=0}^{s} a_j e^{-ij\lambda} = (1-\psi(e^{-i\lambda}))^{-1}$.]
   (b) Show that the spectral density of the $ARMA(p,q)$ process, under the hypothesis of Theorem 3.11(a), is $|1+\theta(e^{-i\lambda})|^2|1-\psi(e^{-i\lambda})|^{-2}$.
8. Show that the hypotheses of Theorems 3.10 and 3.11(a) are satisfied if $\sum_{j=0}^{p-1} |\beta_j| < 1$.
9. Use (3.73) to derive the identity $4\sum_{n=1}^{\infty} \frac{\sin^2(n\pi t)}{n^2\pi^2} = t(1-t)$, $0 \leq t \leq 1$.
10. Estimate $\mathbb{E}|X_t - X_s|^4$ for the sum on the right in (3.73), and prove that X_t has continuous sample paths (a.s.). Then deduce the process defined by the series in (3.74) has, with probability one, continuous sample paths.
11. Let $\{Y_j : j \in \mathbb{Z}\}$ denote the fractional Gaussian noise. Show that the finite dimensional distributions of the partial sum process $\{n^{-h}S_{[nt]} : 0 \leq t \leq 1\}$ converge to those of the corresponding fractional Brownian motion $\{B_t^{(h)} : 0 \leq t \leq 1\}$.
12. (*Poisson Kernel*) Show that in $D(0:1)$ the real part of $h(z,\lambda) = (e^{i\lambda} + z)/(e^{i\lambda} - z)$ is the Poisson kernel $P(r, \theta - \lambda)$. [*Hint*: $h(z,\lambda) = [1 - r^2 + 2ir\sin(\theta-\lambda)]/|1 - ze^{-i\lambda}|^2$].]
13. Show how to modify the proof of "Necessity" of Theorem 3.15 for $g(0) = 0$. [*Hint*: Use Jensen's formula for the function $\tilde{g}(z) = g(z)/a_k z^k$, where the first nonzero term in the expansion of $g(z)$ is $a_k z^k (k > 0)$. Apply Jensen's formula to $\tilde{g}(z)$, with $\tilde{g}(0) = 1$. Then $\int_{[-\pi,\pi]} |\ln|\tilde{g}(r\exp i\theta)||d\theta < \infty$. Use this to prove the desired assertion.]
14. (*Mean value property of the Poisson Integral*) Let f be a continuous function on the unit circle $\mathbb{T}$, and $\Psi f(x,y) = \mathbb{E}_{x,y} f(B_\tau)$, where (i) B is a standard Brownian motion on $\mathbb{R}^2$, starting at (x,y) in the unit disc $D(0:1)$, and τ is the first time the Brownian motion reaches $\mathbb{T}$.

   (a) Show that $\Psi f(x,y)$ is harmonic, and it has the mean value property: $\Psi f(x,y) = \int \Psi f(u,v)\mu_a(d(u,v))$, where (u,v) is (a random) point on the circle with center (x,y) and radius a contained in $D(0:1)$, and $\mu_a(d(u,v))$ is the uniform distribution on this circle. [*Hint*: Use the strong

Markov property of Brownian motion, with stopping time η, which is the first time it reaches this circle.]

(b) Using the property that $\Delta\Psi f = 0$ in $D(0:1)$, and Ψf has the boundary value f on $\mathbb{T}$, check that the (unique) solution of this equation is the Poisson integral $\Psi f(x, y)$ of f with density function $\lambda \to \frac{1}{2\pi}P(r, \theta - \lambda)$, where $r = (x^2 + y^2)^{1/2}$. That is, $\Psi f(z) = (1 - r^2)\frac{1}{2\pi}\int_{[-\pi,\pi]} \tilde{f}(\lambda)P(r, \theta - \lambda)d\lambda$, $(z = (x, y) = r(\cos\theta, \sin\theta), \tilde{f}(\lambda) = f(\cos\lambda, \sin\lambda))$.

15. Using the expansion of g in (3.62), calculate $a_1 = b_1$, and thereby, the step two prediction error $\mathbb{E}|X_n - \mathbb{P}_{M_{n-2}}X_n|^2$.

16. Let f be an analytic function in the disc $D(0 : R) = \{z : |z| < R\}$. Let $0 < r < R$, and let $a_1, \ldots, a_m$ be an enumeration (with possible multiplicities) of the zeros of f in $D(0 : r)$ and $a_{m+1}, \ldots, a_N$ be the zeros, if any, on $\{z : |z| = r\}$. Assume also that $f(0) \neq 0$.

(a) Show that the function $g(z) = f(z)\prod_{1\le n\le m}(r^2 - \overline{a}_n z)/(r(a_n - z))\prod_{n+1\le n\le N} a_n/(a_n - z)$ is analytic in $D(0 : r + \epsilon)$ for a sufficiently small $\epsilon > 0$ and has no zeros in $D(0 : r + \epsilon)$. [*Hint*: The denominators of g cancel out by factorization of f; also $\epsilon > 0$ may be chosen so that f has no further zeros in $D(0 : r + \epsilon)$.]

(b) Show that $\log|g(z)|$ is harmonic in $D(0 : r + \epsilon)$ and therefore has the mean value property $\log|g(0)| = \int_{[-\pi,\pi]} \log|g(r\exp\{i\theta\})|d\theta$. [*Hint*: There exists an analytic function $w(z) = u(z) + iv(z)$ on $D(0 : r + \epsilon)$ such that $g(z) = \exp\{w(z)\}$. Now $|g| = \exp\{u\}$, $\log|g| = u = \mathrm{Re}w$.]

(c) (i) Show that $|g(0)| = |f(0)|\prod_{1\le n\le m}\frac{r}{|a_n|}$, and (ii) $\log|g(r\exp\{i\theta\})| = \log|f(r\exp\{i\theta\})| - \sum_{m+1\le n\le N}\log|1 - \exp\{i(\theta - \theta_n)\}|$ where $a_n = r\exp\{i\theta n\}$, for $m + 1 \le n \le N$. [*Hint*: (i) follows from definition of g. (ii) holds because on $\{z : |z| = r\}$, the modulus of each of the factors in the first product in the expression for $g(z)$ is one.]

(d) Prove that $\int_{[-\pi,\pi]} \log|1 - \exp\{i\theta\}|d\theta = 0$. [*Hint*: This is proved by contour integration in Rudin (1974), Lemma 15.17. Another idea is the following. For $0 < b < 1$,

$$\begin{aligned}
&\int_{[-\pi,\pi]} \log|1 - b\exp\{i\theta\}|d\theta \\
&= \frac{1}{2}\int_{[-\pi,\pi]} \log|1 - b\exp\{i\theta\}|^2 d\theta \\
&= \frac{1}{2}\int_{[-\pi,\pi]} \log(1 - b\exp\{i\theta\})(1 - b\exp\{-i\theta\})d\theta \\
&= -\frac{1}{2}\int_{[-\pi,\pi]} \sum_{k\ge 1}[(bk\exp\{ik\theta\}/k) + (bk\exp\{-ik\theta\}/k)]d\theta \\
&= 0.
\end{aligned}$$

Now let $b \uparrow 1$. Unfortunately, the justification for the interchange of integration and limit is not clear here.]

(e) (Jensen's formula). Prove that

$$\begin{aligned} &|f(0)| \prod_{1 \le n \le N} (r/|a_n|) \\ &\quad = |f(0)| \prod_{1 \le n \le m} (r/|a_n|) \\ &\quad = \exp\{(1/2\pi) \int_{[-\pi,\pi]} \log |f(r \exp\{i\theta\}|d\theta. \end{aligned}$$

[*Hint*: For the first equality, note that $|r/a_n| = 1$ for $m + 1 \le n \le N$. For the second, use (b), (c), (d).]

Chapter 4
Birkhoff's Ergodic Theorem

In the context of stochastic processes, ergodic theory relates the long-run "time-averages" such as the sample mean of an evolving strictly stationary process $X_0, X_1, \ldots$ to a "phase-average" computed as an expected value with respect to a probability distribution on the state space. This is the perspective developed in this chapter.

In addition to the examples of the previous chapters, Markov processes having an invariant probability π also provide a broad class of examples in this regard. The following is a useful re-formulation of the notion of stationarity given in Definition 2.1 in the case of processes indexed by the nonnegative integers, as the invariance of the distribution of the process under a time shift map T on the (path) space S^∞.

Definition 4.1. A discrete parameter stochastic process $\{X_n : n \geq 0\}$ on $(\Omega, \mathcal{F}, P)$ with values in a measurable space $(S, \mathcal{S})$ is said to be *(strictly) stationary* if the distribution of $\mathbf{X} := (X_0, X_1, X_2, \ldots)$ is the same as that of $T^m\mathbf{X} \equiv X_m^+ := (X_m, X_{m+1}, X_{m+2}, \ldots)$ for all $m \geq 0$, where the transformation $T = T^1$, called the *shift transformation*, is defined on S^∞ into S^∞ as $T\mathbf{x} = (x_1, x_2, \ldots)$ for $\mathbf{x} = (x_0, x_1, x_2, \ldots) \in S^\infty$.

Note that the shift transformation is a measurable map on the sequence space S^∞ with respect to the product sigma-field $\mathcal{S}^\infty$ generated by finite dimensional projections.

The law of large numbers for stationary processes refers to a.s. limits of sample averages $\frac{1}{n}\sum_{m=0}^{n-1}\varphi(X_m)$ to a limit $l(\varphi)$ as $n \to \infty$ for suitable functions $\varphi : S \to$

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_4

$\mathbb{R}$. A far reaching generalization[1] of the law of large numbers is the main topic of this chapter. For this, we take an alternative perspective on the above average and express it as follows:

$$\frac{1}{n}\sum_{m=0}^{n-1}\varphi(X_m) = \frac{1}{n}\sum_{m=0}^{n-1} f(T^m\mathbf{X}), \tag{4.1}$$

where $f : S^\infty \to \mathbb{R}$ is defined by $f(\mathbf{x}) = \varphi(x_0), \mathbf{x} = (x_0, x_1, \ldots)$, and $T : S^\infty \to S^\infty$ is the shift transformation. Recall that the distribution of $(X_m, X_{m+1}, \ldots)$ is the probability measure $P \circ (T^m\mathbf{X})^{-1}$ induced on $(S^\infty, \mathcal{S}^{\otimes\infty})$ by the map $\omega \to (X_m(\omega), X_{m+1}(\omega), X_{m+2}(\omega), \ldots)$. From the perspective of dynamical systems, where iterates of T furnish the evolution, "stationarity" of $\mathbf{X}$ means that the probability (distribution) $P \circ \mathbf{X}^{-1}$ on sequence space S^∞ is preserved under the dynamics T. We will return to this perspective in Chapter 6.

Denote the σ-field generated by $\{X_n : n \geq 0\}$ by $\mathcal{G}$. That is, $\mathcal{G}$ is the class of all events of the form $G = [\mathbf{X} \in C] \equiv \mathbf{X}^{-1}C = \{\omega \in \Omega : \mathbf{X}(\omega) \in C\}, C \in \mathcal{S}^{\otimes\infty}$.

Definition 4.2. For an event $G = [\mathbf{X} \in C] \in \mathcal{G}$, write $T^{-1}G := \{\omega \in \Omega : T\mathbf{X}(\omega) \in C\} = [(X_1, X_2, \ldots) \in C] = [\mathbf{X} \in T^{-1}C]$. Such an event G is said to be *invariant* if $P(G\Delta T^{-1}G) = 0$, where Δ denotes the symmetric difference defined by $A\Delta B = (A \cap B^c) \cup (A^c \cap B)$.

Note that while $A\Delta B = \emptyset$ if $A = B$, invariant events are allowed to differ by a P-null event.

By iteration, it follows that if $G = [\mathbf{X} \in C]$ is invariant, then $P(G\Delta T^{-m}G) = 0$ for all $m \geq 0$, where $T^{-m}G = [(X_m, X_{m+1}, X_{m+2}, \ldots) \in C]$. Let f be a real-valued measurable function on $(S^\infty, \mathcal{S}^{\otimes\infty})$. Then $\omega \to f(\mathbf{X}(\omega))$ is $\mathcal{G}$-measurable, and, conversely, all $\mathcal{G}$-measurable functions are of this composite form.

Definition 4.3. Let f be a real-valued measurable function on $(S^\infty, \mathcal{S}^{\otimes\infty})$. A $\mathcal{G}$-measurable function $f(\mathbf{X})$ on $(\Omega, \mathcal{F}, P)$ is said to be *invariant* if $f(\mathbf{X}) = f(T\mathbf{X})$ a.s.

Note that $G = [\mathbf{X} \in C]$ is an invariant event if and only if $\mathbf{1}_G = \mathbf{1}_C(\mathbf{X})$ is an invariant function. Again, by iteration, if $f(\mathbf{X})$ is invariant, then $f(\mathbf{X}) = f(T^m\mathbf{X})$ a.s. for all $m \geq 1$.

In connection with the strong law of large numbers, we are interested in the following invariant events and functions. Given any $\mathcal{G}$-measurable real-valued function $f(\mathbf{X})$, the functions (extended real-valued)

$$\begin{aligned}\bar{f}(\mathbf{X}) &:= \overline{\lim}_{n\to\infty} n^{-1}(f(\mathbf{X}) + f(T\mathbf{X}) + \cdots + f(T^{n-1}\mathbf{X})),\\ \underline{f}(\mathbf{X}) &:= \underline{\lim}_{n\to\infty} n^{-1}(f(\mathbf{X}) + \cdots + f(T^{n-1}\mathbf{X}))\end{aligned} \tag{4.2}$$

[1] This result, which was motivated by considerations of the relationship between "time-averages" and "phase-averages" in statistical physics and dynamical systems, is due to Birkhoff (1931).

are invariant, and the event $[\bar{f}(\mathbf{X}) = \underline{f}(\mathbf{X})]$ is invariant. The class $\mathcal{I}$ of all invariant events (in $\mathcal{G}$) is easily seen to be a σ-field.

Definition 4.4. The class $\mathcal{I}$ of all invariant events (in $\mathcal{G}$) is called the *invariant σ-field*. The invariant σ-field $\mathcal{I}$ is said to be *trivial* if $P(G) = 0$ or 1 for every $G \in \mathcal{I}$.

Notice that a non-degenerate invariant event or function cannot depend on a finite segment of the process.

Definition 4.5. The process $\{X_n : n \geq 0\}$ and the shift transformation T are said to be *ergodic* if $\mathcal{I}$ is trivial.

Example 1. Let $\{X_n : n \geq 0\}$ be an i.i.d. sequence of real-valued random variables. By Kolmogorov's zero-one law,[2] the tail σ-field $\mathcal{T} := \cap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \dots)$ is trivial. Since the invariant σ-field is contained in $\mathcal{T}$, the sequence is therefore ergodic.

Example 2. Suppose that $\mathbf{Y} = (Y_0, Y_1, \dots)$ and $\mathbf{Z} = (Z_0, Z_1, \dots)$ are two i.i.d. sequences of Bernoulli $0-1$ (coin tossing) random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, with $P(Y_n = 1) = \alpha$ and $P(Z_n = 1) = \beta$, for $\alpha, \beta \in (0, 1)$. Suppose $A \in \mathcal{F}$ with $P(A) = p \in (0, 1)$. Define another process $\mathbf{X}$ by $\mathbf{X}(\omega) = \mathbf{Y}(\omega)$, $\omega \in A$, and $\mathbf{X}(\omega) = \mathbf{Z}(\omega)$, $\omega \in A^c$. Then $\mathbf{X}$ is a stationary process. However, if $\alpha \neq \beta$, then the strong law of large numbers, respectively, applied to $\mathbf{Y}$ and $\mathbf{Z}$, implies that the invariant event $G = [\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} X_m = \alpha]$ has probability $p \in (0, 1)$. Thus $\mathbf{X}$ is not ergodic.

As noted in the previous chapter, the implementation of some of the representation theory there may require estimates of expected values for which $f(\mathbf{x}) = x_0$, or variances/covariances for which, after centering, involves $f(\mathbf{x}) = x_0 x_k$ (k fixed) in the result below. In essence this is a generalization of the classical strong law of large numbers.

Theorem 4.1 *(Birkhoff's Ergodic Theorem).* Let $\{X_n : n \geq 0\}$ be a stationary sequence on the state space S (having σ-field $\mathcal{S}$). Let $f(\mathbf{X})$ be a real-valued $\mathcal{G}$-measurable function such that $\mathbb{E}|f(\mathbf{X})| < \infty$. Then

a $n^{-1} \sum_{m=0}^{n-1} f(T^m \mathbf{X})$ converges a.s. and in L^1 to an invariant random variable $g(\mathbf{X})$
b $g(\mathbf{X}) = \mathbb{E} f(\mathbf{X})$ a.s. if $\mathcal{I}$ is trivial.

We first need an inequality.[3] Write

$$\begin{aligned} M_n(f) &:= \max\{0, f(\mathbf{X}), f(\mathbf{X}) + f(T\mathbf{X}), \dots, f(\mathbf{X}) + \cdots + f(T^{n-1}\mathbf{X})\}, \\ M_n(f \circ T) &= \max\{0, f(T\mathbf{X}), f(T\mathbf{X}) + f(T^2\mathbf{X}), \dots, f(T\mathbf{X}) + \cdots + f(T^n\mathbf{X})\}, \\ M(f) &:= \lim_{n\to\infty} M_n(f) = \sup_{n\geq 1} M_n(f). \end{aligned} \tag{4.3}$$

[2] See BCPT, p. 87.

[3] The derivation presented here follows Garcia (1965).

Proposition 4.2 (Maximal Ergodic Theorem). Under the hypothesis of Theorem 4.1,

$$\int_{[M(f)>0]\cap G} f(\mathbf{X})dP \geq 0 \qquad \text{for all } \; G \in \mathcal{I}. \tag{4.4}$$

Proof. Note that $f(\mathbf{X}) + M_n(f \circ T) = M_{n+1}(f)$ on the event $[M_{n+1}(f) > 0]$. Since $M_{n+1}(f) \geq M_n(f)$ and $[M_n(f) > 0] \subset [M_{n+1}(f) > 0]$, it follows that $f(\mathbf{X}) \geq M_n(f) - M_n(f \circ T)$ on$[M_n(f) > 0]$. Also, $M_n(f) \geq 0$, $M_n(f \circ T) \geq 0$. Therefore,

$$\begin{aligned}\int_{[M_n(f)>0]\cap G} f(\mathbf{X})dP &\geq \int_{[M_n(f)>0]\cap G} (M_n(f) - M_n(f \circ T))dP \\ &\geq \int_G M_n(f)dP - \int_{[M_n(f)>0]\cap G} M_n(f \circ T)dP \\ &\geq \int_G M_n(f)dP - \int_G M_n(f \circ T)dP \\ &= 0,\end{aligned}$$

where the last equality follows from the invariance of G and the stationarity of $\{X_n : n \geq 0\}$. Thus, (4.4) holds with $[M_n(f) > 0]$ in place of $[M(f) > 0]$. Now let $n \uparrow \infty$. ■

Now consider the quantities

$$\begin{aligned}A_n(f) &:= \max\left\{f(\mathbf{X}), \frac{1}{2}(f(\mathbf{X}) + f(T\mathbf{X})), \ldots, \frac{1}{n}\sum_{m=0}^{n-1} f(T^m\mathbf{X})\right\}, \\ A(f) &:= \lim_{n\to\infty} A_n(f) = \sup_{n\geq 1} A_n(f).\end{aligned}$$

The following is a consequence of Proposition 4.2.

Corollary 4.3 *(Ergodic Maximal Inequality).* Under the hypothesis of Theorem 4.1, one has, for every $c \in \mathbb{R}$,

$$\int_{[A(f)>c]\cap G} f(\mathbf{X})dP \geq cP([A(f) > c] \cap G) \qquad \text{for all } \; G \in \mathcal{I}. \tag{4.5}$$

Proof. Apply Proposition 4.2 to the function $f - c$ to get

$$\int_{[M(f-c)>0]\cap G} f(\mathbf{X})dP \geq cP([M(f - c) > 0] \cap G).$$

But $[M_n(f-c)>0] \subset [A_n(f-c)>0] = [A_n(f)>c]$, and $[M(f-c)>0] \subset [A(f)>c]$. ■

We are now ready to prove Theorem 4.1, using (4.5).

Proof of Theorem 4.1. Write

$$\bar{f}(\mathbf{X}) := \overline{\lim}_{n\to\infty} \tfrac{1}{n}\sum_{r=0}^{n-1} f(T^r\mathbf{X}), \qquad \underline{f}(\mathbf{X}) := \underline{\lim}_{n\to\infty} \tfrac{1}{n}\sum_{r=0}^{n-1} f(T^r\mathbf{X}),$$

$$G_{c,d}(f) := [\bar{f}(\mathbf{X}) > c,\ \underline{f}(\mathbf{X}) < d] \qquad (c,d \in \mathbb{R}). \tag{4.6}$$

Since $G_{c,d}(f) \in \mathcal{I}$ and $G_{c,d}(f) \subset [A(f)>c]$, (4.5) leads to

$$\int_{G_{c,d}(f)} f(\mathbf{X})dP = \int_{[A(f)>c]\cap G_{c,d}(f)} f(\mathbf{X})dP \geq cP(G_{c,d}(f)). \tag{4.7}$$

Now take $-f$ in place of f and note that $\overline{(-f)} = -\underline{f}$, $\underline{(-f)} = -\bar{f}$, $G_{-d,-c}(-f) = G_{c,d}(f)$ to get from (4.7) the inequality $-\int_{G_{c,d}(f)} f(\mathbf{X})dP \geq -dP(G_{c,d}(f))$, i.e.,

$$\int_{G_{c,d}(f)} f(\mathbf{X})dP \leq dP(G_{c,d}(f)). \tag{4.8}$$

Now if $c > d$, then (4.7) and (4.8) cannot both be true unless $P(G_{c,d}(f)) = 0$. Thus, if $c > d$, then $P(G_{c,d}(f)) = 0$. Apply this to all pairs of rationals $c > d$ to get $P(\bar{f}(\mathbf{X}) > \underline{f}(\mathbf{X})) = 0$. In other words, $(1/n)\sum_{r=0}^{n-1} f(T^r\mathbf{X})$ converges a.s. to $h(\mathbf{X}) := \bar{f}(\mathbf{X})$. To complete the proof of part (a), it is enough to assume $f \geq 0$, since $n^{-1}\sum_0^{n-1} f^+(T^r\mathbf{X}) \to \overline{f}^+(\mathbf{X})$ a.s. and $n^{-1}\sum_0^{n-1} f^-(T^r\mathbf{X}) \to \overline{f}^-(\mathbf{X})$ a.s., where $f^+ = \max\{f,0\}$, $-f^- = \min\{f,0\}$. Assume then $f \geq 0$. First, by Fatou's lemma and stationarity of $\{X_n\}$,

$$\mathbb{E}\bar{f}(\mathbf{X}) = \mathbb{E}\underline{f}(X) \leq \underline{\lim}_{n\to\infty}\mathbb{E}\left(\frac{1}{n}\sum_{r=0}^{n-1} f(T^r(\mathbf{X}))\right) = \mathbb{E}f(\mathbf{X}) < \infty.$$

To prove the L^1-convergence, it is enough to prove the uniform integrability of the sequence $\{(1/n)S_n(f) : n \geq 1\}$, where[4] $S_n(f) := \sum_{m=0}^{n-1} f(T^m\mathbf{X})$. Now since $f(\mathbf{X})$ is nonnegative and integrable, given $\epsilon > 0$, there exists a constant N_ϵ such that $\|f(\mathbf{X}) - f_\epsilon(\mathbf{X})\|_1 < \epsilon$, where $f_\epsilon(\mathbf{X}) := \min\{f(\mathbf{X}), N_\epsilon\}$. Then

[4] See BCPT, p. 17 for this L^1-convergence criteria.

$$\begin{aligned}\int_{[\frac{1}{n}S_n(f)>\lambda]} \frac{1}{n}S_n(f)dP &\leq \int_{\Omega} \frac{1}{n}S_n(f-f_\epsilon)dP + \int_{[\frac{1}{n}S_n(f)>\lambda]} \frac{1}{n}S_n(f_\epsilon)dP \\ &\leq \epsilon + N_\epsilon P\left(\frac{1}{n}S_n(f) > \lambda\right) \\ &\leq \epsilon + N_\epsilon \mathbb{E}f(\mathbf{X})/\lambda. \qquad (4.9)\end{aligned}$$

It follows that the left side of (4.9) goes to zero as $\lambda \to \infty$, uniformly for all n. Therefore $S_n(f)/n$ converges in L^1 to $h(\mathbf{X})$. To show that $h(\mathbf{X}) = \mathbb{E}(f(\mathbf{X})|\mathcal{I})$, let $G := [\mathbf{X} \in C] \in \mathcal{I}$ for some $C \in \mathcal{S}^{\otimes\infty}$, and observe that if we write $\mathbf{1}_G = \mathbf{1}_C(\mathbf{X}) = h(\mathbf{X})$, then $h(T^m\mathbf{X}) = g(\mathbf{X})$ a.s. for every $m = 1, 2, \ldots$. Hence, using the measure-preserving property of T,

$$\mathbb{E}\mathbf{1}_G\frac{1}{n}\sum_{m=0}^{n-1} f(T^m\mathbf{X}) = \frac{1}{n}\sum_{m=0}^{n-1} \mathbb{E}g(T^m\mathbf{X})f(T^m\mathbf{X}) = \mathbb{E}g(\mathbf{X})f(\mathbf{X}) = \mathbb{E}\mathbf{1}_G f(\mathbf{X}). \qquad (4.10)$$

Letting $n \to \infty$, we get the desired relation $\int_G h(\mathbf{X})dP = \int_G f(\mathbf{X})dP$. Part (b) is an immediate consequence of part (a). ■

Corollary 4.4. If $\{X_n : n \geq 0\}$ is a stationary process with state space $(S, \mathcal{S})$ and f is a real-valued measurable function on S such that $\mathbb{E}|f(X_0)| < \infty$, then a.s. and in L^1

$$\frac{1}{n}\sum_{m=0}^{n-1} f(X_m) \to \mathbb{E}(f(X_0)|\mathcal{I})$$

as $n \to \infty$.

Example 3 *(The Classical Strong Law of Large Numbers).* Let $\{X_n : n \geq 0\}$ be an i.i.d. sequence of real-valued random variables. As observed in Example 1, this process is ergodic. If $\mathbb{E}|X_0| < \infty$, then it follows from the ergodic theorem that $\frac{1}{n}\sum_{m=0}^{n-1} X_m \to \mathbb{E}X_0$ as $n \to \infty$ a.s. and in L^1.

Example 4 *(Exchangeable Sequences of Random Variables).* Suppose that $\{X_n : n \geq 0\}$ is a discrete parameter stochastic process with values in a measurable space $(S, \mathcal{S})$ defined on some probability space $(\Omega, \mathcal{F}, P)$. Assume for every $m = 0, 1, 2, \ldots$ the distributions of $(X_0, X_1, \ldots, X_m)$ and $(X_{n_0}, \ldots, X_{n_m})$ for any set of $m + 1$ distinct indices $n_0, \ldots, n_m$. Such a process (or its distribution) is said to be *exchangeable*. Clearly an exchangeable process is stationary. The "mixture" (convex combination) of two, or any finite number, of i.i.d. sequences is exchangeable. More precisely, let $\mu_1, \ldots, \mu_k$ be probabilities on $(S, \mathcal{S})$ and suppose that $\{X_{n_j} : n \geq 0\}$ is an i.i.d. sequence on $(\Omega, \mathcal{F}, P)$ having distribution μ_j for each $j = 1, 2, \ldots, k$. Let J be a random index, independent of the processes $\{X_{n,j} : n \geq 0\}$ for $j = 1, 2, \ldots, k$, with $P(J = j) = p_j$, with $0 < p_j < 1, \sum_{j=1}^{k} p_j = 1$. Then define $X_n = X_{n,J}, n = 0, 1, 2, \ldots$, for every $m \geq 0$, $B \in \mathcal{S}^{\otimes(m+1)}$, and distinct indices $n_0, \ldots, n_m$,

$$\begin{aligned} P((X_{n_0}, \ldots, X_{n_m}) \in B) &= \mathbb{E}P((X_{n_0}, \ldots, X_{n_m}) \in B|J) \\ &= \sum_{j=1}^{k} p_j \mathbb{E}P(X_{n_0}, \ldots, X_{n_m}) \in B|J = j) \\ &= \sum_{j=1}^{k} p_j P((X_{n_0,j}, \ldots, X_{n_m,j}) \in B) \\ &= \sum_{j=1}^{k} p_j P((X_{0,j}, \ldots, X_{m,j}) \in B) \\ &= P((X_{0,J}, \ldots, X_{m,J}) \in B). \end{aligned} \tag{4.11}$$

Thus $\{X_n : n \geq 0\}$ is exchangeable. If f is a bounded real-valued measurable function on $(S, \mathcal{S})$, then a.s. $\frac{1}{n}\sum_{r=0}^{n-1} f(X_r) = \frac{1}{n}\sum_{r=0}^{n-1} f(X_{r,J}) \to \int_S f d\mu_J$, where $\int_S f d\mu_J$ is a random variable which takes the value $\int_S f d\mu_j$ with probability p_j, $j = 1, \ldots, k$. In particular this shows that $\{X_n : n \geq 0\}$ is not ergodic. More generally, let S be a Polish space and $\mathcal{S}$ its Borel σ-field. Let $\mathcal{P}(S)$ be the set of all probabilities on $(S, \mathcal{S})$ and $\mathcal{B}(\mathcal{P})$ its Borel σ-field for the weak topology.

According to the de Finetti theorem,[5,6] therefore, every exchangeable sequence $\{X_n : n \geq 0\}$ with values in $(S, \mathcal{S})$ may be represented as a *mixture* of i.i.d. sequences: $X_n = X_{n,J}, n \geq 0$, where J is a random index with values in $(\mathcal{P}(S), \mathcal{B}(\mathcal{P}))$, and for each $\nu \in \mathcal{P}(S)$, $\{X_{n\nu} : n \geq 0\}$ is an i.i.d. sequence with common distribution ν. Thus if J is not a.s. constant, i.e., $\{X_n : n \geq 0\}$ is exchangeable but not i.i.d., then $\{X_n : n \geq 0\}$ is not ergodic, and for every measurable $f : S \to \mathbb{R}$ such that $\mathbb{E}|f(X_0)| < \infty$, one has a.s.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} f(X_m) = \int_S f(x)\mu_J(dx). \tag{4.12}$$

The following provides an alternative description of ergodicity in a weak sense of "asymptotic independence" of events of the form $[\mathbf{X} \in A]$ and $[T^m\mathbf{X} \in B]$.

Definition 4.6. A stationary process $\mathbf{X} = \{X_0, X_1, \ldots\}$ with values in a measurable space $(S, \mathcal{S})$ is said to be *weak mixing* if for all $A, B \in \mathcal{S}^{\otimes\infty}$, one has

[5] Bhattacharya and Waymire (2021), p. 162.

[6] An extension of de Finetti's theorem to exchangeable Markov processes was initiated in Diaconis and Freedman (1980) that is worthy of mention here. Especially see James et al. (2008) for inspiring connections to transient random walk.

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1}P([\mathbf{X}\in A]\cap[T^m\mathbf{X}\in B]) = P(\mathbf{X}\in A)P(\mathbf{X}\in B).$$

Proposition 4.5. A stationary process is ergodic if and only if it is weak mixing.

Proof. For sufficiency, suppose $\mathbf{X} = \{X_0, X_1, \ldots\}$ is stationary and weak mixing. Let $[\mathbf{X}\in A] \in \mathcal{I}$. Letting $B = A$, we get $P(\mathbf{X}\in A) = P(\mathbf{X}\in A)P(\mathbf{X}\in A)$ and, therefore, $P(\mathbf{X}\in A) = 0$ or 1. Thus $\mathbf{X}$ is ergodic. For the converse, suppose that $\mathbf{X}$ is ergodic. Let $A, B \in \mathcal{S}^{\otimes\infty}$. Then a.s.

$$\lim_{n\to\infty}\mathbf{1}_A(\mathbf{X})\frac{1}{n}\sum_{m=0}^{n-1}\mathbf{1}_B(T^m\mathbf{X}) = \mathbf{1}_A(\mathbf{X})P(\mathbf{X}\in B). \tag{4.13}$$

Taking expectations on both sides proves weak mixing. ■

One needs stronger conditions than weak mixing to derive a central limit theorem (CLT) for partial sums of a stationary process. We briefly mention one of these strong mixing conditions here and state the corresponding CLT without proof. For a comprehensive account of the vast literature on strong mixing conditions and CLTs under them, we refer to Bradley (2003). Among other references, one may mention Billingsley (1968, Theorem 21.1), Ibragimov and Linnik (1971), and Denker (1986).

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which is defined a stationary process $\{X_n : n = 0, 1, 2, \ldots\}$ with values in some measurable (state) space. Consider the sigma-fields $\mathcal{F}_r^t = \sigma\{X_n : r \le n \le t\}$, $\mathcal{F}_r^\infty = \sigma\{X_n : n \ge r\}$.

Definition 4.7. A stationary sequence $\{X_n\}$ is said to be α-mixing, if

$$\alpha(n) = \sup\{|P(A\cap B) - P(A)P(B)| : A \in \mathcal{F}_0^t, B \in \mathcal{F}_{t+n}^\infty, t \ge 0\} \to 0$$

as $n \to \infty$.

A CLT originally derived by Rosenblatt (1956), with an additional moment condition that was relaxed by Cogburn (1960), can be stated as follows.

Proposition 4.6 (Rosenblatt–Cogburn CLT). Let $\{X_n : n = 0, 1, 2, \ldots\}$ be a real-valued α-mixing stationary sequence such that $\mathbb{E}X_n = 0$, $\mathbb{E}X_n^2 < \infty$. Denote $S_n = \sum_{m=0}^n X_m$. If $\sigma_n^2 = \mathbb{E}S_n^2 \to \infty$ as $n \to \infty$, then $\frac{S_n}{\sigma_n}$ converges in distribution to the standard normal distribution $N(0, 1)$ as $n \to \infty$.

The technique for the proof involves breaking up the sum S_n into consecutive "large" and "small" blocks, such that the large blocks are nearly independent of each other, while the small blocks are negligible. The significance of the condition $\mathbb{E}S_n^2 \to \infty$ may be understood by considering the example in which $X_n = Y_n - Y_{n-1}$, where $\{Y_n\}$ is an i.i.d. sequence.

The main emphasis of the present book in the context of such asymptotic limit theorems is on specific dependence structures such as martingales and Markov processes, as detailed in Chapters 15, 16, and 19. The corresponding results generally do not require stationarity in their formulation.

Exercises

1. Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with $\mathbb{E}|X_1| < \infty$. Let $S_n = X_1 + \cdots + X_n, n \geq 1$. Use the ergodic maximal inequality to prove the following for $\lambda > 0$:

   (a) $P(\max_{1\leq k\leq n} \frac{S_k}{k} \geq \lambda) \leq \frac{\mathbb{E}|X_1|}{\lambda},\ n \geq 1$..

   (b) $P(\lim_{n\to\infty} \max_{1\leq k\leq n} \frac{S_k}{k} \geq \lambda) \leq \frac{\mathbb{E}|X_1|}{\lambda}$.

2. Let X be a random variable on $(\Omega, \mathcal{F}, P)$. Define $X_n = X, n = 0, 1, 2, \ldots$. i. Show that $\{X_n\}$ is a stationary process. ii. Show that $\{X_n\}$ is ergodic if and only if X is almost surely constant.
3. Suppose that $\mathbf{Y} = \{Y_n : n \geq 1\}$ and $\mathbf{Z} = \{Z_n : n \geq 1\}$ are two stationary ergodic sequences of 0-1 valued random variables. Show that the distributions of $\mathbf{Y}$ and $\mathbf{Z}$ are mutually singular if and only if $P(Y_1 = 1) \neq P(Z_1 = 1)$. [*Hint*: Use the ergodic theorem to find a set C such that $P(\mathbf{Y} \in C) = 1$ and $P(\mathbf{Z} \in C) = 0$.]
4. (*Symmetric Difference*) Let $(\Omega, \mathcal{F}, P)$ be a probability space.

   (a) Show that $P(A\Delta B) \leq P(A\Delta C) + P(C\Delta B)$ for any $A, B, C \in \mathcal{F}$, where $A\Delta B = (A \cup B)\backslash(A \cap B)$.

   (b) Suppose that $\mathcal{F} = \sigma(\mathcal{C})$, where $\mathcal{C}$ is a π-system of subsets generating $\mathcal{F}$. Show that for any $B \in \mathcal{F}$ and $\epsilon > 0$, there is a $C \in \mathcal{C}$ such that $P(B\Delta C) < \epsilon$. [*Hint*: Define $\mathcal{C} \subset \mathcal{L} = \{B \in \mathcal{F}$: for any $\epsilon > 0$, there exists $C \in \mathcal{C}$ such that $P(B \cap C) < \epsilon\} \subset \mathcal{F}$ and use Dynkin's $\pi - \lambda$ theorem.[7]]

5. Let $\mathbf{X} = \{X_n : n \geq 0\}$ be a stationary process on a probability space $(\Omega, \mathcal{F}, P)$ with a measurable state space $(S, \mathcal{S})$. Show that $\mathbf{X}$ is an ergodic process if and only if it has the property that the only invariant functions (see Definition 4.3) $f(\mathbf{X}) \in L^2(\Omega, \mathcal{F}, P)$ are almost surely constant functions.
6. (*Shannon Entropy*) Let $S = \{1, 2, \ldots, k\}$ be a finite set. Let $\pi = (\pi_1, \ldots, \pi_k)$ be a probability mass function and $\mathbf{p} = ((p_{ij}))_{i,j\in S}$ a stochastic matrix, i.e., $\sum_{k\in S} p_{ik} = 1, p_{ij} \geq 0$, for all $i, j \in S$. Assume $\sum_{i\in S} \pi_i p_{ij} = \pi_j, \forall\ j \in S$. Apply the Kolmogorov extension theorem[8] to construct a stochastic process $\{X_n : n \geq 0\}$ on the product probability space $(\Omega = S^\infty, \mathcal{S}^{\otimes\infty}, P_\pi)$ such that

[7] See BCPT p.4.

[8] See BCPT p.168.

$P_\pi(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \pi_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}, i_0, i_1, \dots, i_n \in S, n \geq 0$.

(a) Show that $\mathbf{X} = \{X_n\}$ is a stationary process.

(b) Show that the *entropy*, defined by $H(\mathbf{X}) = -\lim_{n\to\infty} \frac{\log(\pi_{X_0} \prod_{m=0}^{n-1} p_{X_m,X_{m+1}})}{n}$, exists P_π-a.s. [*Hint*: View p_{X_0,X_1} as a function of $\mathbf{X}$. Then, $p_{X_m,X_{m+1}}$ is the said function of $T^m\mathbf{X}$, where T is the shift transformation.

(c) Show that if $\mathbf{X}$ can be proven to be ergodic, then $H(\mathbf{X}) = -\sum_{i,j\in S} \pi_i p_{ij} \log p_{ij}$. [Ergodicity will indeed be proven in Chapter 16 in more generality.]

7. Compute $\lim_{n\to\infty} \frac{X_n}{n}$ for the non-ergodic stationary process defined in Example 2.

8. (*Range of Random Walk*)[9] Let $\{X_n : n \geq 1\}$ be i.i.d. $\mathbb{R}^k$-valued random variables and $S_n = X_1 + \cdots + X_n, n \geq 1, S_0 = 0$, and consider the number $R_n = |\{S_0, S_1, \dots, S_n\}|$ of distinct sites visited by time n, i.e., the range of the random walk in time n. Use Birkhoff's ergodic theorem to show $R_n/n \to P(S_j \neq 0, j = 1, 2, \dots)$ a.s. as $n \to \infty$. [*Hint*: Write $S_n(\omega) = \sum_{m=1}^n X_1(T^m\omega), n \geq 1, \omega = (\omega_1, \omega_2, \dots) \in (\mathbb{R}^k)^\infty$ for $X_1(\omega) = \omega_1$ and the shift map T. Check that for arbitrary $k \geq 1$, $R_n(\omega) \leq k + \sum_{j=1}^{n-k} \mathbf{1}_{[S_j\neq 0, j=1,\dots,k]}(T^j\omega), n > k$. In particular, $\limsup_n \frac{R_n}{n} \leq P(S_j \neq 0, \forall j \geq 1|\mathcal{J})$. Similarly, find a (simpler) lower bound for each n to show $\liminf_n \frac{R_n}{n} \geq P(S_j \neq 0, j = 1, 2, \dots|\mathcal{J})$. The result follows by calculating the indicated conditional probability.]

9. (*Doubly Stochastic Poisson Process/Cox Process*) Suppose that Λ is a positive non-degenerate random variable and, conditionally given $\Lambda = \lambda, T_1, T_2, \dots$ is an i.i.d. sequence of exponentially distributed random variables with parameter $\lambda > 0$. Show that $\lim_{n\to\infty} \frac{S_n}{n} = \Lambda$ with probability one, where $S_n = T_1 + \cdots + T_n$.

10. Prove deFinnetti's representation of exchangable 0-1 valued stochastic processes $\{X_n : n \geq 1\}$ by using the Riesz representation theorem[10] by completing the following steps:

(a) Let $\lambda_{h,k} = P(X_1 = 1, \dots, X_k = 1, X_{k+1} = 0, \dots, X_h = 0), 1 \leq k < h, h \geq 2$, with $\lambda_{0,0} = 1, \lambda_{0,k} = P(X_1 = 1, \dots, X_k = 1), \lambda_{1,0} = P(X_1 = 0)$, and check that $\lambda_{h,k} = \lambda_{h+1,k} + \lambda_{h+1,k+1}$.

(b) Define a linear functional ℓ on the dense subspace of polynomials in $C[0, 1]$ by linearity and $\ell(x^k) = \lambda_{0,k}, k \geq 0$. Show that ℓ has a continuous extension to a bounded linear functional on $C[0, 1]$.

(c) Let μ denote the probability measure in the Riesz representation of ℓ, i.e., $\ell(f) = \int_0^1 f(x)\mu(dx), f \in C[0, 1]$. Show that $P(X_1 = 1, \dots, X_k =$

[9] See Spitzer (1964), p. 38., where the result is attributed to Kesten, H., F. Spitzer, and W. Whitman. This result had been obtained for the k-dimensional simple symmetric random walk in an earlier paper by Dvoretzky and Erdos (1951).

[10] See BCPT, p.237.

$1, X_{k+1} = 0) = \int_0^1 x^k(1-x)\mu(dx)$. [*Hint*: $[X_1 = 1, \dots, X_k = 1, X_{k+1} = 0] = X_1 = 1, \dots, X_k = 1]\backslash[X_1 = 1, \dots, X_k = 1, X_{k+1} = 1]$.]

(d) Extend (c) by inclusion–exclusion to show $P(X_1 = \epsilon_1, \dots, X_k \epsilon_k) = \int_0^1 x^{\sum_{j=1}^k \epsilon_j}(1-x)^{k-\sum_{j=1}^k \epsilon_j} dx$ for any $\epsilon_1, \dots, \epsilon_k \in \{0, 1\}^k$.

(e) Consider an i.i.d. sequence $\{Y_n : -\infty < n < \infty\}$, $\mathbb{E}Y_n = 0$, $\mathbb{E}Y_n^2 = 1$. Let $X_n = Y_n - Y_{n-m}$ $(n = 0, 1, \dots)$ for some fixed $m \geq 1$.

(i) Prove that $\{X_n\}$ is m-dependent, i.e., $\mathcal{F}_j^k = \sigma\{X_n : j \leq n < k\}$ is independent, of $\mathcal{F}_{k+m}^\infty$ for every $j < k < \infty$.

(ii) Given any sequence of constants $\beta_n \to \infty$, show that $\frac{S_n}{\beta_n} \to 0$ in probability as $n \to \infty$.

(f) Consider an i.i.d. sequence $\{Y_n\}$ as in Exercise 10, and let $X_n = Y_n + Y_{n+m}(n = 0, 1, 2, \dots)$ for some fixed $m \geq 1$. Prove that the m-dependent sequence $\{X_n\}$ satisfies the hypothesis of Proposition 4.6.

Chapter 5
Subadditive Ergodic Theory

Subadditivity of a sequence of positive real numbers $x_1, \dots$ refers to the property $x_{m+n} \le x_m + x_n,\ n \ge 1$. For such sequences, it is a calculus exercise to verify that $\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\ge 1} \frac{x_m}{m}$. The extension of this notion to almost sure convergence of a corresponding class of stochastic processes is the objective of this chapter.

Subadditivity of sequences of non-negative numbers is easily seen to result in asymptotic stability. Specifically, one has the following solution[1] to the calculus problem raised in the abstract.

Proposition 5.1 *(Fekete).* Suppose that $\{a_n\}_{n=1}^{\infty}$ is a sequence of numbers with the subadditivity property:

$$a_{m+n} \le a_m + a_n, \quad n, m \ge 1.$$

Then

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{m\ge 1} \frac{a_m}{m}.$$

Proof. One has

$$\frac{a_n}{n} \le \frac{m}{n}\frac{a_m}{m} + \frac{n-m}{n}\frac{a_{n-m}}{n-m}.$$

[1] Fekete (1923).

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_5

Thus, $\lim_n \frac{a_n}{n} \in [\inf_{m\geq 1} \frac{a_m}{m}, \sup_{m\geq 1} \frac{a_n}{n}]$ if the limit exists. In particular,

$$\liminf_n \frac{a_n}{n} \geq \inf_{m\geq 1} \frac{a_m}{m}.$$

Conversely, write $n = km + \ell$ for any m, $0 \leq \ell < m$. Then $a_n \leq ka_m + a_\ell$, so that

$$\frac{a_n}{n} \leq \frac{km}{km+\ell}\frac{a_m}{m} + \frac{a_\ell}{n}.$$

Thus, for $m = 1, 2, \ldots,$

$$\limsup_n \frac{a_n}{n} \leq \frac{a_m}{m},$$

and hence

$$\limsup_n \frac{a_n}{n} \leq \inf_{m\geq 1} \frac{a_m}{m}.$$

■

An important stochastic version[2] can be stated as follows.[3] To distinguish the *subadditivity property* $a_{m+n} \leq a_m + a_n$, we will henceforth refer to the condition (a) below as *array subadditivity*.

Theorem 5.2 *(Kingman–Liggett Subadditivity Theorem).* Let $\{Z_{m,n} : 0 \leq m < n, n = 1, 2, \ldots\}$ be a collection of random variables such that

a *(Array Subadditivity):* $Z_{0,n} \leq Z_{0,m} + Z_{m,n}, \quad 0 < m < n, n = 1, 2, \ldots$
b For each $m \geq 0$, $\{Z_{m,m+k} : k \geq 1\}$ has the same distribution as $\{Z_{0,1}, Z_{0,2}, \ldots\}$.
c For each $k \geq 1$,$\{Z_{k,2k}, Z_{2k,3k}, \ldots\}$ is a stationary process.
d $\mathbb{E}Z^+_{0,1} < \infty$, and $\inf_n \mathbb{E}\frac{Z_{0,n}}{n} \in [-\infty, \infty)$.

Then, letting $\gamma = \inf_n \frac{1}{n}\mathbb{E}Z_{0,n}$, one has

i $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}Z_{0,n} = \gamma$
ii $\lim_n \frac{1}{n} Z_{0,n} = \overline{Z}$

exists a.s. for some random variable $\overline{Z}$ and, if $\gamma > -\infty$, then it exists in L^1 with $\mathbb{E}\overline{Z} = \gamma$ as well. If the stationary process $\{Z_{k,2k}, Z_{2k,3k}, \ldots\}$ is ergodic, then $\overline{Z} = \gamma$ a.s.

[2] Kingman (1976) provided the initial breakthrough in exploiting subadditivity for an ergodic theory of stationary processes. Liggett (1985) provided the strengthening given here and finds applications for which the hypothesis of Kingman is too strong. The original version of Kingman contains assumption (d), but he required the conditions that $Z_{m,k}+Z_{k,n} \geq Z_{m,n}$, $m = 2, \ldots, n-1$, and that the distribution of $\{Z_{m+k,n+k} : m = 0, 1, \ldots, n-1\}$ be independent of k. These prove to be too strong for some applications.

[3] The proof here follows Kallenberg (2002) and Durrett (1991).

Proof. First, (a) implies $Z^{+}_{0,n} \le Z^{+}_{0,m} + Z^{+}_{m,n}$. Also, $\mathbb{E}Z^{+}_{m,n} = \mathbb{E}Z^{+}_{0,n-m}$ by (b). Hence $\mathbb{E}Z^{+}_{0,n} \le \mathbb{E}Z^{+}_{0,1} + \mathbb{E}Z^{+}_{1,n} = \mathbb{E}Z^{+}_{0,1} + \mathbb{E}Z^{+}_{0,n-1} \le \cdots \le n\mathbb{E}Z^{+}_{0,1}$. Thus $\mathbb{E}Z_{m,n}$ exists and is finite, and $\{\mathbb{E}Z_{0,n} : n \ge 1\}$ and $\{\mathbb{E}Z^{+}_{0,n} : n \ge 1\}$ are *subadditive* (Exercise 5). In particular, by Proposition 5.1, $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}Z_{0,n} = \gamma < \infty$. This proves (i), with $\gamma \in [-\infty, \infty)$. For (ii), assume $\gamma \in (-\infty, \infty)$ first. The array subadditivity property implies that, for each $k \ge 1$, $Z_{0,n} \le Z_{0,mk} + Z_{mk,n} \le Z_{0,(m-1)k} + Z_{(m-1)k,mk} + Z_{mk,n}$. Iterating repeatedly, with $m = [\frac{n}{k}]$, the integer part of $\frac{n}{k}$, one arrives at

$$\frac{Z_{0,n}}{n} \le \frac{1}{n}\sum_{j=1}^{[\frac{n}{k}]} Z_{(j-1)k,jk} + \frac{1}{n}\sum_{j=[\frac{n}{k}]k+1}^{n} Z_{j-1,j}, \quad k, n = 1, 2, \ldots \tag{5.1}$$

By Birkhoff's ergodic theorem applied to the stationary process $Z_{(j-1)k,jk}$, $j = 1, 2 \ldots$, one has $\frac{1}{n}\sum_{j=1}^{[\frac{n}{k}]} Z_{(j-1)k,jk} \to \frac{\overline{Z}_{0,k}}{k}$ a.s. and in L^1, where $\overline{Z}_{0,k}$ is the conditional expectation given the shift-invariant σ-field $\mathcal{T}_k$. In particular, $\mathbb{E}\overline{Z}_{0,k} = \mathbb{E}Z_{0,k}$. Also, the second term is o(1) a.s. and in L^1 since, by the ergodic theorem, $\frac{1}{n}\sum_{j=1}^{n} Z_{j-1,j}$ converges to $\overline{Z}_{0,1}$ accordingly. Now, since the bound (5.1) converges to $\frac{\overline{Z}_{0,k}}{k}$ for each $k = 1, 2, \ldots$, letting $n \to \infty$ in (5.1), one obtains a.s. and in L^1

$$\limsup_{n\to\infty} \frac{Z_{0,n}}{n} \le \inf_k \frac{\overline{Z}_{0,k}}{k} = \overline{Z} < \infty. \tag{5.2}$$

Since convergence in L^1 implies[4] uniform integrability, it follows that

$$\mathbb{E}\limsup_{n\to\infty} \frac{Z_{0,n}}{n} \le \mathbb{E}\inf_n \frac{\overline{Z}_{0,n}}{n} \le \inf_n \mathbb{E}\frac{\overline{Z}_{0,n}}{n} = \inf_n \mathbb{E}\frac{Z_{0,n}}{n} = \gamma < \infty. \tag{5.3}$$

For the reverse inequality, first note that by subadditivity of the numerical sequence $\mathbb{E}Z_{0,j}$, $j = 1, 2, \ldots$, one has

$$\lim_{n\to\infty} \frac{1}{n}\{\mathbb{E}Z_{0,n+k} - \mathbb{E}Z_{0,k}\} = \inf_n \frac{1}{n}\mathbb{E}Z_{0,n} = \gamma, \quad k = 1, 2, \ldots \tag{5.4}$$

Noting that $\frac{1}{n}\sum_{m=1}^{n} \mathbb{E}Z_{m,m+k} = \mathbb{E}Z_{U_n,U_n+k}$, where U_n is uniformly distributed on $\{1, 2, \ldots n\}$ and independent of $Z_{j,k}$'s, it is convenient to define

$$X_{k,n} = Z_{U_n,U_n+k}, \quad Y_{k,n} = Z_{0,U_n+k} - Z_{0,U_n+k-1}. \tag{5.5}$$

[4] BCPT, p.17.

Then hypothesis (**b**) implies that for each n, $\{X_{1,n}, X_{2,n}, \dots\}$ and $\{Z_{0,1}, Z_{0,2}, \dots\}$ have the same distribution. Moreover, $Y_{k,n}, n \geq 1$, is also uniformly integrable since $Y_{k,n}$ is distributed as $Z_{0,k+1} - Z_{0,k}$ for all $n \geq 1$. It follows from (5.4) that a.s. as $n \to \infty$, one has

$$\mathbb{E}Y_{k,n} = \frac{1}{n}\sum_{m=1}^{n}[\mathbb{E}Z_{0,m+k} - \mathbb{E}Z_{0,m+k-1}] = \frac{1}{n}[\mathbb{E}Z_{0,n+k} - \mathbb{E}Z_{0,k}] \to \gamma$$

(see Exercise 2(d)). In particular, therefore, $\sup_n \mathbb{E}|Y_{k,n}| < \infty$, so that the sequence $Y_{k,n}, n = 1, 2, \dots$, is tight for each k. Extracting a weakly convergent subsequence,[5] one has as $n \to \infty$,

$$(\{X_{k,n}\}_{k=1}^{\infty}, \{Y_{k,n}\}_{k=1}^{\infty}) \Rightarrow (\{X_k\}_{k=1}^{\infty}, \{Y_k\}_{k=1}^{\infty}),$$

for some random variables $X_k, Y_k, k \geq 1$, with $\{X_k\}_{k=1}^{\infty}$ distributed as $\{Z_{0,k}\}_{k=1}^{\infty}$. By array subadditivity, one has

$$Y_{1,n} + \cdots + Y_{k,n} = Z_{0,U_n+k} - Z_{0,U_n} \leq Z_{U_n,U_n+k} = X_{k,n}.$$

Thus, letting $n \to \infty$, one has a.s. that $Y_1 + \cdots Y_k \leq^s X_k$ for each k, and, therefore, $\{Y_k : k \geq 1\}$ is a stationary integrable sequence. Here $V \leq^s W$ denotes that V is *stochastically smaller* than W, i.e., $P(W > t) \geq P(V > t)$ for all $t \in \mathbb{R}$. Also note that the distribution of $(Y_{1,n}, \dots, Y_{k,n})$ is the same as that of $(Y_{j+1,n}, \dots, Y_{j+k,n})$ for all $j \geq 1, k \geq 1$ (and $n \geq 1$). Thus, again using the ergodic theorem, one has a.s. and in L^1, as $n \to \infty$,

$$\frac{1}{n}Z_{0,n} \geq^s \frac{1}{n}\sum_{k=1}^{n} Y_k \to \tilde{Y}, \tag{5.6}$$

for some integrable random variable $\tilde{Y}$. It follows that the negative parts $\frac{1}{n}Z_{0,n}^-, n \geq 1$, and hence $\frac{1}{n}Z_{0,n}, n \geq 1$, are uniformly integrable sequences (Exercise 2). With this and uniform integrability of $Y_{k,n}^+$, one has

$$\begin{aligned}
\gamma &= \lim_{n\to\infty} \mathbb{E}Y_{1,n} = \mathbb{E}Y_1 = \mathbb{E}\tilde{Y} \\
&\leq \mathbb{E}\liminf_n \frac{1}{n}Z_{0,n} \quad \text{(by (5.6))} \\
&\leq \mathbb{E}\limsup_n \frac{1}{n}Z_{0,n} \\
&\leq \mathbb{E}\overline{Z} \leq \gamma \quad \text{(by (5.3))}.
\end{aligned} \tag{5.7}$$

[5] BCPT, pp. 142–145.

In particular, therefore, $\mathbb{E}\overline{Z} = \gamma$. Moreover, it now follows that $0 \leq \mathbb{E}[\limsup_n \frac{1}{n}Z_{0,n} - \liminf_n \frac{1}{n}Z_{0,n}] = 0$, so $\lim_n \frac{Z_{0,n}}{n}$ exists a.s. and, noting uniform integrability (recall γ is assumed finite), also in L^1 (see Exercise 2). Combining this with $0 \leq \mathbb{E}[\overline{Z} - \limsup_n \frac{1}{n}Z_{0,n}] = 0$, one gets $\frac{Z_{0,n}}{n} \to \overline{Z}$ a.s. and in L^1 as $n \to \infty$.

In the case $\gamma = -\infty$, a truncation argument outlined in Exercise 7 shows that a.s. convergence in the previous theorem is still valid, but the uniform integrability arguments for L^1 convergence are not applicable.

Finally, if $\{Z_{k,2k}, Z_{2k,3k}, \dots\}$ is ergodic, then $\overline{Z}_{0,n} = \mathbb{E}Z_{0,n}$ a.s. for each n and, therefore, $\overline{Z} = \gamma$ almost surely. ■

Example 1. For a very special case, consider an i.i.d. sequence $X_1, X_2, \dots$ with finite first moment. Define $Z_{0,0} = 0$ and $Z_{m,n} = X_{m+1} + \cdots + X_n, 0 \leq m < n$. Then one has the additivity property

$$Z_{0,n} = Z_{0,m} + Z_{m,n}.$$

Almost sure convergence follows directly from the strong law of large numbers, i.e., ergodic theorem in the i.i.d. case, and the other conditions provide uniform integrability for convergence in L^1. In this sense, one may view the theorems of Kingman and Liggett as extensions of Birchoff's ergodic theorem to subadditive arrays.

Example 2 *(Range of Random Walk).* For a more substantive illustration, let $\{S_n = \sum_{j=1}^{n} : n = 1, 2, \dots\}$ denote a random walk on $\mathbb{Z}^k$ starting at $S_0 = 0$, and consider the number of lattice sites visited in steps $j = m + 1, \dots, n$ as defined by

$$Z_{m,n} = |\{S_j : m \leq j \leq n\}|.$$

Then one has the subadditivity property

$$Z_{0,n} \leq Z_{0,m} + Z_{m,n}.$$

It follows from the subadditive ergodic theorem that $\lim_{n\to\infty} \frac{|\{S_j : 0 < j \leq n\}|}{n}$ exists, where $|\cdot|$ denotes the cardinality of the enclosed set. To identify the limit, observe that S_k contributes a new point to the range if and only if $S_k \notin \{0, S_1, \dots, S_{k-1}\}$. By the time-reversal symmetries of lattice random walk sums, one has

$$\frac{1}{n}\mathbb{E}\sum_{k=1}^{n} \mathbf{1}[S_k \notin \{0, S_1, \dots, S_{k-1}\}] = \frac{1}{n}\sum_{k=1}^{n} P(S_k - S_j \neq 0 \text{ for all } j \leq k-1)$$
$$= \frac{1}{n}\sum_{k=1}^{n} P\left(\sum_{i=j+1}^{k} X_i \neq 0 \text{ for all } j \leq k-1\right)$$

$$= \frac{1}{n}\sum_{k=1}^{n} P\left(\sum_{i=1}^{j} X_i \neq 0 \text{ for all } j \leq k\right)$$
$$\to P(S_j \neq 0 \text{ for all } j).$$

Example 3 *(Branching Random Walk).* A binary branching random walk is a family of random variables X_v indexed by $v \in \cup_{n=0}^{\infty}\{1,2\}^n$, such that for each $v \in \{1,2\}^{\infty}$, $X_\emptyset, X_{v|0}, X_{v|1}, \ldots,$ is a random walk on $\mathbb{R}$ starting at $X_{v|0} = 0$, say, where $v|0 = \emptyset, v|j = (v_0, v_1, \ldots, v_j), j \geq 1$. The distribution of $X_\emptyset = 0, X_{v|0}, X_{v|1}, \ldots,$ does not depend on the path defined by $v \in \{1,2\}^{\infty}$. Assume further that $\psi(t) := \mathbb{E}e^{-tX_{v|1}} < \infty$ for some $t > 0$. Let us observe here that a.s. and L^1-existence of the limit defining the speed is also assured by subadditivity. Let $Z_{0,n} = \inf_{|v|=n} X_v$ denote the left-most position of a walker in the nth generation of a binary branching random walk. Consider the minimum displacement over the first k generations, and start there to compute the minimum displacement for that subtree over the next $n-k$ generations. The minimum over n generations may not involve the minimum over the first k generations. So, although the shortest path in $n+k$ generations need not overlap the shortest paths in the first n and second k generations, one clearly has

$$Z_{0,n+k} \leq Z_{0,n} + Z_{k,n},$$

where $Z_{k,n}$ is independent of $Z_{0,n-k}$ and identically distributed. It follows from the subadditive ergodic theorem[6] that the speed of the left-most particle

$$\lim_{n\to\infty} \frac{Z_{0,n}}{n} \equiv \lim_{n\to\infty} \inf_{|v|=n} \frac{\mathbb{E}X_{v|n}}{n} = \gamma$$

exists a.s. and in L^1.

Remark 5.1. The speed of an extremal particle in a branching random walk is calculated in Bhattacharya and Waymire (2021), Chapter 21, under the assumption that the limit exists. The result obtained is

$$\gamma = -\inf_t \frac{\psi(t)}{t}, \tag{5.8}$$

where $\psi(t) = \ln(2\mathbb{E}e^{-tX_1})$ (assumed to be finite for some $t > 0$). It is interesting to consider this formula in the context of some of the large deviation rates in the i.i.d. case in Chapter 21. In view of the one-sided nature of the deviation rates, for the left-most particle, it is most reasonable to consider the deviation rate of $-X$, where X is a generic displacement random variable. That is, let

[6] From the perspective of subadditivity, this example illustrates the need for the generalization provided by Liggett (1985,b).

$$I(a) = \sup_h \left\{ah - \ln \mathbb{E}e^{-hX}\right\}. \tag{5.9}$$

Then an alternative formula to (5.8) may be expressed in terms of the large deviation rate of $-X$ via (Exercise 13)

$$\gamma = -\inf\{a : I(a) > \ln 2\}. \tag{5.10}$$

Interesting phenomena involving the speed of extremal particles also naturally arise in the case that branching random walks are replaced by branching Markov chains.[7]

Example 4 *(Products of Random Matrices).* The following theorem[8] is an often cited application[9] of subadditive ergodic theory.

Theorem 5.3 *(Furstenberg–Kesten).* Suppose that $A_n = ((A_{ij}^{(n)})), n = 1, 2, \ldots,$ is an i.i.d. sequence of $k \times k$ matrices with positive entries. Assume $\mathbb{E}|\ln A_{ij}^{(n)}| < \infty$ for all i, j. Then $\lim_{n\to\infty} \frac{1}{n} \ln(A^{(1)} \cdots A^{(n)})_{ij}$ exists a.s. and in L^1. Moreover the limit does not depend on i, j.

Proof. Denote the negative logarithm of the element of the first row and column of the matrix product $\prod_{r=m+1}^{n} A^{(r)}$

$$Z_{m,n} = -\ln((A^{(m+1)} \cdots A^{(n)})_{11}), \quad 0 \le m < n.$$

By hypothesis, $\mathbb{E}|Z_{0,1}| < \infty$. Moreover,

$$(A^{(1)} \cdots A^{(n)})_{11} = \sum_{1 \le j_{n-1}, \ldots, j_1 \le k} A_{1j_1}^{(1)} \cdots A_{j_{n-1}1}^{(n)} \le k^{n-1} \prod_{r=1}^{n} \max_{i,j} A_{ij}^{(r)}.$$

Thus,

$$Z_{0,n} - (n-1)\ln k \le \sum_{r=1}^{n} \max_{i,j} \ln A_{ij}^{(r)} \le \sum_{r=1}^{n} \sum_{i,j} |\ln A_{ij}^{(r)}|.$$

In particular,

$$\frac{1}{n}\mathbb{E}Z_{0,n} \le \ln k + \sum_{i,j} \mathbb{E}|\ln A_{ij}^{(1)}| < \infty.$$

[7] See Dascaliuc et al. (2022a) for related calculations.

[8] Furstenberg and Kesten (1960).

[9] This purely mathematical result has important consequences in physics where it is used to quantify important notions of disorder and localization. Comtet et al. (2013) provide a readable review from this perspective.

Since $Z_{m,n}$ is subadditive and satisfies the stationarity requirements for the subadditive ergodic theorem, it follows that there is an invariant random variable Z such that $\frac{1}{n}Z_{0,n} \to Z$ a.s. and in L^1 as $n \to \infty$. In view of the i.i.d. assumptions, $Z = \gamma$ is an almost sure constant. Next consider $\frac{1}{n}\ln((A^{(2)}\cdots A^{(n+1)})_{ij})$ for arbitrary $1 \le i, j \le k$. One has

$$A_{i1}^{(2)}(A^{(3)}\cdots A^{(n)})_{11}A_{1j}^{(n+1)} \le (A^{(2)}\cdots A^{(n+1)})_{ij} \le (A_{1i}^{(1)}A_{j1}^{(n+2)})^{-1}(A^{(1)}\cdots A^{(n+2)})_{11}.$$

By the strong law of large numbers, it follows that $\frac{1}{n}\ln A_{ij}^{(n)} \to 0$ a.s. and in L^1 as $n \to \infty$. Therefore

$$\ln((A^{(3)}\cdots A^{(n)})_{11}) + o(n) \le \ln((A^{(2)}\cdots A^{(n+1)})_{ij}) \le \ln((A^{(1)}\cdots A^{(n+2)})_{11}) + o(n). \tag{5.11}$$

Thus, one obtains a.s. and in L^1 that

$$\lim_{n\to\infty}\frac{1}{n}\ln(A^{(2)}\cdots A^{(n+1)})_{ij} = \gamma. \tag{5.12}$$

■

Additional illustrative examples are given in the exercises.

Exercises

1. Suppose that the sequence $\{a_n : n \ge 1\}$ is superadditive, i.e., $a_{m+n} \ge a_m + a_n$, $m, n = 1, 2, \ldots$. Show that $\lim_{n\to\infty}\frac{a_n}{n} = \sup_{n\ge1}\frac{a_n}{n}$.
2. (a) Show that the array subadditivity property of $\{Z_{m,n}\}$ implies that of $\{Z_{m,n}^+\}$.
   (b) Prove that, under the hypothesis of Theorem 5.2, $\{\mathbb{E}Z_{0,n}^+\}$ and $\{\mathbb{E}Z_{0,n}\}$ are subadditive sequences (allowing the latter to assume the value $-\infty$).
   (c) Assuming γ to be finite, show that $\mathbb{E}Z_{0,n}$ is finite for all $n \ge 1$, that $\mathbb{E}Z_{0,n} \le n\mathbb{E}Z_{0,1}$, and that the same holds for $\{Z_{0,n}^+$.
   (d) Assuming γ finite, prove that $\{Y_{k,n} : n \ge 1\}$ is uniformly integrable, for each k. [*Hint*: $Y_{k,n}$ has the same distribution as $Z_{0,k+1} - Z_{0,k}$, whatever be n.]
3. Prove that, in the case γ is finite, the sequence $\{Z_{mk,n}\}$, with $m = [\frac{k}{n}]$, is uniformly integrable and $\frac{Z_{mk,n}}{n}$ goes to zero a.s. and in L^1 as $n \to \infty$. [*Hint*: $Z_{mk,n} = Z_{mk,mk+r}$ for some $r = 1, \ldots, k-1$, or $Z_{mk,n} = 0$ if $n = mk$, or if $r = 0$, in which case take $Z_{mk,mk} = 0$]. Now $Z_{mk,mk+r}$ has the same distribution as $Z_{0,r}$ (although r may depend on m). Hence $|Z_{mk,n}|$ is stochastically smaller than $\sum_{r=1}^{k-1}|Z_{0,r}|$, proving uniform integrability. Also, for any given $\epsilon > 0$, $\sum_{m\ge1}P(|Z_{mk,n}| > n\epsilon) \le \sum_{m\ge1}P(|Z_{0,j}| > mk\epsilon) \le \sum_{j=1}^{k-1}\mathbb{E}\frac{|Z_{0,j}|}{k\epsilon} < \infty$.

4. Assume the hypothesis of Theorem 5.2, together with finiteness of $\gamma := \lim_{n\to\infty} \mathbb{E}\frac{Z_{0,n}}{n}$.

   (a) Prove that $Z^+_{0,n}/n, n \geq 1$, is uniformly integrable. [*Hint*: Use the analogue of (5.1) for $\{Z^+_{mk,n}/n\}$ and use the analog of Exercise 3 for this sequence.]

   (b) Using (5.6), prove that the sequence $Z^-_{0,n}/n, n \geq 1$, is uniformly integrable and goes to zero, a.s. and in L^1. Use (a) to prove that $\{Z_{0,n}/n\}$ is uniformly integrable and goes to zero, a.s. and in L^1. [*Hint*: $Z^-_{0,n}/n$ is stochastically smaller than the middle term in (5.6), which converges to integrable Y^- a.s. and in L^1.]

5. Prove subadditivity of $\{\mathbb{E}Z_{0,n} : n \geq 1\}$, $\{\mathbb{E}Z^+_{0,n} : n \geq 1\}$ in part (a) of the proof of Theorem 5.2.

6. (*Gelfand Formula*) Suppose that $T : V \to V$ is a nontrivial bounded linear operator on a normed vector space $V, ||\cdot||$. Define $||T||_{op} = \sup_{||x||=1} ||Tx||$. Show that $\rho(T) = \lim_{n\to\infty} ||T^n||^{\frac{1}{n}}_{op}$ exists and is given by $\inf_{n\geq 1} ||T^n||^{\frac{1}{n}}_{op}$. [*Hint*: Noting that if S and T are bounded linear operators, then $||ST||_{op} = \sup_{||x||=1} ||S(\frac{Tx}{||Tx||})|| \cdot ||Tx|| \leq ||S||_{op} \cdot ||T||_{op}$, and check that $n \to \ln ||T^n||_{op}, n \geq 1$, is subadditive.]

7. Show that a.s. convergence continues to hold in the subadditive ergodic theorem in the case that $\gamma = -\infty$ by verifying the following steps. First define truncations for integer $m \in \mathbb{Z}$

$$W^{(m)}_{k,n} = Z_{k,n} \vee m(n-k), k = 0, 1, \dots, n-1$$

   and $\gamma^{(m)} = \inf_n \frac{1}{n}\mathbb{E}(Z_{0,n} \vee mn) \geq m$.

   (a) Verify that $\frac{1}{n}(Z_{0,n} \vee mn) = (\frac{1}{n}Z_{0,n}) \vee m \to W^{(m)}$ a.s. for some random variable $W^{(m)}$. [*Hint*: Check that the conditions for the subadditivity theorem hold with γ replaced by $\gamma^{(m)}$.]

   (b) Show that $\mathbb{E}W^{(m)} = \gamma^{(m)}$.

   (c) Show that $\frac{1}{n}Z_{0,n} \to \inf_m W^{(m)} = W$, say.

   (d) Show that $\mathbb{E}W \leq \inf_m \gamma^{(m)} = \gamma = -\infty$.

8. (*Self-avoiding lattice path counts*) A connected polygonal directed nearest neighbor path γ in the two-dimensional integer lattice is said to be *self-avoiding* if $\gamma(s) \neq \gamma(t)$ for $s \neq t$. Let κ_n denote the number of such paths with n steps and distinct up to translation. The limit $\mu = \lim_{n\to\infty} \kappa_n^{\frac{1}{n}}$ defines the *connectivity constant*.

   (a) Show that μ exists. [*Hint*: Consider subadditivity of $\ln \kappa_n$.

   (b) Show that $2 < \mu < 3$. [*Hint*: Argue that $4(2^n - 1) \leq \kappa_n \leq 4 \cdot 3^{n-1}$.]

9. (*The Furstenberg–Kesten theorem*)

(a) Show that the Furstenberg–Kesten theorem remains valid for any stationary sequence of $k \times k$ random matrices with almost surely positive elements subject to the moment assumptions $\mathbb{E}|\ln A_{ij}^{(n)}| < \infty$ for all i, j, but with a possibly random limit.

(b) Define $Z_{m,n} = \log ||A^{(m)} \cdots A^{(n)}||_{op}$, where $|| \cdot ||_{op}$ is a matrix operator norm, i.e., $||A||_{op} = \sup_{||x||=1} ||Ax||$ for a given norm $|| \cdot ||$ on $\mathbb{R}^k$. Show that the subadditive ergodic theorem applies to $Z_{m,n}$ as well.

10. (*First Passage Percolation*)[10] Consider the k-dimensional ($k \geq 2$) integer lattice $\mathbb{Z}^k$ to be a graph whose vertices are the lattice points and (undirected) edges $e = e(x, y)$ are assigned to adjacent vertices $x, y, |x - y| = 1$. A fluid is injected at a vertex x and requires time T_e to traverse the edge between x and an adjacent vertex y. Assume that the respective times are i.i.d. with finite first moment. Let $\pi(m, n)$ denote a possible nearest neighbor path from vertex me to vertex ne, $e = (1, 0, \ldots, 0)$, and consider the shortest time

$$Z_{m,n} = \min_{\pi(m,n)} \sum_{e \in \pi(m,n)} T_e,$$

for fluid to travel along this path. Show that (a) $\lim_{n\to\infty} \frac{Z_{0,n}}{n}$ exists a.s., and (b) the limit is a.s. constant.

11. The overall rate at which the dynamics separate infinitesimally close initial points is often measured by the maximal Lyapunov exponent[11] of a continuous transformation $T : S \to S$, defined on a metric space (S, ρ), by the expression $\lambda = \lim_{n\to\infty} \frac{1}{n} \limsup_{\rho(x,y)\to 0} \ln \frac{\rho(T^n x, T^n y)}{\rho(x,y)}$. Show that the limit defining λ exists and is given by $\lambda = \inf_n \frac{1}{n} \limsup_{\rho(x,y)\to 0} \frac{\ln \rho(T^n x, T^n y)}{\rho(x,y)}$. [*Hint*: Check the subadditivity of the sequence $a_n = \frac{1}{n} \limsup_{\rho(x,y)\to 0} \ln \frac{\rho(T^n x, T^n y)}{\rho(x,y)}$.]

12. Compute the speed of the left-most particle in the branching random walk with displacements.

(a) X is normal with mean μ and variance σ^2.

(b) $X = 1$ almost surely.

13. Give a proof of the equivalence of (5.8) and (5.10). [*Hint*: Use the respective meanings of the greatest lower bound and least upper bound to show (5.8)≤(5.10) and (5.8)≥(5.10). For the latter inequality, add $\epsilon > 0$ to (5.8).]

14. Compute the maximal Lyapunov exponent for (a) irrational rotations $x \to (x + \alpha) mod(1)$ and (b) the doubling map $x \to 2x(1)$. [*Hint*: $\rho(x, y) = |x - y| \wedge (1 - |x - y|)$.]

[10] Hammersley and Welsh, (1965). Also see Auffinger et al. (2017) .

[11] See Key (1987) and the references therein for examples and illustrative applications of the maximal Lyapunov exponent.

Chapter 6
An Introduction to Dynamical Systems

Ergodic theory was originally developed to study the long time behavior of dynamical systems, especially arising in statistical mechanics. Our aim in this chapter is to analyze some basic features of dynamical systems, such as attractive and repelling periodic orbits, bifurcations, and chaotic phenomena, via some important families of one-dimensional maps. The logistic, or quadratic, family provides an important example.

We first turn briefly to an alternative view of stationary processes as deterministic dynamical systems in a state of equilibrium. Much of the focus of the chapter is, however, on a dynamical system as a law of evolution of a process in time and on the nature of this evolution.

Definition 6.1. A *dynamical system* is a triple $(T, \Omega, \mathcal{F})$ consisting of a measurable map $T : \Omega \to \Omega$ on the measurable state space $(\Omega, \mathcal{F})$. If μ is a probability measure on $(\Omega, \mathcal{F})$ such that $\mu \circ T^{-1}(B) \equiv \mu(T^{-1}(B)) = \mu(B),$ for all $B \in \mathcal{F}$, then T is said to be *measure-preserving*. A set $G \in \mathcal{F}$ is said to be *invariant* if $\mu(G \Delta T^{-1}G) = 0$. The class $\mathcal{I}$ of invariant sets is said to be *trivial* if $\mu(G) \in \{0, 1\}$ for all $G \in \mathcal{I}$, and, in such a case, the map T and measure μ are said to be *ergodic*.

One may note that the collection $\mathcal{I}$ of invariant sets is always a σ-field. In the language of this definition, Birkhoff's ergodic theorem may be recast as follows.

Theorem 6.1 *(Birkhoff's Ergodic Theorem).* Let T be a measure-preserving map on $(\Omega, \mathcal{F}, \mu)$. Then, for every $f \in L^1(\Omega, \mathcal{F}, \mu)$,

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_6

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1} f(T^m\omega) = g(\omega) \quad \mu - a.s., \quad \text{and } L^1, \tag{6.1}$$

where $g : \Omega \to \mathbb{R}$ has the invariance property $g = g \circ T\mu$—a.s. If in addition T is ergodic, then g is the constant $\int_\Omega f d\mu$.

The dynamical systems version and stationary process version are equivalent in the following senses. Assuming the dynamical systems version Theorem 6.1, one may take $(\Omega, \mathcal{F}, \mu)$ to be a canonical model of a stationary process $\{X_n : n \geq 0\}$ with state space $(S, \mathcal{S})$, i.e., $\Omega = S^\infty, \mathcal{F} = \mathcal{S}^{\otimes\infty}$, and μ a probability on $(S^\infty, \mathcal{S}^{\otimes\infty})$ such that the *shift map* $T\omega := (\omega_1, \omega_2, \dots), \omega = (\omega_0, \omega_1, \omega_2, \dots) \in \Omega = S^\infty$ is measure-preserving, and $X_n(\omega) = \omega_n, n \geq 0$. Then the canonical version of Theorem 4.1 becomes a special case of Theorem 6.1, and a non-canonical version follows immediately since its assertion depends only on the distribution of $\mathbf{X}$.

Conversely, given a measure-preserving transformation T on $(\Omega, \mathcal{F}, \mu)$, define the stationary process $X_n(\omega) := T^n\omega (n \geq 1), X_0(\omega) = T^0(\omega) := \omega$, for $\omega \in \Omega$. Let $P = \mu \circ \mathbf{X}^{-1}$ be the probability induced by μ under the map $\mathbf{X} : \Omega \to \Omega^\infty$, where $\mathbf{X}(\omega) := (X_0(\omega), X_1(\omega), \dots), \omega \in \Omega$. Then (i) $\{X_n : n \geq 0\}$ is a stationary sequence, i.e., the shift map $\tilde{T}\omega := (\omega_1, \omega_2, \dots), \omega = (\omega_0, \omega_1, \dots) \in \Omega^\infty$, is measure-preserving for $(\Omega^\infty, \mathcal{F}^{\otimes\infty}, P)$, (ii) $\mathcal{G} \equiv \sigma(\mathbf{X}) = \sigma(X_0) \equiv \mathcal{F}$, since $\omega = X_0(\omega)$ determines $\mathbf{X}(\omega) = (X_0(\omega), X_1(\omega), \dots) = (\omega, T\omega, T^2\omega, \dots)$, and (iii) the invariant σ-field $\mathcal{I}$ of the dynamical system equals the invariant σ-field, say $\tilde{\mathcal{I}}$, of the stationary process $\{X_n : n \geq 0\}$. Hence Theorem 6.1 follows from Theorem 4.1. Thus the two versions of Birkhoff's ergodic theorem are equivalent. In particular, the previous representation theory developed for stationary processes is also available for the analysis of certain classes of dynamical systems.

Example 1 *(Rotation Maps).* Let $\Omega = [0, 1)$ and $\mathcal{F}$ its Borel σ-field. Fix $c \in [0, 1)$ and define $Tx = x + c\ mod 1, x \in \Omega$. Then T preserves the Lebesgue measure μ on Ω. One may also think of Ω as parameterizing the unit circle S^1 in the complex plane by $S^1 := \{e^{2\pi i x} : x \in [0, 1)\}$. This gives T a geometric interpretation of rotation U of the circle in the counterclockwise direction by an angle $2\pi c$, i.e., $U(e^{2\pi i x}) = e^{2\pi i(x+c)} = e^{2\pi i c}e^{2\pi i x}$, and μ induces the normalized arc length measure ν, say, on the circle S^1.
Case i: *(Rational Rotations).* Assume that c is a rational number. First consider $c = 1/q$, where $q \geq 2$ is a positive integer. In this case, T is not ergodic since the set $A_r := \cup_{j=1}^q [\frac{j-1}{q}, \frac{j-1}{q} + \frac{r}{q}), 0 < r < 1$, is invariant, but $\mu(A_r) = r$. Note that on the circle S^1, the set A_r represents the union of q cyclically placed arcs each of length $r/q (< 1/q)$, and the rotation map moves one arc on to the next. Next consider $c = p/q, 1 < p < q$, with p, q relatively prime, positive integers. This time the first interval is moved by T onto the p-th interval, the second onto the $p + 1-$th, ..., the $j-$th interval onto the $p + j$ mod$q$-th interval, $1 \leq j \leq q$. Note that $\{p + jmod(q) : j = 1, 2, \dots q\} = \{1, 2, \dots, q\}$. To see that T is not ergodic

in this case, simply note that every x has a *periodic orbit* $\{x + (j - 1)\frac{p}{q} mod(1) : j = 1, 2, \dots, q\}$ of period q. That is to say, q is the smallest positive integer such that $T^q x = x$, and each x has the periodic orbit $\{x, Tx, T^2 x, \dots, T^{q-1} x\}$. It is now easily checked that for every real-valued function g on Ω,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} g(T^m x) = \frac{1}{q} \sum_{j=0}^{q-1} g\left(x + \frac{jp}{q}\right), \qquad x \in \Omega. \tag{6.2}$$

Case ii: *(Irrational Rotations).* Let $0 < c < 1$ be an irrational number. We will see in this case that T is ergodic. For this, let A be an invariant set. Then, $\mathbf{1}_A(x) = \mathbf{1}_A(x + c)$ $\mu-$a.s. Expand each of these in a Fourier series in $L^2([0, 1), \mu)$; μ being Lebesgue measure.

$$\mathbf{1}_A(x) = \sum_{n\in\mathbb{Z}} a_n e^{2\pi i n x}, \qquad \mathbf{1}_A(x + c) = \sum_{n\in\mathbb{Z}} a_n e^{2\pi i n c} e^{2\pi i n x}, \tag{6.3}$$

where $a_n := \int_{[0,1)} \mathbf{1}_A(x) e^{-2\pi i n x} dx$ ($n \in \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$). In view of the invariance of A, one has $a_n = a_n e^{2\pi i n c}$ for every $n \in \mathbb{Z}$. Since c is irrational, $e^{2\pi i n c} \neq 1$ unless $n = 0$. Thus $a_n = 0$, for all $n \neq 0$, and hence $\mathbf{1}_A(x) = a_0 = \mu(A)$ a.s., i.e., $\mu(A) = 0$ or 1. This proves T is ergodic. By the ergodic theorem, one therefore has for every $f \in L^1([0, 1), \mu)$ and irrational $0 < c < 1$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} g(T^m x) = \lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} g(x + mc) = \int_{[0,1)} g(y) dy \quad \mu-a.s. \tag{6.4}$$

In the following we will often write f instead of T and $f^{(n)}$ instead of T^n, for the nth iterate of f, $f^{(2)} = f \circ f$, $f^{(3)} = f \circ f \circ f, \dots$. This is especially convenient when Ω is an interval or a rectangle S. We first consider the class of contracting dynamical systems. Such systems, even under this apparently stringent hypothesis, are among the most widely applicable.

Theorem 6.2.

(a). *(Contraction Mapping Theorem).* If (S, ρ) is a compact metric space and T is a strict contraction, i.e., $\rho(Tx, Ty) < \rho(x, y), x, y \in S, x \neq y$, then there is a unique fixed point x^* of T.

(b). *(Banach Fixed Point Theorem).* Let (S, ρ) be a complete metric space and T a uniformly strict contraction, i.e., $\rho(Tx, Ty) < c\rho(x, y)x, y \in S, x \neq y$, where $c \in [0, 1)$. Then T has a unique fixed point x^*, say, and $T^n x \to x^*$ as $n \to \infty$ for every $x \in S$.

Proof. (a) Let x^* be a minimizer of the function $x \to \rho(x, Tx)$. Note that T is continuous, and a real-valued continuous function on a compact metric space attains its infimum. Then $Tx^* = x^*$; otherwise, $\rho(Tx^*, T2x^*) < \rho(x^*, Tx^*)$, which contradicts the fact that x^* is a minimizer of $\rho(x, Tx)$. (b) Choose $x \in S$ arbitrarily.

Then $\rho(T^n x, T^{n+1}x) < c\rho(T^{n-1}x, T^n x) < \cdots < c^n \rho(x, Tx) \to 0$ as $n \to \infty$. Thus $\{T^n x\}_{n\geq 0}$ is a Cauchy sequence (Exercise) and, by the completeness of S, the sequence converges to a limit x^*, say. Clearly, $Tx^* = x^*$; for $T^{n+1}x = TT^n x \to Tx^*$, but $T^{n+1}x \to x^*$ as well. Next, let $y \neq x$. Then $y^* \equiv \lim_{n\to\infty} T^n y$ is a fixed point of T. However, $\rho(y*, x^*) = \rho(Ty^*, Tx^*) < c\rho(y^*, x^*)$, which is impossible if $y^* \neq x^*$. ■

The following result is simply a rewriting of Theorem 6.2(b), $T = f$, $T^n = f^{(n)}$, the n-th iterate of f.

Corollary 6.1. Let (S, ρ) be a complete metric space and f a continuous function on S such that $T^k = f^{(k)}$ is a uniformly strict contraction for some $k \geq 1$. Then the conclusion of Theorem 6.2(b) holds.

Example 2 *(Contracting Quadratic Map).* Let $S = [0, 1]$, $f(x : \theta) = \theta x(1 - x)$, $0 \leq \theta < 1$. Then f is a uniformly strict contraction, since $|f'(x)| = \theta|1 - 2x| < 1$ on $[0, 1]$ and Theorem 6.2(b) applies: $f^{(n)}(x)$ converges to the unique fixed point $x^* = 0$ as $n \to \infty$, uniformly on $[0, 1]$.

Remark 6.1. Although the hypotheses of Theorem 6.2 and Corollary 6.1 appear rather strong, they have many important applications. We will later consider applications of the corollary to monotone Markov processes $\{X_n : n = 0, 1, \ldots\}$ on a closed subset C of R^d, with S as the space of probabilities on C given an appropriate metric, and f defined on S by $f(\mu)(B) = \int p(z, B)\mu(dz)$, where $p(z, B)$ is the transition probability of the process landing in the set B in one step if it starts in state z:

$$p(z, B) = P(X_1 \in B | X_0 = z); \tag{6.5}$$

see Theorem 19.1 for a detailed treatment.

Remark 6.2. There are also important applications of fixed point theorems to nonlinear partial differential equations, including the celebrated problem of the existence and uniqueness of global solutions to the incompressible Navier–Stokes equations in 3d for *small* initial data.[1]

Moving away from contracting maps, there are interesting examples of dynamical systems (f, S), which have multiple fixed points as well as periodic points.

Definition 6.2. A point x^* is said to be periodic with period k, if there are k distinct points $x_1 = x^*, x_2 = f(x_1), x_j = f(x_{j-1}), j = 2, \ldots, k$, such that $f(x_k) = x^*$. In this case $(x_1, \ldots, x_k)$ is said to be a period k orbit. A fixed point is a periodic point with period $k = 1$. A fixed point x^* of f on a metric space (S, ρ) is said to be attracting (or locally stable) if there exists an open set U containing x^* such that $f^{(n)}(x) \to x^*$ as $n \to \infty$, for every $x \in U$. More generally, a periodic point

[1] See, for example, Bhattacharya and Waymire (2021) and the references therein.

x^* of f, or its orbit, is said to be attracting if there is an open set U containing the orbit of x^* such that, as $n \to \infty$, $f^{(n)}(x)$ converges to the orbit of x^*, that is, $\rho(f^{(n)}(x), \{\text{orbit of } x^*\}) \to 0$, for all $x \in U$. A periodic point x^* (including a fixed point), or its orbit, is said to be repelling, if there exists an open set U containing the orbit of x^* such that if $x \in U \backslash \{\text{orbit of } x^*\}$, then $f^{(k)}(x)$ does not belong to U for some $k = k(x) \geq 1$.

Proposition 6.4. Suppose S is an interval of the real line, and x^* is a fixed point of a continuously differentiable function f on S into S. If $|f'(x^*)| < 1$, then x^* is an attractive fixed point, and if $|f'(x^*)| > 1$, then x^* is repelling. More generally, a periodic point x^* of period k, or its orbit, is attracting if $|f^{(k)}(x^*))'| < 1$ and repelling if $|(f^{(k)}(x^*))'| > 1$.

Proof. First consider the case of a fixed point x^* such that $|f'(x^*)| < c < 1$. Let $\delta > 0$ be such that $|f'(x)| < c$ on $[x^* - \delta, x^* + \delta] \cap S = \mathbf{I}$, say. Then for all $x \in \mathbf{I}$, $|f(x) - x^*| = |f(x) - f(x^*)| < c|x - x^*| < \delta$, so that $f(x) \in \mathbf{I}$, implying by iteration that $f^{(n)}(x) \in \mathbf{I}$ for all n. Also, $|f^{(n)}(x) - x^*| = |f^{(n)}(x) - f^{(n)}(x^*)| = |f \circ f^{(n-1)}(x) - f \circ f^{(n-1)}(x^*)| < c|f^{(n-1)}(x) - f^{(n-1)}(x^*)| < \cdots < c^n|x - x^*| \to 0$ as $n \to \infty$, uniformly on $\mathbf{I}$. If, on the other hand, x^* is a fixed point such that $|f'(x^*)| > 1$, and c is such that $|f'(x^*)| > c > 1$, then there exists $\delta > 0$ such that $|f'(x)| > c > 1$ for all $x \in [x^* - \delta, x^* + \delta] \cap S = \mathbf{I}$, and $S \backslash \mathbf{I}$ has a nonempty interior. Then the inequalities for $|f^{(n)}(x) - x^*| = |f^{(n)}(x) - f^{(n)}(x^*)|$ above get reversed, and for some $n = n(x)$, $f^{(n(x))}(x)$ lies outside $\mathbf{I}$. It may be noted that if for some $n > n(x)$, $y = f^{(n)}(x) \in \mathbf{I}$ again, then there is an integer $n(y)$ such that $f^{(n(y))}(y)$ lies outside $\mathbf{I}$, so that $f^{(n(x)+n(y))}(x)$ lies outside $\mathbf{I}$. It follows that $f^{(n)}(x)$ lies outside $\mathbf{I}$ for infinitely many n. Next suppose x^* is a periodic point with period $k > 1$. Let $\{x^* = x_1, x_2 = f(x_1), \dots, x_k = f(x_{k-1})\}$ be the periodic orbit of x^*. Note that $f^{(k)}(x^*) = f(x_k)$ and, more generally, each point in the orbit is a fixed point of $f^{(k)}$. By the rule for differentiation of composite functions, one has, with $x_1 = x^*$,

$$\begin{aligned}(f^{(k)})'(x_1) &= f'(f^{(k-1)}(x_1)) \cdot f'(f^{(k-2)}(x1)) \cdot f'(f^{(k-3)}(x_1)) \cdots f'(x_1) \\ &= f'(x_k)f'(x_{k-1}) \cdots f'(x_2)f\prime(x_1). \end{aligned} \tag{6.6}$$

It follows from this that $(f^{(k)'})(x_j)$ is the same for all $j = 1, \dots, k$. Assume now that $|(f^{(k)}(x^*))'| < 1$. Then $|(f^{(k)}(x_j))'| < 1$ for all $j = 1, \dots, k$. Since $x_1, \dots, x_k$ are then attractive fixed points of $f^{(k)}$, there exists an open interval V_j around x_j such that $f^{(n_k)}(x) \to x_j$ as $n \to \infty$, for every $x \in V_j (j = 1, \dots, k)$. Let $V = \cup_{j=1}^k V_j$. Let $[n/k]$ denote the integer part of n/k. If $x \in V$, say $x \in V_j$, then the distance between $f^{(n)}(x) = f^{(n-[n/k]k)} \circ f^{([n/k]k)}(x)$ and the orbit of x^* goes to zero as $n \to \infty$. For $f^{([n/k]k)}(x) \to x_j$ as $n \to \infty$, and $f^{(n-[n/k]k)} \in \{f^{(0)}, f, f^{(2)}, \dots, f^{(k-1)}\}$, where $f^{(0)}$ is the identity map. The proof that x^* is repelling if $|(f^{(k)}(x^*))'| > 1$ may be similarly fashioned and is left as Exercise 10. ■

For an important example, we consider again the quadratic family of maps, but over a wider range of the parameter θ.

Remark 6.3. Below we often make use of the fact that if a subinterval $\mathbf{I}$ of $S = [0,1]$ is left invariant by f, i.e., $f(\mathbf{I}) \subset \mathbf{I}$, and f is monotone increasing on $\mathbf{I}$, then $f^{(n)}$ is increasing on $\mathbf{I}$ for all n. If, in addition, $f(x) \leq x$ for all $x \in \mathbf{I}$, then $f^{(n+1)}(x) \leq f^{(n)}(x)$ for all $n = 1, 2, \ldots$, i.e., $f^{(n)}(x)$ is a decreasing sequence for $x \in \mathbf{I}$. If, instead, $x \leq f(x)$ on $\mathbf{I}$, then $f^{(n-1)}(x) \leq f^{(n)}(x)$ for all $n = 1, 2, \ldots$, i.e., $f^{(n)}(x)$ is an increasing sequence for $x \in \mathbf{I}$.

Example 3 *(Logistic Map and The Quadratic Family).* In studying the mechanism of population growth from one generation to the next, in units of the so-called *carrying capacity* of the environment, biologists have considered the model $f(x) \equiv f(x,\theta) = \theta x(1-x)$ for $x \in [0,1] = S$. For $0 \leq \theta \leq 4$, the map defines a dynamical system. In Chapter 18 on discrete time Markov processes, we will consider random perturbations[2] of this model to take into account the effect of external factors influencing the population size or concentration at time $n+1$ given that at time n.

Proposition 6.5. (a) Let $0 \leq \theta \leq 1$. Then $f^{(n)}(x,\theta) \to 0$, as $n \to \infty$, whatever be $x \in [0,1]$. Thus 0 is an attracting fixed point, and it is the only fixed point of $f(x,\theta)$. (b) Let $1 < \theta \leq 3$. Then $f^{(n)}(x,\theta) \to p_\theta \equiv 1 - 1/\theta$ for all $x \in (0,1)$. Hence there are two fixed points $0, p_\theta$, of which 0 is repelling and p_θ is attracting. (c) For $3 < \theta \leq 1 + \sqrt{5}$, $f(\cdot,\theta)$ has an attracting period-two orbit and repelling fixed points $0, p_\theta = 1 - 1/\theta$.

Proof. First note that a fixed point of f is a solution of the quadratic equation $\theta x(1-x) = x$, provided this solution lies in $[0,1]$. The two solutions are $x = 0$ and $x = p_\theta = 1 - 1/\theta$. The second solution is in $[0,1]$ only if $\theta \geq 1$, and for $\theta = 1$ it is 0. Thus the quadratic family $f(x) = f(x,\theta)$ has two fixed points $0, p_\theta = 1 - 1/\theta$ for all $\theta \in (1,4]$, and only one fixed point for $\theta \in [0,1]$.

(a) For $\theta = 0$, $f(x,\theta) = 0$ for all x, so that 0 is an attractive fixed point. Let $0 < \theta \leq 1$. Then $0 < f(x,\theta) < x$ for all $x \in (0,1]$, and $x \to f(x,\theta)$ is increasing on $[0,1/2]$ and decreasing on $[1/2,1]$ ($f(\cdot,\theta)$ is symmetric about $1/2$). Also, $[0,1/2]$ is an invariant interval: $f([0,1/2],\theta) \subset [0,1/2]$. Hence $0 < f^{(n+1)}(x,\theta) < f^{(n)}(x,\theta)$ for all $x \in (0,1/2]$. Thus $f^{(n)}(x,\theta) \downarrow$ as $n \uparrow$, and $x^* = \lim_n f^{(n)}(x,\theta)$ is a fixed point of $f(\cdot,\theta)$. But the only fixed point of $f(x,\theta)$ is 0. Hence $x^* = 0$. If $x \in [1/2,1]$, then $f(x,\theta) \in [0,1/2]$. Hence, by the preceding argument, $f^{(n)}(x,\theta) \to 0$ for all $x \in [0,1]$ and 0 is attracting.

(b) First let $1 < \theta \leq 2$. Then $f(x,\theta) < p_\theta$ for all $x \in (0,p_\theta)$, and $x \to f(x,\theta)$ is increasing on $(0,p_\theta)$, since $p_\theta \leq 1/2$. Hence $f(x,\theta) < f^{(2)}(x,\theta) < p_\theta$ for all $x \in (0,p_\theta)$, and iterating, one obtains $f^{(n)}(x,\theta) < f^{(n+1)}(x,\theta) <$

[2] Also see Peckham et al. (2018) and the references therein for related applications to sustainability of a biological population subject to random disturbances.

p_θ for all $x \in (0, p_\theta)$. Let x^* be the limit of the increasing sequence $f^{(n)}(x, \theta)$. But the only positive fixed point of $f(x, \theta)$ is p_θ, so that $x^* = p_\theta$. Now let $x \in [p_\theta, 1/2]$. Then $f(x, \theta)$ is increasing and $p_\theta \le f(x, \theta) \le x \le 1/2$. Iterating, one gets $p_\theta \le f^{(n+1)}(x, \theta) \le f^{(n)}(x, \theta) \le 1/2$ for all n. Hence the sequence $f^{(n)}(x, \theta)$ decreases to a limit y^*, $p_\theta \le y* \le 1/2$. Since y^* is a fixed point, it must equal p_θ. For $x \in (1/2, 1)$, $f(x, \theta) \in (0, 1/2)$. Hence, by the convergence of $f^{(n)}(x, \theta)$ to p_θ on $(0, 1/2]$, it follows that $f^{(n)}(x, \theta) \to p_\theta$ for all $x \in (0, 1)$. This shows that p_θ is an attractive fixed point and 0 is a repelling fixed point.

Next, for $2 < \theta \le 1 + \sqrt{5}$, $[1/2, \theta/4]$ is an invariant interval for $f(\cdot, \theta)$. To see this, note that $f(x, \theta) \le \theta/4$ for all x. Hence it is enough to show that $f(x, \theta) \ge 1/2$ on $[1/2, \theta/4]$. Since f is decreasing on $[1/2, \theta/4]$, one needs to show that $f(\theta/4, \theta) \ge 1/2$. To establish this, use the fact that the function $\theta \to g(\theta) = f(\theta/4, \theta) = \theta(\theta/4)(1 - \theta/4)$ has the derivative $\frac{\theta}{2}(1 - \frac{3\theta}{8})$, which is positive on $[2, 8/3)$. Therefore, $g(\theta)$ increases on $[2, 8/3)$ with $g(2) = 1/2$. Also, $g(\theta)$ decreases for $\theta > 8/3$, and it is simple to check that it attains the value $1/2$ at $\theta = 1 + \sqrt{5}$. This establishes the invariance of $[1/2, \theta/4]$. [Note: This is the role of $1 + \sqrt{5}$ in part (c)]. Since $x \to f(x, \theta)$ is decreasing on $[1/2, \theta/4]$, $x \to f^{(2)}(x, \theta)$ is increasing on $[1/2, \theta/4]$, which is invariant under $f(\cdot, \theta)$ (and, therefore, under $f^{(2)}$). It follows that $f^{(2n+2)}(\theta/4, \theta) = f^{(2n)}(f^{(2)}(\theta/4, \theta), \theta) \le f^{(2n)}(\theta/4, \theta)$ (since $f^{(2)}(\theta/4, \theta) \le \theta/4$). That is, $f^{(2n)}(\theta/4, \theta)$ is a decreasing sequence. Let $f^{(2n)}(\theta/4, \theta) \downarrow q_1$. Similarly, $f^{(2n)}(1/2, \theta)$ is an increasing sequence and has a limit q_0. Clearly, q_0 and q_1 are fixed points of $f^{(2)}(x, \theta)$. Because $1/2 \le p_\theta \le \theta/4$, $f^{(2n)}(1/2, \theta) \le f^{(2n)}(p_\theta, \theta)(= p_\theta) \le f^{(2n)}(\theta/4, \theta)$, one gets $1/2 \le q_0 \le p_\theta \le q_1 \le \theta/4$.

Now let $2 < \theta \le 3$. It may be shown that the fourth degree polynomial $f^{(2)}(x, \theta)$ has no fixed points other than 0, p_θ (Exercise 13). Hence $q_0 = q_1 = p_\theta$ if $2 \le \theta \le 3$. If $1/2 \le x \le \theta/4$, then $f^{(2n)}(1/2, \theta) \le f^{(2n)}(x, \theta) \le f^{(2n)}(\theta/4, \theta)$. Hence $f^{(2n)}(x, \theta) \to p_\theta$ as $n \to \infty$ for every $x \in [1/2, \theta/4]$. Next note that $f^{(2n+1)}(x, \theta) = f(f^{(2n)}(x, \theta)) \to f(p_\theta, \theta) = p_\theta$ as $n \to \infty$. We have proved that $f^{(n)}(x, \theta) \to p_\theta$ as $n \to \infty$, for every $x \in [1/2, \theta/4]$. One can easily check that $1/2 < p_\theta < \theta/4$ for $2 < \theta \le 3$. Hence p_θ is an attracting fixed point if $2 < \theta \le 3$, since $(1/2, \theta/4)$ is an open neighborhood of p_θ. To prove that $f^{(n)}(x, \theta) \to p_\theta$ as $n \to \infty$ for every $x \in (0, 1)$, let $x \in (0, 1/2]$. Since $f(\cdot, \theta)$ is increasing and $f(x, \theta) > x$ on $(0, 1/2]$, $f^{(n)}(x, \theta)$ strictly increases with n as long as the sequence remains in $(0, 1/2]$. This is a finite sequence; otherwise, the limit would be a fixed point in $(0, 1/2]$. Hence there exists $n = n(x)$ such that $1/2 < f^{(n)}(x, \theta) \le \theta/4$. It follows from above that $f^{(n)}(x, \theta) \to p_\theta$ *as* $n \to \infty$ *for all* $x \in (0, \theta/4]$. Since $f(x, \theta) \in (0, \theta/4]$ for all $x \in (0, 1)$, the last assertion in italics is proved.

(c) Finally, let $3 < \theta \le 1 + \sqrt{5}$. Since $f'(x, \theta) = \theta - 2x\theta$, $f'(0, \theta) = \theta > 1$, and $f'(1 - 1/\theta, \theta) = 2 - \theta < -1$, both fixed points are repelling. The argument in the second paragraph of the proof of (b) shows that $f^{(2)}(x, \theta)$ has attractive fixed points q_0 and q_1 satisfying $1/2 \leq q_0 \leq p_\theta \leq q_1 \leq \theta/4$. Since p_θ is repelling, one must have $1/2 \leq q_0 < p_\theta < q_1 \leq \theta/4$. ■

Remark 6.4. One says that a bifurcation occurs at $\theta = 1$; for a new fixed point p_θ appears when θ crosses 1 ($\theta > 1$) and, further, this fixed point is attractive and the previous fixed point 0 becomes repelling. It follows from Proposition 6.5 that the threshold $\theta = 3$ is also a *bifurcation point* of the quadratic family, since an attractive period 2 orbit appears for the first time when θ crosses the value 3. It is known that the next bifurcation occurs at $\theta = 1 + \sqrt{6}$, with a period 4 attracting orbit occurring under each θ in $(1 + \sqrt{6}, 1 + \sqrt{6} + \epsilon]$ for some $\epsilon > 0$, all other periodic points, including fixed points, being repelling. This period doubling cascade continues, and after its accumulation point, cascades of intervals of θ appear giving rise to new attractive periodic orbits whose periods are odd multiples of a power of 2, in decreasing order of powers. The odd multiples (excluding 1) of the power of 2 in such a cascade also appear in a decreasing order and end with an interval of θ under which the quadratic map has an attractive orbit of period 3 times the power of 2 (following one with orbits of period 5 times the power of 2). The last such cascade of intervals of θ, corresponding to the smallest power of 2, namely 2, ends with one having attractive orbits of period 3×2. Finally, a cascade of intervals with attractive orbits of odd periods, other than 1, appears, again in decreasing order, ending with one of period 3. After the appearance of this attractive period 3 orbit, the maps with higher values of θ cannot give rise to orbits with new periods since all periods have already appeared in this scheme known as Sarkovskii's theorem.[3] Here is Sarkovskii's scheme for the appearance of new periods, in reverse order:

$$3 \rhd 5 \rhd 7 \rhd \cdots \rhd 2\cdot3 \rhd 2\cdot5 \rhd 2\cdot7 \rhd \cdots \rhd 2^2\cdot3 \rhd 2^2\cdot5 \rhd 2^2\cdot7 \rhd \cdots \rhd 2^3\cdot3 \rhd 2^3\cdot5 \rhd 2^3\cdot7 \rhd \cdots \rhd 2^3 \rhd 2^2 \rhd 1.$$

Each new period appears with an attractive periodic orbit which remains attractive for a range of θ, while all previous periodic orbits become repelling. Sarkovskii's scheme is actually *universal*. That is, if a continuous map f has a periodic orbit of period k, and if $k \rhd m$, then f must have an orbit of period m. Later in this chapter, we will consider the case of $\theta = 4$, where a new phenomenon appears. This new phenomenon known as *chaos* is our next topic for discussion. It is a remarkable fact due to Graczyk and Swiatek (1997) that the set of all θ with attractive periodic orbits is dense.

The notion that deterministic dynamical systems may "appear" to be random is further quantified by the following notion without explicit reference to a probability space.

Definition 6.3. Suppose that Ω is a metric space with metric ρ and Borel σ-field $\mathcal{F}$. A dynamical system $(T, \Omega, \mathcal{F})$ with the following properties is said to be *chaotic:*

[3] For a simple proof of the theorem, we refer to Devaney (1989). A comprehensive treatment of this complex and rich phenomenon is given in Collet and Eckmann (1980).

1. *Sensitive Dependence on Initial Conditions.* There is a $\delta > 0$ such that for any given $\omega \in \Omega$ and neighborhood V of ω, there is an $\omega' \in V$ and $n \geq 0$ for which $\rho(T^n\omega, T^n\omega') > \delta$.
2. *Periodic Points Are Dense in Ω.* The subset of points defined by $\mathrm{Per}(T) := \{\omega \in \Omega : T^n\omega = \omega$ for some $n \geq 1\}$ is dense in Ω.
3. *Topological Transitivity.* For each pair of nonempty open sets V_1 and V_2, there is an $n > 0$ such that $(T^n V_1) \cap V_2 \neq \emptyset$.

Intuitively, sensitivity to initial conditions limits predictability of the evolution in the sense that an arbitrarily small perturbation ω' from ω will still lead to a separation of states $T^n\omega$ and $T^n\omega'$ in time n by a fixed positive amount δ. Topological transitivity is a type of irreducibility of the evolution in the sense that the evolution cannot settle into disjoint invariant subregions. The density of periodic points provides some degree of repetitiveness in the evolution. Variations on the precise definition of chaotic dynamics can be found throughout the mathematics and physics literature.

Example 4 *(Rotation Maps (Continued)).* Let $S = [0, 1]$ and fix $f(x) = x + c$ for a constant $c \in [0, 1)$. This dynamical system is not chaotic, whether c is rational or not, since every rotation leaves the distance (arc length) between points invariant.

Example 5 *(Infinite Convolution with Shift as a Chaotic Dynamical System).* Let $S = \{0, 1\}$ be a two-point set and consider the space $\Omega := S^\infty$ of binary sequences. Then Ω is a compact metric space for the metric defined by $\rho(\omega, \omega') := \sum_{n=1}^{\infty} 2^{-n}|\omega_n - \omega'_n|$, $\omega = (\omega_1, \omega_2, \dots)$, $\omega' = (\omega'_1, \omega'_2, \dots)$. Let $\mathcal{F}$ be the Borel σ-field; equivalently $\mathcal{F}$ is the σ-field generated by finite dimensional sets of the form $A = \{(\omega_1, \dots, \omega_n)\} \times S^\infty, \omega_j \in S, 1 \leq j \leq n$. Let T be the shift transformation $T\omega = (\omega_2, \omega_3, \dots), \omega = (\omega_1, \omega_2, \dots) \in \Omega = S^\infty$. Let us see that this dynamical system $(T, \Omega, \mathcal{F})$ is chaotic. To check sensitive dependence on initial conditions, simply take $\delta = 1/4$. Let $\omega = (\omega_1, \omega_2, \dots) \in \Omega$, and let $B_m = \{\omega' \in \Omega : \omega'_j = \omega_j$ for all $j \leq m\}$ denote the open (and closed) ball of radius 2^{-m} centered at ω. Any open set containing ω must contain B_m for suitably large m. Then choose $\omega' \in B_m$ with $\omega'_{m+1} = 1 - \omega_{m+1}$ and $n = m$, so $T^n\omega = (\omega_{m+1}, \omega_{m+2}, \dots)$. Then $\rho(T^n\omega, T^n\omega') \geq 1/2 > 1/4$. To prove density of periodic points, let $\omega \in \Omega$ and $\epsilon > 0$. Choose n such that $\sum_{j=n+1}^{\infty} 2^{-j} < \epsilon$ and define ω' to be the infinite periodic extension of $(\omega_1, \omega_2, \dots, \omega_n)$. Then $\rho(\omega, \omega') < \epsilon$. Finally, to prove topological transitivity, we will show a stronger property known as *topological mixing*. Namely, let us observe that there is an $\omega^* \in \Omega$ whose orbit $\{\omega^*, T\omega^*, T^2\omega^*, \dots\}$ is dense in Ω. To construct ω^* proceed as follows: the first two bits of ω^* are 0 and 1, followed by all possible two bit blocks 0, 0, 0, 1, 1, 0, 1, 1 in some order, and then all three bit blocks in some order, and so on. Then given any $\omega \in \Omega$, the first m digits will eventually appear as a block of ω^*, say, starting from n. Thus, the first m digits of $T^n\omega^*$ will match $(\omega_1, \dots, \omega_m)$, and hence $\rho(T^n\omega^*, \omega) < 2^{-m}$. Topological transitivity follows from this (Exercise 7). Finally let us note that there are uncountably many mutually singular ergodic invariant probabilities (having infinite support) for T since, for each $p \in [0, 1]$, the infinite Bernoulli product measure μ_p is an invariant probability; i.e., for finite dimensional cylinder sets $\mu_p(\{(\omega_1, \dots, \omega_n)\} \times S^\infty) = p^{\sum_{j=1}^n \omega_j}(1-p)^{\sum_{j=1}^n (1-\omega_j)}$, $\omega_j \in \{0, 1\}$. The distribution of an i.i.d. coin tossing sequence (infinite product measure) is invariant under a shift in the sequence. That the support of each μ_p is infinite for $0 < p < 1$ and that the measures are mutually singular can be seen from the strong law of large numbers, i.e., the support of μ_p is contained in $\{\omega \in \Omega : \lim_{n\to\infty} \frac{\sum_{j=1}^n \omega_j}{n} = p\}$, which is a closed set.

Remark 6.5. It is known that in the quadratic family $f(x; \theta) = \theta x(1-x)$, $0 \le x \le 1$, chaos first occurs at the end of the period doubling cascade at $\theta = 3.56995\dots$ and last on the Sarkovskii order. This marks the first appearance of a period 3 orbit and the onset of chaos.[4] From here on until $\theta = 3.82843$, known as the Pomeau–Manneville region, the dynamics are marked by stable periodic phases interrupted by bursts of chaos, which have applications to semiconductor devices.[5]

The following definition is very useful for showing that two dynamical systems are equivalent, one being the relabeling of the other.

Definition 6.4. Let S and $\tilde{S}$ be two metric spaces and f and g two maps on them, respectively. The dynamical systems (f, S) and $(g, \tilde{S})$ are *topologically conjugate* if there is a homeomorphism $h : S \to \tilde{S}$ such that $h \circ f = g \circ h$ or, equivalently, $g = h \circ f \circ h^{-1}$.

We now provide examples of a number of important one-dimensional chaotic systems, especially the quadratic system (Example 7) with $\theta = 4$, known as the *Ulam–von Neumann map*.

Example 6. Consider the map $g : S^1 \to S^1$ given by $g(\theta) = 2\theta(mod 2\pi)$. We will show that this dynamical system is chaotic. For this, first note that $0 = 2\pi(mod 2\pi)$ is a fixed point, and a period n $(n \ge 1)$ orbit (including orbits of periods that are factors of n) is given by the solutions of $g^{(n)}(\theta) \equiv 2n\theta(mod 2\pi) = \theta(mod 2\pi)$, i.e., $2n\theta = \theta + 2k\pi$ for some nonnegative integer k, so that $\theta = 2k\pi/(2n-1)(k = 0, 1, 2, \dots, 2n-2)$. Note that this is an equidistant set of $2n-1$ points in $[0, 2\pi)$. This being true for all $n \ge 1$, it follows that the set of all periodic points is dense in S^1. To establish topological transitivity, note that if θ belongs to an open arc C in S^1, and θ' $(\neq \theta)$ lies in an open arc C', there exists $n \ge 1$ such that $g^{(n)}(C') = S^1$, since after each iteration the length of C' doubles (or equals S^1). In particular, $g^{(n)}(C')$ intersects C. Finally, in order to prove sensitive dependence on initial conditions, let θ belong to an open arc C in S^1, and let n be such that $g^{(n)}(C) = S^1$, which implies that there exists a point $\theta' \neq \theta$ in C such that the distance between $g^{(n)}(\theta)$ and $g^{(n)}(\theta')$ will be as large as half the arc length of S^1. Hence the dynamical system is chaotic.

[4] See May (1976).

[5] See Jeffries and Perez (1982).

Remark 6.6. As the above arguments show, if for every nonempty open set U, there exists a positive integer n such that $f^{(n)}(U) = S$, and then the dynamical system (f, S) is topologically transitive and has sensitive dependence on initial conditions.

Example 7 *(Ulam–von Neumann map).* Let $g : S^1 \to S^1$ by $g(\theta) = 2\theta(mod 2n)$. Using Example 6, we will show that the Ulam–von Neumann map $f(x) \equiv f(x, 4) = 4x(1 - x)$ on $[0, 1]$ is chaotic. For this, first note the map $g_1(x) = 2x^2 - 1$ on $[-1, 1]$ is topologically conjugate to $f(x, 4)$ via the homeomorphism $h : [-1, 1] \to [0, 1]$ given by $h(x) = (1/2)(1 - x)$, since $f \circ h(x) = 2(1 - x)(1 - \frac{1}{2}(1 - x)) = 1 - x^2 = h \circ g_1(x)$. Next consider the map $\varphi : S^1 \to [-1, 1]$ given by $\varphi(\theta) = \cos\theta$, which is the projection of the circle $S^1 = \{(x, y) : x^2 + y^2 = 1\}(x = \cos\theta, y = \sin\theta)$ onto the x-axis. Let $\psi = h \circ \varphi$. This map is not a homeomorphism because it is two-to-one (except at $\theta = 0$): $\cos\theta = \cos(2\pi - \theta)$. But one still has the relation $\varphi \circ g(\theta) = \cos 2\theta = 2\cos^2\theta - 1 = g_1 \circ \varphi(\theta)$. (Such systems ψ are sometimes called *semiconjugate*). In the following discussion, we make use of the relation $f \circ h \circ \varphi(\theta)(= h \circ g_1 \circ \varphi(\theta)) = h \circ \varphi \circ g(\theta)$. Hence $f \circ \psi = f \circ h \circ \varphi = h \circ g_1 \circ \varphi = h \circ \varphi \circ g = \psi \circ g$. Iterating, one obtains $f^{(2)} \circ \psi = f \circ f \circ h \circ \varphi = f \circ \psi \circ g = \psi \circ g \circ g = \psi \circ g^{(2)}, \ldots f^{(n)} \circ \psi = \psi \circ g^{(n)}, n = 1, 2, \ldots$. To show that the Ulam–von Neumann system $(f, [0, 1])$ is topologically transitive, let U, V be nonempty open intervals in $[0, 1]$. Then $\tilde{U} = h^{-1}(U)$ and $\tilde{V} = h^{-1}(V)$ are open intervals in $[-1, 1]$. There exist open arcs $\hat{U}$ and $\hat{V}$ in the upper half of the circle S^1 such that $\varphi : \hat{U} \to \tilde{U}$ is one-to-one and onto $\varphi(\hat{U}) = \tilde{U}$, and the same is true of $\varphi : \hat{V} \to \tilde{V}$, $\varphi(\hat{V}) = \tilde{V}$. Then $h \circ \varphi(\hat{U}) = U$ and $h \circ \varphi(\hat{V}) = V$. Using $f^{(n)} \circ (h \circ \varphi) = (h \circ \varphi) \circ g^{(n)}$, it follows that $f^{(n)}(U) = h \circ \varphi \circ g^{(n)}(\hat{U})$, $f^{(n)}(V) = h \circ \varphi \circ g^{(n)}(\hat{V})$. Choose $n \geq 1$ such that $g^{(n)}(\hat{V}) = S^1$ (Example 6). Then $f^{(n)}(V) = [0, 1]$ because $f^{(n)}(V) = h \circ \varphi(S^1) = [0, 1]$.

From the above argument and Remark 6.6, it follows that $(f(x, 4), [0, 1])$ is topologically transitive and has sensitive dependence on initial conditions, with δ any positive number smaller than $1/2$. Finally, let V be a nonempty open interval in $[0, 1]$ and $\theta_0 \in \hat{V}$ a periodic point of some period n for (g, S^1), where $\hat{V}$ is an open arc in the upper half of S^1 as described above. Then $g^{(n)}(\theta_0) = \theta_0$. Therefore, the projection $h \circ \varphi(\theta_0) = x_0$, say, of θ_0 on $[0, 1]$ satisfies $f^{(n)}(x_0) = f^{(n)} \circ h \circ \varphi(\theta_0) = h \circ \varphi \circ g^{(n)}(\theta_0) = h \circ \varphi(\theta_0) = x_0$. That is, x_0 is a periodic point of f of period n. Since the set of periodic points of g in S^1 is dense in S^1 , so is the set of periodic points of g in the upper half of S^1. Hence the set of periodic points of $f = f(\cdot, 4)$ is dense in $[0, 1]$.

Example 8 *(The Tent Map).* The tent map $(t, [0, 1])$ is defined by

$$t(u) = \begin{cases} 2u & \text{if } 0 \leq u < 1/2 \\ 2 - 2u & \text{if } 1/2 \leq u \leq 1. \end{cases} \tag{6.7}$$

We will see that the tent map, and the Ulam–von Neumann map $f(x) = f(x, 4) = 4x(1 - x), x \in [0, 1]$, are topologically conjugate. Consider the change of variable

$u \to x$, given by $x = \sin^2 \frac{\pi u}{2} = h(u)$, say, of [0, 1] onto [0, 1]. Note that $h'(u) > 0$ on (0, 1), and $h'(u) = 0$ for $u = 0, 1$. Hence h is strictly increasing (and continuous) on [0, 1]; also $h^{-1}(x) = \frac{2}{\pi} \arcsin \sqrt{x}$. One has $f \circ h(u) = 4 \sin^2 \frac{\pi u}{2} \cos^2(\frac{\pi u}{2}) = \sin^2(\pi u)$ and $h^{-1} \circ f \circ h(u) = \frac{2}{\pi} \arcsin(\sin \pi u)$. For $1/2 < u \leq 1$, formally this has two values $2u$ and $2 - 2u$ (since $\sin \pi u = \sin(\pi - \pi u)$), of which $2u$ falls outside [0, 1]. Hence

$$h^{-1} \circ f \circ h(u) = \begin{cases} \frac{2}{\pi} \arcsin(\sin \pi u) = 2u & \text{if } 0 \leq u < 1/2 \\ \frac{2}{\pi}(\pi - \pi u) = 2 - 2u & \text{if } 1/2 \leq u \leq 1. \end{cases} \tag{6.8}$$

That is, $t = h^{-1} \circ f \circ h$, so that f and t are topologically conjugate. By Example 7, it follows that the tent map is chaotic as well.

Finally, it is simple to check that the uniform distribution on [0, 1] is invariant under the tent map t (Exercise 11). Hence the density of the corresponding invariant distribution of the Ulam–von Neumann map is given by $m(x) = du/dx = dh^{-1}(x)/dx = d\frac{2}{\pi} \arcsin(\sqrt{x})/dx$. Thus,

$$m(x) = \frac{1}{\pi \sqrt{x(1-x)}}, \quad 0 < x < 1. \tag{6.9}$$

Proposition 6.6. Let t be the tent map and let X_0 have the uniform distribution λ on [0, 1]. The stationary process $X_n = t^n X_0 (n = 0, 1, \dots)$ is ergodic with respect to the shift transformation $TX = (X_1, X_n, \dots)$, $X = (X_0, X_1, \dots)$.

Proof. First observe that $X = (X_0, X_1, \dots)$ is a Markov process with transition probability $p(x, dy) = \delta_{t(x)}(dy)$. Hence for every shift-invariant event $F = [X \in C]$, C a Borel subset of $[0, 1]^\infty$, there exists a t-invariant Borel subset B of [0, 1] (i.e., almost surely, $\mathbf{1}_B(X_0) = \mathbf{1}_B(X_1) = \mathbf{1}_B(t(X_0)))$ such that $\mathbf{1}_B(X_0) = \mathbf{1}_C(X)$, almost surely (Exercise 12). Therefore, to prove ergodicity, it is enough to show that $\lambda(B) \in \{0, 1\}$. Suppose $\lambda(B) > 0$, and consider the Fourier expansion

$$\mathbf{1}_B(x) = \sum_{k=-\infty}^{\infty} c_k \exp\{2\pi i k x\}, \left(c_k = \int_{[0,1]} \mathbf{1}_B(x) \exp\{-2\pi i k x\} dx, \quad \text{for all} \quad k \right), \tag{6.10}$$

the convergence of the series being in $L^2([0, 1], \lambda)$. For $0 < x < 1/2$, $\mathbf{1}_B(x) = \mathbf{1}_B(2x)$, λ-almost surely. Hence, λ-almost surely

$$\sum_{k=-\infty}^{\infty} c_k \exp\{2\pi i k x\} = \sum_{k=-\infty}^{\infty} c_k \exp\{4\pi i k x\}, \quad (0 < x < 1/2), \tag{6.11}$$

This implies $c_k = c_{2k}$ for all k. By iteration, one obtains $c_k = c_{2mk}$ for every positive integer m. But for $k \neq 0$, $c_{2mk} \to 0$ as $m \to \infty$. Thus,

$$c_k = 0 \ \text{ for all } \ k \neq 0. \tag{6.12}$$

Since the Fourier series for $\mathbf{1}_B$ is thus identified, namely, $\mathbf{1}_B = c_0$, and the proof of the proposition is complete, one could also use a similar argument for $x > 1/2$, $t(x) = 2(1 - x)$, to obtain

$$\sum_{k=-\infty}^{\infty} c_k \exp\{2\pi ikx\} = \sum_{k=-\infty}^{\infty} c_k \exp\{4\pi ik - 4\pi ikx\} = \sum_{k=-\infty}^{\infty} c_k \exp\{-4\pi ikx\}, \tag{6.13}$$

This implies $c_k = c_{-2k} = \cdots = c_{-2mk}$ for all integers $m > 0$, and $c_{2mk} \to 0$ as $m \to \infty$, for $k \neq 0$. Hence again $\mathbf{1}_B(x) = c_0$ for $1/2 < x < 1$ (λ-a.s.). Therefore, $\mathbf{1}_B(x) = \lambda(B)$ for all x (λ-a.s.); that is, $\lambda(B) = 1$. ■

In view of the conjugacy between the tent map and the Ulam–von Neumann map, one obtains the following result.

Corollary 6.7. Let f be the Ulam–von Neumann map $f(x) = 4x(1 - x)$ on [0, 1], and let X_0 have the distribution with density (6.9). Then the stationary process $X_n = f^{(n)}X_0(n = 0, 1, \ldots)$ is ergodic; that is, the shift-invariant sigma-field is trivial.

Remark 6.7. The Ulam–von Neumann map or some modifications that are still chaotic have many practical uses[6,7] in science and engineering, including random number generation, cryptography, economics, population biology, social networks, etc. Dynamical systems in continuous time are called *flows* which are governed by smooth autonomous ordinary differential equations or systems of them in the case of multidimension. In view of the Poincare´-Bendixson theorem,[8] chaotic phenomena cannot arise in two dimensions (or less). Roughly, this theorem says that, apart from fixed points which may be attractive, repelling, or saddle points, a flow $(x(t), y(t))$ that remains in a bounded domain for all times $t \geq 0$ is either cyclic, i.e., lies on a closed curve and moves periodically on it, or approaches such a cycle as $t \to \infty$.[9]

Exercises

1. Let T be a measure-preserving transformation on the probability space $(\Omega, \mathcal{F}, \mu)$. Define $X_0(\omega) = \omega$, $X_n(\omega) = T^n\omega$, $\omega \in \Omega$, $n \geq 1$.

[6] Among many publications on the subject, we refer to the articles by Ulam and von Neumann (1947), Derrida and Flyvbjerg (1987), and Yu et al. (1990).

[7] Many applications of dynamical systems to economic theory may be found in Bhattacharya and Majumdar (2007), Chapter 1, which also contains an expository account of the elements of chaos theory in discrete time. The classic work of Samuelson (1947) (enlarged edition published in 1983), based on his 1941 Harvard thesis, is a pioneering study of optimization of economic phenomena governed by systems of differential equations, their equilibrium, and stability, as well as what would now be called bifurcations.

[8] See, e.g., Hurewicz (1958).

[9] The first example of a chaotic flow in dimension three is due to Lorenz (1963).

(a) Show that $\{X_n : n \geq 0\}$ is a stationary process on $(\Omega, \mathcal{F}, \mu)$.
(b) Show that $\sigma(X_0, X_1, \dots) = \mathcal{F}$ and that the invariant σ-fields of T and of $\{X_n : n \geq 0\}$ coincide.

2. Suppose T is an ergodic transformation on a probability space $(\Omega, \mathcal{F}, \mu)$. For $B \in \mathcal{F}$, show $\lim_{n\to\infty} \frac{|\{m \leq n : T^m \omega \in B\}|}{n} = \mu(B)$, where $|\cdot|$ denotes cardinality.
3. (*Normal numbers*)

   (a) Show that any irrational real number $x \in [0, 1]$ has a unique expansion, $x = \sum_{j=1}^{\infty} x_j b^{-j}$, to a base $b \in \mathbb{N}$.
   (b) Show that for any $k \in \{0, 1, \dots, b-1\}$, $\lim_{n\to\infty} \frac{|\{j \leq n : x_j = k\}|}{n} = \frac{1}{b}$, for Lebesgue a.e. $x \in [0, 1]$, i.e., each digit in the base b expansion occurs with equal frequency. Such numbers are said to be *normal.*[10]

4. Let $\Omega = [0, 1)$ with Borel σ-field $\mathcal{F}$ and Lebesgue measure μ. Consider the map $Tx = x + c\text{mod}(1)$, with $c = p/q$ where p and q are relatively prime positive integers. Show that (a) each $x \in [0, 1)$ has a periodic orbit $O_x = \{x, Tx, \dots, T^{q-1}x\}$ of period q, and (b) T is measure-preserving and ergodic with respect to the uniform distribution on O_x (which assigns zero probability on $\Omega \backslash O_x$).
5. If T is an ergodic measure-preserving transformation on $(\Omega, \mathcal{F}, \mu_i), i = 1, 2$, then either $\mu_1 = \mu_2$ or μ_1 and μ_2 are mutually singular. [*Hint*: Let A be such that $\mu_1(A) \neq \mu_2(A)$. Use the ergodic theorem for $\mathbf{1}_A$.]
6. Let $\Omega = \{x_1, x_2, \dots, x_m\}$ be a finite set and $\mathcal{F}$ the power set of Ω. Let T be a permutation, i.e., a one-to-one map of Ω onto itself. Let μ be the uniform distribution on Ω, i.e., $\mu(\{x_i\}) = 1/m, i = 1, 2, \dots, m$.

   (a) Show that T is measure-preserving on $(\Omega, \mathcal{F}, \mu)$.
   (b) Show that T is ergodic if and only if it has no periodic point of period less than m; a point x is a periodic point of period q if q is the smallest integer such that $T^q x = x$, and $\{x, Tx, \dots, T^{q-1}x\}$ is a periodic orbit.
   (c) Suppose that T has two cycles, i.e., there exists $x \neq y$ and $q < m$ such that $T^q x = x, T^{m-q} y = y$. Let μ_1 be the uniform distribution on $\Omega_1 = \{x, Tx, \dots, T^{q-1}x\}$ and μ_2 that on $\Omega_2 = \{y, Ty, \dots, T^{m-q-1}y\}$.

      (i) Show that T is ergodic on Ω_1 and on Ω_2.
      (ii) If $\overline{\mu}_i$ denotes the extension of μ_i to Ω defined by $\overline{\mu}_i = \mu_i$ on Ω_i and $\overline{\mu}_i(\Omega \backslash \Omega_i) = 0$, then show T is measure-preserving and ergodic on $(\Omega, \mathcal{F}, \overline{\mu}_i)$.
      (iii) Let $\nu = \alpha\overline{\mu}_1 + (1-\alpha)\overline{\mu}_2$ for some $0 < \alpha < 1$. Show that T is measure-preserving on $(\Omega, \mathcal{F}, \nu)$ but not ergodic.

[10] It is conjectured that $\sqrt{2}, e, \pi, \ln 2$ are normal numbers, but not proven. An example of a normal number was computed by Sierpinski (1917). Also see Becher and Figueira (2002).

7. Show that topological mixing implies its topological transitivity for a dynamical system (T, Ω) on a metric space Ω, provided that (i) Ω is connected or, more generally, (ii) every nonempty open subset of Ω has infinitely many points.
8. Show that irrational rotations of the circle are topologically transitive but do not have sensitive dependence on initial conditions.
9. Show that the chaotic shift map T on $\Omega = \{0, 1\}^\infty$ is *expansive* in the sense that there is a $\delta > 0$ such that for any two distinct points $\omega, \omega' \in \Omega$ there is an n such that $|T^n\omega - T^n\omega'| > \delta$. Show that, in general, an expansive dynamical system has the property of sensitive dependence on initial conditions.
10. Show that x^* in Proposition 6.4 is repelling if $|f^{(k)}(x^*))'| > 1$.
11. Verify that the uniform distribution on [0, 1] is an invariant measure for the tent map.
12. In reference to the proof of Proposition 6.6, show that there is a t-invariant Borel set $B \subset [0, 1]$ such that $\mathbf{1}_C(X) = \mathbf{1}_B(X_0)$ a.s. as asserted. [*Hint*: $X = (X_0, X_1, \dots) = (X_0, tX_0, t^2X_0, \dots)$ depends only on X_0.]
13. Consider the quadratic map $f(x; \theta) = \theta x(1 - x), x \in [0, 1]$ $(0 < \theta \leq 4)$. Show that $f^{(2)}(x; \theta)$ has only two fixed points, 0 and $p_\theta = 1 - \frac{1}{\theta}$.

Chapter 7
Markov Chains

This chapter focusses on a construction and parameterization of discretely indexed Markov chains on a finite or countably infinite state space.

This chapter begins the theory of discrete parameter/time Markov processes which is the subject matter of the rest of the book. In general, the (one-step) transition probability $p(x, dy)$, x belonging to the state space S, contains all the information about the process. Each transition probability corresponds to a unique Markov process. The asymptotic behavior of a wide variety of these Markov processes is explored. The discrete parameter Markov processes on finite or countable state spaces, called Markov chains, provide a fairly complete asymptotic theory to be considered in this and a few later chapters.

Suppose that $\{X_n : n = 0, 1, 2, \dots\}$ is a discrete parameter stochastic process defined on a probability space $(\Omega, \mathcal{F}, P)$ and having a countable, i.e., finite or countably infinite state space S. Throughout this chapter the σ-field of measurable events $\mathcal{S}$ is naturally the *power set* 2^S consisting of all subsets of S. Think of $X_0, X_1, \dots, X_{n-1}$ as "the past," X_n as "the present," and $X_{n+1}, X_{n+2}, \dots$ as "the future" of the process relative to time n. The law of evolution of a stochastic process is often viewed in terms of the conditional distribution of the future given the present and past states of the process. In the case of a sequence of independent random variables or of a random walk, for example, this conditional distribution does not depend on the past. Various illustrative models with this structure are described in the Exercises 13–20.

To simplify notation explicit reference to the underlying probability space $(\Omega, \mathcal{F}, P)$ will often be suppressed.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_7

Definition 7.1. A stochastic process $\{X_0, X_1, \dots, X_n, \dots\}$ with countable state space S has the *Markov property* if, for each n and m, the conditional distribution of $X_{n+1}, \dots, X_{n+m}$ given $X_0, X_1, \dots, X_n$ is the same as its conditional distribution given X_n alone. Such a process having the Markov property is called a *discrete parameter Markov chain.*

From the definition of conditional probability it is enough to check that for each $B \in \mathcal{S}^{\otimes m}$, the conditional probability $P((X_{n+1}, \dots, X_{n+m}) \in B|\sigma(X_0, \dots, X_n))$ is a function of X_n alone; i.e., if so, the function must be $P((X_{n+1}, \dots, X_{n+m}) \in B|\sigma(X_n))$, Exercise 4. An equivalent formulation of the Markov property may be cast as the conditional independence of the future $\{X_{n+1}, \dots, X_{n+m}\}$ and the past $\{X_0, \dots, X_{n-1}\}$ given the present X_n, for each $n, m = 1, 2, \dots$, see Exercise 4. The denumerability of S makes $[(X_{n+1}, \dots, X_{n+m}) \in B] = \cup_{(j_1,\dots,j_m)\in B}[X_{n+1} = j_1, \dots, X_{n+m} = j_m]$ a countable disjoint union for each $B \subset S^m$, i.e., $B \in \mathcal{S}^{\otimes m}$. Thus the Markov property may be expressed: For any $j_1, \dots, j_m \in S, m \ge 1$,

$$\begin{aligned} &P(X_{n+1} = j_1, \dots, X_{n+m} = j_m|\sigma(X_0, \dots, X_n)) \\ &\quad = P(X_{n+1} = j_1, \dots, X_{n+m} = j_m|\sigma(X_n)) = f_n(X_n). \end{aligned} \tag{7.1}$$

Here the dependence of the function f_n on the *fixed* quantities $j_1, \dots, j_m$ and m is suppressed.

The property (7.1) may be stated in elementary terms as

$$\begin{aligned} &P(X_{n+1} = j_1, \dots, X_{n+m} = j_m|X_0 = i_0, \dots, X_n = i_n) \\ &\quad = P(X_{n+1} = j_1, \dots, X_{n+m} = j_m|X_n = i_n), j_1, \dots, j_m \in S. \end{aligned} \tag{7.2}$$

In view of the next proposition, it is actually enough to take $m = 1$ in any of these formulations.

Proposition 7.1. A stochastic process $X_0, X_1, X_2, \dots$ with countable state space S has the Markov property if and only if for each n the conditional distribution of X_{n+1} given $X_0, X_1, \dots, X_n$ is a function only of X_n.

Proof. The necessity of the condition is obvious. For sufficiency we use induction on m. Observe that if $P(X_0 = i_0, \dots, X_n = i_n) > 0$, then

$$\begin{aligned} &P(X_{n+1} = j_1, \dots, X_{n+m} = j_m \mid X_0 = i_0, \dots, X_n = i_n) \\ &\quad = P(X_{n+1} = j_1 \mid X_0 = i_0, \dots, X_n = i_n) \\ &\qquad \times P(X_{n+2} = j_2 \mid X_0 = i_0, \dots, X_n = i_n, X_{n+1} = j_1) \\ &\qquad \times P(X_{n+3} = j_3 \mid X_0 = i_0, \dots, X_n = i_n, X_{n+1} = j_1, X_{n+2} = j_2) \\ &\qquad \times \cdots \times P(X_{n+m} = j_m \mid X_0 = i_0, \dots, X_n = i_n, X_{n+1} = j_1, \dots, X_{n+m-1} = j_{m-1}) \\ &\quad = P(X_{n+1} = j_1 \mid X_n = i_n)P(X_{n+2} = j_2 \mid X_{n+1} = j_1) \\ &\qquad \times P(X_{n+3} = j_3 \mid X_{n+2} = j_2) \times \cdots \times P(X_{n+m} = j_m \mid X_{n+m-1} = j_{m-1}). \end{aligned}$$

The first equality follows from general rules of probability, while the second (last) uses the Markov property with only one immediate future state in place of an arbitrary number of future states (i.e., the case $m = 1$ in Definition 7.1). Notice that the last expression does not depend on the past states, but only on the "present" state, namely that of X_n, proving the general Markov property given in Definition 7.1. ■

Yet another equivalent formulation of the Markov property will be given in the next chapter (cf. Proposition 8.3).

A Markov chain $\{X_0, X_1, \ldots\}$ is said to have a *homogeneous* or *stationary transition law* if the conditional distribution of $X_{n+1}, \ldots, X_{n+m}$ given X_n depends on the *state* at time n, namely X_n, but not on the *time* n. Otherwise, the transition law is called *nonhomogeneous*. An *i.i.d. sequence* $\{X_n\}_{n\geq 1}$ on the integers $\mathbf{Z}$ or k-dimensional integer lattice $\mathbf{Z}^k$, and its associated *(unrestricted) general random walk* on $\mathbf{Z}$ or $\mathbf{Z}^k$, $\{S_n := S_0 + \sum_{j=1}^{n} X_j\}_{n\geq 0}$, possess time-homogeneous transition laws, while an independent nonidentically distributed sequence $\{X_n\}_{n\geq 1}$ and its associated random walk have nonhomogeneous transitions. Unless otherwise specified, by a Markov chain we shall mean a Markov chain with a *homogeneous* transition law from here on.

It is convenient to introduce a *matrix* $\mathbf{p}$ to describe the probabilities of transition between successive states in the evolution of the process. As a notational convention, when the meaning is unambiguous, p_{ij} is often used in place of $p_{i,j}$ to denote the ith row and jth column matrix entry.

Definition 7.2. A *transition probability matrix* or a *stochastic matrix* is a square matrix $\mathbf{p} = ((p_{ij}))$, where i and j vary over a countable set S, satisfying

1. $p_{ij} \geq 0$ for all i and j,
2. $\sum_{j\in S} p_{ij} = 1$ for all i.

The set S is called the *state space* and its elements are *states*.

Informally, think of a particle that moves from point to point in the state space according to the following scheme. At time $n = 0$ the particle is set in motion either by starting it at a fixed state i_0, called the *initial state*, or by randomly locating it in the state space according to a probability distribution μ on S, called the *initial distribution*. In the former case, μ is the distribution concentrated at the state i_0, i.e., $\mu_j = 1$ if $j = i_0$, $\mu_j = 0$ if $j \neq i_0$. In the latter case, the probability is μ_i that at time zero the particle will be found in state i, where $0 \leq \mu_i \leq 1$ and $\sum_i \mu_i = 1$. Given that the particle is in state i_0 at time $n = 0$, a random trial is performed, with probabilities $p_{i_0 j'}$ of the respective states $j' \in S$. If the outcome of the trial is the state i_1, then the particle moves to state i_1 at time $n = 1$. A second trial is performed with probabilities $p_{i_1 j'}$ of states $j' \in S$. If the outcome of the second trial is i_2, then the particle moves to state i_2 at time $n = 2$, and so on.

A typical sample point of this experiment is a sequence of states, say $(i_0, i_1, i_2, \ldots, i_n, \ldots)$, representing a *sample path*. The set of all such sample paths may be taken as the sequence space *sample space* $\Omega = S^\infty$. The position X_n at time n is a random variable whose value is given by $X_n(\omega) = i_n$ if the sample path is $\omega = (i_0, i_1, \ldots, i_n, \ldots) \in \Omega$. The precise specification of the probability P_μ on Ω for the above experiment is given by

$$\begin{aligned} P_\mu(\{i_0\} \times \cdots \times \{i_n\} \times S^\infty) &\equiv P_\mu(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \\ &:= \mu_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}. \end{aligned} \tag{7.3}$$

More generally, for finite-dimensional events of the form

$$\begin{aligned} A &= [(X_0, X_1, \ldots, X_n) \in B] \\ &\equiv B \times S^\infty \subset \Omega = S^\infty, \end{aligned} \tag{7.4}$$

where B is an arbitrary subset of $(n+1)$-tuples of elements of S, the probability of A is specified by

$$P_\mu(A) \equiv P((X_0, \ldots, X_n) \in B) := \sum_{(i_0, i_1, \ldots, i_n) \in B} \mu_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}. \tag{7.5}$$

By Kolmogorov's existence theorem, P_μ extends uniquely as a probability measure on the smallest σ-field $\mathcal{F} = \mathcal{S}^{\otimes\infty}$ containing the class of all events of the form (7.4).

Definition 7.3. The probability space $(\Omega, \mathcal{F}, P_\mu)$ with $X_n(\omega) = \omega_n, \omega \in \Omega$, is referred to as the *canonical model* for the *Markov chain with (homogeneous) transition probabilities* $((p_{ij}))$ *and initial distribution* μ. A stochastic process $\{X_n : n = 0, 1, \ldots\}$ defined on a general probability space is said to be *Markov with transition probability* $p(x, dy)$ and *initial distribution* $\mu(dx)$ if its distribution is P_μ as defined by the canonical model.

The Markov property will be established below at (7.9)–(7.12). In the case of a Markov chain starting in state i, that is, $\mu_i = 1$, we write P_i in place of P_μ in the canonical model.

To specify various joint distributions and conditional distributions associated with this Markov chain, it is convenient to use the notation of matrix multiplication, with $\mathbf{p} = ((p_{ij}))$. By definition the (i, j) element of the matrix $\mathbf{p}^2$ is given by

$$p_{ij}^{(2)} = \sum_{k \in S} p_{ik} p_{kj}. \tag{7.6}$$

The elements of the matrix $\mathbf{p}^n$ are defined recursively by $\mathbf{p}^n = \mathbf{p}^{n-1}\mathbf{p}$ so that the (i, j) element of $\mathbf{p}^n$ is given by

$$p_{ij}^{(n)} = \sum_{k \in S} p_{ik}^{(n-1)} p_{kj} = \sum_{k \in S} p_{ik} p_{kj}^{(n-1)}, \qquad n = 2, 3, \ldots. \tag{7.7}$$

It is easily checked by induction on n that the expression for $p_{ij}^{(n)}$is given directly in terms of the elements of $\mathbf{p}$ according to

$$p_{ij}^{(n)} = \sum_{i_1,\ldots,i_{n-1}\in S} p_{ii_1} p_{i_1i_2} \cdots p_{i_{n-2}i_{n-1}} p_{i_{n-1}j}. \tag{7.8}$$

Now let us check the Markov property of this probability model. Using (7.3) and summing over unrestricted coordinates, the joint distribution of $X_0, X_{n_1}, X_{n_2}, \ldots, X_{n_k}$, with $0 = n_0 < n_1 < n_2 < \cdots < n_k$, is given by

$$\begin{aligned} &P(X_0 = i, X_{n_1} = j_1, X_{n_2} = j_2, \ldots, X_{n_k} = j_k) \\ &\quad = \sum_1 \sum_2 \cdots \sum_k (\mu_i p_{ii_1} p_{i_1i_2} \cdots p_{i_{n_1-1}j_1})(p_{j_1 i_{n_1+1}} p_{i_{n_1+1}i_{n_1+2}} \cdots p_{i_{n_2-1}j_2}) \\ &\quad \times \cdots \times (p_{j_{k-1}i_{n_{k-1}+1}} p_{i_{n_{k-1}+1}i_{n_{k-1}+2}} p_{i_{n_k-1}j_k}), \end{aligned} \tag{7.9}$$

where $\sum_r$ is the sum over the rth block of indices $i_{n_{r-1}+1}, \ldots, i_{n_r}$ $(r = 1, 2, \ldots, k)$. The sum $\sum_k$, keeping indices in all other blocks fixed, yields the factor $p_{j_{k-1}j_k}^{(n_k-n_{k-1})}$ using (7.8) for the last group of terms. Next sum successively over the $(k-1)$st, …, second, and first blocks of factors to get

$$P(X_0 = i, X_{n_1} = j_1, X_{n_2} = j_2, \ldots, X_{n_k} = j_k) = \mu_i p_{ij_1}^{(n_1)} p_{j_1j_2}^{(n_2-n_1)} \cdots p_{j_{k-1}j_k}^{(n_k-n_{k-1})}. \tag{7.10}$$

Now sum over $i \in S$ to get

$$P(X_{n_1} = j_1, X_{n_2} = j_2, \ldots, X_{n_k} = j_k) = \left(\sum_{i\in S} \mu_i p_{ij_1}^{(n_1)}\right) p_{j_1j_2}^{(n_2-n_1)} \cdots p_{j_{k-1}j_k}^{(n_k-n_{k-1})}. \tag{7.11}$$

Using (7.10) and the elementary definition of conditional probabilities the following proposition is obtained.

Proposition 7.2. Let $\{X_n : n = 0, 1 \ldots\}$ be a Markov chain with arbitrary initial distribution μ and transition probability matrix $\mathbf{p}$. The conditional distribution of X_{n+m} given $X_0, X_1, \ldots, X_n$ is given by

$$\begin{aligned} &P(X_{n+m} = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i) \\ &\quad = p_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i) = P(X_m = j \mid X_0 = i), \qquad m \geq 1, \quad j \in S. \end{aligned}$$

Although by Proposition 7.1 the case $m = 1$ would have been sufficient to prove the Markov property, Proposition 7.2 justifies the terminology that $\mathbf{p}^m := ((p_{ij}^{(m)}))$ is the *m-step transition probability matrix*. Note that $\mathbf{p}^m$ is a stochastic matrix for all $m \geq 1$ (in the sense of Definition 7.2).

The calculation of the distribution of X_m follows from (7.10), (7.11). We have

$$\begin{aligned} P_\mu(X_m = j) = \sum_i P_\mu(X_m = j, X_0 = i) &= \sum_i P_\mu(X_0 = i) P_\mu(X_m = j \mid X_0 = i) \\ &= \sum_i \mu_i p_{ij}^{(m)} = (\mathbf{p}^{(m)\prime}\mu)_j, \end{aligned} \tag{7.12}$$

where $\mathbf{p}^{(m)'}$ denotes the transpose matrix, and $(\mathbf{p}^{(m)'}\mu)_j$ is the jth element of this column vector.

Example 1. $S = \{0, 1\}$. $\mathbf{p} = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix}$, $0 \le a, b \le 1$. Excluding the case $a + b = 2$ in which $\mathbf{p} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, with, say, $a < 1$, one may check that $\mathbf{p}$ has two distinct eigenvalues $\lambda_1 = 1, \lambda_2 = a + b - 1 < 1$ with corresponding eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ \frac{b-1}{1-a} \end{bmatrix}$; notice that a stochastic matrix will always have $\lambda = 1$ as an eigenvalue, and, since $\mathbf{px}$ averages $\mathbf{x}$, the magnitude of the others cannot exceed 1, (Exercise 7). Now, writing A for the matrix whose columns are these eigenvectors, one has

$$A = \begin{bmatrix} 1 & 1 \\ 1 & \frac{b-1}{1-a} \end{bmatrix}, \qquad A^{-1} = \frac{1}{2-a-b} \begin{bmatrix} 1-b & 1-a \\ 1-a & a-1 \end{bmatrix},$$

$$\mathbf{p}A = A \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \qquad \mathbf{p} = A \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} A^{-1}, \qquad \mathbf{p}^n = A \begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix} A^{-1}.$$

Thus

$$\mathbf{p}^{(n)} = (2-a-b)^{-1} \begin{bmatrix} 1-b+(1-a)(a+b-1)^n & 1-a-(1-a)(a+b-1)^n \\ 1-b-(1-b)(a+b-1)^n & 1-a+(1-b)(a+b-1)^n \end{bmatrix}.$$

It is interesting to notice the behavior of $\mathbf{p}^{(n)}$, as a function of n, in each of the distinct cases $a + b = 1, 1 < a + b \le 2$, and $0 \le a + b < 1$.

Exercises

1. Consider the simple symmetric random walk on $\{0, 1, 2\}$ with reflecting boundaries at 0 and 2.
   (a) Show $\frac{1}{n}\sum_{r=1}^{n} \mathbf{p}^{(r)}$ converges to the matrix whose rows are identically $(1/4, 1/2, 1/4)$.
   (b) Show that $\pi_0 = \pi_2 = 1/4, \pi_2 = 1/2$ is the unique invariant probability.
   (c) Start $\{X_n : n \ge 0\}$ with the invariant initial distribution. Show that $\{X_n : n \ge 0\}$ is ergodic, but $\{X_{2n} : n \ge 0\}$ is not ergodic.
2. Consider the Markov chain $\{X_n : n \ge 0\}$ on $S = \{1, 2, 3\}$ with transition probabilities $p_{11} = 1, p_{23} = p_{32} = 1$, and $p_{ij} = 0$ otherwise. Determine

the extremal ergodic invariant probabilities and the collection of all invariant probabilities. Calculate $\lim_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1}X_m$ for the initial distribution $\pi = (1/4, 3/8, 3/8)$.

3. (Random Walk on Integer Lattice) Show that any discrete parameter stochastic process $\{X_n : n = 0, 1, \ldots\}$, having independent increments on the integer lattice state space $\mathbb{Z}^k$ is a Markov chain.
4. (a) Show that if $P(A|\sigma(X_0, \ldots, X_n)) = f_n(X_n)$, then $f_n(X_n) = P(A|\sigma(X_n))$.
   (b) Let A, B, C be events with $C, B \cap C$ having positive probabilities. Verify that the following are equivalent versions of the conditional independence of A and B given C : $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$ if and only if $P(A \mid B \cap C) = P(A \mid C)$.
   (c) Prove that the Markov property (in Definition 7.1) is equivalent to the property: Conditionally given X_n, the past $\{X_0, \ldots, X_{n-1}\}$ and the future $\{X_{n+1}, X_{n+2}, \ldots\}$ are independent.
5. (a) Let $\{X_n : n = 0, 1, 2, \ldots\}$ be a sequence of random variables with countable state space S. Call $\{X_n : n \geq 0\}$ rth *order Markov-dependent* if for $i_0, \ldots, i_n, j \in S, n \geq r$

$$P(X_{n+1} = j \mid X_0 = i_0, \ldots, X_n = i_n)$$
$$= P(X_{n+1} = j \mid X_{n-r+1} = i_{n-r+1}, \ldots, X_n = i_n).$$

   Show that $Y_n = (X_n, X_{n+1}, \ldots, X_{n+r-1}), n = 0, 1, 2, \ldots$ is a (first-order) Markov chain under these circumstances.
   (b) Take $S = \mathbb{Z}^k$, and let $V_n = X_{n+1} - X_n, n = 0, 1, 2, \ldots$. Show that if the *position process* $\{X_n : n \geq 0\}$ is a Markov chain, then so is the *position-velocity process* $\{(X_n, V_n) : n \geq 0\}$. [*Hint*: Consider first $\{(X_n, X_{n+1})\}$ and then apply a one-to-one transformation.]
6. Let $\{Y_n : n \geq 0\}$ be an i.i.d. sequence of ± 1-valued Bernoulli random variables with parameter $0 < p < 1$. Define a new stochastic process by $X_n = (Y_n + Y_{n-1})/2$, for $n = 1, 2, \ldots$. Show that $\{X_n : n \geq 1\}$ does *not* have the Markov property.
7. Suppose that S is a finite state space and $\mathbf{p}$ a stochastic matrix indexed by $S \times S$. Show that $|\lambda| \leq 1$ for all eigenvalues of $\mathbf{p}$, and that $\lambda = 1$ must be an eigenvalue. [*Hint*: View p as a matrix operator on a real vector space $V := \{x = (x_j)_{j \in S} : x_j \in \mathbb{R}\}$ with norm defined by $\|x\| = \max\{|x_j| : j \in S\}$.]
8. Calculate $p_{ij}^{(n)}$ for the unrestricted simple random walk on $\mathbb{Z}$ (see Exercise 8). [*Hint*: Add up probabilities of sample paths by counting.]
9. Let $\mathbf{p} = ((p_{ij}))$ denote the transition matrix for the unrestricted general random walk on $\mathbb{Z}$.

   (a) Calculate $p_{ij}^{(2)}$ in terms of the increment distribution Q.
   (b) Show that $p_{ij}^{(n)} = Q^{*n}(j - i)$, where the *n-fold convolution* is defined recursively by

$$Q^{*n}(j) = \sum_k Q^{*(n-1)}(k)Q(j-k), \qquad Q^{*(1)} = Q.$$

10. Let $\{Z_n : n = 0, 1, 2, \ldots\}$ be i.i.d. ±1-valued with $P(Z_n = 1) = p \in (0, 1)$. Define $X_n = Z_n Z_{n+1}$, $n = 0, 1, 2, \ldots$.

    (a) Show that for $k \le n - 1$, $P(X_{n+1} = j \mid X_k = i) = P(X_{n+1} = j)$, i.e., X_{n+1} and X_k are independent for each $k = 0, \ldots, n-1$, $n \ge 1$. Note that $k \le n - 1$.

    (b) Show that $\{X_n : n \ge 0\}$ is a Markov chain if and only if $p = 1/2$.

11. (a) Show that the transition matrix for a sequence of independent and identically distributed (*i.i.d.*) integer-valued random variables is characterized by the property that its rows are identical; i.e., $p_{ij} = p_j$ for all $i, j \in S$.

    (b) Under what further condition is the Markov chain an i.i.d. sequence? [*Hint*: Consider the initial distribution.]

12. (a) Let $\{Y_n : n \ge 0\}$ be a Markov chain with a one-step transition matrix $\mathbf{p}$. Suppose that the process $\{Y_n : n \ge 0\}$ is viewed only at every mth time step (m fixed) and let $X_n = Y_{nm}$, for $n = 0, 1, 2, \ldots$. Show that $\{X_n : n \ge 0\}$ is a Markov chain with *one-step* transition law given by $\mathbf{p}^m$.

    (b) Suppose $\{X_n : n \ge 0\}$ is a Markov chain with transition probability matrix $\mathbf{p}$. Let $n_1 < n_2 < \cdots < n_k$. Prove that

$$P(X_{n_k} = j \mid X_{n_1} = i_1, \ldots, X_{n_{k-1}} = i_{k-1}) = p_{i_{k-1}j}^{(n_k - n_{k-1})}.$$

13. (*Random Walks on a Group*) Let G be a finite group with group operation denoted by $\oplus$ and e its identity element. If $\oplus$ is commutative, i.e., $x \oplus y = y \oplus x$ for all $x, y \in G$, then G is called *Abelian*. Let $X_1, X_2, \ldots$ be i.i.d. random variables taking values in G and having a common probability distribution Q with $Q(\{g\}) = P(X_n = g)$, $g \in G$.

    (a) Show that the *random walk on* G defined by $S_n = X_0 \oplus X_1 \oplus \cdots \oplus X_n$, $n \ge 0$, is a Markov chain and calculate its transition probability matrix. Note that it is *not* necessary for G to be abelian for $\{S_n\}$ to be Markov.

    (b) (*Top-In Card Shuffle*) Construct a model for *card shuffling* as a Markov chain on a (nonabelian) permutation group on N symbols in which the top card of the deck is inserted at a randomly selected location in the deck at each shuffle.

    (c) Calculate the transition probability matrix for $N = 3$. [*Hint*: Shuffles are of the form $(c_1, c_2, c_3) \to (c_2, c_1, c_3)$ or (c_2, c_3, c_1) only.]

14. (*One-Dimensional Ising Ferromagnet*) The one-dimensional Ising model may be viewed as a doubly indexed stationary Markov chain $\{\eta_m : m = 0, \pm 1, \pm 2, \ldots\}$ on *spin values* $S = \{-1, +1\}$ with invariant marginal distribution $(1/2, 1/2)$, and having transition probabilities $\begin{pmatrix} p & q \\ q & p \end{pmatrix}$, where

$p_{11} = p_{-1-1} = p = \frac{e^{\beta J}}{2\cosh(\beta J)}$ is the Ising model parameterization with the so-called *inverse temperature parameter* $\beta > 0$ and *coupling constant* J. The ferromagnetic case in which neighboring spins are more likely to align (i.e., $p > 1/2$) is defined by $J > 0$.

(a) Show that if $\{-1, +1\}$ is viewed as a multiplicative Abelian group, then the Markov chain $\{\eta_m : m = 0, \pm 1, \pm 2, \dots\}$ is simply a random walk on $\{-1, +1\}$; that is, the increments $\eta_{m+1}\eta_m^{-1} \equiv \eta_{m+1}\eta_m, m \in \mathbb{Z}$ are i.i.d. [*Hint*: Let $\sigma_m, m \in \mathbb{Z}$ be i.i.d. ± 1-Bernoulli with parameter $p = P(\sigma_m = 1)$ and show that the random walk defined by $\eta_{m+1} = \eta_m \sigma_{m+1}, m \in \mathbb{Z}$, has the appropriate transition probabilities, and invariant marginal distributions $(1/2, 1/2)$.]

(b) Show that in the ferromagnetic case the distribution, say P, of the Markov chain with invariant distribution $\pi_{+1} = \pi_{-1} = 1/2$ is infinitely divisible[1] as a probability measure on the multiplicative Abelian product group $G = \{-1, +1\}^{\mathbb{Z}}$ with coordinatewise multiplication; that is, for any positive integer m there is a probability measure Q_m on G such that $P = Q_m^{*m}$, where ^{m*} denotes the m-fold convolution. [*Hint*: Show that the (coordinatewise) product of two independent Markov chains with parameters, p_1, p_2 is also a Markov chain with parameter $p_3 = 2p_1p_2 - p_1 - p_2 + 1 \in (1/2, 1)$ for $p_1, p_2 \in (1/2, 1)$.]

15. A random number of individuals with a highly contagious disease enter an infinite population of healthy individuals. During each subsequent period, either one of these carriers will surely infect a new person or be discovered and removed by public health officials. An unremoved infected individual becomes a carrier. Each carrier is discovered and removed, independently of the others, with probability q = 1-p at each unit of time. The time evolution of the number of infected individuals in the population is a Markov chain $\{X_n : n = 0, 1, 2, \dots\}$. What are its transition probabilities?
16. Suppose that at each unit of time each particle located in a fixed region of space has probability p, independently of the other particles present, of leaving the region. Also, at each unit of time a random number of new particles having Poisson distribution with parameter λ enter the region independently of the number of particles already present at time n. Let X_n denote the number of particles in the region at time n. Calculate the transition matrix of the Markov chain $\{X_n : n \geq 0\}$.
17. We are given two boxes A and B containing a total of N labeled balls. A ball is selected at random (all selections being equally likely) at time n from among the N balls and then a box is selected at random. Box A is selected with probability p and B with probability $q = 1 - p$ independently of the ball selected. The selected ball is moved to the selected box, unless the ball is

[1] This property was investigated in Waymire (1984) and Glaffig and Waymire (1987) for Ising models.

already in it. Consider the Markov evolution of the number X_n of balls in box A. Calculate its transition matrix.

18. Each cell of a certain organism contains N particles, some of which are of type A and the others type B. The cell is said to be in state j if it contains exactly j particles of type A. Daughter cells are formed by cell division as follows: Each particle replicates itself and a daughter cell inherits N particles chosen at random from the $2j$ particles of type A and the $2N - 2j$ particles of type B present in the parental cell. Calculate the transition matrix of this Markov chain.
19. (*Length of a Queue*) Suppose that items arrive at a shop for repair on a daily basis but that it takes one day to repair each item. New arrivals are put on a waiting list for repair. Let A_n denote the number of arrivals during the nth day. Let X_n be the length of the waiting list at the end of the nth day. Assume that $A_1, A_2, \dots$ is an i.i.d. nonnegative integer-valued sequence of random variables with $a(x) = P(A_n = x)$, $x = 0, 1, 2, \dots$. Assume that A_{n+1} is independent of $X_0, \dots, X_n$ $(n \geq 0)$. Calculate the transition probabilities for $\{X_n : n \geq 0\}$.
20. (*A Renewal Process*) A system requires a certain device for its operation that is subject to failure. Inspections for failure are made at regular points in time, so that an item that fails during the nth period of time between $n - 1$ and n is replaced at time n by a device of the same type having an independent service life. Let p_n denote the probability that a device will fail during the nth period of its use. Let X_n be the age (in number of periods) of the item in use at time n. A new item is started at time $n = 0$, and $X_n = 0$ if an item has just been replaced at time n. Calculate the transition matrix of the Markov chain $\{X_n : n \geq 0\}$.

Chapter 8
Markov Processes with General State Space

The focus of this chapter is that of discrete parameter Markov processes on a general (measurable) state space S. Some general considerations for the existence and uniqueness of invariant probabilities are provided.

Consider a discrete parameter stochastic process $\{X_n\}_{n\geq 0}$ with general state space S equipped with a σ-field $\mathcal{S}$ of subsets of S. Here the measurable space $(S, \mathcal{S})$ may be a countable set and $\mathcal{S}$ its power set as in the previous chapter, or more generally, S may be a metric space and $\mathcal{S}$ its (Borel) σ-field generated by the open subsets of S, for example. Unless stated otherwise, a map f on S into a metric space M is (implicitly) referred to as *measurable* (or more explicitly, *Borel measurable*) if it is measurable with respect to $\mathcal{S}$ on S and the Borel σ-field on M, i.e., if $f^{-1}(B) \in \mathcal{S}$ $\forall$ Borel sets B in M. Also, for any family $\{Y_\lambda : \lambda \in \Lambda\}$ of random variables on a probability space $(\Omega, \mathcal{F}, P)$, $\sigma\{Y_\lambda : \lambda \in \Lambda\}$ is the σ*-field generated by the family* i.e., it is the smallest σ-field ($\subset \mathcal{F}$) with respect to which Y_λ is measurable for each $\lambda \in \Lambda$.

Definition 8.1. A stochastic process $\{X_0, X_1, \ldots, X_n, \ldots\}$ having state space S equipped with a σ-field $\mathcal{S}$ has the *Markov property with regular transition probabilities* if for each $n \geq 0$,

$$P(X_{n+1} \in B|\sigma\{X_0, X_1, \ldots, X_n\}) = p_n(X_n, B), \quad B \in \mathcal{S}, \tag{8.1}$$

where for each n,

1. For each $B \in \mathcal{B}$, $x \to p_n(x, B)$ is a measurable function on S.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_8

2. For each $x \in S$, $B \to p_n(x, B)$ is a probability measure on $\mathcal{S}$.

A stochastic process having the Markov property is called a *Markov process* with *transition probabilities* $p_n(x, dy)$, $n \geq 0$. The Markov process is said to be *homogeneous* if the transition probabilities $p_n(x, dy)$ are the same for all $n = 1, 2, \dots$, say $p(x, dy)$.

For the most part we will consider Markov processes with homogeneous transition probabilities, and unless explicitly stated otherwise, by a Markov process we will mean Markov process with homogeneous conditional probabilities from here out.

The special case in which there is a σ-finite measure ν on $(S, \mathcal{S})$ and a nonnegative measurable function $p(x, y)$ on $S \times S$ such that $\int_S p(x, y)\nu(dy) = 1$ and $p(x, B) = \int_B p(x, y)\nu(dy)$, for all $x \in S$, occurs often. In this case, (8.7) and (8.12) below are iterated integrals of $p(x_0, x_1)p(x_1, x_2)\dots p(x_{n-1}, x_n)$. This was the case, for example, in Chapter 7, where S is finite or countable and ν is the counting measure.

The first task is to establish the following construction.

Proposition 8.1. Given an initial distribution μ and a transition probability $p(x, dy)$, there is a unique probability measure P_μ on the canonical space $(S^\infty, \mathcal{S}^{\otimes\infty})$ with the property (8.1) for the coordinate projections $X_n(\mathbf{x}) = x_n (n = 0, 1, 2, \dots)$, where $\mathbf{x} = (x_0, x_1, x_2, \dots) \in S^\infty \equiv S^{\{0,1,2,\dots\}}$, and $\mathcal{S}^{\otimes\infty}$ is the product σ-field; i.e., the smallest σ-field on S^∞ for which each of the respective coordinate projection maps is measurable. Moreover,

$$P_\mu(X_{n+1} \in B_{n+1} | \sigma(X_0, \dots, X_n)) = p(X_n, B_{n+1}), \quad B \in \mathcal{S}. \tag{8.2}$$

Proof. To begin, one constructs a collection of probability measures $P_{\mu,n}, n = 0, 1, 2, \dots$, on the finite dimensional product spaces $(S^n, \mathcal{S}^{\otimes n})$, where $S^n = S \times \cdots \times S$, and $\mathcal{S}^{\otimes n} = \mathcal{S} \otimes \cdots \otimes \mathcal{S}$, ($n$-fold), to represent the respective distributions of $(X_0, X_1, \dots, X_n), n = 0, 1, 2 \dots$.

To this end, for a bounded $\mathcal{S}^{\otimes n}$-measurable function f on S^n, define integrations on the product spaces iteratively, beginning with the innermost integral (with respect to $p(x_{n-1}, dx_n)$), keeping all variables except the last one, namely x_n, fixed; that is,

$$\begin{aligned}
&\int_S \cdots \int_S f(x_0, x_1, \dots, x_{n-1}, x_n)p(x_{n-1}, dx_n) \cdots p(x_0, dx_1)\mu(dx_0) \\
&= \int_S \cdots \int_S f_1(x_0, x_1, \dots, x_{n-1})p(x_{n-2}, dx_{n-1}) \cdots p(x_0, dx_1)\mu(dx_0) \\
&= \int_S \cdots \int_S f_2(x_0, x_1, \dots, x_{n-2})p(x_{n-3}, dx_{n-2}) \cdots p(x_0, dx_1)\mu(dx_0) \\
&\vdots
\end{aligned}$$

$$
\begin{aligned}
&= \int_S \int_S f_{n-1}(x_0, x_1) p(x_0, dx_1) \mu(dx_0) \\
&= \int_S f_n(x_0) \mu(dx_0), \qquad (8.3)
\end{aligned}
$$

where

$$f_1(x_0, x_1, \ldots, x_{n-1}) = \int_S f(x_0, x_1, \ldots, x_{n-1}, x_n) p(x_{n-1}, dx_n),$$

$$f_2(x_0, x_1, \ldots, x_{n-2}) = \int_S f_1(x_0, x_1, \ldots, x_{n-1}) p(x_{n-2}, dx_{n-1}),$$

and so on. To justify this integration, first observe that $y \to f(x_0, \ldots, x_{n-1}, y)$ is $\mathcal{S}$-measurable for any fixed $(x_0, \ldots, x_{n-1}) \in S^n$; namely, it is clear for indicators of measurable rectangles $C = B_0 \times \cdots \times B_n$, ($B_i \in \mathcal{S}, i = 0, 1, \ldots, n$), therefore, by the $\pi - \lambda$ theorem for all $C \in \mathcal{S}^{\otimes n}$. Since every bounded measurable function is a pointwise (uniform) limit of a sequence of simple functions, the measurability of $y \to f(x_0, \ldots, x_{n-1}, y)$ follows for all bounded $\mathcal{S}^{\otimes n}$-measurable functions f on S^n. The $\mathcal{S}^{\otimes n}$-measurability of the map $(x_0, \ldots, x_{n-1}) \to f_1(x_0, \ldots, x_{n-1}) = \int_S f(x_0, x_1, \ldots, x_{n-1}, y) p(x_{n-1}, dy)$ follows.

Now, to define $P_{\mu,n}$, take $f = \mathbf{1}_C, C \in \mathcal{S}^{\otimes n+1}$. Writing $C_{x_0,\ldots,x_{n-1}} = \{y \in S : (x_0, \ldots, x_{n-1}, y) \in C\}$, and $f_1(x_0, x_1, \ldots, x_{n-1}) = p(x_{n-1}, C_{x_0,\ldots,x_{n-1}})$, define

$$P_{\mu,0} = \mu, \;\; P_{\mu,1}(C) = \int_S \int_S \mathbf{1}_{C_{x_0}}(x_1) p(x_0, dx_1) \mu(dx_0) \; C \in \mathcal{S}^{\otimes 2}. \qquad (8.4)$$

More generally, define for $n \geq 2$

$$
\begin{aligned}
P_{\mu,n}(C) &= \int_S \cdots \int_S \mathbf{1}_{C_{x_0,\ldots,x_{n-1}}}(x_n) p(x_{n-1}, dx_n) p(x_{n-2}, dx_{n-1}) \cdots p(x_0, dx_1) \mu(dx_0) \\
&= \int_S \cdots \int_S p(x_{n-1}, C_{x_0,\ldots,x_{n-1}}) p(x_{n-2}, dx_{n-1}) \cdots p(x_0, dx_1) \mu(dx_0). \qquad (8.5)
\end{aligned}
$$

In particular, for a measurable rectangle $C = B_0 \times \cdots \times B_n$, (8.5) reduces to

$$P_{\mu,1}(B_0 \times B_1) = \int_{B_0} p(x_0, B_1) \mu(dx_0),$$

and

$$P_{\mu,n}(B_0 \times \cdots \times B_n) = \int_{B_0} \cdots \int_{B_{n-1}} p(x_{n-1}, B_n) p(x_{n-2}, dx_{n-1}) \cdots p(x_0, dx_1) \mu(dx_0).$$

Finite additivity of $P_{\mu,n}$ follows by writing $\mathbf{1}_C = \sum_{j=1}^{m} \mathbf{1}_{C_j}$ for disjoint $C_j \in \mathcal{S}^{\otimes n}$, and countable additivity then follows by the monotone convergence theorem. This completes the construction of the finite dimensional distributions. It is simple to check that this collection of probability measures is *consistent* in the sense of the Kolmogorov existence theorem.[1] Thus, by the Kolmogorov existence theorem in the case that S is also a Borel subset of a Polish space or, more generally by Tulcea's theorem[2] for an arbitrary measurable space $(S, \mathcal{S})$, one arrives at a probability measure P_μ on $(S^\infty, \mathcal{S}^{\otimes\infty})$ such that the distribution of the projection $(X_0, X_1, \dots, X_n) : S^\infty \to S^{n+1}$ is $P_{\mu,n}$, $n = 0, 1, 2, \dots$. For this, let $B_{n+1} \in \mathcal{S}$. If $C \in \mathcal{S}^{\otimes(n+1)}$, then it follows from (8.5) that

$$\begin{aligned} &P_{\mu,n+1}(C \times B_{n+1}) \\ &= \int_S \cdots \int_S \mathbf{1}_C(x_0, \dots, x_n) \mathbf{1}_{B_{n+1}}(x_{n+1}) P_{\mu,n+1}(dx_0 \times \cdots \times dx_{n+1}) \\ &= \int_{C \times S} p(x_n, B_{n+1}) P_{\mu,n}(dx_0 \times \cdots \times dx_n). \end{aligned} \tag{8.6}$$

If one takes $B_{n+1} = S$, then the consistency is checked: $P_{\mu,n+1}(C \times S) = P_{\mu,n}(C)$. This completes the construction of P_μ on $(S^\infty, \mathcal{S}^{\otimes\infty})$. Finally, note that (8.6) also expresses the Markov property in the form:

$$P_\mu(X_{n+1} \in B_{n+1} | \sigma(X_0, \dots, X_n)) = p(X_n, B_{n+1}). \qquad \blacksquare$$

With the construction carried out in the proof of Proposition 8.1, given a transition probability $p(x, dy)$ and an initial distribution $\mu(dx)$, one may make the following definition.

Definition 8.2. A stochastic process $\{X_0, X_1, \dots\}$ on an arbitrary probability space $(\Omega, \mathcal{F}, P)$ is Markov with transition probability $p(x, dy)$ and initial distribution $\mu(dx)$ if its distribution is P_μ on $(S^\infty, \mathcal{S}^{\otimes\infty})$. If $\mu = \delta_x$, then we write P_x for P_{δ_x}.

Note: To avoid a clutter of symbols, we will often abuse notation in probabilities associated with Markov processes $X = \{X_n : n = 0, 1, \dots\}$ defined on a (possibly non-canonical probability space) $(\Omega, \mathcal{F}, P)$ as $P_\mu(X \in B)$ to indicate that X_0 has distribution μ. That is, we may use the expression for the corresponding probability in canonical space, where X_n is the nth coordinate projection on S^∞.

Proposition 8.2. If $\{X_n\}_{n\geq 0}^\infty$ has the Markov property, then one may obtain the distribution at $m \geq 1$ time points into the future inductively for $B_1, B_2, \dots, B_m \in \mathcal{S}$, as

$$P(X_{n+1} \in B_1, \dots, X_{n+m} \in B_m | \sigma\{X_0, X_1, \dots, X_n\})$$

[1] See BCPT p. 236 or Billingsley (1968), p. 235.

[2] see BCPT p. 168 or Billingsley (1986), pp. 510–511.

$$\begin{aligned}
&= P(X_{n+1} \in B_1, \ldots, X_{n+m} \in B_m | \sigma\{X_n\}) \\
&= \int_{B_1} \cdots \int_{B_m} p(x_{m-1}, dx_m) \ldots p(x_1, dx_2) p(X_n, dx_1) \\
&= P_x(X_1 \in B_1, \ldots, X_m \in B_m)|_{x=X_n}, \qquad (8.7)
\end{aligned}$$

where the integration is an iterated integral.

Proof. The first equality follows from the Markov property. The second follows from (8.6), as does the last. To prove this in detail, simply condition the left hand side of (8.7) on the larger σ-field $\sigma(X_0, \ldots, X_{n+m-1})$ and use the smoothing property of conditional expectations. Since

$$\int_{B_m} p(x_{m-1}, dx_m) = p(x_{m-1}, B_m),$$

the first integration yields

$$f_1(X_{n+1}, \ldots, X_{n+m-1}) = \mathbf{1}[X_{n+1} \in B_1, \ldots, X_{n+m-1} \in B_{m-1}] \cdot p(X_{m-1}, B_m),$$

in the notation introduced earlier. That is, the left side of (8.7) equals

$$\mathbb{E}(f_1(X_{n+1}, \ldots, X_{n+m-1})|\sigma(X_0, \ldots, X_n)). \qquad (8.8)$$

Next taking the conditional expectation of (8.8), given $\sigma(X_0, \ldots, X_{n+m-2})$, one has

$$\begin{aligned}
&\mathbb{E}(f_2(X_{n+1}, \ldots, X_{n+m-2})|\sigma(X_0, \ldots, X_n)) \\
&\quad = \mathbf{1}[X_{n+1} \in B_1, \ldots, X_{n+m-2} \in B_{m-2}] \int_{B_{m-1}} p(X_{m-1}, B_m) p(X_{m-2}, dx_{m-1}).
\end{aligned}$$

Continuing in this way, the probability is ultimately given by a function of X_n of the form:

$$\begin{aligned}
&P(X_{n+m} \in B_m, \ldots, X_{n+1} \in B_1 | \sigma\{X_0, X_1, \ldots, X_n\}) \\
&\quad = \mathbb{E}(f_{m-1}(X_{n+1})|\sigma(X_0, \ldots, X_n) \\
&\quad = \mathbb{E} f_{m-1}(X_{n+1}|\sigma(X_n)) = \int_{B_1} f_{m-1}(x_1) p(X_n, dx_1). \qquad (8.9)
\end{aligned}$$

■

Observe that taking $B_m = B \in \mathcal{S}$, $B_1 = B_2 = \cdots = B_{m-1} = S$ in (8.7), one has

$$P(X_{n+m} \in B|\sigma\{X_0, \ldots, X_n\}) = p^{(m)}(X_n, B), \qquad (8.10)$$

where the *m-step transition probabilities* are recursively given by

$$p^{(m+1)}(x, B) = \int_S p^{(m)}(y, B)p(x, dy), \quad B \in \mathcal{S},\ x \in S, \tag{8.11}$$

with $p^{(1)}(x, B) \equiv p(x, B)$. Notice that by taking successive conditional expectations of $\mathbf{1}[X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}, X_n \in B_n]$ given $\sigma(X_0, \dots, X_{n-1}), \sigma(X_0, \dots, X_{n-2}), \dots$, one obtains

$$\begin{aligned} &P_\mu(X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}, X_n \in B_n) \\ &\quad = \int_{B_0}\int_{B_1} \cdots \int_{B_n} p(x_{n-1}, dx_n)p(x_{n-2}, dx_{n-1}) \cdots p(x_0, dx_1)\mu(dx_0) \\ &\quad = P_{\mu,n}(B_0 \times B_1 \times \cdots \times B_n), \end{aligned}$$

where μ is the *initial distribution* of the process as defined by

$$\mu(B) = P(X_0 \in B), \quad B \in \mathcal{S}. \tag{8.12}$$

The following is another equivalent version of the Markov property that is commonly used.

Proposition 8.3. Let $\{X_n\}_{n\geq 1}$ be a Markov process with a transition probability $p(x, dy)$ and some initial distribution μ. Then the conditional distribution of the *after-n process* defined by $X_n^+ := (X_n, X_{n+1}, \dots)$, given $\sigma\{X_0, \dots, X_n\}$, is P_{X_n}. That is, it equals P_x on the event $[X_n = x] \subset \Omega$.

Proof. In view of (8.7), one has for finite dimensional events of the form $C = B_0 \times B_1 \times B_m \times S^\infty$, $B_i \in \mathcal{S}, 0 \leq i \leq m$, that $P(X_n^+ \in C|\sigma\{X_0, \dots, X_n\}) = P_{X_n}((X_0, X_1, \dots) \in C)$. Now observe that the collection of sets $C \in \mathcal{S}^{\otimes\infty}$ such that this equation holds is a λ-system, which contains a π-system of finite dimensional events. The assertion thus follows from an application of the $\pi - \lambda$ theorem. ■

Situations in which there is an initial probability π for the Markov process which is invariant under the evolution are of particular interest, especially because when unique it represents the long term behavior of the process regardless of how it is initiated.

Definition 8.3. A probability π on $\mathcal{S}$ is said to be an *invariant probability* or *steady state distribution* for a Markov process $\{X_n\}_{n\geq 0}$ with transition probabilities $p(x, dy)$ if

$$\int_S p(x, B)\pi(dx) = \pi(B) \quad \text{for all } B \in \mathcal{S}. \tag{8.13}$$

Notice that if π is an invariant initial probability for $\{X_n\}_{n\geq 0}$, i.e., X_0 has the invariant distribution π, then the left side of (8.13) is $P(X_1 \in B)$. One has that $P(X_n \in B) = \pi(B)$, $B \in \mathcal{S}$, that is,

$$P(X_{n+1} \in B) = \int_S p(x, B) P(X_n \in dx) = \int_S p(x, B)\pi(dx) = \pi(B), \qquad (8.14)$$

which for $n = 0$ is true by definition, and the general case follows by the Markov property and induction. In addition to questions of

(i) *Existence*
(ii) *Uniqueness*

of invariant probabilities, one also seeks

(iii) *Basins of Attraction*

i.e., initial distributions under which convergence to a given invariant probability will hold, and

(iv) *Rates of Convergence*

to name a few of the central topics of the theory. In view of (8.7),

$$P(X_{n+1} \in B_1, \ldots, X_{n+m} \in B_m) = P(X_1 \in B_1, \ldots, X_m \in B_m), \quad B_i \in \mathcal{S}. \qquad (8.15)$$

In particular, in this context the distribution of the after-n process $X_n^+ \equiv \{X_{n+m} : m = 0, 1, 2, \ldots\}$ coincides with that of $X_0^+ = \{X_m : m = 0, 1, 2, \ldots\}$ for each $n = 1, 2, \ldots$, a property referred to as *stationarity* of the process $\{X_0, X_1, \ldots\}$, as discussed in the earlier chapters of the text. From this perspective, theorems providing

(v) *Law of Averages*

and

(vi) *Fluctuation Law*

in the forms of a strong law of large numbers (ergodic theorem) and a central limit theorem for averages of the form $\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)$ in the presence of a unique invariant initial distributions are also essential to a complete theory.

For the next definition and elsewhere in the book, $B(S)$ denotes *the set of all real-valued bounded measurable functions on* S, equipped with the sup-norm $\|f\| = \sup_{x\in S}|f(x)|$.

Definition 8.4. Given a transition probability $p(x, dy)$ on $(S, \mathcal{S})$, the *transition operator* T is the map on $B(S)$ (into $B(S)$) defined by

$$Tf(x) = \int_S f(y)p(x, dy), \quad f \in B(S). \qquad (8.16)$$

Note that the measurability of $x \to p(x, B)$ for every $B \in \mathcal{S}$ implies the measurability of Tf (Exercise 1). The transition operator is a positive linear contraction on $B(S)$ such that $T\mathbf{1} = \mathbf{1}$, where $\mathbf{1} \in B(S)$ is the constant function on S with constant numerical value one (Exercise 1). In particular, given any such transition operator, one may recover $p(x, B) = T\mathbf{1}_B(x), x \in S, B \in \mathcal{S}$.

Note that for a Markov process $\{X_n : n = 0, 1, 2, \dots\}$ with transition probability $p(x, dy)$, one has $Tf(x) = \mathbb{E}(f(X_1)|X_0 = x) \equiv \mathbb{E}(f(X_1)|\sigma(X_0))|_{X_0=x}$, which may be expressed as $\mathbb{E}_x f(X_1)$. In fact, by the Markov property and induction,

$$\begin{aligned}
\mathbb{E}_x f(X_{n+1}) &= \mathbb{E}_x\mathbb{E}_x[f(X_{n+1}) \mid \sigma(X_0, \dots, X_n)] \\
&= \mathbb{E}_x\mathbb{E}_{X_n} f(X_{n+1}) \\
&= \mathbb{E}_x Tf(X_n) \\
&= \mathbb{E}_x\mathbb{E}(Tf(X_n)|\sigma(X_{n-1})) \\
&= \mathbb{E}_x TTf(X_{n-1}) = \mathbb{E}_x T^2 f(X_{n-1}) = \cdots = \mathbb{E}T^n f(X_1) \\
&= T^{n+1} f(x), \quad x \in S, n \geq 0,
\end{aligned}$$

i.e., T^n is the transition operator defined by the n-step transition probability $p^{(n)}(x, dy)$. The relation (8.13) implies (and is, therefore, equivalent to)

$$\int_S (Tf)(x)\pi(dx) = \int_S f(x)\pi(dx) \quad \text{for all } f \in B(S). \tag{8.17}$$

Thus (8.17) also defines the convenient notation "$\pi(dy) = \int_S p(x, dy)\pi(dx)$." Just as (8.13) implies (8.17), (8.14) implies

$$\int_S (T^n f)(x)\pi(dx) = \int_S f(x)\pi(dx) \quad \text{for all } f \in B(S), \text{ for all } n \geq 1. \tag{8.18}$$

One approach to obtain invariant probabilities is by consideration of long time steady state distributions. In particular, one might anticipate that if for some $x \in S$, the distribution $p^{(n)}(x, dy)$ of the state X_n converges weakly as $n \to \infty$ to some limit probability distribution π_x, then π_x should be invariant under continued evolution. However, as the following example shows, one must be careful.

Example 1 *(Liggett).* Let $S = \{0, 1, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots, \frac{m}{m+1}, \dots\}$, and let $\mathcal{S}$ be the power set of S. Then S is a metric space with the usual distance on the line, $d(x, y) = |x - y|$, and $\mathcal{S}$ is the Borel σ-field. Fix $0 < \theta \leq 1$ and define

$$\begin{aligned}
&p\left(0, \left\{\tfrac{1}{2}\right\}\right) = \theta, &&p\left(0, \left\{\tfrac{2}{3}\right\}\right) = 1 - \theta \\
&p\left(\tfrac{m}{m+1}, \left\{\tfrac{m+1}{m+2}\right\}\right) = \theta, &&p\left(\tfrac{m}{m+1}, \left\{\tfrac{m+2}{m+3}\right\}\right) = 1 - \theta, \quad m = 1, 2, \dots . \\
&p(1, \{0\}) = \theta, &&p\left(1, \left\{\tfrac{1}{2}\right\}\right) = 1 - \theta.
\end{aligned}$$

Then $p^{(n)}(x, dy)$ converges weakly to $\delta_{\{1\}}$, but this is clearly not an invariant probability.

Note that weak convergence requires convergence of the sequences of integrals $\int_S f(y)p^{(n+1)}(x,dy) = T^{n+1}f(x) = T^n(Tf)(x)$, for all $f \in C_b(S)$, as $n \to \infty$; recall $C_b(S) \subset B(S)$ denotes the subset of all bounded continuous functions on the metric space S. As usual whenever S is a metric space, we take $\mathcal{S}$ to be the Borel σ-field on S for the uniform norm on $C_b(S)$.

Definition 8.5. A transition probability $p(x,dy)$ on a metric space S is said to be *Feller continuous*, or *weakly Feller continuous*, if for every $f \in C_b(S)$, $Tf \in C_b(S)$. In this case one also says $p(x,dy)$ has the *(weak) Feller property*.

Observe that Feller continuity of $p(x,dy)$ means that the map $x \to p(x,dy)$, on S into the set $\mathcal{P}(S)$ of all probability measures on $(S,\mathcal{S})$, is *weakly continuous*. Moreover, T is a positive, linear contraction operator on $C_b(S)$ with $T1 = 1$. Conversely, if S is a compact metric space, then any such operator on $C_b(S)$ uniquely determines Feller transition probabilities $p(x,dy)$ by applying the Riesz Representation Theorem[3] from analysis to the bounded linear functional $f \to Tf(x)$, $f \in C_b(S)$, for each $x \in S$ (Exercise 6).

Notice that since $C_b(S)$ is measure-determining,[4] the condition (8.17) defining an invariant probability may be restricted to $f \in C_b(S)$.

Another obstacle to this approach to the determination of invariant probabilities is evident in the simple two-state example $p_{01} = p_{10} = 1$, $p_{00} = p_{11} = 0$. In this case, $\pi_0 = \pi_1 = 1/2$ is the unique invariant probability, but $p_{01}^{(n)}$ oscillates between 1 and 0 as a function of n. However, these oscillations can be averaged out by considerations of $\frac{1}{2n+1}\sum_{r=0}^{2n} p_{0j}^{(r)} \to 1/2$, as $n \to \infty$, for $j = 0, 1$. So this example suggests that time averaging may be required, and Liggett's Example 1 shows that the hypothesis of Feller continuity in Proposition 8.4 cannot in general be dispensed with in the weak convergence approach to invariant probabilities (Exercise 5). From a probabilistic perspective, note also that

$$\frac{1}{m}\sum_{r=0}^{m-1} p^{(r)}(x,B) = \frac{1}{m}\sum_{r=0}^{m-1} \mathbb{E}_x \mathbf{1}[X_r \in B] = \mathbb{E}_x\left(\frac{\sum_{r=0}^{m-1}\mathbf{1}[X_r \in B]}{m}\right) \tag{8.19}$$

is the expected proportion of visits to the set $B \in \mathcal{S}$ in time 0 to $m-1$, starting from x.

Proposition 8.4. Suppose S is a metric space and $p(x,dy)$ is a Feller continuous transition probability on $(S,\mathcal{S})$. (a) If for some $x \in S$ there is a sequence of integers $1 \le n_1 < n_2 < \dots$ such that, as $k \to \infty$,

$$\frac{1}{n_k}\sum_{r=0}^{n_k-1} p^{(r)}(x,dy) \text{ converges weakly to } \pi_x(dy) \tag{8.20}$$

[3] BCPT p 237.

[4] See BCPT, p. 11.

for some probability measure π_x, then π_x is an invariant for $p(x, dy)$.

(b) If, for some sequence $1 \le n_1 < n_2 \ldots$, (8.20) holds for every $x \in S$ with the same limit $\pi_x = \pi$ for all x, then π is the unique invariant probability.

Proof.

(a) The relation (8.20) says that

$$\frac{1}{n_k} \sum_{r=0}^{n_k-1} (T^r f)(x) \longrightarrow \int_S f(y)\pi_x(dy) \quad \text{for all } f \in C_b(S). \tag{8.21}$$

Replacing f by Tf (which belongs to $C_b(S)$ by hypothesis), one gets

$$\frac{1}{n_k} \sum_{r=1}^{n_k} T^r f(x) \longrightarrow \int_S Tf(y)\pi_x(dy) \quad \text{for all } f \in C_b(S). \tag{8.22}$$

But the difference between the left sides of (8.21) and (8.22) equals in magnitude $|(T^{n_k} f)(x) - f(x)|/n_k \le 2\sup\{|f(x)| : x \in S\}/n_k$, which goes to zero as $k \to \infty$. Hence the limits in (8.21) and (8.22) are the same. Thus, since $C_b(S)$ is measure-determining, one has $\pi_x(dz) = \int_S p(y, dz)\pi_x(dy)$; see Lemma 1 below.

(b) By (a), π is invariant. Suppose that, under the hypothesis of part (b), π' is another invariant probability, and then integrating the two sides of (8.21) with respect to π', one obtains

$$\frac{1}{n_k} \sum_{r=0}^{n_k-1} \int_S (T^r f)(x)\pi'(dx) \longrightarrow \int_S \left[\int_S f(y)\pi(dy)\right] \pi'(dx). \tag{8.23}$$

By invariance of π', the left side equals $\int f d\pi'$ (see (8.18)), while the right side is $\int f d\pi$. Thus $\int f d\pi' = \int f d\pi$ for every $f \in C_b(S)$, implying $\pi' = \pi$ since $C_b(S)$ is measure-determining. ■

Lemma 1. If Q_1 and Q_2 are probability measures on the Borel σ-field of a metric space S such that $\int_S f dQ_1 = \int_S f dQ_2$ for all bounded continuous real-valued functions f on S, then $Q_1 = Q_2$.

Proof. Let $\mathcal{C}$ be the collection of Borel sets B such that $Q_1(B) = Q_2(B)$. Then it is simple to check that $\mathcal{C}$ is a σ-field. Since $\mathcal{B}$ the Borel σ-field is the smallest σ-field containing all closed sets, it is sufficient to show that $\mathcal{C}$ contains all closed sets. For this, it is enough to show that for each (closed) $F \subset S$, there exists a sequence of nonnegative functions $\{f_n\} \subset C_b(S)$ such that $f_n \downarrow \mathbf{1}_F$ as $n \uparrow \infty$. Since F is closed, one may view $x \in F$ in terms of the equivalent condition that $\rho(x, F) = 0$, where $\rho(x, F) := \inf\{\rho(x, y) : y \in F\}$. Let $h_n(r) = 1 - nr$ for $0 \le r \le 1/n, h_n(r) = 0$ for $r \ge 1/n$. Then take $f_n(x) = h_n(\rho(x, F))$. In particular, $\mathbf{1}_F(x) = \lim_n f_n(x), x \in S$, and Lebesgue's monotone convergence theorem applies to show $F \in \mathcal{C}$. ■

As an immediate corollary, one gets the following corollary:

Corollary 8.5. If a transition probability $p(x, dy)$ on a metric space has the (weak) Feller property and there exists a Caesaro limit in the weak topology, namely,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r=0}^{n-1} p^{(r)}(x, dy) = \pi(dy) \tag{8.24}$$

such that the probability measure π does not depend on x, then π is the unique invariant probability.

Many important Markov processes do not admit an invariant probability, such as the case, for example, of a random walk on $\mathbb{R}^k$ with an arbitrary step size distribution $Q \neq \delta_{\{0\}}$ (Exercise 8). There is one case, namely that of a compact state space, where every Feller transition probability admits at least one invariant probability.

Proposition 8.6. Let S be a compact metric space and $\mathcal{S}$ its Borel σ-field. If $p(x, dy)$ is a Feller transition probability on $(S, \mathcal{S})$, then it admits an invariant probability.

Proof. Fix $x \in S$ and consider the sequence of probability measures μ_n, $n \geq 1$, given by $\mu_n(B) = (1/n)\sum_{m=1}^{n} p^{(m)}(x, B)$, $B \in \mathcal{S}$. Since $\mathcal{P}(S)$ is a weakly compact metric space,[5] there exists a subsequence $\{\mu_{n_k} : k = 1, 2, \ldots\}$, which converges weakly to a probability measure π_x, say. By Proposition 8.4, π_x is invariant. ■

As a simple corollary, we get the following result for finite Markov chains.

Corollary 8.7. A Markov chain on a finite state space S has at least one invariant probability.

Proof. This follows from Proposition 8.6 by making S a compact metric space with the metric $d(x, y) = 1$ if $x \neq y$, $d(x, x) = 0$. Then $\mathcal{S} \equiv \mathcal{B}(S)$ is the class of all subsets of S, and every real-valued function on S is continuous. ■

Remark 8.1. A direct proof of Corollary 8.7, which does not use Proposition 8.6 will be given later (see Corollary13.8 in Chapter 13).

The next approach to invariant probabilities is based on *symmetries*. For the definition below, consider a stationary Markov process $\{X_n\}_{n\geq 0}$ on a state space $(S, \mathcal{S})$ with transition probability $p(x, dy)$. Since the distribution of such a process is invariant under time shift, i.e., $\{X_n\}_{n\geq 0}$ and $\{X_n\}_{n\geq k}$ have the same distribution, one may use Kolmogorov's existence theorem to construct a stationary Markov

[5] See BCPT, p.142.

process $\{Z_n\}_{-\infty<n<\infty}$ having the same transition probability p and the same invariant probability π.

Definition 8.6. A Markov process with a transition probability p and an invariant probability π is said to be *time-reversible* if the stationary Markov process $\{Z_n\}_{-\infty<n<\infty}$, with this transition probability and this invariant distribution, has the same distribution as the *time-reversed* process $\{Y_n\}_{-\infty<n<\infty}$, where $Y_n := Z_{-n}$ $(-\infty < n < \infty)$. We refer to $\{Z_n\}_{-\infty<n<\infty}$ as the *double-sided version.*

In the context of the "movie metaphor," the statistics of a stationary data stream does not depend on when viewing begins, while a time-reversible data stream sequence is the same whether it is viewed forward or backward.

For the propositions below, assume that the transition probability $p(x, dy)$ has a density $p(x, y)$ with respect to a σ-finite measure μ on $(S, \mathcal{S})$, with $(x, y) \to p(x, y)$ measurable (on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ into $([0, \infty), \mathcal{B}_{[0,\infty)})$).

Proposition 8.8 *(Detailed Balance Condition).* Let $\pi(dy)$ be a probability measure on $(S, \mathcal{S})$ with a density $\pi(y)$ with respect to μ. (a) If

$$\pi(x)p(x, y) = \pi(y)p(y, x) \quad \text{a.e. } (\mu \times \mu), \tag{8.25}$$

then π is a time-reversible invariant probability for the Markov process. (b) For a Markov process with transition probability density $p(x, y)$ and invariant probability density $\pi(y)$, (8.25) is necessary for the process to be time-reversible.

Proof.

(a) Let p and π satisfy (8.25). Then for every Borel measurable f,

$$\begin{aligned}\int_S Tf(x)\pi(dx) &= \int_S Tf(x)\pi(x)\mu(dx) = \int_S \left(\int_S f(y)p(x, y)\mu(dy)\right)\pi(x)\mu(dx)\\ &= \int_S\int_S f(y)p(x, y)\pi(x)\mu(dy)\mu(dx)\\ &= \int_S \left(\int_S f(y)p(y, x)\mu(dx)\right)\pi(y)\mu(dy)\\ &= \int_S f(y)\pi(y)\mu(dy) = \int_S f(y)\pi(dy),\end{aligned}$$

implying π is invariant. Let $\{X_n\}_{n\geq 0}$ be a stationary Markov process with transition probability p and invariant initial distribution π. Then, by the Markov property, the joint density of $(X_n, X_{n+1}, \dots, X_{n+k})$, with respect to $\mu \times \cdots \times \mu$, at $(y_0, y_1, \dots, y_k) \in S^{k+1}$ is

$$g(y_0, y_1, \dots, y_k) := \pi(y_0)p(y_0, y_1)p(y_1, y_2)\dots p(y_{k-1}, y_k), \tag{8.26}$$

while the joint density of $(X_{n+k}, X_{n+k-1}, \dots, X_n)$, at the same point $(y_0, y_1, \dots, y_k) \in S^{k+1}$, is

$$\begin{aligned}
h(y_0, y_1, \ldots, y_k) &:= g(y_k, y_{k-1}, \ldots, y_1, y_0) \\
&= \pi(y_k) p(y_k, y_{k-1}) p(y_{k-1}, y_{k-2}) \cdots p(y_1, y_0) \\
&= \pi(y_{k-1}) p(y_{k-1}, y_k) p(y_{k-1}, y_{k-2}) \cdots p(y_1, y_0) \\
&= p(y_{k-1}, y_k) \pi(y_{k-1}) p(y_{k-1}, y_{k-2}) \cdots p(y_1, y_0) \\
&= p(y_{k-1}, y_k) \pi(y_{k-2}) p(y_{k-2}, y_{k-1}) \cdots p(y_1, y_0) \\
&\vdots \\
&= p(y_{k-1}, y_k) p(y_{k-2}, y_{k-1}) \cdots p(y_1, y_2) \pi(y_1) p(y_1, y_0) \\
&= p(y_{k-1}, y_k) p(y_{k-2}, y_{k-1}) \cdots p(y_1, y_2) \pi(y_0) p(y_0, y_1) \\
&= g(y_0, y_1, \ldots, y_k).
\end{aligned}$$

Since this is true for all $k \geq 1$, the finite dimensional distributions of the double-sided version $\{Z_n\}_{-\infty<n<\infty}$ and $\{Y_n\}_{-\infty<n<\infty}$ with $Y_n := Z_{-n}$ $(-\infty < n < \infty)$, described in Definition 8.6, coincide. Thus, using the $\pi - \lambda$ theorem,[6] it follows that the process $\{Z_n\}_{-\infty<n<\infty}$ and its time-reversal $\{Y_n\}_{-\infty<n<\infty}$ have the same distribution.

(b) For the stationary Markov process $\{Z_n\}_{-\infty<n<\infty}$ to be time-reversible, it is necessary that the distribution of (Z_0, Z_1) is the same as that of $(Y_0, Y_1) \equiv (Z_0, Z_{-1})$. But the latter has the same distribution as (Z_1, Z_0). The left side of (8.25) is the p.d.f. of (Z_0, Z_1) at (x, y), while the right side is the p.d.f. of (Z_1, Z_0) at (x, y). Thus for time-reversibility, (8.25) must hold. ■

If π is an invariant probability, then by Jensen's inequality,

$$\int_S \Big(\int_S |f(y)| p(x, dy)\Big)^2 \pi(dx) \leq \int_S \int_S f^2(y) p(x, dy) \pi(dx) = \int_S f^2(y) \pi(dy),$$

so that one may extend the transition operator T to $L^2(S, \pi) \supseteq B(S)$. We will now show that, in analytical terms, time-reversibility of a Markov process means that the transition operator T is *self-adjoint* on $L^2(S, \pi)$, i.e.,

$$\langle Tf, g\rangle = \langle f, Tg\rangle \qquad \text{for all } f, g \in L^2(S, \pi). \tag{8.27}$$

Here $\langle \quad \rangle$ is the *inner product* on the Hilbert space $L^2(S, \pi)$,

$$\langle g, h\rangle = \int_S g(y) h(y) \pi(dy).$$

We will denote by $\|g\|$ the L^2–*norm:* $\|g\|^2 = \langle g, g\rangle$.

Proposition 8.9. Let π be an invariant probability of a Markov process. (a) The transition operator T is a contraction on $L^2(S, \pi)$. (b) If π is a time-reversible invariant probability, then T is self-adjoint.

[6] See BCPT, p. 4.

Proof.

(a) This is proved in (8.28).

(b) Let $f, g \in L^2(S, \pi)$, and assume π is time-reversible in the sense of Definition 8.6. Then, conditioning on X_0, one has

$$\begin{aligned}\langle Tf, g\rangle &= \int_S \left(\int_S f(y)p(x, dy)\right) g(x)\mu(dx)\\ &= \mathbb{E}\big(\mathbb{E}(f(X_1)|\sigma(X_0))g(X_0)\big) = \mathbb{E}f(X_1)g(X_0)\\ &= \mathbb{E}f(X_0)g(X_1) = \langle f, Tg\rangle.\end{aligned}$$

■

Recall in the case of finite S that, given an initial distribution μ for X_0, the distribution μ_1 of X_1 may be obtained by the transformation $\mu \to \mu_1 = \mathbf{p}'\mu$, where $\mathbf{p}'$ is the transpose matrix, see (7.12). More generally, one may define an adjoint operator as follows.

Definition 8.7. Given a transition probability $p(x, dy)$ on $(S, \mathcal{S})$, the *adjoint* linear operator T^* is defined on the linear space $\mathcal{M}(S)$ of all finite signed measures on $(S, \mathcal{S})$ by

$$(T^*\mu)(B) = \int_S p(x, B)\mu(dx) \qquad (B \in \mathcal{S}, \mu \in \mathcal{M}(S)). \tag{8.28}$$

In general, if μ is a probability measure, then $T^*\mu$ is the distribution of X_1 where X_0 has distribution μ. In particular, π is an invariant probability if and only if

$$T^*\pi = \pi. \tag{8.29}$$

To see the connection between the $L^2(S, \pi)$-adjoint of T and this more general operator, then, irrespective of (8.27), identify $f \in L^2(S, \pi)$ with the signed measure $fd\pi$, and note that $T^*(fd\pi)(dy)$ is given by

$$\begin{aligned}\int_S g(y)T^*(fd\pi)(dy) &= \int_S\int_S g(y)p(x, dy)f(x)\pi(dx) = \int_S Tg(x)f(x)\pi(dx)\\ &= \langle Tg, f\rangle = \langle g, T^*f\rangle, g \in L^2(S, \pi),\end{aligned} \tag{8.30}$$

where, by an obvious abuse of notation, $T^*f \in L^2(S, \pi)$ is given by the $L^2(S, \pi)$-adjoint operator to T. In the interpretation of T^* as an operator on $L^2(S, \pi)$, 1 is an *eigenvalue* of T^* with the constant *eigenvector* $f(\cdot) \equiv 1$. For the adjoint operator on $\mathcal{M}(S)$, this eigenvector corresponds to the invariant measure $\pi(dy) = 1 \cdot \pi(dy)$.

Define $T^{*n}\mu = T^*(T^{*n-1}\mu)$ iteratively on $\mathcal{M}(S)$. In the case μ is a probability measure, and X_0 has distribution μ, $T^*\mu$ is the distribution of X_1, $T^{*2}\mu$ is the distribution of $X_2, \ldots, T^{*n}\mu$ is the distribution of X_n. Thus, whereas iterates of the transition operator T govern the "evolution of states" via $T^n f(x) = \mathbb{E}_x f(X_n)$, the iterates of the adjoint T^* govern the "evolution of probability distributions" of the Markov process.

Now, T^{*n} is a linear operator on $\mathcal{M}(S)$, as is T^n on $C_b(S)$. The term *adjoint operator* given to T^* (or, T^{*n}) is more fully justified in terms of the basic identities

$$\int_S Tf(x)\mu(dx) = \int_S f(x)(T^*\mu)(dx), \quad \int_S T^n f(x)\mu(dx) = \int_S f(x)(T^{*n}\mu)(dx). \tag{8.31}$$

The first equality in (8.31) follows from (8.28), first for simple functions f and then by approximating $f \in B(S)$ uniformly by simple functions. The second equality in (8.31) follows by induction on n. When μ is a probability measure, then the second equality says $\mathbb{E}_\mu[\mathbb{E}(f(X_n) \mid X_0)] = \mathbb{E}_\mu f(X_n)$, with X_0 having distribution μ.

Example 2. $S = \{0, 1\}$, $\mathbf{p} = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix}$, $0 \le a,\ b \le 1$. S is a metric space with the discrete metric $d(0,0) = d(1,1) = 0$, $d(1,0) = d(0,1) = 1$, $\mathcal{S}$ is the power set, and every function $f : S \to \mathbb{R}$ is a bounded, continuous function. By the computation at the end of the previous chapter, one has for $a + b < 2$, with $a + b \neq 0$,

$$\lim_{n\to\infty} \mathbf{p}^{(n)} = \begin{bmatrix} \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \\ \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \end{bmatrix}.$$

In particular, the invariant probability (vector) $\pi = (\pi_0, \pi_1)'$ is given by

$$\begin{aligned} \pi_0 &= \lim_{n\to\infty} p_{i0}^{(n)} = \tfrac{1-b}{2-a-b} \\ \pi_1 &= \lim_{n\to\infty} p_{i1}^{(n)} = \tfrac{1-a}{2-a-b}\ , \qquad i = 0, 1. \end{aligned}$$

Alternatively, one may determine π_0, π_1 from time-reversibility via the detailed balance and total probability one equations. In this case one could then use the $L^2(S, \pi)$ theory for self-adjoint operators to obtain convergence (Exercise 7).

In the cases $a + b = 2$ and $a + b = 0$, one has $\mathbf{p} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{p} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, respectively. In the former case every probability on $S = \{0, 1\}$ is an invariant probability, and hence there are infinitely many invariant probabilities. In the latter case there is a unique invariant probability given by $\pi_0 = \pi_1 = \frac{1}{2}$; but $\mathbf{p}^n$ does not have a limit, although (8.24) holds.

Example 3. Suppose $S = \mathbb{R}$ with Borel σ-field $\mathcal{S}$. Let $\varepsilon_1, \varepsilon_2, \ldots$ be an i.i.d. sequence of standard normal random variables and let $b \in \mathbb{R}$. Consider the sequence of random variables

$$X_{n+1} = bX_n + \varepsilon_{n+1}, \quad n = 0, 1, 2, \ldots .$$

Iterating the recursion, one has that

$$X_{n+1} = b^{n+1} X_0 + \sum_{j=0}^{n} b^j \varepsilon_{n+1-j}.$$

In particular, the m-step transition probability $p^{(m)}(x, dy)$ is given by the Gaussian distribution with mean $b^m x$ and variance $\sum_{j=0}^{m-1} b^{2j} = \frac{b^{2m}-1}{b^2-1}$ if $|b| \neq 1$. In the case $b = \pm 1$, $p^{(m)}(x, dy)$ is Gaussian with mean $(\pm 1)^m x$ and variance m. In any case $p(x, dy)$ clearly has the (weak) Feller property, and in particular, if $|b| < 1$, then $p^{(m)}(x, dy)$ converges (weakly) as $m \to \infty$ to the invariant distribution $\pi(dy) = \frac{1}{\sqrt{2\pi(1-b^2)^{-1}}} e^{-\frac{1-b^2}{2} y^2} dy$, $y \in S = (-\infty, \infty)$, possessing a Gaussian density with respect to Lebesgue measure. Note that for $|b| < 1$, if X is $N(0, \sigma^2)$ and Z is standard normal and independent of X, then $X =^{\text{dist}} bX + Z$ (equality in distribution) if and only if $\sigma^2 = (1 - b^2)^{-1}$. One may check that π is also a time-reversible invariant probability (Exercise 13).

Example 4 *(Random Walk on a Finite Graph).* A finite graph consists of a finite set $S = \{v_1, \ldots, v_k\}$ of k *vertices* together with a relation $\mathcal{E} \subset \{1, \ldots, k\} \times \{1, \ldots, k\}$, with the property that $(i, j) \in \mathcal{E}$ if and only if $(j, i) \in \mathcal{E}$, defining *edges* as follows: there is an edge e_{ij} connecting vertices v_i and v_j if and only if $(i, j) \in \mathcal{E}$, denoted by means of the obvious abuse of notation $e_{ij} \in \mathcal{E}$. The graph is said to be *connected* if for any pair of distinct vertices v_i, v_j there is a path of $m \geq 1$ edges $e_{ii_1}, e_{i_1 i_2}, \ldots e_{i_{m-1} i_m}$ with $i_m = j$. For a fixed vertex v_i, the integer $d_i := card\{j : e_{ij} \in \mathcal{E}\}$ is called the *degree* of v_i. A *random walk on a finite connected graph* $(S, \mathcal{E})$ may be defined as a Markov chain with state space S and transition probabilities given by $p_{v_i, v_j} = 1/d_i$ if and only if $e_{ij} \in \mathcal{E}$, else $p_{v_i, v_j} = 0$. It is straightforward to check that up to normalization, the vertex degrees define the unique time-reversible invariant probability for a random walk on a finite connected graph (see Exercise 12).

An approach to the construction of invariant probabilities, similar to that in Proposition 8.4 but which is valid without the Feller property, is given below.

Proposition 8.10. Let $p(x, dy)$ be a transition probability on a state space $(S, \mathcal{S})$.

a If for some $x \in S$ there is a sequence of integers $1 \leq n_1 < n_2 < \cdots$ such that, as $k \to \infty$,

$$\frac{1}{n_k} \sum_{r=0}^{n_k - 1} p^{(r)}(x, B) \longrightarrow \pi_x(B) \quad \text{for all } B \in \mathcal{S} \tag{8.32}$$

for some probability measure π_x on $(S, \mathcal{S})$, then π_x is an invariant probability for p.

b If, for some sequence $\{n_k : k \geq 1\}$, (8.32) holds for every $x \in S$ with the same limit $\pi_x = \pi$ for all x, then π is the unique invariant probability for p.

Proof. The proof is essentially the same as that of Proposition 8.4, with $C_b(S)$ replaced by $B(S)$—the space of real-valued bounded measurable functions on S. Note that (8.32) implies

$$\frac{1}{n_k}\sum_{r=0}^{n_k-1}(T^r f)(x) \longrightarrow \int_S f(y)\pi_x(dy) \quad \text{for all } f \in B(S). \qquad \blacksquare$$

To close this chapter, let us note that a large class of examples of Markov chains occur as functions of a given, perhaps more primitive Markov chain. Of course, one-to-one functions are simply relabeling of the states and do not affect the dependence structure. A more general class of functions can be obtained as follows[7].

Definition 8.8. A measurable function φ on (S, $\mathcal{S}$) to a measurable space (S′, $\mathcal{S}'$), is said to be an invariant function of a group G of transformations on S if (i) $\varphi(gx) = \varphi(x)$ for all $g \in G$, $x \in S$. If, in addition, (ii) every measurable invariant function is a measurable function of φ, then φ is said to be a maximal invariant.

Example 5.

1. For each $x \in S$, the *orbit* of x under G is defined by $o(x) = \{gx : g \in G\}$. Note that each invariant function is constant on orbits. Let S and S' be metric spaces and $\varphi : S \to S'$ a measurable surjection such that (a) φ is constant on orbits, (b) $\varphi(x) \neq \varphi(y)$ if $o(x) \neq o(y)$, i.e., φ is a relabeling of $o(x)$, and S' may be viewed as a relabeling of the space of orbits. Then φ is a maximal invariant since (i) invariance is obvious by (a), and (ii) if $\varphi(x) = \varphi(y)$, then $\rho(x) = \rho(y)$ for any invariant function ρ by (b), i.e., ρ is a function of φ.
2. $\varphi(x) = |x|$ is a maximal invariant of the reflection group $\{e, -e\}$, where $ex = x$, $(-e)x = -x$, $x \in S = \mathbb{R}$, $S' = [0, \infty)$.

Proposition 8.11. Suppose $X = \{X_n\}$ is a Markov process on S whose transition probabilities are invariant under the group G of transformations from S to S', i.e.,

$$p(gx, g(B)) = p(x, B), \quad \forall\, x \in S, g \in G, B \in \mathcal{S}.$$

If φ is a maximal invariant, then $\{\varphi(X_n)\}$ is Markov.

Proof. Take conditional expectations with respect to the larger $\sigma(X_m : m \leq n)$, followed by the smaller $\sigma(\varphi(X_m) : m \leq n)$, $n = 1, 2, \ldots$, to get

$$P(\varphi(X_{n+1}) \in B|\varphi(X_m), m \leq n) = \mathbb{E}\{p(X_n, \varphi^{-1}(B))|\varphi(X_m), m \leq n\}. \tag{8.33}$$

By invariance of φ, $\varphi \circ g = \varphi$, and one has from (i),

[7] The continuous parameter version of this result is given in Bhattacharya and Waymire (1990, 2009), pp. 502–503, for non-injective functions of a Markov process. Also see Bhattacharya and Waymire (1990).

$$\begin{aligned} p(x,\varphi^{-1}(B)) &= p(g^{-1}x, g^{-1}(\varphi^{-1}(B))) \\ &= p(g^{-1}x, \varphi^{-1}(B)). \end{aligned} \tag{8.34}$$

That is, the function $x \to p(x,\varphi^{-1}(B))$ is invariant. By (ii), therefore, it is a function $q(\varphi(x), B)$, say, of φ. Thus,

$$\begin{aligned} P(\varphi(X_{n+1}) \in B|\varphi(X_m), m \le n) &= \mathbb{E}\{p(X_n, \varphi^{-1}(B))|\varphi(X_m), m \le n\} \\ &= \mathbb{E}\{q(\varphi(X_n), B)|\varphi(X_m), m \le n\} \\ &= q(\varphi(X_n), B). \end{aligned} \tag{8.35}$$

Thus, $\{\varphi(X_n)\}$ is Markov with one-step transition probabilities $q(\varphi(x), B)$. ■

Example 6 *(Reflecting Simple Symmetric Random Walk).* Consider the unrestricted simple symmetric random walk on $\mathbb{Z}$ starting at the origin, defined by $S_n := \sum_{j=1}^n X_j, n \ge 1,\ S_0 = 0$, where the displacements $X_n : n \ge 1$ are i.i.d. ±1-valued symmetric Bernoulli random variables. Then, since $p_{ij} = \frac{1}{2}\delta_{i-1}(j) + \frac{1}{2}\delta_{i+1}(j), i, j \in \mathbb{Z}$ is invariant under the reflection group $G = \{e, -e\}$ where $e(i) = i, i \in \mathbb{Z}$, it follows that $\{R_n := |S_n|\}_n$ is a Markov chain. (Also see Example 11.1.)

Exercises

1. Prove that for a transition probability p the measurability $x \to p(x,B)$ for all $B \in \mathcal{S}$ implies that $x \to Tf(x)$ is measurable for all $f \in \mathbb{B}(S)$. Show that (i) T is a linear operator on $\mathbb{B}(S)$, (ii) $||Tf|| \le ||f||, f \in \mathbb{B}(S), ||f|| = \sup_{x\in S} |f(x)|$, (iii)$T\mathbf{1} = \mathbf{1}$, where $\mathbf{1}(x) = 1,$ for all $x \in S$, and (iv) $Tf \ge 0$ on S if $f \in \mathbb{B}(S)$ is a nonnegative function.
2. Let $p(x, dy)$ be a transition probability on $(S, \mathcal{S})$.
   (a) Show that $x \to P_x(B)$ is $\mathcal{S}$-measurable for all $B \in \mathcal{S}^{\otimes\infty}$, and letting $\mathbf{X}$ denote the identity map on S^∞, the function $y \to \mathbb{E}_y f(\mathbf{X})$ is $\mathcal{S}$-measurable for all bounded measurable $f : S^\infty \to \mathbb{R}$.
   (b) For every bounded $\mathcal{S}^{\otimes n}$-measurable function f on S^n, show that the function $(x_0, x_1, \ldots, x_{n-1}) \to \int_S f(x_0, x_1, \ldots, x_{n-1}, y)p(x_{n-1}, dy)$ is $\mathcal{S}^{\otimes n}$-measurable.
3. Express time-reversibility and detailed balance without requiring densities. [*Hint*: Detailed balance may be stated as $\pi(dx)p(x,dy) = \pi(dy)p(y,dx)$, suitably interpreted, and the consequent time-reversibility compares the joint distribution of $(X_0, \ldots, X_k)$ with that of $(X_k, X_{k-1}, \ldots, X_0)$.]
4. Prove that $T^n f(x) = \int_S f(y)p^{(n)}(x,dy) = \mathbb{E}(f(X_n)|X_0 = x), x \in S$, for all bounded, measurable functions f on S, and $n \ge 1$.

5. Show that the transition probability p in Example 1 (i) is not Feller continuous, and (ii) does not satisfy the hypothesis of Proposition 8.10. [*Hint*: (i) has the relative topology of $[0, 1]$ so that all points $m/(m + 1), m = 0, 1, \ldots,$ are isolated and only 1 is a point of accumulation. Hence a real-valued function on S is continuous if and only if $f(\frac{m}{m+1}) \to f(1)$ as $m \to \infty$. For (ii), $p^{(n)}(x, dy) \to \delta(dy)$ weakly as $n \to \infty$ for all $x \in S$.]
6. Let S be a locally compact separable metric space. Suppose that T is a positive linear contraction operator on $C_b(S)$ with $T1 = 1$. Use the Riesz Representation Theorem to show that T uniquely determines Feller transition probabilities $p(x, dy)$ such that $Tf(x) = \int_S f(y)p(x, dy)$.
7. In the case of the two-state Markov chain in Example 2, use time-reversibility to compute the invariant probability π and establish convergence by an appeal to the spectral theorem for self-adjoint linear operators on $L^2(S, \pi)$.
8. Let $X_0, X_n := X_0 + Z_1 + \cdots + Z_n$ $(n \geq 1)$ be a (general) random walk $S = \mathbb{R}^k$ with step size distribution Q, i.e., $\{Z_n : n \geq 1\}$ is an i.i.d. sequence with common distribution Q on $(\mathbb{R}^k, \mathcal{B}^k)$, independent of X_0. Show that (i) $\{X_n : n \geq 0\}$ is Markov and (ii) no invariant probability exists if $Q \neq \delta_{\{0\}}$.
9. (*Birth–Death Chain with Two Reflecting Boundaries*) Let $S = \{0, 1, 2, \ldots, d\}$ $(d > 1)$, $p(x, x + 1) = \beta_x$, $p(x, x - 1) = \delta_x \equiv 1 - \beta_x$, with $0 < \beta_x < 1$ $(x = 1, 2, \ldots, d - 1)$, $\beta_0 \equiv p(0, 1) = 1$, $\delta_d \equiv p(d, d - 1) = 1$. Prove that there exists a unique invariant probability π, and the Markov process with this initial distribution is time-reversible. [*Hint*: There is a unique probability π for which (8.25) holds (with μ as counting measure). To solve for π, note that (8.25) implies $\pi(x + 1)/\pi(x) = \beta_x/\delta_{x+1}$ $(x = 0, 1, \ldots, d - 1)$. Check that this must be true of the ratios for any invariant probability.]
10. (*Time-Reversed Stationary Markov Process*) Let $\{X_n : n \geq 0\}$ be a stationary Markov process on $(S, \mathcal{S})$ with a transition probability density $p(x, y)$ (w.r.t. a σ-finite measure μ) and an invariant probability density $\pi(y)$, $y \in S$. Let $\{Z_n : -\infty < n < \infty\}$ be a stationary Markov process such that $\{Z_n : n \geq 0\}$ has the same distribution as $\{X_n : n \geq 0\}$. Show that $\{Y_n := Z_{-n} : n \in \mathbb{Z}\}$ is (i) stationary and (ii) Markov with the transition probability density $q(x, y) := (\pi(y)/\pi(x))p(y, x)$; note that this is simply "Bayes formula" for (X_n, X_{n+1}). [*Hint*: Check that π is an invariant probability for q, and then construct a stationary double-sided Markov process $\{R_n\}_{-\infty<n<\infty}$ with these transition probabilities and invariant probability. Then check that the processes $\{Y_n : n \in \mathbb{Z}\}$ and $\{R_n : n \in \mathbb{Z}\}$ have the same distribution by considering finite dimensional events.]
11. Let $\{X_n : n \geq 0\}$ be a Markov chain on the countable state space S. Assume that for any $i, j \in S$, one has $p_{ij}^{(n)} > 0$ for some $n \geq 1$. We say that $p = ((p_{ij}))$ is *irreducible* in this case. Define $Y_n = (X_n, X_{n+1})$, $n = 0, 1, 2, \ldots$.

    (a) Show that $\{Y_n : n \geq 0\}$ is a Markov chain on $S' = \{(i, j) \in S \times S : p_{ij} > 0\}$.
    (b) Show that if $\{X_n : n \geq 0\}$ is irreducible, then so is $\{Y_n : n \geq 0\}$.

(c) Show that if $\{X_n : n \geq 0\}$ has invariant distribution $\boldsymbol{\pi} = (\pi_i)$, then $\{Y_n : n \geq 0\}$ has invariant distribution $(\pi_i p_{ij})$.

(d) Show that an irreducible Markov chain on a state space S with an invariant initial distribution $\boldsymbol{\pi}$ is time-reversible if and only if (*Kolmogorov Condition*):

$$p_{ii_1} p_{i_1 i_2} \cdots p_{i_k i} = p_{ii_k} p_{i_k i_{k-1}} \cdots p_{i_1 i} \qquad \text{for all } i, i_1, \ldots, i_k \in S, \quad k \geq 1.$$

(e) If there is a $j \in S$ such that $p_{ij} > 0$ for all $i \neq j$ in (d), then for time-reversibility it is both necessary and sufficient that $p_{ij} p_{jk} p_{ki} = p_{ik} p_{kj} p_{ji}$ for all i, j, k.

12. (*A General Finite State Space Graph*) Let $\{X_n : n \geq 0\}$ be an *irreducible* Markov chain on a finite state space S; i.e., for each $i, j \in S$, there is an $n \geq 1$ such that $p_{ij}^{(n)} > 0$. Define a *graph G* having states of S as vertices with edges joining i and j if and only if either $p_{ij} > 0$ or $p_{ji} > 0$.

(a) Show that G is connected; i.e., for any two sites i and j, there is a path of edges from i to j.

(b) Show that if $\{X_n : n \geq 0\}$ has an invariant distribution $\boldsymbol{\pi}$, then for any $A \subset S$,

$$\sum_{i \in A} \sum_{j \in S \setminus A} \pi_i p_{ij} = \sum_{i \in A} \sum_{j \in S \setminus A} \pi_j p_{ji}$$

i.e., the net probability flux across a *cut* of S into complementary subsets $A, S \setminus A$ is in balance. [*Hint*: Notice that $\sum_{i \in A} \sum_{j \in S} \pi_i p_{ij} = \sum_{i \in A} \sum_{j \in S} \pi_i p_{ji}$.]

(c) Show that if G contains no cycles of three or more vertices, i.e., $m = 3$ or more distinct vertices $v_1, \ldots, v_m$ such that v_i and v_{i+1} are joined by an edge for $i = 1, \ldots, m$ and $v_{m+1} = v_1$, then the process is time-reversible started with π. A connected graph without cycles is called a *tree graph*. [*Hint*: Proceed inductively on the number of states.]

(d) Give a graphical proof that an invariant probability for a birth–death Markov chain on $\{0, 1, \ldots, N\}$ with reflecting boundaries at 0, N must be time-reversible.

13. Prove the time-reversibility of Example 3 when $|b| < 1$ and $\{\epsilon_n : n \geq 1\}$ is an i.i.d. standard normal sequence.

14. Consider a Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ on a metric space S defined recursively by $X_{n+1} = g(X_n, \epsilon_{n+1})$, $n \geq 0$, where (i) $\{\epsilon_n : n \geq 1\}$ is a sequence of i.i.d. random variables with values in a metric space U, and independent of X_0, and (ii) $g : S \times U \to S$ is continuous. Show that the Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ has the Feller property.

Chapter 9
Stopping Times and the Strong Markov Property

Given a stopping time τ, the Markov property for discrete parameter Markov processes is extended to the conditional distribution of the process "after" time τ given the σ-field generated by the process up to time τ. This is referred to as a strong Markov property.

One of the most useful general properties of discrete time Markov processes is that the Markov property holds even when the "past" is given up to certain types of random times. Indeed, we have tacitly used it in proving that the simple symmetric random walk reaches every state infinitely often with probability 1. This argument is more generally revisited below in Example 1.

A class of special random times, called *stopping times* or *Markov times*, may be defined as follows. Let $\{X_n : n = 0, 1, 2, \ldots\}$ be a stochastic process having state space S and defined on some probability space $(\Omega, \mathcal{F}, P)$. A positive (possibly infinite) integer-valued random variable τ is a stopping time if and only if $[\tau \leq n] \in \sigma(\{X_0, \ldots, X_n\}), n = 0, 1, \ldots$. Intuitively, whether or not to stop by time $\tau = m$ can be decided by observing the stochastic process up to time m. For an example, consider the first time τ_B the process $\{X_n : n \geq 0\}$ reaches B $(\in \mathcal{S})$, defined by

$$\tau_B(\omega) = \inf\{n \geq 0 : X_n(\omega) \in B\}. \tag{9.1}$$

If ω is such that $X_n(\omega) \notin B$ whatever be n (i.e., if the process never reaches B), then take $\tau_B(\omega) = \infty$. Observe that

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_9

$$[\tau_B \leq m] := \{\omega : \tau_B(\omega) \leq m\} = \bigcup_{n=0}^{m} \{\omega : X_n(\omega) \in B\}. \tag{9.2}$$

Thus $[\tau_B \leq m] \in \mathcal{F}_m := \sigma\{X_0, \ldots, X_m\}$, $m \geq 0$. Hence τ_B is a stopping time, as are the *rth return times* $\tau_B^{(r)}$ *to* B defined recursively by

$$\begin{aligned} &\tau_B^{(1)}(\omega) = \inf\{n \geq 1 : X_n(\omega) \in B\}, \\ &\tau_B^{(r)}(\omega) = \inf\{n > \tau_B^{(r-1)}(\omega) : X_n(\omega) \in B\}, \end{aligned} \qquad \text{for } r = 2, 3, \ldots \tag{9.3}$$

(Exercise 1). Once again, the infimum over an empty set is to be taken as ∞. If B is a singleton $\{y\}$ we will often write τ_y for $\tau_{\{y\}}$, and $\tau_y^{(r)}$ instead of $\tau_{\{y\}}^{(r)}$.

In view of Proposition 8.3, the Markov property may be expressed that given the "past" and "present" $\mathcal{F}_m := \sigma\{X_0, \ldots, X_m\}$ up to time m, the conditional distribution of the "after-m" stochastic process $X_m^+ = \{(X_m^+)_n : n \geq 0\} := \{X_{m+n} : n = 0, 1, \ldots\}$ is P_{X_m}. In other words, if the process is re-indexed after time m with $m + n$ being regarded as time n, then this stochastic process is conditionally distributed as a Markov chain having transition probability $p(x, dy)$ and initial state X_m.

Suppose now that τ is a stopping time. Given the "past up to time τ" means given the values of τ and $X_0, X_1, \ldots, X_\tau$; that is, conditionally given the pre-τ σ-field $\mathcal{F}_\tau$ defined by the collection of events $G \in \sigma\{X_n : n \geq 0\}$ such that

$$G \cap [\tau = m] \in \mathcal{F}_m \quad \text{for all } \ 0 \leq m < \infty. \tag{9.4}$$

Equivalently for discrete parameter processes, this is the σ-field generated by the *stopped process,* i.e., $\mathcal{F}_\tau = \sigma(X_{\tau \wedge n} : n = 0, 1, 2, \ldots)$, (see Exercise 2).

By the *after-τ process* we now mean the stochastic process

$$X_\tau^+ = \{(X_\tau^+)_n = X_{\tau+n} : n = 0, 1, 2, \ldots\},$$

which is well defined only on the set $[\tau < \infty]$. Observe that (9.4) is equivalent to

$$G \cap [\tau \leq m] \in \mathcal{F}_m \qquad \text{for all } 0 \leq m < \infty. \tag{9.5}$$

The use of the natural filtration $\mathcal{F}_m = \sigma\{X_0, \ldots, X_m\}, m = 0, 1, 2, \ldots$ in defining stopping times above can be easily generalized as follows.

Definition 9.1. Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \mathcal{F}_m \subset \cdots, m = 0, 1, 2, \ldots$ be an arbitrary filtration of a probability space $(\Omega, \mathcal{F}, P)$. An extended real-valued random variable $\tau : \Omega \to [0, \infty]$ is called a $\{\mathcal{F}_m : m \geq 0\}$-stopping time if $[\tau \leq m] \in \mathcal{F}_m$ for each $m = 0, 1, \ldots$. A Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ on a state space $(S, \mathcal{S})$ is said to be $\{\mathcal{F}_n\}$-adapted if (i) X_n is $\mathcal{F}_n$-measurable for all n, and (ii) the conditional distribution of the after-n process X_n^+, given $\mathcal{F}_m$ is P_{X_m}.

Definition 9.2. A $\{\mathcal{F}_n\}$-adapted Markov process $\{X_n : n \geq 0\}$ has the *strong Markov property* if for every $\{\mathcal{F}_n\}_{n=0}^{\infty}$–stopping time τ the conditional distribution of the after-τ process X_τ^+, given the pre-τ σ-field $\mathcal{F}_\tau$, is P_{X_τ} on the set $[\tau < \infty]$. That is, for every $C \in \mathcal{S}^{\otimes\infty}$ and $G \in \mathcal{F}_\tau$,

$$P([X_\tau^+ \in C] \cap G \cap [\tau < \infty]) = \mathbb{E}(\mathbf{1}_{G\cap[\tau<\infty]} P_{X_\tau}(C)). \tag{9.6}$$

Theorem 9.1. Every discrete parameter $\{\mathcal{F}_n\}$-adapted Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ has the strong Markov property.

Proof. Choose and fix a positive integer k along with k time points $0 \leq m_1 < m_2 < \cdots < m_k$, and $B_1, \ldots, B_k \in \mathcal{S}$. Let $G \in \mathcal{F}_\tau$. Then,

$$\begin{aligned}
&P([(X_\tau^+)_{m_i} \in B_i, 1 \leq i \leq k] \cap G \cap [\tau < \infty]) \\
&= \sum_{m=0}^{\infty} P([X_{m+m_i} \in B_i, 1 \leq i \leq k] \cap G \cap [\tau = m]) \\
&= \sum_{m=0}^{\infty} \mathbb{E}\,\mathbb{E}(\mathbf{1}(G \cap [\tau = m])\mathbf{1}([X_{m+m_i} \in B_i, 1 \leq i \leq k])|\mathcal{F}_m) \\
&= \sum_{m=0}^{\infty} \mathbb{E}\,\big(\mathbf{1}(G \cap [\tau = m])\mathbb{E}(\mathbf{1}([X_{m+m_i} \in B_i, 1 \leq i \leq k])|\mathcal{F}_m)\big) \\
&= \sum_{m=0}^{\infty} \mathbb{E}\,\{\mathbf{1}(G \cap [\tau = m])h(X_m)\} = \mathbb{E}\{\mathbf{1}(G \cap [\tau < \infty])h(X_\tau)\},
\end{aligned} \tag{9.7}$$

where $h(x) = P_x(X_{m_i} \in B_i, 1 \leq i \leq k)$. The desired Theorem 9.1 follows from (9.7) by the $\pi - \lambda$ theorem. ■

Remark 9.1. We will sometimes omit the term $\{\mathcal{F}_n\}$-adapted for a Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ if $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$, or if the context makes the filtration clear. There are many examples where the natural and more convenient filtration is larger than that defined by the sequence itself. This is especially true for function of Markov chains such as (a) $\{|S_n|\}$, the modulus of simple symmetric random walk (Exercise 6, Chapter 7), or (b) the residual life $\{R_n\}$ in a renewal process for i.i.d. nonnegative integer valued random variables considered in Proposition 8.4, Chapter 8, Bhattacharya and Waymire (2021).

Example 1. Consider the case of the simple symmetric random walk on $\mathbb{Z}$ defined by $X_n = X_0 + Z_1 + \cdots + Z_n, n \geq 1$, where $Z_1, Z_2, \ldots$ are i.i.d. symmetrically distributed ±1 random variables, and X_0 is an integer-valued random variable independent of $Z_1, Z_2, \ldots$. One wishes to prove that $P(\tau_y < \infty) = 1$ for $y \in \mathbb{Z}$. This may be obtained from the (ordinary) Markov property applied to $\varphi(x) := P(\tau_y < \tau_a|X_0 = x), a \leq x \leq y$, by conditioning on $\sigma(X_0, X_1)$ as follows: For $a < x < y$, conditioning on $\sigma(X_0, X_1)$, i.e., by $\sigma(S_1^x)$,

$$\begin{aligned}\varphi(x) &= P([X_1^+ \text{ reaches } y \text{ before reaching } a] \cap [X_1 = x+1] | X_0 = x)\\ &\quad + P([X_1^+ \text{ reaches } y \text{ before reaching } a] \cap [X_1 = x-1] | X_0 = x)\\ &= \frac{1}{2}\varphi(x+1) + \frac{1}{2}\varphi(x-1) \qquad (9.8)\end{aligned}$$

with boundary values $\varphi(y) = 1, \varphi(a) = 0$. Solving one obtains $\varphi(x) = (x - a)/(y-a)$, for all $x < y$. Thus $P(\tau_y < \infty | X_0 = x) = 1$ follows by letting $a \to -\infty$. Similarly, or by symmetry, $P(\tau_y < \infty | X_0 = x) = 1$ for all $x > y$. If $x = y$, then $\tau_y^{(1)} = 1 + \tau_y$ and, noting that $[\tau_y^{(1)} < \infty] = [\tau_y(X_1^+) < \infty]$, condition on $\sigma(X_0, X_1)$ to get (by Proposition 8.1)

$$\begin{aligned}P_x(\tau_y^{(1)} < \infty) &= \mathbb{E}_x P_{X_1} \tau_y^{(1)} < \infty)\\ &= \frac{1}{2}P_{x-1}(\tau_y < \infty) + \frac{1}{2}P_{x+1}(\tau_y < \infty) = 1.\end{aligned}$$

Since $\tau_y = \tau_y^{(1)}$ for $x \neq y$, we have shown that $P(\tau_y^{(1)} < \infty | X_0 = x) = 1,$ for all x. While this calculation only required the Markov property, next consider the problem of showing that the process will return to y infinitely often. One would like to argue that conditioning on the process up to its return to y, it merely starts over. This of course is the strong Markov property. So let us examine carefully the calculation to show that $\tau_y^{(r)} < \infty$ a.s. for every $r = 1, 2, \ldots$. Now let $x \neq y$. Then $\tau_y = \tau_y^{(1)}$, and $\tau_y^{(1)} := \inf\{n \geq 1 : (X_{\tau_y}^+)_n = y\}$, the first return time to y of the process $X_{\tau_y}^+$, one has

$$P_x(\tau_y^{(2)} < \infty)\mathbb{E}_x P_{X_{\tau_y^{(1)}}}(\tau_y^{(1)} < \infty) = \mathbb{E}_x P_y(\tau_y^{(1)} < \infty) = 1. \qquad (9.9)$$

The second equality uses Theorem 9.1, the last equality follows from (9.8). Now this argument remains valid if one replaces $\tau_y^{(1)}$ by $\tau_y^{(r-1)}$ and $\tau_y^{(2)}$ by $\tau_y^{(r)}$ and assumes that $\tau_y^{(r-1)} < \infty$ almost surely. Hence, by induction, $P(\tau_y^{(r)} < \infty | X_0 = x) = 1$ for all positive integers r. This is equivalent to the recurrence of the state y in the sense that

$$P(X_n = y \text{ for infinitely many } n | X_0 = x) = P(\cap_{r=1}^{\infty} [\tau_y^{(r)} < \infty] | X_0 = x) = 1.$$

The importance of the strong Markov property will be amply demonstrated throughout the text.

Exercises

1. (*r-th Passage Time*)

   (a) Let τ and η be stopping times with respect to a filtration $\mathcal{F}_n, n \geq 0$. Which of the following are stopping times? (i) $\tau \vee \eta = \max\{\tau, \eta\}$, (ii) $\tau \wedge \eta = \min\{\tau, \eta\}$, (iii) $\tau + \eta$, (iv) τ^2, (v) $\tau^{\frac{1}{2}}$.
   (b) Show that for any $r \geq 1$, the r-th passage time to a set $B \in \mathcal{S}$ defined by (9.3) is a stopping time.

2. (*Stopped Process*) Suppose that $\{X_n : n \geq 0\}$ is a discrete parameter Markov process and τ is a stopping time. Let $\{Y_n = X_{\tau \wedge n} : n \geq 0\}$ denote the stopped process, i.e., $Y_n = X_n, n \leq \tau$, and $Y_n = X_\tau, n \geq \tau$.

   (a) Let $\tau \equiv \tau_B$ be the *first passage* (or *hitting time*) of $B \in \mathcal{S}$ defined at (9.1). Then, τ is a stopping time (recall (9.2)). Show that $\{Y_n : n \geq 0\}$ is a (homogeneous) Markov process.
   (b) Give an example to show that, in general, the stopped process is *not* a Markov process. [*Hint*: Consider the time τ of the second visit to b in a two–state Markov chain on $S = \{a, b\}$.]

3. For the stopped process defined in the preceding Exercise 2, show that $\mathcal{F}_\tau = \sigma(Y_0, Y_1, \dots)$, where $\mathcal{F}_n = \sigma(X_0, \dots, X_n), n \geq 1$.
4. A balanced six-sided die is rolled repeatedly. Let Z denote the smallest number of rolls for the occurrence of all six possible faces. Let Z_j = smallest number of tosses to obtain the jth new face after $j - 1$ distinct faces have occurred. Then $Z = Z_1 + \cdots + Z_6$.

   (a) Use the strong Markov property to give a proof that $Z_1, \dots, Z_6$ are independent random variables. [*Hint*: Let $X_1, X_2, \dots$ be the respective outcomes on the successive tosses. Check that each $\tau_j, j \geq 2$, denoting the first time after τ_{j-1} that X_n is not among $X_1, \dots, X_{\tau_{j-1}}$, with $\tau_1 = 1$, defines a stopping time. Then $Z_j = \tau_j - \tau_{j-1}, j \geq 2$.]
   (b) Calculate the distributions of $Z_2, \dots, Z_6$.
   (c) Calculate EZ and Var Z.

5. (*Coupon Collector's Problem*) A box contains N balls labeled $0, 1, 2, \dots, N - 1$. Let $T \equiv T_N$ be the number of selections (at random with replacement) required until each ball is sampled at least once. Let T_j be the number of selections required to sample j *distinct* balls.

   (a) Show that if X_n is the outcome of the nth draw, i.e., $X_n = j$ if the ball labeled j is selected at the nth draw, then T_j is a stopping time with respect to the filtration $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}, n \geq 1$.
   (b) $T = (T_N - T_{N-1}) + (T_{N-1} - T_{N-2}) + \cdots + (T_2 - T_1) + T_1$, where $T_1 = 1$, $T_2 - T_1, \dots, T_{j+1} - T_j, \dots, T_N - T_{N-1}$ are independent geometrically distributed with parameters $(N - j)/N$, respectively.

(c) Let τ_j be the number of selections to get ball j. Then τ_j is geometrically distributed.
(d) $P(T > m) \leq Ne^{-m/N}$. [*Hint*: $P(T > m) \leq \sum_{j=1}^{N} P(\tau_j > m)$.]
(e) $P(T > m) = \displaystyle\sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}\left(1 - \frac{k}{N}\right)^m$. [*Hint*: Use inclusion–exclusion on $[T > m] = \cup_{j=1}^{N}[\tau_j > m]$.]
(f) Let $X_1, X_2, \ldots$ be the successive labels on the balls selected. Is T a stopping time for $\{X_n : n \geq 0\}$?

6. (*Independent Coupling Process*) Let $\{X_n : n \geq 0\}$ and $\{Y_n : n \geq 0\}$ be independent Markov chains with common transition probability matrix $\mathbf{p}$.

(a) Show that $\{(X_n, Y_n) : n \geq 0\}$ is a Markov chain on the state space $S \times S$.
(b) Calculate the transition law of $\{(X_n, Y_n) : n \geq 0\}$.
(c) Let $T = \inf\{n : X_n = Y_n\}$. Show that T is a stopping time for the process $\{(X_n, Y_n) : n \geq 0\}$.
(d) Let $\{Z_n : n \geq 0\}$ be the process obtained by watching $\{X_n : n \geq 0\}$ up until time T and then switching to $\{Y_n : n \geq 0\}$ after time T; i.e., $Z_n = X_n$, $n < T$, and $Z_n = Y_n$, for $n \geq T$. Show that $\{Z_n : n \geq 0\}$ is a Markov chain and calculate its transition law.

7. (*Record Times*) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of nonnegative random variables having a continuous distribution (so that the probability of a tie is zero). Define $R_1 = 1$, $R_k = \inf\{n \geq R_{k-1} + 1 : X_n \geq \max(X_1, \ldots, X_{n-1})\}$, for $k = 2, 3, \ldots$.

(a) Show that $\{R_n : n \geq 1\}$ has the Markov property and calculate its transition probabilities. [*Hint*: All $i_k!$ rankings of $(X_1, X_2, \ldots, X_{i_k})$ are equally likely. Consider the event $[R_1 = 1, R_2 = i_2, \ldots, R_k = i_k]$ and count the number of rankings of $(X_1, X_2, \ldots, X_{i_k})$ that correspond to its occurrence.]
(b) Let $T_n = R_{n+1} - R_n$. Is $\{T_n : n \geq 0\}$ a Markov chain? [*Hint*: Compute $P(T_3 = 1 \mid T_2 = 1, T_1 = 1)$ and $P(T_3 = 1 \mid T_2 = 1)$.]

8. (*Record Values*) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of nonnegative integer-valued random variables. Define the record times $R_1 = 1, R_2, R_3, \ldots$ as in Exercise 7. Define the *record values* by $V_k = X_{R_k}$, $k = 1, 2, \ldots$.

(a) Show that each R_k is a stopping time for $\{X_n : n \geq 1\}$.
(b) Show that $\{V_k\}$ is a Markov chain and calculate its transition probabilities.
(c) Extend (b) to the case when the distribution function of X_k is continuous.

9. Let $S_0^x = x$, $S_n^x = x + Z_1 + \cdots + Z_n$, $n \geq 1$, be a simple symmetric random walk starting at an integer x, i.e., Z_j, $j \geq 1$, is an i.i.d. sequence of ±1- valued random variables with equal probabilities. Let $\mathcal{F}_n = \sigma\{Z_1, \ldots, Z_n\}$, $n \geq 1$, and let $\mathcal{F}_0$ be the trivial sigmafield $\{\emptyset, \Omega\}$.

(a) Prove that $Q_n = (S_n^x)^2 - n, n = 0, 1, 2, \dots$ is a martingale[1] with respect to the filtration $\mathcal{F}_n, n \geq 0$.

(b) Let $\tau = \inf\{n \geq 0 : S_n^x \in \{a, b\}\}$ be the first time the random walk reaches integers a or b starting at x, $a \leq x \leq b$. Compute $\mathbb{E}\tau$ using the optional stopping theorem.[2]

10. (*Lazy Random Walk*) Consider the *lazy symmetric simple random walk starting at x* obtained by allowing the distribution of displacements $Z_j, j \geq 1$, in Exercise 9 to be $\pm 1, 0$ with $P(Z_j = 1) = P(Z_j = -1) = \delta$, and $P(Z_j = 0) = 1 - 2\delta$, $j \geq 1$, for fixed $\delta \in (0, 1)$.

(a) Show that the corresponding lazy random walk $S_n^x = x + Z_1 + \cdots + Z_n, n \geq 1$, $S_0^x = x$, is a martingale, and use this to compute the probability $\psi(x)$ that it reaches a before b starting from x, $a \leq x \leq b$. Show that $\psi(x)$ does not depend on δ and, in particular, coincides with the probability obtained in the case $\delta = 1/2$.

(b) Prove that $Q_n = (S_n^x)^2 - n\mathbb{E}Z_n^2, n = 0, 1, 2, \dots$, is a martingale with respect to the filtration $\mathcal{F}_n, n \geq 0$.

(c) Define the escape time τ as in Exercise 9 and compute $\mathbb{E}\tau$.

11. (*Asymmetric Simple Random Walk*) Consider the *asymmetric simple random walk starting at x* obtained by allowing the distribution of displacements $Z_j, j \geq 1$, in Exercise 9 to be ± 1 with $P(Z_j = 1) = p, P(Z_j = -1) = q = 1 - p, j \geq 1$, for fixed $p \in (0, 1), p \neq q$. Show that $\mathcal{E}_n = (\frac{q}{p})^{S_n^x}, n \geq 0$, is a martingale and apply the optional stopping theorem, with $\tau = \inf\{n : S_n^x \in \{a, b\}\}$, to compute the probability that $S_n^x, n \geq 0$, reaches a before b.

[1] BCPT p. 53.

[2] BCPT p. 61.

Chapter 10
Transience and Recurrence of Markov Chains

Two fundamental long term properties of states of Markov chains on a finite or countably infinite state space are those of *transience* and *recurrence*, respectively. The former refers to a class of unstable states in the sense that the process will eventually no longer visit them, while the latter are sure to be visited infinitely often. The recurrent states are further classified in terms of the average time required to return. Those for which the expected return time is finite are referred to as positive-recurrent, or ergodic states.

The unrestricted simple random walk $\{S_n\}_{n\geq 0}$ is an example in which any state $i \in S$ can be reached from every state j in a finite number of steps with positive probability. If $\mathbf{p}$ denotes its transition probability matrix, then $\mathbf{p}^2$ is the transition probability matrix of $\{Y_n\}_{n\geq 0} := \{S_{2n} : n = 0, 1, 2, \ldots\}$. However, for the Markov chain $\{Y_n\}_{n\geq 0}$, transitions in a finite number of steps are possible from odd to odd integers and from even to even, but not otherwise. For $\{S_n\}_{n\geq 0}$ one says that there is one class of *"essential"* states and for $\{Y_n\}_{n\geq 0}$ that there are two classes of essential states.

A different situation occurs when the random walk has two absorbing boundaries on $S = \{c, c+1, \ldots, d-1, d\}$, i.e., the nonzero transition probabilities are specified by $p_{cc} = p_{dd} = 1$, $p_{x,x+1} = p$, $p_{x,x-1} = q = 1 - p$, $c + 1 \leq x \leq d - 1$. The states c, d can be reached (with positive probability) from $c+1, \ldots, d-1$. However, $c + 1, \ldots, d - 1$ cannot be reached from c or d. In this case $c + 1, \ldots, d - 1$ are called *"inessential"* states while $\{c\}$, $\{d\}$ form two classes of essential states. The "inessential" will not play a role in the long-run behavior of the process. If a chain has several essential classes, the process restricted to each class can be analyzed separately.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_10

Definition 10.1. Write $i \rightarrow j$ and read it as "*j is accessible from i*" if $p_{ij}^{(n)} > 0$ for some $n \geq 1$. Write $i \leftrightarrow j$ and read "*i and j communicate*" if $i \rightarrow j$ and $j \rightarrow i$. Say "*i is essential*" if $i \rightarrow j$ implies $j \rightarrow i$ (i.e., if any state j is accessible from i, then i is accessible from that state). We shall let $\mathcal{E}$ denote the set of all essential states. States that are not essential are called *inessential.*

Since

$$p_{ij}^{(n)} = \sum_{i_1, i_2, \ldots, i_{n-1} \in S} p_{ii_1} p_{i_1 i_2} \cdots p_{i_{n-1} j}, \tag{10.1}$$

$i \rightarrow j$ if and only if there exists a path of states $i, i_1, i_2, \ldots, i_{n-1}, j$ such that p_{ii_1}, $p_{i_1 i_2}, \ldots, p_{i_{n-1} j}$ are strictly positive.

Proposition 10.1.

a For every i there exists (at least one) j such that $i \rightarrow j$.
b $i \rightarrow j$, $j \rightarrow k$ imply $i \rightarrow k$.
c "i essential" implies $i \leftrightarrow i$.
d i essential, $i \rightarrow j$ imply "j is essential" and $i \leftrightarrow j$.
e On $\mathcal{E}$ the relation "$\leftrightarrow$" is an equivalence relation (i.e., reflexive, symmetric, and transitive).

Proof.

(a) For each i, $\sum_{j \in S} p_{ij} = 1$. Hence there exists at least one j for which $p_{ij} > 0$; for this j one has $i \rightarrow j$.

(b) $i \rightarrow j, j \rightarrow k$ means that there exist $m \geq 1, n \geq 1$ such that $p_{ij}^{(m)} > 0$, $p_{jk}^{(n)} > 0$. Hence,

$$p_{ik}^{(m+n)} = \sum_{l \in S} p_{il}^{(m)} p_{lk}^{(n)} = p_{ij}^{(m)} p_{jk}^{(n)} + \sum_{l \neq j} p_{il}^{(m)} p_{lk}^{(n)} \geq p_{ij}^{(m)} p_{jk}^{(n)} > 0. \tag{10.2}$$

Thus, $i \rightarrow k$. Note that the first equality is a consequence of the relation $\mathbf{p}^{m+n} = \mathbf{p}^m \mathbf{p}^n$.

(c) Suppose i is essential. By (a) there exists j such that $p_{ij} > 0$. Since i is essential, this implies $j \rightarrow i$, i.e., there exists $m \geq 1$ such that $\mathbf{p}_{ji}^{(m)} > 0$. But then

$$p_{ii}^{(m+1)} = \sum_{l \in S} p_{il} p_{li}^{(m)} = p_{ij} p_{ji}^{(m)} + \sum_{l \neq j} p_{il} p_{li}^{(m)} > 0. \tag{10.3}$$

Hence $i \rightarrow i$ and, therefore, $i \leftrightarrow i$.

(d) Suppose i is essential, $i \to j$. Then there exist $m \geq 1, n \geq 1$ such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Hence $i \leftrightarrow j$. Now suppose k is any state such that $j \to k$, i.e., there exists $m' \geq 1$ such that $p_{jk}^{(m')} > 0$. Then, by (b), $i \to k$. Since i is essential, one must have $k \to i$. Together with $i \to j$ this implies (again by (b)) $k \to j$. Thus, if any state k is accessible from j, then j is accessible from that state k, proving that j is essential.

(e) If $\mathcal{E}$ is empty (which is possible, as, for example, in the case $p_{i,i+1} = 1, i = 0, 1, 2, \ldots$), then there is nothing to prove. Suppose $\mathcal{E}$ is nonempty. Then: (i) On $\mathcal{E}$ the relation "$\leftrightarrow$" is reflexive by (c). (ii) If i is essential and $i \leftrightarrow j$, then (by (d)) j is essential and, of course, $i \leftrightarrow j$ and $j \leftrightarrow i$ are equivalent properties. Thus "$\leftrightarrow$" is symmetric (on $\mathcal{E}$ as well as on S). (iii) If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \to j$ and $j \to k$. Hence $i \to k$ (by (b)). Also, $k \to j$ and $j \to i$ imply $k \to i$ (again by (b)). Hence $i \leftrightarrow k$. This shows that "$\leftrightarrow$" is transitive (on $\mathcal{E}$ as well as on S).

■

From the proof of (e) the relation "$\leftrightarrow$" is seen to be symmetric and transitive on all of S (and not merely $\mathcal{E}$). However, it is not generally true that $i \leftrightarrow i$ (or, $i \to i$) for all $i \in S$. In other words, reflexivity may break down on S.

Example 1 *(One-Dimensional Simple Random Walk).* $S = \{0, \pm 1, \pm 2, \ldots\}$. Assume $0 < p_{i,i+1} = p < 1$, $p_{i,i-1} = 1 - p$. Then $i \to j$ for all states $i \in S$, $j \in S$. Hence $\mathcal{E} = S$.

Example 2 *(Simple Random Walk with Two Absorbing Boundaries).* Here $S = \{c, c+1, \ldots, d\}$. $p_{cc} = p_{dd} = 1$, $p_{i,i+1} = p$, $p_{i,i-1} = 1 - p$, $c + 1 \leq i \leq d - 1$. Then $\mathcal{E} = \{c, d\}$. Note that c is not accessible from d, nor is d accessible from c.

Example 3 *(Simple Random Walk with Two Reflecting Boundaries).* Here $S = \{c, c+1, \ldots, d\}$. $p_{c,c+1} = p_{d,d-1} = 1$, $p_{i,i+1} = p \in (0, 1)$, $p_{i,i-1} = 1-p$, $c+1 \leq i \leq d - 1$. Then, $\mathcal{E} = S$.

Definition 10.2. A transition probability matrix $\mathbf{p}$ having one essential class and no inessential states is called *irreducible*. A Markov chain with an irreducible transition probability matrix is also called *irreducible*.

Since $\leftrightarrow$ is an equivalence relation on $\mathcal{E}$, i.e., it is reflexive, symmetric, and transitive, as a general rule it may be decomposed into disjoint equivalence classes of the form $\mathcal{E}(i) = \{j : j \leftrightarrow i\}$ (Exercise 1).

Our last item of bookkeeping concerns the role of possible cyclic motions within an essential class. In the unrestricted simple random walk example, note that $p_{ii} = 0$ for all $i = 0, \pm 1, \pm 2, \ldots$, but $p_{ii}^{(2)} = 2pq > 0$. In fact $p_{ii}^{(n)} = 0$ for all odd n, and $p_{ii}^{(n)} > 0$ for all even n. In this case, we say that the period of i is 2. More generally, if $i \to i$, then the *period of* i is the greatest common divisor of the integers in the set $A = \{n \geq 1 : p_{ii}^{(n)} > 0\}$. If $d = d_i$ is the period of i, then $p_{ii}^{(n)} = 0$ whenever n is not a multiple of d and d is the largest integer with this property.

Proposition 10.2.

a If $i \leftrightarrow j$, then i and j possess the same period. In particular "period" is constant on each equivalence class.

b Let $i \in \mathcal{E}$ have a period $d = d_i$. For each $j \in \mathcal{E}(i)$ there exists a unique integer $r_j, 0 \leq r_j \leq d-1$, such that $p_{ij}^{(n)} > 0$ implies $n = r_j(\mod d)$ (i.e., $n = 0 \text{mod} d$, or $n = sd + r_j$ with $s \geq 0$, an integer, and $1 \leq r_j \leq d-1$).

Proof.

(a) Clearly,

$$p_{ii}^{(a+m+b)} \geq p_{ij}^{(a)} p_{jj}^{(m)} p_{ji}^{(b)} \tag{10.4}$$

for all positive integers a, m, b. Choose a and b such that $p_{ij}^{(a)} > 0$ and $p_{ji}^{(b)} > 0$. If $p_{jj}^{(m)} > 0$, then $p_{jj}^{(2m)} \geq p_{jj}^{(m)} p_{jj}^{(m)} > 0$, and

$$p_{ii}^{(a+m+b)} \geq p_{ij}^{(a)} p_{jj}^{(m)} p_{ji}^{(b)} > 0, \qquad p_{ii}^{(a+2m+b)} \geq p_{ij}^{(a)} p_{jj}^{(2m)} p_{ji}^{(b)} > 0. \tag{10.5}$$

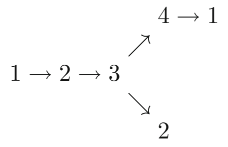
Therefore, d (the period of i) divides $a+m+b$ and $a+2m+b$, so that it divides the difference $m = (a+2m+b) - (a+m+b)$. Hence, the period of i does not exceed the period of j. By the same argument (since $i \leftrightarrow j$ is the same as $j \leftrightarrow i$), the period of j does not exceed the period of i. Hence the period of i equals the period of j.

(b) Choose a such that $p_{ji}^{(a)} > 0$. If $p_{ij}^{(m)} > 0$, $p_{ij}^{(n)} > 0$, then $p_{ii}^{(m+a)} \geq p_{ij}^{(m)} p_{ji}^{(a)} > 0$, and $p_{ii}^{(n+a)} \geq p_{ij}^{(n)} p_{ji}^{(a)} > 0$. Hence d, the period of i, divides $m+a, n+a$ and, therefore, $m - n = m + a - (n + a)$. Since this is true for all m, n such that $p_{ij}^{(m)} > 0$, $p_{ij}^{(n)} > 0$, it means that the difference between any two integers in the set $A\{n : p_{ij}^{(n)} > 0\}$ is divisible by d. This implies that there exists a *unique* integer r_j, $0 \leq r_j \leq d-1$, such that $n = r_j(\mod d)$ for all $n \in A$ (i.e., $n = sd + r_j$ for some integer $s \geq 0$ where s depends on n). ■

It is generally *not true* that the period of an essential state i is $\min\{n \geq 1 : p_{ii}^{(n)} > 0\}$. To see this consider the chain with state space $\{1, 2, 3, 4\}$ and transition matrix

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Schematically, only the following one-step transitions are possible.



Thus $p_{11}^{(2)} = 0$, $p_{11}^{(4)} > 0$, $p_{11}^{(6)} > 0$, etc., and $p_{11}^{(n)} > 0$ for all odd n. The states communicate with each other and their common period is 2, although $\min\{n : p_{11}^{(n)} > 0\} = 4$. Note that $\min\{n \geq 1 : p_{ii}^{(n)} > 0\}$ is a multiple of d_i since d_i divides all n for which $p_{ii}^{(n)} > 0$. Thus, $d_i \leq \min\{n \geq 1 : p_{ii}^{(n)} > 0\}$.

Proposition 10.3. Let $i \in \mathcal{E}$ have period $d > 1$. Let C_r be the set of $j \in \mathcal{E}(i)$ such that $r_j = r$, where r_j is the remainder term as defined in Proposition 10.2(b). Then

a $C_0, C_1, \ldots, C_{d-1}$ are disjoint, $\bigcup_{r=0}^{d-1:} C_r = \mathcal{E}(i)$.
b If $j \in C_r$, then $p_{jk} > 0$ implies $k \in C_{r+1}$, where we take $r + 1 = 0$ if $r = d - 1$.

Proof.

(a) Follows from Proposition 10.2(b).
(b) Suppose $j \in C_r$ and $p_{ij}^{(n)} > 0$. Then $n = sd + r$ for some $s \geq 0$. Hence, if $p_{ij} > 0$, then

$$p_{ik}^{(n+1)} \geq p_{ij}^{(n)} p_{jk} > 0, \tag{10.6}$$

which implies $k \in C_{r+1}$ (since $n + 1 = sd + r + 1 = r + 1(\mod d)$), by Proposition 10.2(b). ■

Here is what Proposition 10.3 means. Suppose i is an essential state and has a period $d > 1$. In one step (i.e., one time unit) the process can go from $i \in C_0$ only to some state in C_1 (i.e., $p_{ij} > 0$ only if $j \in C_j$). From states in C_1, in one step the process can go only to states in C_2. This means that in two steps the process can go from i only to states in C_2 (i.e., $p_{ij}^{(2)} > 0$ only if $i \in C_2$), and so on. In d steps the process can go from i only to states in $C_{d+1} = C_0$, completing one cycle (of d steps). Again in $d + 1$ steps the process can go from i only to states in C_1, and so on. In general, in $sd + r$ steps the process can go from i only to states in C_r.

Example 4. In the case of the unrestricted simple random walk, the period is 2 and all states are essential and communicate with each other. Fix $i = 0$. Then $C_0 = \{0, \pm 2, \pm 4, \ldots\}$, $C_1 = \{\pm 1, \pm 3, \pm 5, \ldots\}$. If we take i to be any even integer, then C_0, C_1 are as above. If, however, we start with i odd, then $C_0 = \{\pm 1, \pm 3, \pm 5, \ldots\}$, $C_1 = \{0, \pm 2, \pm 4, \ldots\}$.

Let $\{X_n : n \geq 0\}$ be a Markov chain with countable state space S and transition probability matrix $\mathbf{p} = ((p_{ij}))$. As in the case of random walks, the frequency of returns to a state is an important feature of the evolution of the process.

Definition 10.3. A state j is said to be *recurrent* if

$$P_j(X_n = j \text{ i.o.}) = 1, \tag{10.7}$$

and *transient* if

$$P_j(X_n = j \text{ i.o.}) = 0. \tag{10.8}$$

In view of the Borel–Cantelli Lemma (Part II), an easy example of a recurrent state j is provided by an i.i.d. (hence Markov) sequence $X_0, X_1, X_2, \dots$ such that $P(X_1 = j) > 0$. In particular, the second part of the Borel–Cantelli lemma $\sum_{n=1}^{\infty} P(X_n = j) = \infty$ holds. However, in what follows we will see this condition does continue to be necessary and sufficient under Markov dependence.

Recall the successive *return times* to the state j defined by

$$\tau_j^{(0)} = 0, \quad \tau_j^{(1)} := \inf\{n > 0 : X_n = j\}, \quad \tau_j^{(r)} = \inf\{n > \tau_j^{(r-i)} : X_n = j\}, \tag{10.9}$$

for $r = 1, 2, \dots$, with the convention that $\tau_j^{(r)} = \infty$ if there is no $n > \tau_j^{(r-1)}$ for which $X_n = j$. Write

$$\rho_{ij} = P_i(X_n = j \text{ for some } n \geq 1) = P_i(\tau_j^{(1)} < \infty). \tag{10.10}$$

Using the strong Markov property (Theorem 9.1) we get by the same calculation as in the example of the previous section that

$$\begin{aligned} P_i(\tau_j^{(r)} < \infty) &= P_i(\tau_j^{(r-1)} < \infty \text{ and } X_{\tau_j^{(r-1)}+n} = j \text{ for some } n \geq 1) \\ &= \mathbb{E}_i(\mathbf{1}_{[\tau_j^{(r-1)}<\infty]} P_{X_{\tau_j^{(r-1)}}}(X_n = j \text{ for some } n \geq 1)) \\ &= \mathbb{E}_i(\mathbf{1}_{[\tau_j^{(r-1)}<\infty]})\rho_{jj} = P_i(\tau_j^{(r-1)} < \infty)\rho_{jj}. \end{aligned} \tag{10.11}$$

Therefore, by iteration,

$$P_i\big(\tau_j^{(r)} < \infty\big) = P_i\big(\tau_j^{(1)} < \infty\big)\rho_{jj}^{r-1} = \rho_{ij}\rho_{jj}^{r-1} \qquad (r = 2, 3, \dots). \tag{10.12}$$

In particular, with $i = j$,

$$P_j\big(\tau_j^{(r)} < \infty\big) = \rho_{jj}^{r} \qquad (r = 1, 2, 3, \dots). \tag{10.13}$$

Now

$$P_j(X_n = j \text{ i.o.}) = P_j(\cap_{r=1}^{\infty}[\tau_j^{(r)} < \infty]) = \lim_{r\to\infty} P_j(\tau_j^{(r)} < \infty) = \begin{cases} 1 & \text{if } \rho_{jj} = 1. \\ 0 & \text{if } \rho_{jj} < 1. \end{cases} \tag{10.14}$$

Further, write $N(j) \equiv \sum_{n=0}^{\infty} \mathbf{1}_{[X_n=j]}$ for the *number of visits to the state* j by the Markov chain $\{X_n\}_{n\geq 0}$, and denote its expected value by

$$G(i, j) = \mathbb{E}_i N(j) = \sum_{n=0}^{\infty} p_{ij}^{(n)} = (I - \mathbf{p})^{-1}. \tag{10.15}$$

$G(i, j)$ is also referred to as the (discrete parameter) *Green's function* of the Markov chain, (see Examples 5 and 6 below). Now using (10.12)

$$\mathbb{E}_i N(j) = \sum_{r=0}^{\infty} P_i(N(j) > r) = \delta_{ij} + \sum_{r=0}^{\infty} P_i\big(\tau_j^{(r+1)} < \infty\big) = \delta_{ij} + \rho_{ij} \sum_{r=0}^{\infty} \rho_{jj}^r, \tag{10.16}$$

where δ_{ij} is 1 or 0 according to $i = j$ or $i \neq j$. Thus,

$$G(i, j) = \begin{cases} \delta_{ij} & \text{if } i \not\to j, \text{ i.e., } \rho_{ij} = 0, \\ \delta_{ij} + \rho_{ij}/(1 - \rho_{jj}) & \text{if } i \to j \text{ and } \rho_{jj} < 1, \\ \infty & \text{if } i \to j \text{ and } \rho_{jj} = 1. \end{cases} \tag{10.17}$$

This calculation provides two useful characterizations of recurrence; one is in terms of the long-run expected number of returns and the other in terms of the probability of eventual return.

Theorem 10.4.

a Every state is either recurrent or transient. A state j is recurrent iff $\rho_{jj} = 1$ iff $G(j, j) = \infty$, and transient iff $\rho_{jj} < 1$ iff $G(j, j) \equiv (1 - \rho_{jj})^{-1} < \infty$. If j is transient $p_{ij}^{(n)} \to 0$ as $n \to \infty$ for all i.

b If i is recurrent, $i \to j$, then j is recurrent, and $\rho_{ij} = \rho_{ji} = 1$. Thus, recurrence (or transience) is a class property. In particular, if all states communicate with each other, then either they are all recurrent, or they are all transient.

c Let i be recurrent, and $S(i) := \{j \in S : i \to j\}$ be the class of states which communicate with i. Let $\bar{\pi}$ be a probability distribution on $S(i)$. Then

$$P_{\bar{\pi}}(X_n \text{ visits every state in } S(i) \text{ i.o.}) = 1. \tag{10.18}$$

Proof. Part (a) follows from (10.14), (10.15), (10.17). For part (b), suppose i is recurrent and $i \to j (j \neq i)$. Let A_r denote the event that the Markov chain visits j between the r-th and $(r + 1)$st visits to state i. Then under P_i, $A_r (r \geq 0)$ are independent events and have the same probability θ, say. Now $\theta > 0$. For if $\theta = 0$, then $P_i(X_n = j \text{ for some } n \geq 1) = P_i(\bigcup_{r \geq 0} A_r) = 0$, contradicting $i \to j$. It now follows from the second half of the Borel–Cantelli Lemma that $P_i(A_r \text{ i.o.}) = 1$. This implies $G(i, j) = \infty$ and hence, by (10.17), $\rho_{jj} = 1$. Hence j is recurrent. Also, $\rho_{ij} \geq P_i(A_r \text{ i.o.}) = 1$. By the same argument, $\rho_{ji} = 1$. Note that $G(j, j) = 1 + \rho_{jj}/(1 - \rho_{jj}) = 1/(1 - \rho_{jj})$ for transient states $j \in S$.

To prove part (c) use part (b) to get for arbitrary $i \in S(j)$,

$$P_{\bar{\pi}}(X_n \text{ visits } i \text{ i.o.}) \sum_{k \in S(j)} \bar{\pi}_k P_k(X_n \text{ visits } i \text{ i.o.}) \sum_{k \in S(j)} \bar{\pi}_k = 1. \tag{10.19}$$

Hence

$$P_{\bar{\pi}}\left(\bigcap_{i \in S(j)} [X_n \text{ visits } i \text{ i.o.}] \right) = 1. \tag{10.20}$$

■

Theorem 10.4 shows that the difference between recurrence and transience is quite dramatic. If j is recurrent, then $P_j(N(j) = \infty) = 1$. If j is transient, not only is it true that $P_j(N(j) < \infty) = 1$ but also $\mathbb{E}_j(N(j)) < \infty$.

Corollary 10.4. Every inessential state is transient.

Proof. If j is inessential, then there exist $i \in S$ and $m \geq 1$ such that

$$p_{ji}^{(m)} > 0 \qquad \text{and} \qquad p_{ij}^{(n)} = 0 \qquad \text{for all } n \geq 1. \tag{10.21}$$

Hence, using the Markov property,

$$\begin{aligned} P_j(N(j) < \infty) &\geq P_j(X_m = i, X_n \neq j \text{ for } n > m) \\ &= p_{ji}^{(m)} P_i(X_n \neq j \text{ for } n > 0) = p_{ji}^{(m)} > 0. \end{aligned} \tag{10.22}$$

By Proposition 10.4, j is transient, since (10.22) says j is not recurrent. ■

Corollary 10.5. Assume there exists an invariant probability $\boldsymbol{\pi} = \{\pi_j : j \in S\}$ for $\mathbf{p}$. *(a)* If $\pi_j > 0$, then j is recurrent. *(b)* If the chain is irreducible, then all states are recurrent.

Proof. (a) Suppose $\pi_j > 0$. If j is transient, then $\pi_j = \sum_{i \in S} \pi_i \cdot p_{ij}^{(n)} \to 0$ as $n \to \infty$, since $p_{ij}^{(n)} \to 0$, a contradiction. (b) From (a) and the fact that there exists some j such that $\pi_j > 0$, it follows that j is recurrent. All states are then recurrent by Theorem 10.4(b). ■

Example 5. The Green's function for the (transient) simple random walk with $p > \frac{1}{2}$ can be computed from the first passage time probabilities (Exercise 18) as

$$G(i,j) = \begin{cases} \left(\frac{q}{p}\right)^{i-j} \Big/ (2p-1) & \text{for } i > j, \\ 1/(2p-1) & \text{for } i \le j. \end{cases} \tag{10.23}$$

Example 6 (Polya's Theorem). Suppose $\mathbf{p}$ is the transition probability matrix for the simple symmetric random walk $\{X_n\}_{n=0}^{\infty}$ on the k-dimensional integer lattice $\mathbb{Z}^k$. The i.i.d. increments of the random walk take $2k$ values $\pm\mathbf{e}_i, i = 1, 2, \ldots, k$, with equal probability; here $\mathbf{e}_i$ denotes the standard unit vector with 1 in the i-th coordinate and 0 otherwise. The rather straightforward proof of recurrence in one-dimension (Exercise 14) extends to two dimensions by a 45 degree rotation of coordinate axes to render the coordinates as independent one-dimensional random walks (Exercise 15). However, the transience of the random walk for $k \ge 3$ is a distinct phenomenon. Although somewhat cumbersome, one can use combinatorial arguments to show that $p_{0,0} \le cn^{-\frac{k}{2}}$ for a positive constant[1] c. By irreducibility one has $G(x, y) < \infty$ for all $x, y \in \mathbb{Z}^k, k \ge 3$. Moreover, since $G(x, y) = \rho_{xy}/(1 - \rho_{xx}) \le (1 - \rho_{xx})^{-1} = (1 - \rho_{00})^{-1}$, where the last equality uses the translation invariance $\rho_{xx} = \rho_{00}$, it follows that $G(x, y)$ is uniformly bounded for all $x, y \in \mathbb{Z}^k, k \ge 3$. An alternative proof of transience for the cases $k \ge 3$ can be obtained as follows. Write $S_n = Z_1 + \cdots + Z_n, n \ge 1$, where $Z_1, Z_2, \ldots$ are i.i.d. $\pm e_j-$ valued random vectors with equal probabilities $\frac{1}{2k}, j = 1, 2, \ldots k$, and the ith component of e_j is the Kronecker delta δ_{ij}. Since S_n takes values in the integer lattice, its characteristic function

$$\varphi^n(\xi) = Ee^{i\xi\cdot S_n} = (Ee^{i\xi\cdot Z_1})^n \tag{10.24}$$

is periodic and its Fourier coefficients may be alternatively computed from

$$\varphi^n(\xi) = Ee^{i\xi\cdot S_n} = \sum_y e^{i\xi\cdot y} P(S_n = y) \tag{10.25}$$

according to (Exercise 5):

$$p_{xy}^{(n)} = P(S_n = y - x) = (2\pi)^{-k} \int_{(-\pi,\pi]^k} e^{-i\xi\cdot(y-x)} \varphi^n(\xi) d\xi. \tag{10.26}$$

Thus

$$G(x, y) \le (2\pi)^{-k} \sum_{n=0}^{\infty} \int_{(-\pi,\pi]^k} |\varphi(\xi)|^n d\xi = (2\pi)^{-k} \int_{(-\pi,\pi]^k} \frac{d\xi}{1 - |\varphi(\xi)|}. \tag{10.27}$$

[1] See, e.g., Bhattacharya and Waymire (1990, 2009), pp. 13–15, or Bhattacharya and Majumdar (2007), pp. 156–157.

By definition of Z_1 one easily has $|\varphi(\xi)| = |\frac{1}{2k}\sum_{m=1}^{k}(e^{i\xi_m} + e^{-i\xi_m})| = |\frac{1}{k}\sum_{m=1}^{k}\cos(\xi_m)|$. For convergence of the integral in (10.27) it is enough to check convergence of $\int_U \frac{d\xi}{1-|\varphi(\xi)|}$ for neighborhoods $U \subset (-\pi, \pi]^k$ of the singularities $(0, \dots, 0)$, $(\pm\pi, \dots, \pm\pi)$ of the integrand. Since $(1-\cos(x))/x^2 \to 1/2$ as $x \to 0$, by continuity for sufficiently small $\epsilon > 0$, if $|\xi_m| < \epsilon$, then $0 < \cos\xi_m \le 1 - \frac{\xi_m^2}{4}$ and thus for $\xi \in U = (-\epsilon, \epsilon)^k$, $|\varphi(\xi)| = \varphi(\xi) \le 1 - \frac{1}{4k}\sum_{m=1}^{k}\xi_m^2$, and hence for $k \ge 3$,

$$\int_U \frac{d\xi}{1-|\varphi(\xi)|} \le 4k\int_U \frac{d\xi}{\sum_{m=1}^{k}\xi_m^2} = c_k(\epsilon)\int_0^1 \frac{r^{k-1}dr}{r^2} < \infty, \tag{10.28}$$

for a positive constant $c_k(\epsilon)$ by a polar coordinate change of variables. Similarly one may check convergence at the other singularities for $k \ge 3$ (Exercise 6). Note that in addition to finiteness of $G(x, y)$ one again sees from (10.27) that $G(x, y)$ is uniformly bounded for all $x, y \in \mathbb{Z}^k$, $k \ge 3$.

Remark 10.1. An interesting notion to capture highly transient phenomena was introduced by James and Peres (1997), and explored further in James et al (2007), in which there would be *cut points* j such that for some m, one has $X_m = j$ and the set $\{X_0, \dots, X_m\}$ is disjoint from the set $\{X_{m+1}, X_{m+2}, \dots\}$. Lawler (1991) proved that the simple symmetric random walk on $\mathbb{Z}^k$, $k \ge 4$, has infinitely many cut points almost surely. This, and more, was extended to the case of random walks on $\mathbb{Z}^3$ by Blachère (2003).

Exercises

1. Let S be a countable set and $\leftrightarrow$ an equivalence relation on it. Prove that S is the disjoint union of equivalence classes of the form $\mathcal{E}(i) = \{j \in S : j \leftrightarrow i\}$. [*Hint*: $\mathcal{E}(i) = \mathcal{E}(j)$ or $\mathcal{E}(i) \cap \mathcal{E}(j) = \emptyset$.]
2. Construct a finite state Markov chain such that
   (a) There is only one inessential state.
   (b) The set $\mathcal{E}$ of essential states decomposes into two equivalence classes with periods $d = 1$ and $d = 3$.
3. (a) Give an example of a transition matrix for which all states are inessential.
   (b) Show that if S is finite, then there is at least one essential state.
4. Classify all states for $\mathbf{p}$ given below into essential and inessential subsets. Decompose the set of all essential states into equivalence classes of communicating states.

$$\begin{bmatrix} \frac{1}{3} & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ \frac{2}{5} & 0 & 0 & 0 & \frac{3}{5} & 0 & 0 \\ 0 & 0 & 0 & \frac{5}{6} & 0 & 0 & \frac{1}{6} \\ 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{3}{4} & 0 \end{bmatrix}.$$

5. Derive the formula (10.26). [*Hint*: Multiply by $e^{-i\xi\cdot z}$, $z \in \mathbb{Z}^k$, and integrate (as iterated integrals) with respect to ξ. Note that for integral z, $\int_{-\pi}^{\pi} e^{-iz\xi} d\xi = (2\pi)^{-1}\delta_{z0}$.]
6. Check convergence of the integral $\int_U \frac{d\xi}{1-|\varphi(\xi)|}$ for neighborhoods U of $(\pm\pi, \dots, \pm\pi)$.
7. Let $\mathbf{p}$ be the transition matrix on $S = \{0, 1, 2, 3\}$ defined below.

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

Show that S is a single class of essential states of period 2 and calculate $\mathbf{p}^n$ for all n.

8. Show by induction on N that all states communicate in the Top-In Card Shuffling example of Exercises 13(b) of Chapter 7.
9. Prove that $\{R_n\} = \{|S_n|\}$ is Markov, where S_n is the simple symmetric random walk on $\mathbb{Z}$. Classify all states of $\{R_n : n \geq 0\}$.
10. (*A Birth or Collapse Model*) Consider the two cases:

(a) $p_{i,i+1} = \frac{1}{i+1}, \quad p_{i,0} = \frac{i}{i+1}, \quad i = 0, 1, 2, \dots.$

(b) $p_{i,0} = \frac{1}{i+1}, \quad p_{i,i+1} = \frac{i}{i+1}, \quad i \geq 1, \quad p_{0,1} = 1.$

Determine in each case whether the Markov chain is transient, null recurrent, or positive recurrent. Can you generalize this to birth–collapse with $p_{01} = 1$, and $p_{12} \dots p_{n,n+1} \to 0$ as $n \to \infty$?

11. Let $p_{i,i+1} = p$, $p_{i,0} = q$, $i = 0, 1, 2, \dots$. Classify the states of $S = \{0, 1, 2, \dots\}$ as transient or recurrent $(0 < p < 1, q = 1 - p)$.
12. Fix $i, j \in S$. Write

$$r_n = P_i(X_n = j) \equiv p_{ij}^{(n)} \quad (n \geq 1), \quad r_0 = 1,$$
$$f_n = P_i(X_m \neq j \text{ for } m < n, X_n = j) \quad (n \geq 1).$$

(a) Using the strong Markov property show that $r_n = \sum_{m=1}^{n} f_m r_{n-m}$ $(n \geq 1)$.

(b) Sum (a) over n to give an alternative proof of (10.17).

(c) Use (a) to indicate how one may compute the distribution of the time of the first visit to state j (after time zero), starting in state i, in terms of $p_{ij}^{(n)}$ $(n \geq 1)$. [*Hint*: Consider generating functions $\hat{f}(t) = \sum_n f_n t^n, \hat{r}(t) = \sum_n r_n t^n$.]

13. An *invariant measure* for a transition matrix $((p_{ij}))$ is a sequence of nonnegative numbers (m_i) such that $\sum_i m_i p_{ij} = m_j$ for all $j \in S$. An invariant measure may or may not be normalizable to a probability distribution on S.

    (a) Let $p_{i,i+1} = p_i$ and $p_{i,0} = 1 - p_i$ for $i = 0, 1, 2, \ldots$. Show that there is a unique invariant measure (up to multiples) if and only if $\lim_{n\to\infty} \prod_{k=1}^{n} p_k = 0$; i.e., if and only if the chain is recurrent, since the product is the probability of no return to the origin.

    (b) Show that invariant measures exist for the unrestricted simple random walk but are not unique in the transient case and are unique (up to multiples) in the recurrent case.

    (c) Let $p_{00} = p_{01} = \frac{1}{2}$ and $p_{i,i-1} = p_{i,i} = 2^{-i-2}$, and $p_{i,i+1} = 1 - 2^{-i-1}$, $i = 1, 2, 3, \ldots$. Show that the probability of not returning to 0 is positive (i.e., transience), but that there is a unique invariant measure.

14. For the one-dimensional simple symmetric random walk prove that (i) $p_{00}^{(n)} = 0$ for all odd n, and $p_{00}^{(2n)} = cn^{-\frac{1}{2}}$ for a positive constant c, (ii) the random walk is recurrent. [*Hint*: (i) $p_{00}^{(2n)} = \binom{2n}{n} 2^{-2n}$, (Stirling's formula): $n! \sim \sqrt{2\pi n} n^n e^{-n}$ in the sense that the ratio is one in the limit as $n \to \infty$, (ii) Check that the random walk is irreducible and apply Theorem 10.4(b).]

15. Let $Z_n, n \geq 0$, be i.i.d. two-dimensional random vectors, with $P(Z_n = (\pm 1, 0)) = P(Z_n = (0, \pm 1) = 1/4$, respectively, for the four cases. Define the *two-dimensional simple symmetric random walk*,, starting at x on the integer lattice $\mathbb{Z}^2 \equiv \mathbb{Z} \times \mathbb{Z}$ by $S_n^x = x + Z_1 + \cdots + Z_n, n \geq 1$, $S_0^x = x \in \mathbb{Z}^2$.

    (a) Show that the two-dimensional simple symmetric random walk is irreducible.[*Hint*: Let $(i, j), (k, l) \in \mathbb{Z}^2$. There is a polygonal path of finite length $m = |i - k| + |l - j|$ from (i, j) to (k, l), in increments of $(\pm 1, 0)$ and/or $(0, \pm 1)$, and the random walk can move along this with positive probability, i.e., $p_{(i,j),(k,l)}^{(m)} \geq (1/4)^m > 0$.]

    (b) Consider the transformed random vectors $W_n = \sqrt{2} Z_n A, n \geq 0$, where A is the matrix (rotation) $A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$. Show that the coordinates of $W_n = (W_n^{(1)}, W_n^{(2)}$ are symmetrically distributed independent Bernoulli ± 1-valued random variables, and use the estimate $p_{0,0}^{(2n)}$ for the one-dimensional simple symmetric random walk to show that the two-dimensional simple symmetric random walk is also recurrent.

16. Show that the two-dimensional random walk with increments $W_n, n \geq 0$, in the previous exercise, is *not* irreducible. [*Hint*: Check that the random walk with increments W_n partitions the state space into two equivalence classes of states (j, k) in which j and k are of the same parity, and another in which (j, k) have opposite parity.] Show that every state is recurrent regardless of which equivalence class it belongs.
17. Let $k \geq 3$ be an integer, and suppose $W_n = (W_n^{(1)}, \ldots, W_n^{(k)}), n \geq 1$, is a sequence of i.i.d. k-dimensional random vectors with independent, symmetric Bernoulli ±1-valued coordinates. Show that $T_n^x = z + W_1 + \cdots + W_n, n \geq 1$, $T_0^x = x \in \mathbb{Z}^k$, is transient, but not irreducible having 2^{k-1} equivalence classes. Show that for $k \geq 3$ there is no one-to-one map between $T_n^x, n \geq 0$ and the k-dimensional symmetric simple random walk $S_n^x, n \geq 0$, defined by the 2^k possible independent equally likely displacements of the form $(0, \ldots, 0 \pm 1, 0, \ldots 0)$
18. Consider the one-dimensional simple asymmetric random walk, $p \neq q = 1-p$, $0 < p < 1$. Let $p > 1/2$.

    (a) Show that $\rho_{xy} = 1$ if $x < y$, and $\rho_{xy} = (q/p)^{x-y}$ if $x > y$.
    (b) Calculate ρ_{xx}.
    (c) Calculate ρ_{xy} for the case $p < 1/2$. Determine the Greens' function for this random walk.

Chapter 11
Birth–Death Chains

Birth–death Markov chains comprise a special class of Markov processes on the integers which move to nearest neighbor states to the left or right, or stay put, in single transitions. Simple random walks provide examples for which the one-step transition probabilities do not depend on the states from which transitions are made.

The birth–death Markov chains, which include simple random walk, may be regarded as space-time discretizations of diffusion processes such as Brownian motion. The transition law defining a birth–death Markov chain has the form

$$p_{ij} = \begin{cases} \beta_i & \text{if } j = i + 1 \\ \delta_i & \text{if } j = i - 1 \\ \alpha_i & \text{if } j = i \\ 0 & \text{otherwise,} \end{cases} \tag{11.1}$$

where $\alpha_i + \beta_i + \delta_i = 1$. In particular, the *displacement* probabilities may depend on the state in which the process is located. However, it has a special "pseudo-continuity type", or skip-free, property, in that it cannot skip over states in its evolution.

For the unrestricted birth–death chain, i.e., $S = \mathbb{Z}$, one assumes $\beta_i > 0, \delta_i > 0,$ for all $i \in S$. Hence the Markov chain is irreducible. The state space for a birth–death chain is either all of $\mathbb{Z}$ or a finite or semi-infinite "interval" of contiguous states. The transition probabilities at left or right boundaries of the state space are permitted to be zero or one. In any case, the long-run behavior of a birth–death chain depends on the nature of its (local) transition probabilities $p_{i,i+1} = \beta_i$, $p_{i,i-1} = \delta_i$

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_11

at interior states i, as well as on its transitions at boundaries when present. In this section our aim is to obtain explicit conditions on these parameters under which, according to the more general theory of the preceding sections, various long-run behaviors may occur, e.g., transience, recurrence, positive recurrence, and convergence to steady state distributions.

Proposition 11.1 *(Unrestricted Birth–Death Chain).* Let $S = \mathbb{Z} \equiv \{0, \pm 1, \pm 2, \ldots\}$ and assume that $0 < \beta_i, \delta_i < 1, \beta_i + \delta_i \leq 1,$ for all $i \in S$. Let $c, d \in S$, $c < d$, and define for $c \leq i \leq d, i \in S$,

$$\begin{aligned}\psi(i) &:= P_i(X_{\tau_{\{c,d\}}} = c, \tau_{\{c,d\}} < \infty) = P_i(\{X_n\} \text{ reaches } c \text{ before } d \mid X_0 = i)\\ &= P_i(\tau_c < \tau_d),\end{aligned} \tag{11.2}$$

where τ_j denotes the first time the chain reaches j, and $\tau_{\{c,d\}} = \tau_c \wedge \tau_d$ is the first hitting time of the set $\{c, d\}$. Then

$$\psi(y) = \frac{\sum_{x=y}^{d-1} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}}}{1 + \sum_{x=c+1}^{d-1} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}}} \qquad (c + 1 \leq y \leq d - 1). \tag{11.3}$$

Proof. One has

$$\psi(i) = (1 - \beta_i - \delta_i)\psi(i) + \beta_i \psi(i + 1) + \delta_i \psi(i - 1),$$

or equivalently,

$$\beta_i(\psi(i + 1) - \psi(i)) = \delta_i(\psi(i) - \psi(i - 1)) \qquad (c + 1 \leq i \leq d - 1). \tag{11.4}$$

The *boundary conditions* for ψ are

$$\psi(c) = 1, \qquad \psi(d) = 0. \tag{11.5}$$

Rewrite (11.4) as

$$\psi(i + 1) - \psi(i) = \frac{\delta_i}{\beta_i}(\psi(i) - \psi(i - 1)), \tag{11.6}$$

for $(c + 1 \leq i \leq d - 1)$. Iteration now yields

$$\psi(x + 1) - \psi(x) = \frac{\delta_x}{\beta_x} \frac{\delta_{x-1}}{\beta_{x-1}} \cdots \frac{\delta_{c+1}}{\beta_{c+1}}(\psi(c + 1) - \psi(c)) \tag{11.7}$$

for $c + 1 \leq x \leq d - 1$. Summing (11.7) over $x = y, y + 1, \ldots, d - 1$, one obtains

$$\psi(d) - \psi(y) = \sum_{x=y}^{d-1} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}}(\psi(c + 1) - \psi(c)). \tag{11.8}$$

Let $y = c + 1$ and use (11.5) to obtain

$$\psi(c+1) = \frac{\sum_{x=c+1}^{d-1} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}}}{1 + \sum_{x=c+1}^{d-1} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}}} \ . \tag{11.9}$$

Using this in (11.8) (and using $\psi(d) = 0$, $\psi(c) = 1$) one gets the desired result. ■

Corollary 11.2. Let $\{X_n : n \geq 0\}$ be an unrestricted birth–death chain on $S = \mathbb{Z}$ under the conditions of the Proposition 11.1. For $c < y$ with $y, c \in \mathbb{Z}$, define

$$\rho_{yc} := P_y(\tau_c < \infty) \equiv P_y(X_n = c \text{ for some } n \geq 1). \tag{11.10}$$

Then

$$\begin{aligned} \rho_{yc} &= 1 \qquad \text{for all } y > c \text{ iff } \sum_{x=1}^{\infty} \frac{\delta_1 \delta_2 \cdots \delta_x}{\beta_1 \beta_2 \cdots \beta_x} = \infty, \\ &< 1 \qquad \text{for all } y > c \text{ iff } \sum_{x=1}^{\infty} \frac{\delta_1 \delta_2 \cdots \delta_x}{\beta_1 \beta_2 \cdots \beta_x} < \infty. \end{aligned} \tag{11.11}$$

Similarly for $y < d$, $y, d \in \mathbb{Z}$,

$$\begin{aligned} \rho_{yd} &= 1 \qquad \text{for all } y < d \text{ iff } \sum_{x=-\infty}^{0} \frac{\beta_x \beta_{x+1} \cdots \beta_0}{\delta_x \delta_{x+1} \cdots \delta_0} = \infty, \\ &< 1 \qquad \text{for all } y < d \text{ iff } \sum_{x=-\infty}^{0} \frac{\beta_x \beta_{x+1} \cdots \beta_0}{\delta_x \delta_{x+1} \cdots \delta_0} < \infty. \end{aligned} \tag{11.12}$$

Proof. Observe that (Exercise 1),

$$\begin{aligned} \rho_{yc} = \lim_{d \uparrow \infty} \psi(y) &= 1 \qquad \text{if } \sum_{x=c+1}^{\infty} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}} = \infty, \\ &< 1 \qquad \text{if } \sum_{x=c+1}^{\infty} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}} < \infty \ (c < y). \end{aligned} \tag{11.13}$$

Since, for $c + 1 \leq 0$,

$$\begin{aligned} \sum_{x=c+1}^{\infty} \frac{\delta_x \delta_{x-1} \cdots \delta_{c+1}}{\beta_x \beta_{x-1} \cdots \beta_{c+1}} &= \sum_{x=c+1}^{0} \frac{\delta_{c+1} \delta_{c+2} \cdots \delta_x}{\beta_{c+1} \beta_{c+2} \cdots \beta_x} \\ &\quad + \frac{\delta_{c+1} \delta_{c+2} \cdots \delta_0}{\beta_{c+1} \beta_{c+2} \cdots \beta_0} \sum_{x=1}^{\infty} \frac{\delta_1 \delta_2 \cdots \delta_x}{\beta_1 \beta_2 \cdots \beta_x} \end{aligned} \tag{11.14}$$

and a similar equality holds for $c + 1 > 0$. By relabeling the states i as $-i$ ($i = 0, \pm 1, \pm 2, \ldots$), one gets (11.12) (Exercise 2). ■

Corollary 11.3. Let $\{X_n : n \geq 0\}$ be an unrestricted birth–death chain on $S = \mathbb{Z}$ under the conditions of the Proposition 11.1. (i) If both sums in (11.11) and (11.12) diverge, then all states are recurrent, i.e.,

$$\rho_{yy} = 1, \text{ for all } y \in S. \tag{11.15}$$

(ii) If one of the sums (11.11) or (11.12) is convergent, then all states are transient, i.e.,

$$\rho_{yy} < 1, \qquad y \in S. \tag{11.16}$$

Proof. By the Markov property, conditioning on X_1,

$$\rho_{yy} = \delta_y \rho_{y-1,y} + \beta_y \rho_{y+1,y} + (1 - \delta_y - \beta_y)\rho_{y,y} \tag{11.17}$$

$\rho_{y-1,y} = 1$, $\rho_{y+1,y} = 1$, so that (11.17) implies (11.15). For (ii), say (11.11) is convergent, then by (11.17) we get (11.16). ■

Natural restrictions on birth–death chains may occur in the form of *boundary conditions* at endpoints of finite or semi-infinite intervals. Possibilities include absorption, and pure or partial reflection as illustrated in the following example.

Example 1 *(The Bernoulli–Laplace Model).* A simple model to describe the mixing of two incompressible liquids in possibly different proportions can be obtained by the following considerations. Consider two containers labeled box I and box II, respectively, each having N balls. Among the total of $2N$ balls, there are $2r$ red and $2w$ white balls, $1 \leq r \leq w$. At each instant of time, a ball is randomly selected from each of the boxes, and moved to the other box. The state at each instant is the number of red balls in box 1. In this example, the state space is $S = \{0, 1, \ldots, 2r\}$ and the evolution is a Markov chain on S with transition probabilities given for $1 \leq i \leq 2r - 1$,

$$\begin{aligned} p_{i,i+1} &= \frac{(w+r-i)(2r-i)}{(w+r)^2} \\ p_{ii} &= \frac{i(2r-i)}{(w+r)^2} + \frac{(w+r-i)(w-r+i)}{(w+r)^2} \\ p_{i,i-1} &= \frac{i(w-r+i)}{(w+r)^2} \end{aligned} \tag{11.18}$$

and

$$p_{00} = p_{2r,2r} = \frac{w-r}{w+r} \quad p_{01} = p_{2r,2r-1} = \frac{2r}{w+r}. \tag{11.19}$$

Definition 11.1. A state a will be called an *absorbing* boundary for the birth–death chain if $\alpha_a = 1 - \beta_a - \delta_a = 1$. If $\delta_a = 0$ and $\beta_a > 0$, then we will say that a is a (left side) *reflecting* boundary. If $\beta_a = 0$ and $\delta_a > 0$, then we will say that a is a (right side) *reflecting* boundary. A reflecting boundary with $\beta_a = 1$ or $\delta_a = 1$ is said to be *purely reflecting*.

Remark 11.1. Let us note that a birth–death Markov chain $\{X_n : n = 0, 1, \dots\}$ on $S = \{0, 1, 2, \dots\}$ with absorbing boundary at zero may be constructed from an unrestricted birth–death chain $\{Y_n : n = 0, 1, \dots\}$ on $\mathbb{Z}$, and having the same birth–death probabilities on S, but otherwise arbitrary, by starting on S and defining $X_n = Y_{n\wedge\tau_0}, n = 0, 1, \dots$, where $\tau_0 = \inf\{n \geq 0 : Y_n = 0\} \leq \infty$ (Exercise 13). A more interesting case is that of constructing a birth–death Markov chain $\{X_n : n = 0, 1, \dots\}$ on $S = \{0, 1, 2, \dots\}$, with reflecting boundary at zero, from an unrestricted birth–death chain $\{Y_n : n = 0, 1, \dots\}$ on $\mathbb{Z}$. For this one extends the birth and death probabilities by $\beta_{-i} = \delta_i, \delta_{-i} = \beta_i, i = 1, 2, \dots$, leaving β_0, δ_0 arbitrary, and defines $X_n = |Y_n|, n = 0, 1, 2, \dots$. One may apply Proposition 8.11 to show that $\{X_n : n = 0, 1, 2, \dots\}$ is Markov (Exercise 14).

Proposition 11.4 *(Two Reflecting Boundaries).* Let $S = \{0, 1, 2, \dots, N\}$, and suppose that 0 and N are reflecting boundaries, and $0 < \beta_i, \delta_i < 1, 1 \leq i \leq N - 1$. Then all states are recurrent.

Proof. Take $c = 0$, $d = N$ in (11.2). Then $\psi(y)$ gives the probability that the process starting at y reaches 0 before reaching N. The probability $\varphi(y)$, for the process to reach N before 0 starting at y, may be obtained in the same fashion by changing the boundary conditions (11.5) to $\varphi(0) = 0, \varphi(N) = 1$ to get that $\varphi(y) = 1 - \psi(y)$. Alternatively, check that $\varphi(y) \equiv 1 - \psi(y)$ satisfies the equation (11.6) (with φ replacing ψ) and the boundary conditions $\varphi(0) = 0$, $\varphi(N) = 1$. We leave it as an exercise to argue that such a solution is necessarily unique (Exercise 4). All states are recurrent, by Corollaries 8.7, 10.5 (see Exercise 5 for an alternative proof). ■

Proposition 11.5 *(One Absorbing Boundary).* Let $S = \{0, 1, 2, \dots\}$, and suppose that 0 is an absorbing boundary, but $0 < \beta_i, \delta_i < 1$, for $i = 1, 2, \dots$. Then 0 is recurrent and each $i \geq 1$ is transient.

Proof. For $c, d \in S$, the probability $\psi(y)$ is given by (11.3) and the probability ρ_{y0}, which is also interpreted as the *probability of eventual absorption* starting at $y > 0$, is given by

$$\begin{aligned}\rho_{y0} &= \lim_{d\uparrow\infty} \frac{\sum_{x=y}^{d-1} \frac{\delta_x\delta_{x-1}\cdots\delta_1}{\beta_x\beta_{x-1}\cdots\beta_1}}{1 + \sum_{x=1}^{d-1} \frac{\delta_x\delta_{x-1}\cdots\delta_1}{\beta_x\beta_{x-1}\cdots\beta_1}} \\ &= 1 \qquad \text{iff } \sum_{x=1}^{\infty} \frac{\delta_1\delta_2\cdots\delta_x}{\beta_1\beta_2\cdots\beta_x} = \infty \qquad (\text{for } y > 0). \end{aligned} \tag{11.20}$$

Whether or not the last series diverges,

$$\rho_{y0} \geq \delta_y \delta_{y-1} \cdots \delta_1 > 0, \qquad \text{for all } y > 0 \tag{11.21}$$

and

$$\begin{aligned} &\rho_{yd} \leq 1 - \delta_y \delta_{y-1} \cdots \delta_1 < 1, \text{ for } d > y > 0, \\ &\rho_{0d} = 0, \qquad\qquad\qquad\qquad\quad \text{for all } d > 0. \end{aligned} \tag{11.22}$$

By (11.17), $\rho_{y,y} = (\delta_y \rho_{y-1,y} + \beta_y \rho_{y+1,y})/(\beta_y + \delta_y)$. Since (11.21) implies $\rho_{y-1,y} < 1$, one has

$$\rho_{yy} < 1 \qquad (y > 0). \tag{11.23}$$

Thus, all nonzero states y are *transient*. ■

Proposition 11.6 *(One Reflecting Boundary).* Let $S = \{0, 1, 2, 3, \ldots\}$ and suppose that 0 is a reflecting boundary, but $\beta_i > 0$ for all i, and $\delta_i > 0$ for $i \geq 1$, $\beta_i + \delta_i \leq 1$. Then all states are recurrent if and only if the infinite series (11.20) diverges, i.e., if and only if $\rho_{y0} = 1$.

Proof. First assume that the infinite series in (11.20) diverges, i.e., $\rho_{y0} = 1$ for all $y > 0$. Then condition on X_1 to get

$$\rho_{00} = (1 - \beta_0)\rho_{00} + \beta_0 \rho_{10}, \tag{11.24}$$

so that

$$\rho_{00} = 1. \tag{11.25}$$

By Theorem 10.4, the chain is recurrent. On the other hand, if the series in (11.20) converges, then $\rho_{y0} < 1$ for all $y > 0$. In particular, from (11.24), we see $\rho_{00} < 1$. Again, by Theorem 10.4, the chain is transient. ■

The determination of conditions for recurrence and transience, under various other boundary possibilities such as two absorbing, or one absorbing and one reflecting boundary, are left to the exercises.

We next turn to the question of existence and uniqueness of invariant probabilities of birth–death chains. Recall that a Markov chain with transition probability matrix $\mathbf{p}$ an invariant probability is a probability $\boldsymbol{\pi} = (\pi_j : j \in S)$ on S such that

$$\mathbf{p}'\boldsymbol{\pi} = \boldsymbol{\pi}. \tag{11.26}$$

Then $\mathbf{p}'^{(n)}\boldsymbol{\pi} = \boldsymbol{\pi}$ for each time $n = 1, 2, \ldots$. That is, $\boldsymbol{\pi}$ is *invariant* under the transition law $\mathbf{p}$. Recall also that if $\{X_n : n \geq 0\}$ is started with an invariant initial distribution $\boldsymbol{\pi}$, then $\{X_n : n \geq 0\}$ is *stationary*.

Proposition 11.7 *(Two Reflecting Boundaries).* Let $S = \{0, 1, 2, \ldots, N\}$ and assume that 0, N are both reflecting boundaries. If $0 < \beta_i, \delta_i < 1$, $1 \le i \le N-1$, then all states are recurrent and the unique invariant probability $\boldsymbol{\pi}$ is given by

$$\pi_j = \frac{\beta_0 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_j}\pi_0 \qquad (1 \le j \le N),$$

$$\pi_0 = \left(1 + \sum_{j=1}^{N} \frac{\beta_0\beta_1 \cdots \beta_{j-1}}{\delta_1\delta_2 \cdots \delta_1}\right)^{-1}. \tag{11.27}$$

Proof. All states are recurrent by Proposition 11.4. The unique invariant probability is easily obtained by solving

$$\begin{aligned}&\pi_0(1-\beta_0) + \pi_1\delta_1 = \pi_0,\\ &\pi_{j-i}\beta_{j-1} + \pi_j(1-\beta_j-\delta_j) + \pi_{j+1}\delta_{j+1} = \pi_j \quad (j = 1, 2, \ldots, N-1),\end{aligned} \tag{11.28}$$

or

$$\pi_{j-1}\beta_{j-1} - \pi_j(\beta_j + \delta_j) + \pi_{j+1}\delta_{j+1} = 0, \tag{11.29}$$

subject to $\pi_j \ge 0$ for all j and $\sum_j \pi_j = 1$. The solutions are easily checked to be given by (11.29). The solution as a probability measure is unique. ■

Example 2 *(Equilibrium for the Bernoulli–Laplace Model).* For the Bernoulli–Laplace model the invariant distribution $\boldsymbol{\pi} = (\pi_i : i = 0, 1, \ldots, 2r)$ is the hypergeometric distribution calculated from (11.27) as

$$\begin{aligned}\pi_j &= \frac{\beta_0 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_j}\pi_0 = \frac{2r(w+r)}{j(w-r+j)}\prod_{i=1}^{j-1}\frac{(w+r-i)(2r-i)}{i(w-r+i)}\pi_0\\ &= \frac{\binom{2r}{j}\binom{2w}{w+r-j}}{\binom{2w+2r}{w+r}}\pi_0.\end{aligned} \tag{11.30}$$

Proposition 11.8 *(One Reflecting Boundary).* Let $S = \{0, 1, 2, \ldots\}$ with 0 as a reflecting boundary and $0 < \beta_1, \delta_i < 1, i \ge 1$. Then there exists an invariant probability given by

$$\pi_j = \frac{\beta_0\beta_1 \cdots \beta_{j-1}}{\delta_1\delta_2 \cdots \delta_j}\pi_0 \qquad (j \ge 1) \tag{11.31}$$

if and only if

$$\sum_{j=1}^{\infty} \frac{\beta_0\beta_1\cdots\beta_{j-1}}{\delta_1\delta_2\cdots\delta_j} < \infty, \tag{11.32}$$

in which case $\pi_0 = (1 + \sum_{j=1}^{\infty} \frac{\beta_1\cdots\beta_{j-1}}{\delta_1\cdots\delta_j})^{-1}$. The invariant probability is unique under the condition (11.32).

Proof. The system of equations $\mathbf{p}'\boldsymbol{\pi} = \boldsymbol{\pi}$ are

$$\begin{aligned} \pi_0(1-\beta_0) + \pi_1\delta_1 &= \pi_0, \qquad (11.33)\\ \pi_{j-1}\beta_{j-1} + \pi_j(1-\beta_j-\delta_j) + \pi_{j+1}\delta_{j+1} &= \pi_j \quad (j \geq 1). \end{aligned}$$

The solution in terms of π_0 is given by (11.31). In order that this may be a probability distribution one must have (11.32). In this case one must take

$$\pi_0 = \left(1 + \sum_{j=1}^{\infty} \frac{\beta_0\beta_1\cdots\beta_{j-1}}{\delta_1\delta_2\cdots\delta_j}\right)^{-1}.$$

■

Proposition 11.9 *(Unrestricted Birth–Death Chain).* Let $S = \{0, \pm1, \pm2, \ldots\}$ and assume $0 < \beta_i, \delta_i < 1$, for all $i \in S$. Then the recurrent and an invariant probability exists and is given by

$$\pi_j = \begin{cases} \frac{\beta_0\beta_1\cdots\beta_{j-1}}{\delta_1\delta_2\cdots\delta_j}\pi_0 & (j \geq 1),\\ \frac{\delta_{j+1}\delta_{j+2}\cdots\delta_0}{\beta_j\beta_{j+1}\cdots\beta_{-1}}\pi_0 & (j \leq -1), \end{cases} \tag{11.34}$$

if and only if

$$\sum_{j\leq -1} \frac{\delta_{j+1}\delta_{j+2}\cdots\delta_0}{\beta_j\beta_{j+1}\cdots\beta_{-1}} < \infty, \qquad \sum_{j\geq 1} \frac{\beta_0\beta_1\cdots\beta_{j-1}}{\delta_1\delta_2\cdots\delta_j} < \infty, \tag{11.35}$$

in which case

$$\pi_0 = \left(1 + \sum_{j\leq -1} \frac{\delta_{j+1}\delta_{j+2}\cdots\delta_0}{\beta_j\beta_{j+1}\cdots\beta_{-1}} + \sum_{j\geq 1} \frac{\beta_0\beta_1\cdots\beta_{j-1}}{\delta_1\delta_2\cdots\delta_j}\right)^{-1}. \tag{11.36}$$

The invariant probability is unique.

Proof. The equations $\mathbf{p}'\boldsymbol{\pi} = \boldsymbol{\pi}$ are

$$\pi_{j-1}\beta_{j-1} + \pi_j(1-\beta_j-\delta_j) + \pi_{j+1}\delta_{j+1} = \pi_j \qquad (j = 0, \pm1, \pm2, \ldots) \tag{11.37}$$

which are uniquely solved in terms of π_0 under the conditions as asserted. ■

Remark 11.2. One may notice that the convergence of the series in (11.35) explicitly implies the divergence of the series in (11.11), (11.12). In other words, one explicitly sees that the existence of an equilibrium distribution for the chain implies its recurrence. The same remark applies to the birth–death chain with one or two reflecting boundaries.

Example 3 *(The Ehrenfest Model of Heat Exchange).* The Ehrenfest model illustrates the process of heat exchange between two bodies that are in contact and insulated from the outside. The temperatures are assumed to change in steps of one unit and are represented by the numbers of balls in two boxes. The two boxes are marked I and II and there are $2d$ balls labeled $1, 2, \ldots, 2d$. Initially some of these balls are in box I and the remainder in box II. At each step a ball is chosen at random (i.e., with equal probabilities among ball numbers $1, 2, \ldots, 2d$) and moved from its box to the other box. If there are i balls in box I, then there are $2d - i$ balls in box II. Thus there is no overall heat loss or gain. Let X_n denote the number of balls in box I after the nth trial. Then $\{X_n : n = 0, 1, \ldots\}$ is a Markov chain with state space $S = \{0, 1, 2, \ldots, 2d\}$ and transition probabilities

$$\begin{aligned} p_{i,i-1} = \frac{i}{2d}, \qquad p_{i,i+1} &= 1 - \frac{i}{2d}, \qquad \text{for } i = 1, 2, \ldots, 2d-1, \\ p_{01} &= 1, \qquad p_{2d,2d-1} = 1, \\ p_{ij} &= 0, \qquad \text{otherwise.} \end{aligned} \tag{11.38}$$

This is a *birth–death chain with two reflecting boundaries* at 0 and $2d$. The transition probabilities are such that the mean change in temperature, in box I, say, at each step is proportional to the negative of the existing temperature gradient, or temperature difference, between the two bodies. We will first see that the model yields *Newton's law of cooling* at the level of the evolution of the averages. Assume that initially there are i balls in box I. Let $Y_n = X_n - d$, the excess of the number of balls in box I over d. Writing $e_n = \mathbb{E}_i(Y_n)$, the expected value of Y_n given $X_0 = i$, one has

$$\begin{aligned} e_n &= \mathbb{E}_i(X_n - d) = \mathbb{E}_i[X_{n-1} - d + (X_n - X_{n-1})] \\ &= \mathbb{E}_i(X_{n-1} - d) + \mathbb{E}_i(X_n - X_{n-1}) = e_{n-1} + \mathbb{E}_i\left(\frac{2d - X_{n-1}}{2d} - \frac{X_{n-1}}{2d}\right) \\ &= e_{n-1} + \mathbb{E}_i\left(\frac{d - X_{n-1}}{d}\right) = e_{n-1} - \frac{e_{n-1}}{d} = \left(1 - \frac{1}{d}\right) e_{n-1}. \end{aligned}$$

Note that in evaluating $\mathbb{E}_i(X_n - X_{n-1})$ we first calculated the conditional expectation of $X_n - X_{n-1}$ given X_{n-1} and then took the expectation of this conditional mean. Now, by successive applications of the relation $e_n = (1 - 1/d)e_{n-1}$,

$$e_n = \left(1 - \frac{1}{d}\right)^n e_0 = \left(1 - \frac{1}{d}\right)^n \mathbb{E}_i(X_0 - d) = (i - d)\left(1 - \frac{1}{d}\right)^n. \tag{11.39}$$

Suppose in the physical model the frequency of transitions is τ per second. Then in time t there are $n = t\tau$ transitions. Write $\nu = \big(-\log(1-\frac{1}{d})\big)\tau$. Then

$$e_n = (i-d)e^{-\nu t}, \tag{11.40}$$

which is *Newton's law of cooling*. The *equilibrium distribution* for the Ehrenfest model is easily seen, using (11.27), to be

$$\pi_j = \binom{2d}{j}2^{-2d}, \qquad j = 0, 1, \dots, 2d. \tag{11.41}$$

That is, $\boldsymbol{\pi} = (\pi_j : j \in S)$ is binomial with parameters $2d, \frac{1}{2}$. Note that $d = \mathbb{E}_\pi X_n$ is the (constant) mean temperature under equilibrium in (11.40).

The physicists P. and T. Ehrenfest in 1907, and later Smoluchowski in 1916, used this model in order to explain an apparent paradox that at the turn of the century threatened to wreck Boltzmann's *kinetic theory of matter*. In the kinetic theory, heat exchange is a random process, while in *thermodynamics* it is an orderly irreversible progression toward equilibrium. In the present context, thermodynamic equilibrium would be achieved when the temperatures of the two bodies became equal, or at least approximately or macroscopically equal. But if one uses a kinetic model such as the one described above, from the state $i = d$ of thermodynamical equilibrium the system will eventually pass to a state of extreme disequilibrium (e.g., $i = 0$) owing to *recurrence*. This would contradict irreversibility of thermodynamics. However, one of the main objectives of kinetic theory was to explain thermodynamics, a largely phenomenological macroscopic-scale theory, starting from the molecular theory of matter.

Historically it was Poincaré who first showed that statistical-mechanical systems have a *recurrence property*. A scientist named Zermelo then forcefully argued that recurrence contradicted irreversibility. Although Boltzmann rightly maintained that the time required by the random process to pass from the equilibrium state to a state of macroscopic nonequilibrium would be so large as to be of no physical significance, his reasoning did not convince other physicists. The Ehrenfest and Smoluchowski finally resolved the dispute by demonstrating how large the passage time may be from $i = d$ to $i = 0$ in the present model.

It follows from (11.27) that the expected return times for the Ehrenfest model are

$$\mathbb{E}(\tau_j(1)|X_0 = j) = \frac{1}{\pi_j} \equiv 2^{2d}\Big/\binom{2d}{j}, \ (j = 0, \dots, 2d). \tag{11.42}$$

To compute the expected time to reach one state from another (e.g., from the "equilibrium state" d to extreme disequilibrium 0), we consider more generally a birth–death-chain on $S = \{0, 1, \dots, N\}$, $N > 2$, with reflecting boundaries $\{0, N\}$. Thus we take $\beta_j + \delta_j = 1$ for all j, $0 < \beta_j < 1$ for $j = 1, \dots, N-1, \beta_0 = 1, \delta_N = 1$. We now turn to the computation of expected times of reaching one state from another. Let $\tau_y = \inf\{n \geq 0 : X_n = y\}$. $m(i) := \mathbb{E}(\tau_0 | X_0 = i)$ for $0 < j \leq N$. For $0 < j < N$, one has $m(j) = 1 + \beta_j m(j+1) + \delta_j m(j-1)$. For $1 < j < N$, this may be expressed as $\beta_j(m(j+1) - m(j)) - \delta_j(m(j) - m(j-1)) = -1$. Multiplying through by $\beta_1 \cdots \beta_{j-1}/\delta_1 \cdots \delta_j$ one obtains

$$\frac{\beta_1 \cdots \beta_j}{\delta_1 \cdots \delta_j}[m(j+1) - m(j)] - \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}}[m(j) - m(j-1)] = -\frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j}. \tag{11.43}$$

Summing these over $j = N-1, \ldots, i$, $(i > 1)$, one obtains

$$\frac{\beta_1 \cdots \beta_{N-1}}{\delta_1 \cdots \delta_{N-1}}[m(N) - m(N-1)] - \frac{\beta_1 \cdots \beta_{i-1}}{\delta_1 \cdots \delta_{i-1}}[m(i) - m(i-1)] = -\sum_{j=i}^{N-1} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j}, \tag{11.44}$$

or, noting that $m(N) = 1 + m(N-1)$, $\delta_N = 1$,

$$m(i) - m(i-1) = \frac{\delta_1 \cdots \delta_{i-1}}{\beta_1 \cdots \beta_{i-1}} \sum_{i=1}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j}, \quad i > 1. \tag{11.45}$$

In particular,

$$m(2) - m(1) = (\delta_1/\beta_1) \sum_{j=2}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j}. \tag{11.46}$$

On the other hand, one has $m(1) = 1 + \beta_1 m(2) + \delta_1 m(0) = 1 + \beta_1 m(2)$, since $m(0) = 0$, so that $\beta_1(m(2) - m(1)) = \delta_1 m(1) - 1$. Using this in (11.44) one obtains

$$m(1) = (\beta_1/\delta_1)(m(2) - m(1)) + 1/\delta_1 = \sum_{j=2}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j} + \frac{1}{\delta_1} = \sum_{j=1}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_j}. \tag{11.47}$$

Summing (11.45) over $i = k, \ldots, 2$, and using (11.47), one obtains

$$\begin{aligned} m(k) &= m(1) + \sum_{i=2}^{k} \frac{\delta_1 \cdots \delta_{i-1}}{\beta_1 \cdots \beta_{i-1}} \sum_{j=i}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j} \\ &= \sum_{j=1}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_j} + \sum_{i=2}^{k} \frac{\delta_1 \cdots \delta_{i-1}}{\beta_1 \cdots \beta_{i-1}} \sum_{j=1}^{N} \frac{\beta_1 \cdots \beta_{j-1}}{\delta_1 \cdots \delta_{j-1}\delta_j}. \end{aligned} \tag{11.48}$$

Next, for a state $0 < d < N$, we calculate $\mathbb{E}(\tau_d | X_0 = j) \equiv \bar{m}(j)$, say. Then (11.43) holds for $\bar{m}(j)$, $0 < j < d$. Also, note the boundary conditions: $\bar{m}(0) - \bar{m}(1) = 1$, and $\bar{m}(d) = 0$, so that for $j = 2, \ldots, d-1$,

$$\frac{\beta_1\cdots\beta_j}{\delta_1\cdots\delta_j}[\bar{m}(j+1)-\bar{m}(j)]-\frac{\beta_1\ldots\beta_{j-1}}{\delta_1\cdots\delta_{j-1}}[\bar{m}(j)-\bar{m}(j-1)]=-\frac{\beta_1\ldots\beta_{j-1}}{\delta_1\cdots\delta_j}, \tag{11.49}$$

and

$$\frac{\beta_1}{\delta_1}[\bar{m}(2)-\bar{m}(1)]+1=-\frac{1}{\delta_1}=-\frac{\beta_0}{\delta_1}. \tag{11.50}$$

Summing over $j = i, i-1, \ldots, 1$, one has, using $\beta_0 = 1$, to write the sum compactly,

$$\frac{\beta_1\cdots\beta_i}{\delta_1\cdots\delta_i}[\bar{m}(i+1)-\bar{m}(i)]+1=-\sum_{j=1}^{i}\frac{\beta_1\cdots\beta_j}{\delta_1\cdots\delta_{j-1}\delta_j}, \tag{11.51}$$

or

$$\bar{m}(i+1)-\bar{m}(i)=\frac{\delta_1\cdots\delta_i}{\beta_1\cdots\beta_i}[-1-\sum_{j=1}^{i}\frac{\beta_1\ldots\beta_{j-1}}{\delta_1\cdots\delta_{j-1}\delta_j}], \tag{11.52}$$

for $i = 1, \ldots, d-1$. Thus summing this over $i = 1, \ldots, d-1$, and recalling $\bar{m}(d) = 0$, one gets

$$\bar{m}(d)-\bar{m}(1)=-\bar{m}(1)=-\sum_{i=1}^{d-1}\frac{\delta_1\cdots\delta_i}{\beta_1\cdots\beta_i}[1+\sum_{j=1}^{i}\frac{\beta_1\ldots\beta_{j-1}}{\delta_1\cdots\delta_{j-1}\delta_j}]. \tag{11.53}$$

This gives the value of $\bar{m}(1)$. Using this in (11.52) (i.e., summing up from $i = k-1, \ldots, 1$) one may obtain $\bar{m}(k)$. In particular, using the boundary condition at 0, one has

$$\begin{aligned}\bar{m}(0)&=1+\bar{m}(1)\\&=1+\sum_{i=1}^{d-1}\frac{\delta_1\cdots\delta_i}{\beta_1\cdots\beta_i}\left[1+\sum_{j=1}^{i}\frac{\beta_1\cdots\beta_{j-1}}{\delta_1\cdots\delta_j}\right].\end{aligned} \tag{11.54}$$

We now apply these computations to the Ehrenfest model where $S = \{0, 1, \ldots, 2d\}$, $\beta_i = \frac{2d-i}{2d}$, $\delta_i = \frac{i}{2d}$ $(i = 1, 2, \ldots, 2d-1)$, $\beta_0 = 1$, $\delta_{2d} = 1$. Letting $k = d$ in (11.48), we get the first sum as

$$\sum_{i=1}^{2d}[(2d-1)\ldots(2d-i+1)/(2d)^{i-1}]/[j!/(2d)^i]=\sum_{i=1}^{2d}\binom{2d}{j}=2^{2d}-1. \tag{11.55}$$

Similarly, the second (double) sum in (11.48) equals $\sum_{i=2}^{d} \frac{(i-1)!(2d-i)!}{(2d)!} \sum_{j=i}^{2d} \binom{2d}{j} = \sum_{i=2}^{d} \binom{2d}{i-1} \sum_{j=i}^{2d} \binom{2d}{j}$. Therefore,

$$m(d) = \mathbb{E}(\tau_0 | X_0 = d) = 2^{2d} - 1 + \sum_{i=2}^{d} [\binom{2d}{i-1}^{-1} \sum_{j=i}^{2d} \binom{2d}{j} > 2^{2d}. \tag{11.56}$$

For the computation of $\mathbb{E}(\tau_d | X_0 = 0)$, (11.54) yields

$$\begin{aligned} \bar{m}(0) &= \mathbb{E}(\tau_d | X_0 = 0) \\ &= 1 + \sum_{i=1}^{d-1} \frac{\delta_1 \cdots \delta_i}{\beta_1 \cdots \beta_i} \left[1 + \sum_{j=1}^{i} \beta_0 \beta_1 \cdots \beta_{j-1} / \delta_1 \cdots \delta_{j-1} \delta_j \right] \\ &= 1 + \sum_{i=1}^{d-1} i!(2d-i-1)!/(2d-1)! \left[1 + \sum_{j=1}^{i} (2d)!/(2d-j)!j! \right] \\ &= 1 + \sum_{i=1}^{d-1} \left[\sum_{j=0}^{i} \binom{2d}{j} \right] / \binom{2d-1}{i}, \end{aligned} \tag{11.57}$$

which is smaller than $1 + d(d-1)/2$. Indeed, a careful calculation shows yields (Exercise 19)

$$\bar{m}(0) = d + d \log d + 0(1) \text{ as } d \to \infty. \tag{11.58}$$

Remark 11.3. For $d = 10\,000$ balls and rate of transition one ball per second, it follows that

$$\begin{aligned} \bar{m}(0) &\le 102\,215 \text{ seconds} < 29 \text{ hours}, \\ m(d) &\qquad > 10^{6000} \text{ years}. \end{aligned} \tag{11.59}$$

Thus it takes only about a day on the average for the system to reach equilibrium from a state farthest from equilibrium but takes an average time inconceivably large, even compared to cosmological scales, for the system to go back to that state from equilibrium. For $d = 10\,000$ one gets, using Stirling's approximation for the second estimate,

$$\mathbb{E}(\tau_0^{(1)} | X_0 = 0) \simeq 2^{20\,000}, \qquad \mathbb{E}(\tau_d^{(1)} | X_0 = d) \simeq 100\sqrt{\pi}. \tag{11.60}$$

Thus, within time scales over which applications of thermodynamics make sense, one would not observe a passage from equilibrium to a (macroscopic) nonequilibrium state. Although Boltzmann did not live to see it, this vindication of his theory ended a rather spirited debate on its validity and contributed in no small measure to its eventual acceptance by physicists.

The spectral representation for $\mathbf{p}$ is left as an Exercise. The $2d$ eigenvalues that one obtains are given by $\alpha_j = j/d$, $j = \pm 1, \pm 2, \dots, \pm d$ (Exercise 18).

***Example 4** (Random Walk on the Hypercube $\mathbb{Z}_2^m$ and the Ehrenfest Model).* Consider the group generated additively modulo 2 by $m+1$ basis elements given by the column vectors e_i having 1 in the ith coordinate and zeros in the remaining $m-1$ coordinates ($i = 1, \dots, m$), together with the identity element given by e_0 and having zeros in all m coordinates. The group operation is (coordinate wise) Euclidean addition modulo 2. The group has 2^m elements, and can be viewed as the m-dimensional hypercube $G = \mathbb{Z}_2^m$. The uniform distribution H (Haar measure) on this group assigns mass 2^{-m} to each element of G. The nearest neighbor random walk on $\mathbb{Z}_2^m$ is a Markov chain $\{X_n : n \geq 0\}$ defined by

$$P(X_{n+1} = x + e_i \ (\text{mod } 2) | X_n = x) = \frac{1}{m+1} \quad (i = 0, 1, \dots, m), x \in \mathbb{Z}_2^m. \tag{11.61}$$

It is simple to check that the m-step transition probability of this Markov chain satisfies $p^{(m)}(x, y) \geq \frac{1}{m^m}$, for all $x, y \in G$, so that, by Doeblin's theorem, one has (Exercise 20)

$$\sup_{x \in \mathbb{Z}_2^m} ||p^{(n)}(x, \cdot) - H||_{TV} \leq (1 - (2/m)^m)^{[n/m]} \sim \exp\left\{-\frac{n}{m}\left(\frac{2}{m}\right)^m\right\}. \tag{11.62}$$

A much improved *cut-off phenomena*, pioneered by Persi Diaconis[1] and colleagues and students, shows the following (Diaconis (1988), p.28):

$$\sup_{x \in \mathbb{Z}_2^m} ||p^{(k)}(x, \cdot) - H||_{TV} \leq \frac{1}{2}(\exp\{e^{-c}\} - 1), \text{ for } k = (1/4)(m+1)(\log m + c), \tag{11.63}$$

and given any $\epsilon > 0$, there exists $C < 0$ such that for $c < C$, with k as in (11.63) and all sufficiently large m,

$$\inf_{x \in \mathbb{Z}_2^m} ||p^{(kn)}(x, \cdot) - H||_{TV} \geq 1 - \epsilon. \tag{11.64}$$

Note that, with c positive and large, the right side of (11.63) can be made arbitrarily small. But if one takes $c < C(< 0)$, C depending on ϵ, then the left side of (11.64) is greater than $1 - \epsilon$ for all sufficiently large m. Thus there is a sharp *cutoff* at $c = 0$ in (11.63), i.e., for k on either side of $\frac{1}{4}(m+1)\log m$. The proof, given by Diaconis

[1] See Diaconis (1996) for historical background.

(1988) loc. cit, uses group representation theory, or Fourier transform on Abelian groups.

Let us now observe that the Markov chain in this example for $m = 2d$ on the hypercube $\mathbb{Z}_2^d$ is the same as the Ehrenfest model, except that the underlying birth–death chain in the present example has a probability of $1/(2d + 1)$ of staying at its present state; referred to as a *lazy random walk.* This corresponds to the possible occurrence of the identity e_0 of the group as an increment. Thus, instead of (11.38), the parameters are

$$\beta_i = p_{i,i+1} = \frac{2d}{2d+1}\frac{2d-i}{2d}, \quad p_{i,i} = \alpha_i = \frac{1}{2d+1}, 0 \le i \le 2d-1,$$
$$\delta_i = p_{i,i-1} = \frac{2d}{2d+1}\frac{i}{2d}, 1 \le i \le 2d, \quad p_{2d,2d} = \alpha_{2d} = \frac{1}{2d+1}. \tag{11.65}$$

This does not change the invariant probability π (See (11.41)). But the Ehrenfest model is periodic with period 2, and therefore convergence in total variation norm to equilibrium only happens separately on the set of odd integers and on the set of even integers, i.e., $p^{(n)}(x, dy)$ does not converge in total variation norm to π. On the other hand, the (lazy) random walk on $\mathbb{Z}_2^{2d}$ is aperiodic. This variant on the Ehrenfest model is the one treated above.

Along these lines, it is also interesting to compare the expected time to the equilibrium state, defined by the average d of π, starting from the farthest nonequilibrium states 0 (or $2d$), with the exponent k for the speed of convergence in the modified model given by $(.86)^{[n/k]}$, where $k = \frac{1}{4}(2d + 1)(\log 2d + 1)$. The latter is also indicative of the order of time at which the steady state equilibrium is reached (approximately).

Exercises

1. Let A_d be the set $\{\omega : X_0(\omega) = y, \{X_n(\omega) : n \ge 0\}$ reaches c before $d\}$, where $y > c$. Show that $A_d \uparrow A = \{\omega : X_0(\omega) = y, \{X_n(\omega) : n \ge 0\}$ ever reaches $c\}$, as $d \uparrow \infty$.
2. Prove (11.12) by using (11.11) and looking at $\{-X_n : n \ge 0\}$.
3. Prove (11.4), (11.17), and (11.24) by conditioning on X_1 and using the Markov property.
4. Suppose that $\varphi(i)$ $(c \le i \le d)$ satisfy the equations (11.4) and the boundary conditions $\varphi(c) = 0$, $\varphi(d) = 1$. Prove that such a φ is unique.
5. Consider a birth–death chain on $S = \{0, 1, \ldots, N\}$ with both boundaries reflecting.

   (a) Prove that $P_i(T_j > mN) \le (1 - \delta_N\delta_{N-1}\cdots\delta_1)^m$ if $i > j$, and $\le (1 - \beta_0\beta_1\cdots\beta_{N-1})^m$ if $i < j$. Here $T_j = \inf\{n \ge 1 : X_n = j\}$.

   (b) Use (i) to prove that $\rho_{ij} \equiv P_i(T_j < \infty) = 1$ for all i, j.

6. Consider a birth–death chain on $S = \{0, 1, \ldots\}$ with 0 reflecting. Argue as in Exercise 5 to show that $\rho_{0y} = 1$ for all y.
7. Consider a birth–death chain on $S = \{0, 1, \ldots, N\}$ with $0, N$ absorbing. Calculate

$$\lim_{n\to\infty} n^{-1} \sum_{m=1}^{n} p_{ij}^{(m)}, \qquad \text{for all } i, j.$$

8. Let 0 be a reflecting boundary for a birth–death chain on $S = \{\ldots, -3, -2, -1, 0\}$. Derive the necessary and sufficient condition for recurrence.
9. If 0 is absorbing, and N reflecting, for a birth–death chain on $S = \{0, 1, \ldots, N\}$, then show that 0 is recurrent and all other states are transient.
10. Let $\mathbf{p}$ be the transition probability matrix of a birth–death chain on $S = \{0, 1, 2, \ldots\}$ with

$$\beta_j = \frac{j+2}{2(j+1)}, \qquad \delta_j = \frac{j}{2(j+1)}, \qquad j = 0, 1, 2, \ldots.$$

    (a) Are the states transient or recurrent?
    (b) Compute the probability of reaching c before d, $c < d$, starting from state $i, c \le i \le d$.

11. Suppose $\mathbf{p}$ is the transition matrix of a birth–death chain on $S = \{0, 1, 2, \ldots\}$ such that $\beta_0 = 1, \beta_j \le \delta_j = 1 - \beta_j$ for $j = 1, 2, \ldots$. Show that all states must be recurrent.
12. Let $\{X_n : n \ge 0\}$ be the asymmetric simple random walk on $S = \{0, 1, 2, \ldots\}$ with $\beta_j = p < \frac{1}{2}$, $j = 1, 2, \ldots$ and (partial) reflection at 0 with $p_{0,0} = p_{0,1} = \frac{1}{2}$.

    (a) Calculate the invariant initial distribution $\boldsymbol{\pi}$.
    (b) Calculate $\mathbb{E}_{\boldsymbol{\pi}} X_n$ as a function of $p < \frac{1}{2}$.

13. Show that the construction of a birth–death process with absorbing boundary at zero in Example 11.1 is a Markov process with the given birth–death probabilities.
14. Show that the construction of a birth–death process with reflecting boundary at zero in Remark 11.1 is a Markov process with the given birth–death probabilities.[*Hint*: Express the (extended) unrestricted transition probabilities as $p_{ij} = \delta_i \delta_{i-1,j} + \beta_i \delta_{i+1,j}, i, j \in \mathbb{Z}$, where $\delta_{\ell,k}$ is the Kronecker delta. See Example 6.]
15. (*A Birth–Death Queue*) During each unit of time either one customer arrives for service and joins a single line or no customers arrive for service. The probability of one customer arriving is λ, and no customer arrives with probability 1 − λ. Also during each unit of time, independently of new arrivals, a single service is completed with probability p or continues into the next period with probability

$1 - p$. Let X_n be the total number of customers (waiting in line or being serviced) at the nth unit of time.

(a) Show that $\{X_n : n \geq 0\}$ is a birth–death chain on $S = \{0, 1, 2, \ldots\}$.
(b) Discuss transience, recurrence, positive recurrence.
(c) Calculate the invariant initial distribution when $\lambda < p$.
(d) Calculate $\mathbb{E}_\pi X_n$ when $\lambda < p$, where $\boldsymbol{\pi}$ is the invariant initial distribution.

16. Suppose that balls labeled $1, \ldots, N$ are initially distributed between two boxes labeled **I** and **II**. The *state* of the system represents the number of balls in box **I**. Determine the one-step transition probabilities for each of the following rules of motion in the state space.

(a) At each time step a ball is randomly (uniformly) selected from the numbers $1, 2, \ldots, N$. Independently of the ball selected, box **I** or **II** is selected with respective probabilities p_1 and $p_2 = 1 - p_1$. The ball selected is placed in the box selected.
(b) At each time step, if possible, a ball is randomly (uniformly) selected from the numbers in box **I** with probability p_1, or from those in **II** with probability $p_2 = 1 - p_1$. If the box selected is empty, then a ball is selected from the other box. A box is then selected with respective probabilities in proportion to number of balls in it, which could be zero. The ball selected is placed in the box selected.
(c) At each time step a ball is randomly (uniformly) selected from the numbers in box **I** with probability proportional to the current size, of **I**, which could be zero, or from those in **II** with the complementary probability. A box is also selected with probabilities in proportion to current number of balls in it. The ball selected is placed in the box selected.

17. Calculate the invariant distribution for Exercise 16(a) where

$$p_{i,j} = \begin{cases} \frac{N-i}{N}\, p_1, & \text{if} j = i+1, \\ \frac{i}{N} p_1 + \frac{(N-i)}{N}\, p_2, & \text{if } j = i, i = 0, 1, \ldots, N, \\ \frac{i}{N}\, p_2, & \text{if } j = i-1, \\ 0, & \text{otherwise.} \end{cases}$$

Discuss the situation for Exercise 16(b) and (c).

18. (*Ehrenfest Model*)

(a) Compute the unique invariant probability π for the Ehrenfest model.
(b) Show that the transition operator T of a birth–death Markov chain on $S = \{0, 1, \ldots, N\}$ with reflecting boundaries is a self-adjoint operator on $L^2(S, \pi)$.
(c) For the Ehrenfest model show that the eigenvalues of the transition operator are $\alpha_j = \frac{j}{d}$, $j = \pm 1, \ldots, \pm d$.

19. Verify the asymptotic formula (11.58).

20. Verify (11.62).
21. (*A Cut-Off Phenomena*) Suppose that a deck of N cards is shuffled by repeatedly taking the top card and inserting it into the deck at a random location. Let G_N be the (nonabelian) group of permutations on N symbols and let $X_1, X_2, \ldots$ be i.i.d. G_N-valued random variables with

$$P(X_k = \langle i, i-1, \ldots, 1\rangle) = 1/N \qquad \text{for } i = 1, 2, \ldots, N,$$

where $\langle i, i-1, \ldots, 1\rangle$ is the permutation in which the card in the ith location from the top moves to $i-1$, $i-1$ to $i-2$, ..., 2 to 1, and 1 to i. Let S_0 be the identity permutation and let $S_n = X_1 \cdots X_n$, where the group operation of composition of maps is being expressed multiplicatively. Let T denote the first time the original bottom card arrives at the top and is inserted back into the deck. Then

(a) T is a stopping time.
(b) T has the additional property that $P(T = k, S_k = g)$ does not depend on $g \in G_N$.
[*Hint*: Show by induction on N that at time $T-1$ the $(N-1)!$ arrangements of the cards beneath the top card are equally likely.]
(c) Property (ii) is equivalent to $P(S_k = g \mid T = k) = 1/|G_N|$; i.e., the deck is mixed at time T.[2]
(d) Show that

$$\max_A \left| P(S_n \in A) - \frac{|A|}{|G_N|} \right| \le P(T > n) \le N e^{-n/N}.$$

[*Hint*: Write $P(S_n \in A) = P(S_n \in A, T \le n) + P(S_n \in A, T > n)$ and condition.]

[2] This property is referred to as the *strong uniform time property* by Aldous and Diaconis (1986), who introduced this example and approach to cut-offs.

Chapter 12
Hitting Probabilities & Absorption

Absorbing boundary conditions can be imposed on (sets of) states of a discrete parameter Markov chain by simply modifying the transition probabilities to make the states inescapable once reached. Calculations of time to absorption, i.e., hitting times of the absorbed states, will be formulated as boundary value problems.

Let $\{X_n : n \geq 0\}$ denote a discrete parameter Markov chain with countable state space S, starting in state $i \in S$. Consideration of the time at which a state or set of states will be reached may also be made as follows. Suppose that $\mathbf{p}$ is a transition probability matrix for $\{X_n : n \geq 0\}$. Let τ_j denote the time required to reach j,

$$\tau_j = \inf\{n : X_n = j\}. \tag{12.1}$$

To calculate the distribution of τ_j, consider that

$$P_i(\tau_j > m) = \sum\nolimits^* p_{ii_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m}, \tag{12.2}$$

where $\sum^*$ denotes summation over all m-tuples $(i_1, i_2, \ldots, i_m)$ of elements from $S \backslash \{j\}$. Now let $\mathbf{p}^0$ denote the matrix obtained by deleting the jth row and jth column from $\mathbf{p}$,

$$\mathbf{p}^0 = ((p_{ik} : i, k \in S \backslash \{j\})). \tag{12.3}$$

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_12

The matrix $\mathbf{p}^0$ is the transition probability law for the *killed process* $\{X_n : n < \tau_j\}$; i.e., the process can only be observed prior to its arrival in state j, where it is removed from the state space S. Then, by definition of matrix multiplication, the calculation (12.2) may be expressed as

$$P_i(\tau_j > m) = \sum_k p_{ik}^{0(m)}, \tag{12.4}$$

and, therefore,

$$P_i(\tau_j = m) = \sum_k p_{ik}^{0(m-1)} - \sum_k p_{ik}^{0(m)}, \qquad m > 1. \tag{12.5}$$

The proof of the following result is left as Exercise 1.

Proposition 12.1. Let $\mathbf{p}$ be a transition probability matrix for a Markov chain $\{X_n : n \geq 0\}$ starting in state i. Let B be a nonempty subset of S, $i \notin B$. Let

$$\tau_B = \inf\{n \geq 0 : X_n \in B\}. \tag{12.6}$$

Then,

$$P_i(\tau_B \leq m) = 1 - \sum_k p_{ik}^{0(m)}, \qquad m = 1, 2, \ldots, \tag{12.7}$$

where $\mathbf{p}^0$ is the matrix obtained by deleting the rows and columns of $\mathbf{p}$ corresponding to the states in B.

While the matrix $\mathbf{p}^0$ is not a proper transition probability matrix on the state space S, if, instead, each of the rows in $\mathbf{p}$ corresponding to states $j \in B$ is replaced by rows e'_j having 1 in the jth place and 0 elsewhere, then the resulting matrix $\hat{\mathbf{p}}$, say, is a proper (stochastic) transition probability matrix and

$$P_i(\tau_B \leq m) = 1 - \sum_{k \notin B} \hat{p}_{ik}^{(m)}. \tag{12.8}$$

The matrix $\hat{\mathbf{p}}$ is the transition probability matrix of the *stopped process* $\{X_{\tau_B \wedge n} : n = 0, 1, 2, \ldots\}$.

The reason (12.8) holds is that up to the first passage time τ_B the distribution of Markov chains having transition probability matrices $\mathbf{p}^0$ and $\hat{\mathbf{p}}$ (starting at i) are the same. In particular,

$$\hat{p}_{ik}^{(m)} = p_{ik}^{0(m)} \qquad \text{for } i, k \notin B. \tag{12.9}$$

Notice that the states belonging to B are absorbing and hence recurrent under $\hat{\mathbf{p}}$. It will be convenient for what follows to consider the operators

$$Tf(i) = \sum_{j \in S} f(j)p_{ij}, \qquad Af(i) = Tf(i) - f(i) = \mathbf{p}f(i) - f(i),\ i \in S, \tag{12.10}$$

for any bounded, or possibly unbounded nonnegative function f on S.

Corollary 12.2. Let $B \subset S$ and suppose that $\mathbf{p} = \hat{\mathbf{p}}$ such that all states in B^c are transient. Then given a nonnegative function f on S which is superharmonic on B^c, i.e., $Tf \leq f$ on B^c, there is a unique pair $g \geq 0, h \geq 0$ such that $Af = -g$ on B^c, $Af = g = 0$ on B, $Ah = 0$ on S, and $f = Gg + h$ on S, where $G = (I - \mathbf{p})^{-1}$ is the Greens function of the (transient) Markov chain restricted to B^c.

Proof. Note that since $\mathbf{p} = \hat{\mathbf{p}}$, any function is harmonic on B in the sense that $Tf(i) = \sum_{j \in S} f(j)p_{ij} = \sum_{j \in S} f(j)\delta_{ij} = f(i)$, for all $i \in B$. Under the stated conditions $G(i, j) < \infty$ for all $i, j \in S$ (see (10.15), (10.17)). Define $g = 0$ on B and $g = -Af$ on B^c. Then, $Gg(i) = \sum_{j \in B^c} G(i, j)g(j)$ is well-defined and $f = Gg$ on B^c. Define $h = \lim_{n \to \infty} T^n f$. The limit exists, since $T^n f = f$ on B for all n, and $0 \leq T^n f \leq f$ is a decreasing nonnegative sequence on B^c, deceasing to zero. As already noted, since $\mathbf{p} = \hat{\mathbf{p}}$, one has $h = f$ on B. The remainder of the proof is left as part of Exercise 1. ■

Proposition 12.3. Within the framework of the above corollary, let $\tau_B = \inf\{n \geq 0 : X_n \in B\}$ and let $f(i) = \mathbb{E}_i \tau_B, i \in S$. Then

$$f \equiv 0 \text{ on } B, \quad Af = -1 \text{ on } B^c. \tag{12.11}$$

Proof. On B^c one has using the Markov property and the fact that $\tau_B = \tau_B(X_1^+) + 1$

$$\begin{aligned} f(i) &= \mathbb{E}_i \tau_B = \mathbb{E}_i \mathbb{E}_i(\tau_B | \sigma(X_0, X_1)) \\ &= \mathbb{E}_i \mathbb{E}_{X_1} \tau_B + 1 = \mathbb{E}_i f(X_1) + 1 = Tf(i) + 1. \end{aligned} \tag{12.12}$$

Thus $Af = -1$ on B^c. Also $f = 0$ on B since $[\tau_B = 0] \supseteq [X_0 = i]$ for $i \in B$. ■

In addition to the *hitting probability* problem for a subset $B \subset S$, it is natural to consider the *hitting distribution* of the state upon arrival in B, namely

$$f_j(i) := P_i(\tau_B < \infty, X_{\tau_B} = j) \qquad (j \in B, i \in S). \tag{12.13}$$

Of course, if $P_i(\tau_B < \infty) < 1$, then X_{τ_B} is a *defective* random variable under P_i, being defined on the set $[\tau_B < \infty]$ of P_i-probability less than 1. For this probability observe that $f_j(i) = \delta_{ij}, i \in B$. On the other hand, if $i \in B^c$, then

$$f_j(i) = \mathbb{E}_i P_i(\tau_B < \infty, X_{\tau_B} = j) = \mathbb{E}_i f_j(X_1) = Tf_j(i), i \in B^c. \tag{12.14}$$

In other words, for fixed $j \in B$, the function $u = f_j$ solves the following exterior Dirichlet problem with $\varphi = \delta_{\cdot j}$ on B.

Definition 12.1. Let $B \subset S$ and suppose let $\varphi : B \to \mathbb{R}$ is a bounded function. A solution to the *(exterior) Dirichlet problem* for T (or for $A = T - I$) is a bounded function u such that $Au(i) = 0, i \in B^c$, and $u(i) = \varphi(i), i \in B$. The term *exterior* is used when B^c is unbounded.

Example 1. Recall that for the one-dimensional simple symmetric random walk starting at $x \in [a, b]$ this is precisely the boundary value problem solved for $u(x) = P_x(\tau_a < \tau_b) = P_x(\tau_{\{a,b\}} < \infty, X_{\tau_{\{a,b\}}} = a)$, with boundary values $u(a) = 1, u(b) = 0$.

The (exterior) Dirichlet problem stipulates that the function should be "harmonic on B^c" with prescribed values on B. Notice that by the Markov property, (condition on X_1 in (12.13)),

$$f_j(i) = \sum_k \hat{p}_{ik} f_j(k) \qquad (j \in B, i \in S). \tag{12.15}$$

Denoting by $\mathbf{f}_j$ the function $(f_j(i) : i \in S)$, one may express (12.15) as follows: For fixed $j \in B$,

$$\mathbf{f}_j = \hat{\mathbf{p}}\mathbf{f}_j. \tag{12.16}$$

Equivalently, $\mathbf{f}_j$ is harmonic on S with respect to the transition matrix $\hat{\mathbf{p}}$ of the stopped process. We have thus proved part (a) of the following proposition.

Proposition 12.4. Let $\mathbf{p}$ be a transition probability matrix and B a nonempty subset of S.

a Then for each $j \in B$, $f_j(i) = P_i(\tau_B < \infty, X_{\tau_B} = j), i \in S$, is a solution to the exterior Dirichlet problem with $\varphi(i) = \delta_{ij}, i \in B$.
b This is the unique bounded solution if and only if

$$P_i(\tau_B < \infty) = 1 \qquad \text{for all } i \in S. \tag{12.17}$$

Proof. Since (a) has been proven it is enough to establish (b). Let $i \in B^c$. Then for $i, k \in B^c$

$$\hat{p}_{ik}^{(n)} \le \sum_{j \in B^c} \hat{p}_{ij}^{(n)} = P_i(\tau_B > n) \downarrow P_i(\tau_B = \infty) \qquad \text{as } n \uparrow \infty. \tag{12.18}$$

Hence, if (12.17) holds, then

$$\lim_{n\to\infty} \hat{p}_{ik}^{(n)} = 0 \qquad \text{for all } i, k \in B^c. \tag{12.19}$$

On the other hand, if $i \in B^c, k \in B$, then

$$\hat{p}_{ik}^{(n)} = P_i(\tau_B \leq n, X_{\tau_B} = k) \uparrow P_i(\tau_B < \infty, X_{\tau_B} = k) = f_k(i). \qquad (12.20)$$

Moreover, $\hat{p}_{ik}^{(n)} = \delta_{ik} \quad$ for all n, if $i \in B, k \in S$. Now, for fixed $j \in B$, let $\mathbf{a}$ be another solution with same values on $B, a(i) = \delta_{ij}, i \in B$, besides $\mathbf{f}_j$. Then $\mathbf{a}$ satisfies (12.16), which on iteration yields $\mathbf{a} = \hat{\mathbf{p}}^n\mathbf{a}$. Taking the limit as $n \uparrow \infty$, and using (12.19), (12.20), one obtains for each $i \in B^c$, using Scheffé's Theorem,

$$\begin{aligned} a(i) &= \lim_{n\to\infty} \sum_k \hat{p}_{ik}^{(n)} a(k) \\ &= \sum_{k\in B} f_k(i)a(k) = f_j(i) \end{aligned} \qquad (12.21)$$

for all $i \in B^c$, since $a(k) = 0$ for $k \in B\backslash\{j\}$ and $a(j) = 1$. Hence $\mathbf{f}_j$ is the unique solution with $f_j(i) = \delta_{ij}, i \in B$. Conversely, if $P_i(\tau_B < \infty) < 1$ for some $i \in B^c$, then the function $\mathbf{h} = (h(i) : i \in S)$ defined by

$$h(i) := 1 - P_i(\tau_B < \infty) = P_i(\tau_B = \infty) \qquad (i \in S), \qquad (12.22)$$

may be checked to be $\mathbf{p}$-harmonic in B^c with (boundary-) value *zero* on B. The harmonic property is a consequence of the Markov property (Exercise 6),

$$\begin{aligned} h(i) = P_i(\tau_B = \infty) &= \sum_k p_{ik} P_k(\tau_B = \infty) \\ &= \sum_k p_{ik} h(k) = \sum_k \hat{p}_{ik} h(k) \qquad (i \in B^c). \end{aligned} \quad (12.23)$$

Since $P_i(\tau_B = 0) = 1$ for $i \in B$, $h(i) = 0$ for $i \in B$. It follows that both $\mathbf{f}_j$ and $\mathbf{f}_j + \mathbf{h}$ satisfy (12.16). Since $\mathbf{h} \neq \mathbf{0}$, the solution of (12.16) is not unique. ■

Corollary 12.5. Assume that $P_i(\tau_B < \infty) = 1$ for all $i \in S$. Then the exterior Dirichlet problem for B and bounded $\varphi : B \to \mathbb{R}$ has the unique (bounded) solution

$$u(i) := \mathbb{E}_i \varphi(X_{\tau_B}). \qquad (12.24)$$

Proof. One may directly verify that $\mathbb{E}_i\varphi(X_{\tau_B}) = \sum_j \varphi(j)P_i(X_{\tau_B} = j) = \sum_j \varphi(j)f_j(i)$ solves the problem from the corresponding result for f_j. Similar limit arguments used in the proof of Proposition 12.4 may be applied to obtain uniqueness. ■

Remark 12.1. One may notice that $\{u(X_{\tau_B \wedge n}) : n = 0, 1, 2 \ldots\}$ is a bounded martingale if u solves the exterior Dirichlet problem for B. This can be used to give an alternative proof of the corollary as follows. $P_i(\tau_B < \infty) = 1$, then with probability one $X_{\tau_B \wedge n} = X_{\tau_B}$ for all n sufficiently large, so that $u(X_{\tau_B \wedge n}) \to u(X_{\tau_B})$ a.s. as $n \to \infty$. Thus, since martingales have constant expected values $u(i) = \mathbb{E}_i u(X_{\tau_B \wedge n}) \to \mathbb{E}_i u(X_{\tau_B})$ as $n \to \infty$ by the dominated convergence theorem.

Remark 12.2. It may seem somewhat surprising given the transience/recurrence dichotomy for random walk according to dimension, but according to the Choquet–Deny theorem the only bounded harmonic functions for the simple symmetric random walk on $\mathbb{Z}^k$ are the constants. The proof is postponed to Chapter 24 where it is provided as an application of *coupling* methods.

We will conclude this section with an absorption probability calculation for a popular model in population genetics.

Example 2 *(Wright–Fisher and Cannings' Gene Frequency Models).* A general class of gene frequency models was identified by Cannings (1973) that includes some of the more well-known models of mathematical biology, such as the Wright–Fisher model which will serve as motivation here.

To reduce the more technical biological[1] jargon, consider a system of $2N$ individual highly capricious voters. Let X_n denote the number of individuals in favor of the issue at times $n = 0, 1, \ldots$. In the evolution, each one of the individuals will randomly re-decide their position under the influence of the current overall opinion as follows. Let $\theta_n = X_n/2N$ denote the proportion in favor of the issue at time n. Then given $X_0, X_1, \ldots, X_n$, each of the $2N$ individuals, independently of the choices of the others, elects to favor the issue with probability θ_n, or oppose it with probability $1 - \theta_n$. That is,

$$P(X_{n+1} = k \mid X_0, X_1, \ldots, X_n) = \binom{2N}{k} \theta_n^k (1-\theta_n)^{2N-k}, \tag{12.25}$$

for $k = 0, 1, \ldots, 2N$. So $\{X_n : n \geq 0\}$ is a Markov chain with state space $S = \{0, 1, \ldots, 2N\}$ and one-step transition matrix $\mathbf{p} = ((p_{ij}))$, where

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \qquad i, j = 0, 1, \ldots, 2N. \tag{12.26}$$

Notice that $\{X_n : n \geq 0\}$ is an aperiodic Markov chain. The "boundary" states $\{0\}$ and $\{2N\}$ form closed classes of essential states. The set of states $\{1, 2, \ldots, 2N-1\}$ constitute an inessential class. One may easily check that absorption is sure to occur (Exercise 14), so the main objective of this example is the calculation of $P_i(\tau_{\{0,2N\}} > m)$. For this we use Cannings' observation that one may express (Exercise 17)

$$X_{n+1} = \sum_{k=1}^{X_n} Y_k, \tag{12.27}$$

[1] For the biological interpretations and approach presented here see Cannings (1974) and Durrett (2008).

where $(Y_1, \ldots, Y_{2N})$ is a $\{0, 1, \ldots, 2N\}^{2N}$-valued random vector having a symmetric multinomial distribution, independently of X_n. That is,

$$P(Y_1 = y_1, \ldots, Y_{2N} = y_{2N}) = \frac{(2N)!}{y_1! \cdots y_{2N}!}\left(\frac{1}{2N}\right)^{2N} \qquad \sum_{k=1}^{2N} y_k = 2N. \tag{12.28}$$

More generally, Cannings (1973) observed that extensions of this class of models to those in which the distribution of $(Y_1, \ldots, Y_{2N})$ is invariant under permutations of the indices, i.e., exchangeable, will include a number of gene frequency models that occur in mathematical biology.

Proposition 12.6. Assume that the random vector $(Y_1, \ldots, Y_{2N})$ has an exchangeable distribution and that $\sum_{k=1}^{2N} Y_k = 2N$ in the model (12.27). Let $\lambda_1, \lambda_2, \ldots, \lambda_{2N}$ denote the eigenvalues of the transition probability matrix $p = ((p_{ij}))$. Then $\lambda_j = \mathbb{E}\prod_{k=1}^{j} Y_k$, $j = 0, 1, 2, \ldots, 2N$. Moreover, $\lambda_0 = \lambda_1 = 1 < \lambda_2$, and $\lambda_j \geq \lambda_{j+1}$, $j = 3, \ldots, 2N - 1$.

Proof. Define a matrix $V = ((v_{ij}))$ by $v_{ij} = i^j$, $(0^0 = 1)$. Then $(pV)_{ij} = \sum_k p_{ik}k^j = \mathbb{E}(X_{n+1}^j | X_n = i)$. On the other hand, one may write (Exercise 16)

$$\mathbb{E}(X_{n+1}^j | X_n = i) = (VU)_{ij} = \sum_{k=0}^{j} i^k u_{kj}, \tag{12.29}$$

where $U = ((u_{ij}))$ is an upper triangular matrix. Therefore, $pV = VU$ and, hence $p = VUV^{-1}$ since V is invertible (Exercise 15). It follows that since $VUV^{-1} - \lambda I = V(U - \lambda I)V^{-1}$, the matrices p and U have the same eigenvalues. Since U is upper triangular, its eigenvalues are simply its diagonal elements. In view of (12.29) the diagonal elements are the coefficients u_{jj} of the highest powers of i in the polynomial expansion for $\mathbb{E}(X_{n+1}^j | X_n = i)$. But expanding $\mathbb{E}(\sum_{k=1}^{j} Y_k)^i$ and using (exchangeability) permutation invariance of $\mathbb{E}Y_1^{j_1} \cdots Y_i^{j_i}$, as well as $\sum_{k=1}^{2N} Y_k = 2N$, one sees that the coefficient of the highest power of i is $\mathbb{E}\prod_{k=1}^{j} Y_k$. To see that the eigenvalues appear in decreasing order, consider

$$\begin{aligned}\mathbb{E}\prod_{i=1}^{j} Y_i &= \mathbb{E}\prod_{i=1}^{j-1} Y_i\left(2N - \sum_{k \neq j} Y_k\right)\\ &= N\mathbb{E}\prod_{i=1}^{j-1} Y_i - (j-1)\mathbb{E}\left(\prod_{i=1}^{j-2} Y_k\right)Y_{j-1}^2 - (2N - j)\mathbb{E}\prod_{i=1}^{j} Y_i.\end{aligned} \tag{12.30}$$

Therefore

$$\begin{aligned}\mathbb{E}\prod_{i=1}^{j} Y_i &= \frac{N\mathbb{E}\prod_{i=1}^{j-1} -(j-1)\mathbb{E}\prod_{i=1}^{j-2} Y_i Y_{j-1}^2}{2N-j+1}\\ &\leq \mathbb{E}\prod_{i=1}^{j-1} Y_i = \lambda_{j-1}. \qquad (12.31)\end{aligned}$$

■

Corollary 12.7 *(Wright–Fisher Fixation Rate).* For the Wright–Fisher model one has[2] $\lambda_0 = \lambda_1 = 1, \lambda_j = \frac{(2N)_j}{(2N)^{2N}}, j = 2, \ldots, 2N.$

$$P_i(\tau_{\{0,2N\}} > m) \sim \left(1 - \frac{i}{2N}\right)^m \quad \text{as } m \to \infty.$$

Proof. Since $2N = \mathbb{E}\sum_{i=1}^{2N} Y_i = 2N\mathbb{E}Y_1$, it follows that $\lambda_1 = \mathbb{E}Y_1 = 1$. For $2 \leq j \leq 2N$ one has with a change of variable and multinomial theorem from algebra,

$$\begin{aligned}\lambda_j &= \mathbb{E}\prod_{k=1}^{j} Y_k\\ &= (2N)^{-2N} \sum_{\sum_{i=1}^{2N} y_i = 2N} \prod_{i=1}^{j} y_i \frac{(2N)!}{\prod_{i=1}^{2N} y_i!}\\ &= \frac{(2N)!}{(2N)^{2N}} \sum_{\sum_{i=1}^{2N} y_i = 2N-j} \frac{1}{\prod_{i=1}^{2N} y_i!}\\ &= \frac{(2N)!}{(2N)^{2N}} \frac{(2N)^{2N-j}}{(2N-j)!} = \frac{(2N)_j}{(2N)^j}.\end{aligned}$$

Thus,

$$\lambda_j = \frac{(2N)_j}{(2N)^j} = \left(1 - \frac{1}{2N}\right)\cdots\left(1 - \frac{j-1}{2N}\right), \qquad 1 \leq j \leq 2N. \qquad (12.32)$$

Noting that $\lambda_0 = \lambda_1 = 1$, the largest nontrivial eigenvalue is $\lambda_2 = \frac{(2N)_2}{(2N)^2} = 1 - \frac{1}{2N}$. To see that this is the fixation rate, first note that since U is upper triangular, the

[2] This computation had been made previously by Feller (1951) by other methods.

diagonal elements of U^m are $u^m_{jj} = \lambda^m_j$. Therefore, writing $V^{-1} = ((v^{ij}))$, one has

$$P_i\left(\tau_{\{0,2N\}} > m\right) = \sum_{j=1}^{2N-1} \hat{p}^{(m)}_{ij} = \sum_{j=1}^{2N-1} \sum_{k=0}^{2N} \lambda^m_k v_{ik} v^{kj} = \sum_{k=0}^{2N} \left(\sum_{j=1}^{2N-1} v_{ik} v^{kj} \right) \lambda^m_k. \tag{12.33}$$

Since the left side of (12.33) must go to zero as $m \to \infty$, the coefficients of $\lambda^m_0 \equiv 1$ and $\lambda^m_1 \equiv 1$ must be zero. Thus,

$$\begin{aligned} P_i(\tau_{\{0,2N\}} > m) &= \sum_{k=0}^{2N} \sum_{j=1}^{2N-1} v_{ik} v^{kj} \lambda^m_k \\ &= \lambda^m_2 \left[\left(\sum_{j=1}^{2N-1} v_{i2} v^{2j} \right) + \sum_{k=3}^{2N} \left(\sum_{j=1}^{2N-1} v_{ik} v^{kj} \right) \left(\frac{\lambda_k}{\lambda_2} \right)^m \right] \\ &\sim (\text{const.}) \lambda^m_2 \qquad \text{for large } m. \end{aligned} \tag{12.34}$$

Taking logarithms one obtains the asserted rate. ■

***Example 3** (Discrete Gaussian Free Field).* The main purpose of this example is to illustrate a role for the two-dimensional simple symmetric random walk on a finite domain in $\mathbb{Z}^2$ with absorbing boundary for the calculation of the covariance of a Gaussian random field arising from physics,[3] variously referred to as the *(discrete) Gaussian free field* (DGFF), the *discrete massless free field*, or the *harmonic crystal*. The DGFF is defined by a Gaussian random field of mean zero real random variables $\Phi_n = \{\Phi_n(x) : x \in \Lambda_n\}$, indexed by a subset $\Lambda_n = [0, n]_0 \times [0, n]_0$ of the two-dimensional integer lattice, and taking zero values on the boundary $\partial\Lambda_n$. (Here $[0, n]_0 = \{0, 1, \ldots, n-1, n\}$, and the boundary $\partial\Lambda_n$ consists of sites $x \in \Lambda_n$ having a nearest neighbor $x \pm u$ outside of Λ_n. The interior of Λ_n is defined by $\Lambda_n \backslash \partial\Lambda_n$.) The distribution of Φ_n is defined as a Gibbs distribution with energy Hamiltonian given by

$$H_n(\varphi) = \frac{1}{16} \sum_{x,y \in \Lambda_n : |x-y|=1} (\varphi_x - \varphi_y)^2. \tag{12.35}$$

More specifically,

***Definition 12.2** (Discrete Gaussian Free Field).* The discrete Gaussian free field (DGFF) on Λ_n with Dirichlet boundary is the random field $\Phi_n = \{\Phi_n(x) : x \in \Lambda_n\}$, indexed by Λ_n with values in the product space $(\mathbb{R}^{\Lambda_n}, \mathcal{B}^{\otimes\Lambda_n})$, such that $\Phi_n(x) = 0$

[3] For the physics and more general related models defined by random walks on finite graphs see the expository papers by Sheffield (2007), and by Berestycki and Powell (2021).

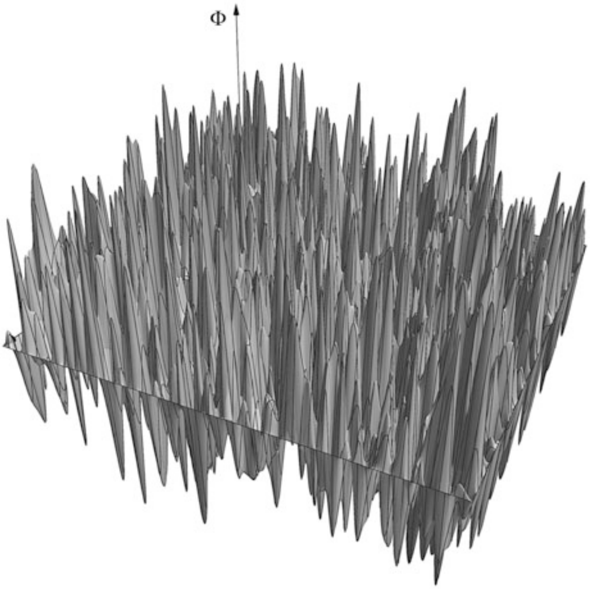


Fig. 12.1 Discrete Gaussian Free Field

for $x \in \partial\Lambda_n$, and having the pdf $Z_n^{-1}\exp\{-H_n(\varphi)\}\mathbf{1}_{\varphi_x=0,x\in\partial\Lambda_n}, \varphi \in \mathbb{R}^{\Lambda_n}$, with respect to Lebesgue measure on $\mathbb{R}^{\Lambda_n}$, where Z_n is the normalization (constant) to a probability.

Remark 12.3. One may picture the discrete Gaussian free field as a model of a pixelated random surface (see Figure[4] 12.1) where $\Phi_n(x)$ is the (signed) height of the pixel location $x \in \Lambda_n$. It serves a role in the analysis of the continuum Gaussian free field model formally analogous to that played by the random walk in the analysis of the limiting Brownian motion.

A simple equivalent definition of the DGFF could be made based on a prescription of the covariance function. For this calculation observe that the energy Hamiltonian may be equivalently expressed in terms of the transition probabilities of a simple symmetric random walk on $\Lambda_n = [0,n]_0 \times [0,n]_0$ with absorbing boundary. These transition probabilities are given by $p(x, x \pm u) = 1/4$ for $x = (i,j), -n < i,j < n, u = (1,0),(0,1)$, and $p(x,x) = 1$ if $x \in \partial\Lambda_n$. Therefore,

$$H_n(\varphi) = \frac{1}{4}\sum_{x,y\in\Lambda_n} p(x,y)(\varphi_x - \varphi_y)^2. \tag{12.36}$$

[4] This figure is a sample realization generated by Eric Roon, using code written by Samuel S. Watson https://math.mit.edu/~sswatson/code.html.

As a consequence the covariance can be obtained as the Greens function G_n of the random walk with Dirichlet boundary as follows.

Proposition 12.8.

$$\mathrm{Cov}(\Phi_n(x), \Phi_n(y)) = \mathbb{E}_x \sum_{m=0}^{\tau_{\partial\Lambda_n}-1} \mathbf{1}_{[S_m=y]} = (I-P)^{-1}_{xy}, \quad x, y \in \Lambda_n \backslash \partial\Lambda_n, \tag{12.37}$$

where $\{S_m : m = 0, 1, \dots\}$ is the random walk, $\tau_{\partial\Lambda_n} = \inf\{m : S_m \in \partial\Lambda_n\}$ denotes the time to reach the (absorbing) boundary, P is the (defective) one-step transition probability matrix for the absorbing random walk with the rows and columns corresponding to absorbing states removed, and I is the identity matrix.

Proof. To see this observe that, ignoring normalization, with a little matrix algebra one may express the quadratic form defining the pdf as

$$\exp\left\{-\frac{1}{16}\sum_{x,y\in\Lambda_n:|x-y|=1}(\varphi_x-\varphi_y)^2\right\} = \exp\left\{-\frac{1}{2}\varphi'(I-P)\varphi\right\}, \tag{12.38}$$

where φ' denotes the transpose of the vector $\varphi = (\varphi_x)_{x\in\Lambda_n}$, P is the (defective) one-step transition probability matrix for the absorbing random walk with the rows and columns corresponding to absorbing states removed, and I is the identity matrix. Thus, $((G_n(x,y))) = (((I-P)^{-1}_{x,y}))$ is the covariance matrix for the (non-normalized) Gaussian pdf (12.38) and, denoting the x, y matrix entry by subscript, one has

$$\begin{aligned}\mathrm{Cov}(\Phi_n(x), \Phi_n(y)) &= (I-P)^{-1}_{x,y}\\ &= \left(\sum_{m=0}^{\infty} P^m\right)_{x,y}\\ &= \sum_{m=0}^{\infty} P(S_m = y, m < \tau_{\partial\Lambda_n} | S_0 = x)\\ &= \mathbb{E}_x \sum_{m=0}^{\tau_{\partial\Lambda_n}-1} \mathbf{1}_{[S_m=y]}.\end{aligned}$$

■

The following is a direct argument that $(I-P)^{-1}$ can serve to define a covariance matrix.

Proposition 12.9. $(I-P)^{-1}$ is symmetric and positive-definite.

Proof. Let $\langle\cdot,\cdot\rangle$ denote the usual Euclidean inner product. While symmetry is obvious from that of $I-P$, for positive-definiteness let $\varphi = (\varphi_x)_{x\in\Lambda_n} \neq 0 \in \mathbb{R}^{\Lambda_n}$, and first note that since

$$\langle (I-P)^{-1}\varphi, \varphi\rangle = \langle (I-P)\tilde{\varphi}, \tilde{\varphi}\rangle,$$

where $\varphi = (I-P)\tilde{\varphi}$, it is sufficient to check positive-definiteness for $I-P$. Now,

$$\langle (I-P)\varphi, \varphi\rangle = |\varphi|^2 - \sum_x \left(\sum_y p(x,y)\varphi_y\right)\varphi_x \tag{12.39}$$

is positive since, by Lyapunov's inequality[5], the double sum is no more than $\sum_x (\sum_y p(x,y)|\varphi_y|^2)^{\frac{1}{2}}|\varphi_x|$ which, in turn, using the Cauchy–Schwarz inequality and symmetry of P, is bounded above by

$$\left(\sum_x \sum_y p(y,x)\varphi_y^2\right)^{\frac{1}{2}} |\varphi| = \left(\sum_y |\varphi_y|^2\right)^{\frac{1}{2}} |\varphi| = |\varphi|^2,$$

since $\sum_x p(y,x) = 1$. ■

An alternative proof is sketched as an Exercise 11.

Remark 12.4. Higher (and lower) dimensional generalizations of this definition are made possible by simply replacing Λ_n by a d-dimensional integer lattice $[0,n]_0^d$, and replacing the two-dimensional simple symmetric random walk accordingly. However, it is also interesting to consider the one-dimensional model obtained by replacing Λ_n by $[0,n]_0$. The *one-dimensional Gaussian free field (discrete)* and its limiting relation to the Brownian bridge[6] is considered in the exercises. (See Exercise 19).

Exercises

1. Prove Proposition 12.1, and complete the proof of Corollary 12.2.
2. Let $\{X_n : n \geq 0\}$ be a two-state Markov chain on $S = \{0, 1\}$ and let τ_0 be the first time $\{X_n : n \geq 0\}$ reaches 0. Calculate $P_1(\tau_0 = n), n \geq 1$, in terms of the parameters p_{10} and p_{01}.
3. Let $\{X_n : n \geq 0\}$ be a three-state Markov chain on $S = \{0, 1, 2\}$ where 0, 1, 2 are arranged counterclockwise on a circle, and at each time a transition occurs one unit clockwise with probability p or one unit counterclockwise with probability $1 - p$. Let τ_0 denote the time of the first return to 0. Calculate $P(\tau_0 > n), n \geq 1$.

[5] BCPT p. 13.

[6] See Bhattacharya and Waymire (2021) for Brownian bridge background.

4. Let τ_0 denote the first time starting in state 2 that the Markov chain in Chapter 6, Exercise 8, reaches state 0. Calculate $P_2(\tau_0 > n)$.
5. Verify that the Markov chains starting at i having transition probabilities $\mathbf{p}$ and $\hat{\mathbf{p}}$, and viewed up to time τ_A have the same distribution by calculating the probabilities of the event $[X_0 = i, X_1 = i_1, \ldots, X_m = i_m, \tau_A = m]$ under each of $\mathbf{p}$ and $\hat{\mathbf{p}}$.
6. Write out a detailed explanation of (12.23).
7. Let $\mathbf{p}$ be the transition probability matrix on $S = \{0, \pm 1, \pm 2, \ldots\}$ defined by

$$p_{ij} = \begin{cases} \frac{1}{i} & \text{if } i > 0, j = 0, 1, 2, \ldots, i-1 \\ \frac{1}{|i|} & \text{if } i < 0, j = 0, -1, -2, \ldots, -i+1 \\ 1 & \text{if } i = 0, j = 0 \\ 0 & \text{if } i = 0, j \neq 0. \end{cases}$$

   (a) Calculate the absorption rate.
   (b) Show that the mean time to absorption starting at $i > 0$ is given by $\sum_{k=1}^{i}(1/k)$.

8. Let $\{X_n : n \geq 0\}$ be the simple branching process on $S = \{0, 1, 2, \ldots\}$ with offspring distribution $\{f_j\}$, $\sum_{j=0}^{\infty} j f_j \leq 1$.

   (a) Show that all nonzero states in S are transient and that $\lim_{n\to\infty} P_1(X_n = k) = 0, k = 1, 2, \ldots$.
   (b) Describe the unique invariant probability distribution for $\{X_n : n \geq 0\}$.

9. Consider the simple symmetric random walk $\{X_n : n \geq 0\}$ on $\mathbb{Z}^k$. Show that for any subset $A \subset \mathbb{Z}^k$ the function $h(x) = P_x(X_n \in A i.o.)$ is a bounded harmonic function (and, hence, constant by the Choquet–Deny theorem 24.1 to be proven in Chapter 24).
10. Let G_n denote the Greens function of the two-dimensional discrete Gaussian free field with Dirichlet boundary. Show, for each fixed $x \in \Lambda_n$, $y \to G_n(x, y)$ solves the Poisson equation

$$\Delta G_n(x, y) = -\delta_{x,y}, y \in \Lambda_n \backslash \partial \Lambda_n \quad G_n(x, y) = 0, y \in \partial \Lambda_n,$$

where Δ denotes the discrete graph-Laplacian $\Delta\varphi(x) = \frac{1}{4}\sum_{y:|x-y|=1}(\varphi_x - \varphi_y)$. [*Hint*: Condition the expected number of visits to y on the first displacement of the random walk and then use a symmetry argument.]

11. Show that the Greens function (matrix) $(I - P)^{-1}$ for the discrete Gaussian free field is positive-definite by defining a weighted inner-product $\langle \varphi, \theta \rangle_x = \sum_y \varphi_y \theta_y p(xy)$ on the real L^2 space of functions φ, θ on Λ_n. Write $(\sum_y p(x, y)\varphi_y)^2 = \langle \varphi, 1 \rangle_x^2$, where 1 is the constant vector on Λ_n with value one. Use the Cauchy–Schwarz inequality to show $(\sum_y p(x, y)\varphi_y)^2 \leq \sum_y \varphi_y^2 p(x, y)$.
12. Consider the generalization of the discrete Gaussian free field with Dirichlet boundary on the d-dimensional lattice with $\Lambda_n = [0, n]_0^d (d \geq 1)$, obtained

by replacing the two-dimensional random walk by the corresponding simple symmetric random walk, as described in Remark 12.4. Show that in the limit as $n \to \infty$, $G_n(x,x) \sim c_1 n$ for $d = 1$, $G_n(x,x) \sim c_2 \ln n$ for $d = 2$, and $G_n(x,x) \sim c_d$ for $d \geq 3$. [*Hint*: Recall Polya's theorem for the simple symmetric random walk.]

13. Show that the Wright–Fisher process $\{X_n : n = 0, 1, \dots\}$ is a nonnegative martingale.
14. Consider the Wright–Fisher model, and let $f_j(i)$ denote the probability of ultimate absorption at $j = 0$ or at $j = 2N$ starting from state $i \in S$.

    (a) Show that f_{2N} satisfies the (exterior) Dirichlet problem: $f_{2N}(i) = \sum_k p_{ik} f_{2N}(k) \qquad \text{for } 0 < i < 2N,$
    $f_{2N}(2N) = 1, \qquad f_{2N}(0) = 0.$
    (b) Derive a similar boundary value problem for f_0.
    (c) Show that $f_{2N}(i) = \frac{1}{2N}$, $f_0(i) = \frac{2N-i}{2N}$, $i = 0, \dots, 2N$,
    (d) Show that $P(\tau_{\{0,2N\}} < \infty) = 1$.

15. Show that the matrix $V = ((i^j))$ is invertible.
16. Show that there is an upper triangular matrix U such that (12.29) holds. Show also that the leading coefficient of the resulting polynomial is $\mathbb{E} \prod_{i=1}^{j} Y_i$. [*Hint*: For a start, compute $\mathbb{E}(X_{n+1}^j | X_n = i)$ in the cases $j = 0, 1, 2$ as a polynomial in i. Use exchangeability and the constraint $\sum_{i=1}^{2N} Y_i = 2N$ to gather together common polynomial coefficients.]
17. Suppose that $(Y_1, \dots, Y_{2N})$ has the symmetric multinomial distribution given by (12.28). Show that for any $1 \leq i \leq 2N$, $\sum_{k=1}^{i} Y_k$ has a binomial distribution with parameters $2N$, $p = \frac{i}{2N}$.
18. (*Alternative Construction of Discrete Gaussian Free Field*) Define a Dirichlet inner product $\langle \varphi_1, \varphi_2 \rangle = \frac{1}{2d} \sum_{x \in \Lambda_n} \nabla\varphi_1(x) \cdot \nabla\varphi_2(x)$, where the ith component of the gradient vector $\nabla\varphi(x)$ is defined as $\varphi(x + u_i) - \varphi(x)$, and u_i the standard unit ith coordinate unit vector. Let $\{h_j : j = 1, 2, \dots, (n+1)^d\}$ be an orthonormal basis for the finite dimensional Hilbert space $H = \{\varphi : \Lambda_n \to \mathbb{R}, \varphi(x) = 0, x \in \partial\Lambda_n\}$ with the Dirichlet inner product. For i.i.d. standard normal random variables $Z_1, Z_2, \dots, Z_{(n+1)^d}$, define $\Phi_n(x) = \sum_{j=1}^{(n+1)^d} Z_j h_j(x)$, $x \in \Lambda_n$. Show that Φ_n is distributed as the discrete Gaussian free field with Dirichlet boundary.
19. (*One-dimensional Gaussian Free Field*) The one-dimensional discrete Gaussian free field with Dirichlet boundary may be defined as the mean zero Gaussian process $\Phi_n = \{\Phi_n(x) : x \in [0,n]_0\}$, where $[0,n]_0 = \{0, 1, 2, \dots, n\}$, such that $\Phi_n(0) = \Phi_n(n) = 0$, and for $x, y \in [0,n]_0 \backslash \{0, n\}$, $\mathrm{Cov}(\Phi_n(x), \Phi_n(y)) = \sum_{m=0}^{\infty} P(S_m = y, m < \tau_{\{0,n\}} | S_0 = x)$, where $\{S_m : m = 0, 1, 2, \dots\}$ is the simple symmetric random walk on $[0,n]_0$ with absorbing boundaries $0, n$.

    (a) Compute the covariance function of Φ_n. [*Hint*: Recall the spectral (eigenvalue) diagonalization in Chapter 11 for simple symmetric random walk with absorption and compute $(I - P)^{-1}_{x,y}$.]

(b) Show that $\{\frac{1}{\sqrt{n}}\Phi_n([xn]) : 0 \le x \le 1\}$ converges weakly to $\sqrt{2}B^{(0)}$, where $B^{(0)}$ is the standard Brownian bridge[7] [*Hint*: Check that conditionally on the event $[\sum_{x=0}^{n-1}(\Phi_n(x+1) - \Phi_n(x)) = 0]$, $\Phi_n(x+1) - \Phi_n(x) : x = 0, 1, \dots, n-1\}$ are i.i.d. Gaussian with mean zero and variance 2.]

(c) Use the Hilbert space construction in Exercise 18 for dimension $d = 1$ to describe the limit in terms of the Lévy-Ciesielski construction[8] of the Brownian bridge in the limit.

(d) Let $\Omega_n = \{\omega : \Lambda_n \to \mathbb{Z} : |\omega_x - \omega_y| = 1, x, y \in \Lambda_n, \omega_x = 0, x \in \partial\Lambda_n\}$. Assume all $\omega \in \Omega_n$ to be equiprobable. Show[9] that in dimensions $d = 1$, the sequence of random fields $\Psi_n(\omega, x) = n^{-\frac{1}{2}}\omega_{[nx]}, x \in [0, 1]_0, \omega \in \Omega_n$ converges in distribution to Brownian bridge.

[7] See Bhattacharya and Waymire (2021).

[8] See BCPT, Chapter IX.

[9] The corresponding weak convergence problem is open for $d = 2$. A limit is expected to exist as a generalized random field on $[0, 1]_0^2$.

Chapter 13
Law of Large Numbers and Invariant Probability for Markov Chains by Renewal Decomposition

The renewal decomposition refers to a decomposition of the sample paths of a countable state Markov chain into i.i.d. recurrent "cycles" of the process between returns to a given state. The long term behavior of the process is then computed in terms of (i.i.d.) averages over these cycles.

For Markov chains with countable state space S, the existence of an invariant distribution is intimately connected with the limiting frequency of returns to recurrent states. For a process in steady state, one expects the equilibrium probability of a state j to coincide with the fraction of time spent on the average by the process in state j. To this effect, a major goal of this chapter is to obtain the invariant distribution as a consequence of a (strong) law of large numbers.

Assume from now on, unless otherwise specified, that S *comprises a single class of (communicating) recurrent states under* $\mathbf{p}$ *for the Markov chain* $\{X_n : n \geq 0\}$. Let f be a real-valued function on S, and define the cumulative sums

$$S_n = \sum_{m=0}^{n} f(X_m) \qquad (n = 1, 2, \ldots). \tag{13.1}$$

For example, if $f(i) = 1$ for $i = j$ and $f(i) = 0$ for $i \neq j$, then $S_n/(n+1)$ is the average number of visits to j in time 0 to n. As in (10.9), let $\tau_j^{(r)}$ denote the time of the rth visit to state j. Write the contribution to the sum S_n from the rth *block of time* $(\tau_j^{(r)}, \tau_j^{(r+1)}]$ as

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_13

$$Z_r = \sum_{m=\tau_j^{(r)}+1}^{\tau_j^{(r+1)}} f(X_m) \qquad (r = 0, 1, 2, \ldots). \tag{13.2}$$

Proposition 13.1. The sequence of random variables $\{Z_1, Z_2, \ldots\}$ is i.i.d., no matter what the initial distribution of $\{X_n : n = 0, 1, 2, \ldots\}$ may be.

Proof. By the strong Markov property, the conditional distribution of the process $\{X_{\tau_j^{(r)}}, X_{\tau_j^{(r)}+1}, \ldots, X_{\tau_j^{(r)}+n}, \ldots\}$, given the past up to time $\tau_j^{(r)}$, i.e., $\mathcal{F}_{\tau_j^{(r)}}$, is P_j, which is the distribution of $\{X_0, X_1, \ldots, X_n, \ldots\}$ under $X_0 = j$. Hence, the conditional distribution of Z_r given the process up to time $\tau_j^{(r)}$ is that of $Z_0 = f(X_1) + \cdots + f(X_{\tau_j^{(1)}})$, given $X_0 = j$. This conditional distribution does not change with the values of X_0, X_1, …, $X_{\tau_j^{(r)}} (= j)$, $\tau_j^{(r)}$. Hence, Z_r is independent of all events that are determined by the process up to time $\tau_j^{(r)}$, i.e., independent of $\mathcal{F}_{\tau_j^{(r)}}$. In particular, Z_r is independent of $Z_1, \ldots, Z_{r-1}$. ■

The decomposition in the preceding proof will be referred to as the *renewal decomposition*. The strong law of large numbers now provides that, with probability 1,

$$\lim_{r\to\infty} \frac{1}{r} \sum_{s=1}^{r} Z_s = \mathbb{E} Z_1, \tag{13.3}$$

provided that $\mathbb{E}|Z_1| < \infty$. In what follows, we will make the stronger assumption that

$$\mathbb{E} \sum_{m=\tau_j^{(1)}+1}^{\tau_j^{(2)}} |f(X_m)| < \infty. \tag{13.4}$$

The objective is to relate the reciprocal of the asymptotic proportion of time spent at a given state j with the average recurrence time of j. We will make use of the following elementary fact along the way.

Lemma 1. If for a sequence of numbers a_r, $(r = 1, 2, \ldots)$, $\frac{1}{N}\sum_{r=1}^{N} a_r$ converges to a finite limit c, say, as $N \to \infty$, then $a_N/N \to 0$ as $N \to \infty$.

Proof. To see this, simply note that

$$\frac{a_N}{N} = \frac{1}{N}\sum_{r=1}^{N} a_r - \frac{1}{N-1}\sum_{r=1}^{N-1} a_r + \frac{1}{N(N-1)}\sum_{r=1}^{N-1} a_r.$$

Clearly, the right side goes to 0 as $N \to \infty$, and thus the left side as well. ■

Theorem 13.2. Assume the condition (13.4), and let N_n denote the *number of visits to state j by time n* given by

$$N_n = \max\left\{r \geq 0 : \tau_j^{(r)} \leq n\right\}. \tag{13.5}$$

Then,

$$\lim_{n\to\infty} \frac{n}{N_n} = \lim_{n\to\infty} \frac{\tau_j^{(N_n)} + n - \tau_j^{(N_n)}}{N_n} = \mathbb{E}\left(\tau_j^{(2)} - \tau_j^{(1)}\right). \tag{13.6}$$

Proof. Start with the decomposition

$$S_n = \sum_{m=0}^{\tau_j^{(1)}} f(X_m) + \sum_{r=1}^{N_n} Z_r - \sum_{m=n+1}^{\tau_j^{(N_n+1)}} f(X_m). \tag{13.7}$$

For each sample path, there are a finite number, $\tau_j^{(1)} + 1$, of summands in the first sum on the right side, except for a set of sample paths having probability zero. Therefore,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{\tau_j^{(1)}} f(X_m) = 0, \qquad \text{with probability 1.} \tag{13.8}$$

The last sum on the right side of (13.7) has at most $\tau_j^{(N_n+1)} - \tau_j^{(N_n)}$ summands, this number being the time between the last visit to j by time n and the next visit to j. Although this sum depends on n, under the condition (13.4), we still have that

$$\left|\frac{1}{n} \sum_{m=n+1}^{\tau_j^{(N_n+1)}} f(X_m)\right| \leq \frac{1}{n} \sum_{m=\tau_j^{(N_n)}+1}^{\tau_j^{(N_n+1)}} |f(X_m)| \to 0 \text{ a.s.} \qquad \text{as } n \to \infty. \tag{13.9}$$

For this, we use Lemma 1. In particular, noting that: (i) $N_n \to \infty$ a.s. as $n \to \infty$, (ii) $\frac{1}{N}\sum_{r=1}^{N}\sum_{m=j^{(r)}+1}^{j^{(r+1)}} |f(X_m)|$ converges to the finite limit given by (13.4), and (iii) $n \geq N_n$, it follows that (13.9) holds. Therefore,

$$\frac{S_n}{n} = \frac{1}{n}\sum_{r=1}^{N_n} Z_r + R_n = \left(\frac{N_n}{n}\right)\frac{1}{N_n}\sum_{r=1}^{N_n} Z_r + R_n, \tag{13.10}$$

where $R_n \to 0$ as $n \to \infty$ with probability 1 under (13.4). Also, for each sample path outside a set of probabilities 0, $N_n \to \infty$ as $n \to \infty$ and therefore by (13.3),

$$\lim_{n\to\infty} \frac{1}{N_n} \sum_{r=1}^{N_n} Z_r = \mathbb{E}Z_1 \tag{13.11}$$

if (13.4) holds. Now, replacing f by the constant function $f \equiv 1$ in (13.11), we have

$$\lim_{n\to\infty} \frac{\tau_j^{(N_n+1)} - \tau_j^{(1)}}{N_n} = \mathbb{E}\left(\tau_j^{(2)} - \tau_j^{(1)}\right), \tag{13.12}$$

assuming that the right side is finite. Similarly, $\frac{\tau_j^{(N_n)}}{N_n} = \frac{N_n-1}{N_n}\frac{1}{N_n-1}\sum_{r=1}^{N_n-1} Z_r + \frac{\tau_j^{(1)}}{N_n} \to \mathbb{E}_j\tau_j^{(1)}$ a.s. as $n \to \infty$. Thus

$$n - \tau_j^{(N_n)} \le \tau_j^{(N_n+1)} - \tau_j^{(N_n)} = o(N_n),$$

a.s. as $n \to \infty$, and one has

$$\lim_{n\to\infty} \frac{n}{N_n} = \lim_{n\to\infty} \frac{\tau_j^{(N_n)} + n - \tau_j^{(N_n)}}{N_n} = \mathbb{E}\left(\tau_j^{(2)} - \tau_j^{(1)}\right). \qquad \blacksquare$$

Remark 13.1. Note that the right side, $\mathbb{E}(\tau_j^{(2)} - \tau_j^{(1)})$, is the average recurrence time $\mathbb{E}_j\tau_j^{(1)}$ of the state j, and the left side is the reciprocal of the asymptotic proportion of time spent at j; here, for simplicity, we write $\mathbb{E}_j(\cdot)$ for $\mathbb{E}(\cdot|X_0 = j)$, without explicitly requiring that the X_j's are coordinate projections of the canonical model.

Definition 13.1. A state j *is positive recurrent if*

$$\mathbb{E}_j\tau_j^{(1)} < \infty. \tag{13.13}$$

We will eventually obtain that positive recurrence is a class property (also see Theorem 13.6). Combining (13.10)–(13.12) gives the following result.

Proposition 13.3. Suppose j is a positive recurrent state under $\mathbf{p}$ and that f is a real-valued function on S such that

$$\mathbb{E}_j\{|f(X_1)| + \cdots + |f(X_{\tau_j^{(1)}})|\} < \infty. \tag{13.14}$$

Then the following are true:

(a) With P_j-probability 1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n} f(X_m) = \mathbb{E}_j(f(X_1)+\cdots+f(X_{\tau_j^{(1)}}))/\mathbb{E}_j\tau_j^{(1)}. \tag{13.15}$$

(b) If S comprises a single class of essential states (irreducible), then the limit in (13.15) holds with probability one regardless of the initial distribution.

Theorem 13.4. Suppose S consists of a single recurrent class of communicating states, and there exists a positive recurrent state j. (a) Then the function π defined by

$$\pi(B) = \mathbb{E}_j\left(\text{Number of visits to } B \text{ during}\left[1,\tau_j^{(1)}\right]\right)/\mathbb{E}_j\tau_j^{(1)}, \quad j\in S, B\subset S, \tag{13.16}$$

is a probability measure on S. Also, whatever the initial state or initial distribution of the Markov chain $\{X_n := 0, 1, \dots\}$, for every function f integrable with respect to π, one has

$$\frac{1}{n}\sum_{m=1}^{n} f(X_m) \to \int_S f d\pi \text{ as } n\to\infty, \tag{13.17}$$

with probability one, and for every bounded f,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}\int f(y)p^{(m)}(i,dy) = \int_S f d\pi, \qquad \text{for all } i\in S. \tag{13.18}$$

(b) This measure π is the unique invariant probability for $\mathbf{p}$.

Proof. (a) First note that π defined by (13.16) is a probability measure on S. Next, it has been shown, see (13.1)–(13.6) and Proposition 13.3, that (13.17) holds for every function f on S that is integrable with respect to π, no matter what the initial state i (or initial distribution) may be. By Lebesgue's dominated convergence theorem, (13.17) implies (13.18) when the initial state is i. (b) Specializing to $f = \mathbf{1}_B$ in (13.18), one has for every $B\subset S$, $i\in S$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} p^{(m)}(i,B) = \pi(B). \tag{13.19}$$

By Proposition 8.10, it now follows that π is the unique invariant probability. ■

Remark 13.2. It will be shown later (Theorem 13.6) that if one state in a recurrent class is positive recurrent, then all states in the class are positive recurrent.

Remark 13.3. In combination with Proposition 16.1, it will follow that the notion of an irreducible positive recurrent Markov chain on a countable state space is equivalent to that of an irreducible stationary ergodic Markov chain, see Corollary 16.2.

Remark 13.4. Under the hypothesis of Theorem 13.4, one has the important formula

$$\pi_j \equiv \pi(\{j\}) = \frac{1}{\mathbb{E}_j \tau_j}, \quad j \in S. \tag{13.20}$$

Also, $\pi_k > 0$ for all k, since the probability of a visit to k starting from j in time $[1, \tau_j^{(1)}]$ is positive.

The next problem is to determine what happens if S comprises a single class of recurrent states that are *not* positive recurrent.

Definition 13.2. A recurrent state j is said to be *null recurrent* if

$$\mathbb{E}_j \tau_j^{(1)} = \infty. \tag{13.21}$$

Proposition 13.5. If j is a null recurrent state, then $\lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} p_{ij}^{(m)} = 0$ for any $i \in S$.

Proof. If $i \nrightarrow j$, then $p_{ij}^{(m)} = 0$ for all m. Assume $i \rightarrow j$. The sequence $\{Z_r : r = 1, 2, \ldots\}$ defined by (13.2) with $f \equiv 1$ is still an i.i.d. sequence of random variables, but the common mean is infinity. It follows from the strong law of large numbers (Exercise 4) that, with P_i-probability 1,

$$\lim_{n\to\infty} \frac{\tau_j^{(N_n)}}{N_n} = \infty.$$

Since $n \geq \tau_j^{(N_n)}$, we have

$$\lim_{n\to\infty} \frac{n}{N_n} = \infty$$

and, therefore,

$$\lim_{n\to\infty} \frac{N_n}{n+1} = 0 \qquad \text{with } P_i - \text{probability } 1. \tag{13.22}$$

Since $0 \leq N_n/(n+1) \leq 1$ for all n, Lebesgue's dominated convergence theorem applied to (13.22) yields

$$\lim_{n\to\infty} \mathbb{E}_i \left(\frac{N_n}{n+1} \right) = 0. \tag{13.23}$$

But

$$\mathbb{E}_i N_n = \mathbb{E}_i \left(\sum_{m=0}^{n} \mathbf{1}_{[X_m = j]} \right) = \sum_{m=1}^{n} p_{ij}^{(m)}, \tag{13.24}$$

and (13.23), (13.24) lead to

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} p_{ij}^{(m)} = 0. \tag{13.25}$$

■

Remark 13.5. Note that (13.18) holds if S comprises a single class of positive recurrent states. Moreover, recall in the case that j is transient, and (10.17) implies that the Green's function $G(i, j) < \infty$, i.e.,

$$\sum_{m=0}^{\infty} p_{ij}^{(m)} < \infty.$$

In particular, therefore,

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0 \qquad (i \in S), \tag{13.26}$$

if j is a transient state.

Example 1. Consider the cyclic two-state Markov chain on $S = \{0, 1\}$ with $p_{01} = p_{10} = 1$, $p_{00} = p_{11} = 0$. If $f : S \to \mathbb{R}$, then observe that almost surely $\tau_j^{(r+1)} - \tau_j^{(r)} = 2$ and $\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n} f(X_m) = \frac{f(0)+f(1)}{2} = \frac{1}{2} f(0) + \frac{1}{2} f(1) = \sum_{i\in S} f(i)\pi_i$, where $\pi_0 = \pi_1 = \frac{1}{2}$ is the unique invariant probability.

The main results of this chapter may be summarized as follows.

Theorem 13.6. Assume that all states communicate with each other. Then one has the following results:

(i) Either all states are recurrent, or all states are transient.

(ii) If all states are recurrent, then they are either all positive recurrent or all null recurrent.

(iii) There exists an invariant distribution if and only if all states are positive recurrent. Moreover, in the positive recurrent case, the invariant distribution $\boldsymbol{\pi}$ is unique and is given by

$$\pi_j = \left(\mathbb{E}_j \tau_j^{(1)} \right)^{-1} \qquad (j \in S). \tag{13.27}$$

(iv) In case the states are positive recurrent and the invariant distribution is π, and $\mathbb{E}_\pi|f(X_1)| < \infty$, then regardless of the initial distribution $\boldsymbol{\mu}$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} f(X_m) = \sum_{i\in S} \pi_i f(i) = \mathbb{E}_\pi f(X_1) \tag{13.28}$$

with $P_{\boldsymbol{\mu}}$-probability 1.

Proof. Part (i) follows from Theorem 10.4. (ii) Let S be a recurrent class of states. If j is positive recurrent, then, by Theorem 13.4, there exists a unique invariant probability π. If possible, suppose k is a null recurrent state in the essential class S, then one has, by Proposition 13.5, $\frac{1}{n}\sum_{m=1}^{n} p_{ik}^{(m)} \to 0$ as $n \to \infty$ for all $i \in S$. Integrating with respect to π, i.e., multiplying by π_i and summing over i, the left side equals π_k for every n, while the right is zero in the limit as $n \to \infty$. This is a contradiction, see Remark 13.4. Hence, all states are positive recurrent if one is. Parts (iii) and (iv) now follow from Theorem 13.4. ■

Remark 13.6. If the assumption that "all states communicate with each other" in Theorem 13.6 is dropped, then S can be decomposed into a set $\mathcal{J}$ of inessential states and (disjoint) classes $S_1, S_2, \ldots, S_t$ of essential states. The transition probability matrix $\mathbf{p}$ may be restricted to each one of the classes $S_1, \ldots, S_t$, and the conclusions of Theorem 13.4 will hold individually for each class. If more than one of these classes is positive recurrent, then more than one invariant distribution exist, and they are supported on disjoint sets. Since any convex combination of invariant distributions is again invariant, an infinity of invariant distributions exist in this case. The following result takes care of the set $\mathcal{J}$ of inessential states in this connection.

Corollary 13.7. Every invariant distribution assigns zero probability to inessential, transient, and null recurrent states.

Proof. Recall that every inessential state is transient, so there is some redundancy in the statement of the corollary. Suppose π is an invariant probability. Use (13.26), (13.25), and invariance of π, and argue as in the proof of part (b) of Theorem 13.6 to conclude that $\pi_j = 0$ if j is either transient or null recurrent. ■

Corollary 13.8. If S is finite, then there exists at least one positive recurrent state and therefore at least one invariant distribution $\boldsymbol{\pi}$. This invariant distribution is unique if and only if all positive recurrent states communicate.

Proof. Suppose that all states are either transient or null recurrent. Then

$$\lim_{n\to\infty} \frac{1}{n+1}\sum_{m=0}^{n} p_{ij}^{(m)} = 0 \qquad \text{for all } i, j \in S. \tag{13.29}$$

Since $(n+1)^{-1}\sum_{m=0}^{n} p_{ij}^{(m)} \leq 1$ for all i, j, and there are only finitely many states j, by Lebesgue's dominated convergence theorem,

$$\begin{aligned}\sum_{j\in S}\lim_{n\to\infty}\left(\frac{1}{n+1}\sum_{m=0}^{n}p_{ij}^{(m)}\right) &= \lim_{n\to\infty}\sum_{j\in S}\left(\frac{1}{n+1}\sum_{m=0}^{n}p_{ij}^{(m)}\right)\\ &= \lim_{n\to\infty}\left(\frac{1}{n+1}\sum_{m=0}^{n}\sum_{j\in S}p_{ij}^{(m)}\right) \qquad (13.30)\\ &= \lim_{n\to\infty}\left(\frac{1}{n+1}\sum_{m=0}^{n}1\right) = \lim_{n\to\infty}\frac{n+1}{n+1} = 1.\end{aligned}$$

But the first term in (13.31) is zero by (13.29). We have reached a contradiction. Thus, there exists at least one positive recurrent state. The rest follows from Theorem 13.3 and the remark following its proof. ■

The approach based on renewal decompositions requires “point recurrence” that is generally not available on general state spaces. So proofs of law of large numbers for Markov processes on general state spaces will require alternative methods. An approach based on Birkhoff’s ergodic theorem will be given in Chapter 16.

Exercises

1. Provide a detailed case-by-case analysis of the summary results given in Theorem 13.6 in the context of the general two-state transition probabilities of Example 1 in Chapter 7.
2. Let **p** be the transition probability matrix for the asymmetric random walk on $S = \{0, 1, 2, \ldots\}$ with 0 absorbing and $p_{i,i+1} = p > \frac{1}{2}$ for $i \geq 1$. Show for fixed $i > 0$,

$$\mu_n(\{j\}) := \frac{1}{n}\sum_{m=1}^{n}p_{ij}^{(m)}, \qquad j \in S,$$

   does *not* converge weakly to the unique invariant probability $\delta_0(\{j\})$ as $n \to \infty$.
3. Let $Y_1, Y_2, \ldots$ be an i.i.d. sequence of nonnegative random variables with $\mathbb{E}Y_1 = \infty$, $S_n = Y_1 + \cdots + Y_n$, and then $S_n/n \to \infty$ a.s. as $n \to \infty$. [*Hint*: Use $S_n \geq Y_n$ and Borel–Cantelli lemma on $P(Y_n > nBi.o.)$ for arbitrary $B > 0$.]
4. Prove

$$\lim_{n\to\infty}\frac{\tau_j^{N_n}}{N_n} = \infty$$

   P_i—a.s. for a null recurrent state j such that $i \leftrightarrow j$.

5. (*General Birth–Collapse*) Let $\mathbf{p}$ be a transition probability matrix on $S = \{0, 1, 2, \ldots\}$ of the form $p_{i,i+1} = p_i$, $p_{i,0} = 1 - p_i$, $i = 0, 1, 2, \ldots, 0 < p_i < 1$, $i \geq 1$, $p_0 = 1$. Show:

   (a) All states are recurrent

$$\text{iff } \lim_{k\to\infty} \prod_{j=1}^{k} p_j = 0 \qquad \text{iff } \sum_{j=1}^{\infty}(1 - p_j) = \infty.$$

   (b) If all states are recurrent, then positive recurrence holds

$$\text{iff } \sum_{k=1}^{\infty} \prod_{j=1}^{k} p_j < \infty.$$

   (c) Calculate the invariant distribution in the case $p_j = 1/(j + 2)$.

6. Calculate the invariant distribution for the *renewal model* of Chapter 7, Exercise 20, in the case that $p_n = p^{n-1}(1 - p)$, $n = 1, 2, \ldots$, where $0 < p < 1$.
7. (*Large Sample Consistency in Statistical Parameter Estimation*) Let $X_n = 1$ or 0 according to whether the nth day at a specified location is *wet* (rain) or *dry*. Assume $\{X_n : n \geq 0\}$ is a two-state Markov chain with parameters $\beta = P(X_{n+1} = 1 \mid X_n = 0)$ and $\delta = P(X_{n+1} = 0 \mid X_n = 1)$, $n = 0, 1, 2, \ldots, 0 < \beta < 1, 0 < \delta < 1$. Suppose that $\{X_n\}$ is in *equilibrium* with the invariant initial distribution $\boldsymbol{\pi} = (\pi_1, \pi_0)$. Define *statistics* based on the sample $X_0, X_1, \ldots, X_n$ to estimate β, π_1, respectively, by $\hat{\pi}_1^{(n)} = S_n/(n + 1)$ and $\hat{\beta}^{(n)} = T_n/n$, where $S_n = X_0 + \cdots + X_n$ is the number of wet days and $T_n = \sum_{k=0}^{n-1} \mathbf{1}_{[(X_k, X_{k+1})=(0,1)]}$ is the number of dry-to-wet transitions. Calculate $\lim_{n\to\infty} \hat{\pi}_1^{(n)}$ and $\lim_{n\to\infty} \hat{\beta}^{(n)}$.
8. (*One-Dimensional Nearest Neighbor Ising Model*) The one-dimensional nearest neighbor Ising model of magnetism consists of a random distribution of ±1-valued random variables (*spins*) at the sites of the integers $n = 0, \pm 1, \pm 2, \ldots$. The parameters of the model are the *inverse temperature* $\beta = \frac{1}{kT} > 0$, where T is the temperature and k is a universal constant called *Boltzmann's constant*, an *external field parameter* H, and an *interaction parameter (coupling constant)* J. The spin variables X_n, $n = 0, \pm 1, \pm 2, \pm 3, \ldots$, are distributed according to a stochastic process indexed by $\mathbb{Z}$ on the state space $\{-1, 1\}$ with the *Markov property* and having stationary transition law given by

$$P(X_{n+1} = \eta \mid X_n = \sigma) = \frac{\exp\{\beta J\sigma\eta + \beta H\eta\}}{2\cosh(\beta H + \beta J\sigma)}$$

   for $\sigma, \eta \in \{+1, -1\}$, $n = 0, \pm 1, \pm 2, \ldots$; by the Markov property is meant that the conditional distribution of X_{n+1} given $\{X_k : k \leq n\}$ does not depend on $\{X_k, k \leq n - 1\}$.

(a) Calculate the unique invariant distribution $\boldsymbol{\pi}$ for $\mathbf{p}$.

(b) Calculate the *large-scale magnetization* (i.e., ability to pick up nails), defined by

$$M_N = [X_{-N} + \cdots + X_N]/(2N + 1),$$

in the so-called bulk (thermodynamic) limit as $N \to \infty$.

(c) Calculate and plot the graph (i.e., *magnetic isotherm*) of $\mathbb{E}X_0$ as a function of H for *fixed temperature*. Show that in the limit as $H \to 0^+$ or $H \to 0^-$, the bulk magnetization $\mathbb{E}X_0$ tends to 0, i.e., there is no (zero) *residual magnetization* remaining when H is turned off at any temperature.

(d) Determine when the process (in equilibrium) is reversible for the invariant distribution.

Chapter 14
The Central Limit Theorem for Markov Chains by Renewal Decomposition

This chapter builds on the renewal decomposition of the previous chapter to obtain a central limit theorem for fluctuations in the i.i.d. cycles under second moment assumptions.

Let us suppose that $\{X_n : n = 0, 1, \ldots\}$ is an irreducible positive recurrent Markov chain on a countable state space S with invariant probability π. For a function $f : S \to \mathbb{R}$ such that $\mathbb{E}_\pi|f(X_0)| < \infty$, one has the SLLN for $S_n/n = \frac{1}{n}\sum_{m=1}^{n} f(X_m) \to \mu = \mathbb{E}_\pi f(X_0)$ a.s. as $n \to \infty$. It is natural to consider the asymptotic fluctuation law for $\sqrt{n}(S_n/n - \mu) = \frac{\bar{S}_n}{\sqrt{n}}$, where $\bar{S}_n = \sum_{m=1}^{n} \bar{f}(X_m) := \sum_{m=1}^{n}(f(X_m) - \mu)$, i.e., f is replaced by $\bar{f} := f - \mu$. We will see that the same "point recurrence (renewal) decomposition" as used in Chapter 12 can be applied to this problem. Specifically, fixing $j \in S$, we will see that a CLT for a (random) sum of i.i.d. random variables $\bar{Z}_1, \bar{Z}_2, \ldots$ may be applied under a finite second moment hypothesis, where

$$\bar{Z}_r = \sum_{m=\tau_j^{(r)}+1}^{\tau_j^{(r+1)}} \bar{f}(X_m), r = 1, 2, \ldots, \tag{14.1}$$

and we assume

$$\sigma_j^2 := \mathbb{E}\bar{Z}_1^2 = \mathbb{E}_j \left|\sum_{m=1}^{\tau_j} f(X_m)\right|^2 < \infty. \tag{14.2}$$

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_14

Observe that $\mathbb{E}\bar{Z}_1 = (\mathbb{E}_j\tau_j^{(1)})\mathbb{E}_\pi \bar{f}(X_0) = 0 \qquad (r = 1, 2, \ldots)$. Thus $\{\bar{Z}_r : r = 1, 2, \ldots\}$ is an i.i.d. sequence with mean zero and finite variance $\sigma_j^2 = E\bar{Z}_1^2$. Now by the classical central limit theorem for i.i.d. sequences with finite second moment, $n \to \infty$, $(1/\sqrt{n})\sum_{r=1}^n \bar{Z}_r$ converges in distribution to the Gaussian law with mean zero and variance σ_j^2. One may easily check that the limit distribution of of $(1/\sqrt{n})\bar{S}_n$ is the same as that of (Exercise 1)

$$\frac{1}{\sqrt{n}}\sum_{r=1}^{N_n}\bar{Z}_r = \left(\frac{N_n}{n}\right)^{1/2}\frac{1}{\sqrt{N_n}}\sum_{r=1}^{N_n}\bar{Z}_r. \tag{14.3}$$

Thus we need an extension of the classical central limit theorem for i.i.d. summands that applies to sums of random numbers of i.i.d. random variables.

Proposition 14.1. Let $\{Y_m : m \geq 1\}$ be i.i.d., $EY_m = 0, 0 < \sigma^2 := EY_m^2 < \infty$. Let $\{\nu_n : n \geq 1\}$ be a sequence of nonnegative integer-valued random variables with

$$\lim_{n\to\infty}\frac{\nu_n}{n} = \alpha \qquad \text{in probability} \tag{14.4}$$

for some constant $\alpha > 0$. Then $\sum_{m=1}^{\nu_n} Y_m/\sqrt{\nu_n}$ converges in distribution to $N(0, \sigma^2)$.

Proof. Without loss of generality, let $\sigma = 1$. Write $S_n := Y_1 + \cdots + Y_n$. Consider that

$$\frac{S_{\nu_n}}{\sqrt{\nu_n}} = \sqrt{\frac{[n\alpha]}{\nu_n}}\left(\frac{S_{[n\alpha]}}{\sqrt{[n\alpha]}} + \frac{S_{\nu_n} - S_{[n\alpha]}}{\sqrt{[n\alpha]}}\right). \tag{14.5}$$

Choose $\epsilon > 0$ arbitrarily. Then, for $0 < a_n \to \infty$ to be determined,

$$\begin{aligned} &P(|S_{\nu_n} - S_{[n\alpha]}| \geq \epsilon([n\alpha])^{1/2}) \\ &\quad \leq P(|\nu_n - [n\alpha]| \geq a_n) + P\left(\max_{\{m:|m-[n\alpha]|<a_n\}}|S_m - S_{[n\alpha]}| \geq \epsilon([n\alpha])^{1/2}\right). \end{aligned}$$

The first term on the right will go to zero as $n \to \infty$ if we take $a_n = cn$, some $c > 0$, by (14.4). The second term is estimated by Kolmogorov's maximal inequality,[1] as being no more than

$$2\left(\epsilon([n\alpha])^{1/2}\right)^{-2} a_n = \epsilon \tag{14.6}$$

[1] See BCPT p.160.

by taking $a_n = \epsilon^3[n\alpha]/2$ (Exercise 1(b)). Thus,

$$\frac{S_{\nu_n} - S_{[n\alpha]}}{([n\alpha])^{1/2}} \to 0 \qquad \text{in probability.} \tag{14.7}$$

Since $S_{[n\alpha]}/([n\alpha])^{1/2}$ converges in distribution to $N(0, 1)$, it follows from (14.7) that so does $S_{\nu_n}/([n\alpha])^{1/2}$. The desired convergence now follows from (14.4). ■

By Proposition 14.1, $N_n^{-1/2}\sum_{r=1}^{N_n}\bar{Z}_r$ is asymptotically Gaussian with mean zero and variance σ_j^2. Since N_n/n converges to $(\mathbb{E}_j\tau_j^{(1)})^{-1} = \pi_j$, it follows that the expression in (14.3) is asymptotically Gaussian with mean zero and variance $(\mathbb{E}_j\tau_j^{(1)})^{-1}\sigma_j^2$. This is then the asymptotic distribution of $n^{-1/2}\bar{S}_n$. Moreover, defining

$$\begin{aligned} W_n(t) &= \frac{\bar{S}_{[nt]}}{\sqrt{n+1}} \\ \tilde{W}_n(t) &= W_n(t) + \frac{1}{\sqrt{n+1}}(nt - [nt]X_{[nt]+1}) \qquad (t \geq 0), \end{aligned} \tag{14.8}$$

all the finite dimensional distributions of $\{W_n(t)\}$, as well as $\{\tilde{W}_n(t)\}$, converge in distribution to those of Brownian motion with zero drift and diffusion coefficient

$$D = \left(\mathbb{E}_j\tau_j^{(1)}\right)^{-1}\sigma_j^2 \tag{14.9}$$

(Exercise 3). In fact, the convergence of the full distribution can also be obtained by consideration of the above renewal argument. The precise form of the FCLT for Markov chains may be stated as follows.

Theorem 14.2 *(Functional Central Limit Theorem (FCLT)).* If S is a positive recurrent class of states and if (14.2) holds, then, as $n \to \infty$, $W_n(1) = (n+1)^{-1/2}\bar{S}_n$ converges in distribution to a Gaussian law with mean zero and variance D given by (14.9). Moreover, the stochastic process $\{W_n(t)\}$ (or $\{\tilde{W}_n(t)\}$) converges in distribution to Brownian motion with zero drift and diffusion coefficient D.

Proof. First consider

$$X_t^{(n)} := \frac{1}{\sigma\sqrt{n}}(\bar{Z}_1 + \cdots + \bar{Z}_{[nt]}).$$

Since $\bar{Z}_1$, $\bar{Z}_2$,... are i.i.d. with finite second moment, the FCLT[2] provides that $\{X_t^{(n)}\}$ converges in distribution to standard Brownian motion. The corresponding result for $\{W_n(t)\}$ follows by an application of the maximal inequality to show

[2] See e.g., BCPT Theorem 11.8.

$$\sup_{0\le t\le 1}\left|X_t^{(n)}-W_{[n\mathbb{E}_j\tau]}(t)\right|\to 0 \qquad \text{in probability as } n\to\infty, \tag{14.10}$$

where τ is the first return time to j. ■

Exercises

1. (a) Let $Y_1, Y_2, \ldots$ be i.i.d. with $\mathbb{E}Y_1^2<\infty$. Show that $\max(Y_1,\ldots,Y_n)/\sqrt{n}\to 0$ a.s. as $n\to\infty$. [*Hint*: Show that $P(Y_n^2>n\epsilon \text{ i.o.})=0$ for every $\epsilon>0$.]
   (b) With the notation (14.1)–(14.8), verify that $n^{-1/2}\bar{S}_n$ has the same limiting distribution as (14.3).
   (c) Use Kolmogorov's maximal inequality to justify (14.6), (14.7) to complete the steps in the proof of Proposition 14.1.
2. Consider the two-state Markov chain $\{X_n : n\ge 0\}$ with $S=\{-1,+1\}$ having transition probabilities $p_{-1,-1}=p_{1,1}=q=1-p$, $p_{-1,1}=p_{1,-1}=p$, with $0<p\le 1$. Calculate the asymptotic variance (as a function of p) for the CLT applied to $\frac{1}{\sqrt{n}}\sum_{m=0}^{n-1}X_m$.
3. Let $\{W_n(t) : t\ge 0\}$ be the path process defined in (14.4). Let $t_1<t_2<\cdots<t_k$, $k\ge 1$, be an arbitrary finite set of time points. Show that $(W_n(t_1,\ldots,W_n(t_k))$ converges in distribution as $n\to\infty$ to the multivariate Gaussian distribution with mean zero and variance–covariance matrix $((D\min\{t_i,t_j\}))$, where D is defined by (14.9).
4. Suppose that $\{X_n : n\ge 0\}$ is a positive recurrent Markov chain with state space $S=\{1,2,\ldots,r\}$ having unique invariant distribution (π_j). Let

$$N_n(i)=\#\{k : X_k=i, 1\le k\le n\}, \qquad i\in S.$$

   (a) Show that

$$\sqrt{n}\left(\frac{N_n(1)}{n}-\pi_1,\ldots,\frac{N_n(r)}{n}-\pi_r\right)$$

   is asymptotically Gaussian under P_π, with mean 0 and variance–covariance matrix $\Gamma=((\gamma_{ij}))$, where

$$\gamma_{ij}=\delta_{ij}\pi_i-\pi_i\pi_j+\sum_{k=1}^{\infty}\left(p_{ij}^{(k)}\pi_i-\pi_i\pi_j\right)$$
$$+\sum_{k=1}^{\infty}\left(p_{ji}^{(k)}\pi_j-\pi_j\pi_i\right), \qquad \text{for } 1\le i,j\le r.$$

[*Hint*: Express $N_n(i) = \sum_{k=0}^{n} \mathbf{1}_{[X_k=i]}$ for the calculations of mean and variance–covariance. For centering the latter, note that $\sum_{k=1}^{n} \sum_{m=1}^{k-1} \pi_i \pi_j = \frac{1}{2} n(n-1)\pi_i \pi_j$.]

(b) Show how this formula reduces in the case of a two-state Markov chain with symmetric transition probabilities. [*Hint*: See Example 1 of Chapter 7 with $a = b = p \in [0, 1)$.]

5. Let $\{X_n\}$ be a Markov chain on S and define $Y_n = (X_n, X_{n+1})$, $n = 0, 1, 2, \ldots$ Let $\mathbf{p} = ((p_{ij}))$ be the transition matrix for $\{X_n\}$.

   (a) Show that $\{Y_n : n \geq 0\}$ is a Markov chain on the state space defined by $S' = \{(i, j) \in S \times S : p_{ij} > 0\}$.
   (b) Show that if $\{X_n : n \geq 0\}$ is irreducible and aperiodic, then so is $\{Y_n : n \geq 0\}$.
   (c) Suppose that $\{X_n : n \geq 0\}$ has invariant distribution $\boldsymbol{\pi} = (\pi_i)$. Calculate the invariant distribution of $\{Y_n : n \geq 0\}$.
   (d) Let $(i, j) \in S'$, and let T_n be the number of one-step transitions from i to j by $X_0, X_1, \ldots, X_n$ started with the invariant distribution $\boldsymbol{\pi}$. Calculate $\lim_{n\to\infty}(T_n/n)$ and describe the fluctuations about the limit for large n.

6. Use the result of Exercise 5 of Chapter 7 to describe an extension of the SLLN and the CLT to certain rth order dependent Markov chains.
7. In reference to Exercise 5, assume that $\{X_n : n \geq 0\}$ is irreducible and positive recurrent with invariant probability π. Let $g \in l^2(S, \pi)$ be arbitrary and consider $f(y) = f(y_1, y_2) := g(y_2) - g(y_1)$. Calculate the asymptotic variance parameter D for $\frac{1}{\sqrt{n+1}} \sum_{m=0}^{n} f(Y_m)$.
8. For the one-dimensional nearest neighbor Ising model of Exercise 8 in Chapter 13, calculate the following:

   (a) The pair correlations $\rho_{n,m} = \mathrm{Cov}(X_n, X_m)$.
   (b) The large-scale variance (*magnetic susceptibility*) parameter $\mathrm{Var}(X_0)$.
   (c) Describe the distribution of the fluctuations in the (bulk limit) magnetization (cf. Chapter 13, Exercise 8(b)).

Chapter 15
Martingale Central Limit Theorem

The martingale central limit theorem provides convergence of suitably centered and scaled sums of martingale difference sequences having finite second moments that encompass a wide range of applications that extend well beyond the classical formulations for i.i.d. summands. The approach is based upon infinitesimal conditions for a stochastic process to be a Gaussian process of interest in their own right.

Let $\{X_{k,n} : 1 \leq k \leq k_n\}$ be, for each $n \geq 1$, a square-integrable martingale difference sequence, with respect to an increasing family of σ-fields $\{\mathcal{F}_{k,n} : 0 \leq k \leq k_n\}$, with $k_n \to \infty$ as $n \to \infty$. Write

$$\begin{aligned}
&\sigma_{k,n}^2 := \mathbb{E}\left(X_{k,n}^2 \mid \mathcal{F}_{k-1,n}\right), \quad s_{k,n}^2 := \sum_{j=1}^{k} \sigma_{j,n}^2, \quad S_{n,k_n} := \sum_{j=1}^{k_n} X_{j,n}, \\
&M_n := \max\left\{\sigma_{k,n}^2; 1 \leq k \leq k_n\right\}, \\
&L_{k,n}(\epsilon) := \sum_{j=1}^{k} \mathbb{E}\left(X_{j,n}^2 \mathbf{1}_{[|X_{j,n}|>\epsilon]} \mid \mathcal{F}_{j-1,n}\right).
\end{aligned} \tag{15.1}$$

Theorem 15.1 *(Brown's Martingale CLT*[1]*).* Assume that, as $n \to \infty$, (i) $s_{k_n,n}^2 \to 1$ in probability and (ii) $L_{k_n,n}(\epsilon) \to 0$ in probability, for every $\epsilon > 0$. Then $S_{k_n,n}$ converges in distribution to $N(0, 1)$.

[1] Brown (1971).

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_15

Proof. Consider the conditional characteristic functions

$$\varphi_{k,n}(\xi) := \mathbb{E}(\exp\{i\xi X_{k,n}\} \mid \mathcal{F}_{k-1,n}), \qquad (\xi \in \mathbb{R}). \tag{15.2}$$

It would be enough to show that:

a $\mathbb{E}\left(\frac{\exp\{i\xi S_{k_n,n}\}}{\prod_1^{k_n}\varphi_{k,n}(\xi)}\right) = 1$, provided $|\prod_1^{k_n}(\varphi_{k,n}(\xi))^{-1}| \leq \delta(\xi)$, a constant.

b $\prod_1^{k_n}\varphi_{k,n}(\xi) \to \exp\{-\xi^2/2\}$ in probability.

Indeed, if $|(\prod_1^{k_n}\varphi_{k,n}(\xi))^{-1}| \leq \delta(\xi)$, then (a), (b) imply

$$\begin{aligned}
&|\mathbb{E}\exp\{i\xi S_{k_n,n}\} - \exp\{-\xi^2/2\}| \\
&\quad = \exp\{-\xi^2/2\}\mathbb{E}\left|\frac{\mathbb{E}\exp\{i\xi S_{k_n,n}\}}{\exp\{-\xi^2/2\}} - \mathbb{E}\left(\frac{\exp\{i\xi S_{k_n,n}\}}{\prod_1^{k_n}\varphi_{k,n}(\xi)}\right)\right| \\
&\quad \leq \exp\{-\xi^2/2\}\mathbb{E}\left|\frac{1}{\exp\{-\xi^2/2\}} - \frac{1}{\prod_1^{k_n}\varphi_{k,n}(\xi)}\right| \to 0.
\end{aligned}$$

Now part (a) follows by taking successive conditional expectations given $\mathcal{F}_{k-1,n}$ $(k = k_n, k_n - 1, \ldots, 1)$, if $\left(\prod_1^{k_n}\varphi_{k,n}(\xi)\right)^{-1}$ is integrable. Note that the martingale difference property is not needed for this. It turns out, however, that in general $\prod_1^{k_n}\varphi_{k,n}(\xi)$ cannot be bounded away from zero. Our first task is then to replace $X_{k,n}$ by new martingale differences $Y_{k,n}$ for which this integrability does hold and whose sum has the same asymptotic distribution as $S_{k_n,n}$. To construct $Y_{k,n}$, first use assumption (b) to check that $M_n \to 0$ in probability. Therefore, there exists a *nonrandom* sequence $\delta_n \downarrow 0$ such that

$$P(M_n \geq \delta_n) \to 0 \qquad \text{as } n \to \infty. \tag{15.3}$$

Similarly, there exists, for each $\epsilon > 0$, a *nonrandom* sequence $\Theta_n(\epsilon) \downarrow 0$ such that

$$P(L_{k_n,n}(\epsilon) \geq \Theta_n(\epsilon)) \to 0 \qquad \text{as } n \to \infty. \tag{15.4}$$

Consider the events

$$A_{k,n}(\epsilon) := \left[\sigma_{k,n}^2 < \delta_n, L_{k,n} < \Theta_n(\epsilon), s_{k,n}^2 < 2\right], \qquad (1 \leq k \leq k_n). \tag{15.5}$$

Then $A_{k,n}(\epsilon)$ is $\mathcal{F}_{k-1,n}$-measurable. Therefore, $Y_{k,n}$ defined by

$$Y_{k,n} := X_{k,n}\mathbf{1}_{A_{k,n}(\epsilon)} \tag{15.6}$$

has zero conditional expectation, given $\mathcal{F}_{k-1,n}$. Although $Y_{k,n}$ depends on ϵ, we will suppress this dependence for notational convenience. Note also that

$$\begin{aligned} P(Y_{k,n} = X_{k,n} \text{ for } 1 \le k \le k_n) &\ge P\left(\bigcap_{k=1}^{k_n} A_{k,n}(\epsilon)\right) \\ &= P(M_n < \delta_n, L_{k_n,n}(\epsilon) < \Theta_n(\epsilon), s^2_{n,k_n} < 2) \to 1. \end{aligned} \tag{15.7}$$

We will use the notation (15.1–15.2) with a "*tilde*" symbol to denote the corresponding quantities for $\{Y_{k,n}\}$. For example, using the fact $\mathbb{E}(Y_{k,n} \mid \mathcal{F}_{k-1,n}) = 0$ and a Taylor expansion,

$$\begin{aligned} &\left|\tilde{\varphi}_{k,n}(\xi) - \left(1 - \tfrac{\xi^2}{2}\,\tilde{\sigma}^2_{k,n}\right)\right| \\ &\quad = \mathbb{E}\left|\mathbb{E}\left[\exp(i\xi Y_{k,n}) - \left(1 + i\xi Y_{k,n} + \frac{(i\xi)^2}{2}\,Y^2_{k,n}\right) \mid \mathcal{F}_{k-1,n}\right]\right| \\ &\quad = \mathbb{E}\left|\mathbb{E}\left[-\xi^2 Y^2_{k,n}\int_0^1 (1-u)(\exp\{iu\xi Y_{k,n}\} - 1)du \mid \mathcal{F}_{k-1,n}\right]\right| \\ &\quad \le \epsilon\,\frac{|\xi|^3}{2}\,\tilde{\sigma}^2_{k,n} + \xi^2\mathbb{E}(Y^2_{k,n}\mathbf{1}_{[|Y_{k,n}|>\epsilon]} \mid \mathcal{F}_{k-1,n}) \\ &\quad \le \epsilon\,\frac{|\xi|^3}{2}\,\tilde{\sigma}^2_{k,n} + \xi^2\mathbb{E}(X^2_{k,n}\mathbf{1}_{[|Y_{k,n}|>\epsilon]} \mid \mathcal{F}_{k-1,n}). \end{aligned} \tag{15.8}$$

Fix $\xi \in \mathbb{R}^1$. Since $\tilde{M}_n < \delta_n$, $0 \le 1 - (\xi^2/2)\tilde{\sigma}^2_{k,n} \le 1$ $(1 \le k \le k_n)$ for all large n. Therefore, using (15.5, 15.8),

$$\begin{aligned} \left|\prod_1^{k_n}\tilde{\varphi}_{k,n}(\xi) - \prod_1^{k_n}\left(1 - \frac{\xi^2}{2}\,\tilde{\sigma}^2_{k,n}\right)\right| &\le \sum_{k=1}^{k_n}\left|\tilde{\varphi}_{k,n}(\xi) - \left(1 - \frac{\xi^2}{2}\,\tilde{\sigma}^2_{k,n}\right)\right| \\ &\le |\xi|^3\epsilon + \xi^2\Theta_n(\epsilon), \end{aligned} \tag{15.9}$$

and

$$\begin{aligned} &\mathbb{E}\left|\prod\nolimits_1^{k_n}\left(1 - \tfrac{\xi^2}{2}\,\tilde{\sigma}^2_{k,n}\right) - \exp\left\{-\tfrac{\xi^2}{2}\,\tilde{s}^2_{k,k_n}\right\}\right| \\ &\quad = \mathbb{E}\left|\prod_1^{k_n}\left(1 - \frac{\xi^2}{2}\,\tilde{\sigma}^2_{k,n}\right) - \prod_1^{k_n}\exp\left\{-\frac{\xi^2}{2}\,\tilde{\sigma}^2_{k,n}\right\}\right| \\ &\quad = \frac{\xi^4}{8}\sum\tilde{\sigma}^4_{k,n} \le \frac{\xi^4}{8}\,\delta_n\tilde{s}^2_{n,k_n} \le \frac{\xi^4}{4}\,\delta_n. \end{aligned} \tag{15.10}$$

Therefore,

$$\left|\prod_1^{k_n}\tilde{\varphi}_{k,n}(\xi) - \exp\left\{-\frac{\xi^2}{2}\,\tilde{s}_{n,k_n}\right\}\right| \le \mathbb{E}|\xi|^3\epsilon + \xi^2\Theta_n(\epsilon) + \frac{\xi^4}{4}\,\delta_n. \tag{15.11}$$

Moreover, (15.11) implies

$$
\begin{aligned}
|\textstyle\prod_1^{k_n}\tilde{\varphi}_{k,n}(\xi)| &\geq \exp\left\{-\frac{\xi^2}{2}\,\tilde{s}^2_{n,k_n}\right\} - \mathbb{E}(|\xi|^3\epsilon + \xi^2\Theta_n(\epsilon) + \frac{\xi^4}{4}\,\delta_n) \\
&\geq \exp\{-\xi^2\} - |\xi|^3\epsilon - \mathbb{E}\left(\xi^2\Theta_n(\epsilon) + \frac{\xi^4}{4}\,\delta_n\right).
\end{aligned} \tag{15.12}
$$

By choosing ϵ sufficiently small, one has for all sufficiently large n (depending on ϵ), $|\prod_1^{k_n}\tilde{\varphi}_{k,n}(\xi)|$ is bounded away from zero (uniformly in n). Therefore, (a) holds for $\{Y_{k,n}\}$, for all sufficiently small ϵ (and all sufficiently large n, depending on ϵ). By using relations as in (15.3) and the inequalities (15.11, 15.12) and the fact that $\tilde{s}^2_{n,k_n} \to 1$ in probability, we get

$$
\begin{aligned}
&\overline{\lim}_{n\to\infty}\left|\mathbb{E}\exp\{i\xi\tilde{S}_{n,k_n}\} - \exp\left\{-\frac{\xi^2}{2}\right\}\right| \\
&\quad\leq \exp\left\{-\frac{\xi^2}{2}\right\}\overline{\lim}_{n\to\infty}\,\mathbb{E}\left|\left(\exp\left\{-\frac{\xi^2}{2}\right\}\right)^{-1} - \left(\prod_1^{k_n}\tilde{\varphi}_{k,n}(\xi)^{-1}\right)\right| \\
&\quad\leq \exp\left\{-\frac{\xi^2}{2}\right\}\exp\left\{\frac{\xi^2}{2}\right\}\left(\exp\left\{-\frac{\xi^2}{2}\right\} - |\xi|^3\epsilon\right)^{-1} \\
&\overline{\lim}_{n\to\infty}\,\mathbb{E}\left|\prod_1^{k_n}\tilde{\varphi}_{k,n}(\xi) - e^{-\xi^2/2}\right| \\
&\quad\leq \left(\exp\left\{-\frac{\xi^2}{2}\right\} - |\xi|^3\epsilon\right)^{-1}|\xi|^3\epsilon.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\overline{\lim}_{n\to\infty}\left|\mathbb{E}\exp\{i\xi S_{k_n,n}\} - \exp\left\{-\frac{\xi^2}{2}\right\}\right| \\
&\quad\leq \overline{\lim}_{n\to\infty}\,|\mathbb{E}\exp\{i\xi S_{k_n,n}\} - \mathbb{E}\exp\{i\xi\tilde{S}_{k_n,n}\}| \\
&\quad+ \overline{\lim}_{n\to\infty}\left|\mathbb{E}\exp\{i\xi\tilde{S}_{k_n,n}\} - \exp\left\{-\tfrac{\xi^2}{2}\right\}\right| \\
&\quad= 0 + \overline{\lim}_{n\to\infty}\left|\mathbb{E}\exp\{i\xi\tilde{S}_{k_n,n}\} - \exp\left\{-\tfrac{\xi^2}{2}\right\}\right| \\
&\quad\leq \left(\exp\left\{-\tfrac{\xi^2}{2}\right\} - |\xi|^3\epsilon\right)^{-1}|\xi|^3\epsilon.
\end{aligned}
$$

The extreme right side of (15.13) goes to zero as $\epsilon \downarrow 0$, while the extreme left does not depend on ϵ. ■

The classical Lindeberg CLT is an immediate consequence of Theorem 15.1 (Exercise 1). Condition (ii) of the theorem is called the *conditional Lindeberg condition*.

We next derive an important versatile theorem[2] whose consequences include (a) an interesting property of a class of Markov processes known as diffusions. What is also remarkable is that, with the verification of its criteria and a tightness condition, Theorem 15.2 also yields (b) the Billingsley–Ibragimov functional central limit theorem (FCLT) for square-integrable martingales with stationary increments over equidistant intervals (Theorem 15.5). The latter result plays an important role in the derivation of the FCLT for functions of ergodic Markov processes in the next chapter. It may be emphasized that Donsker's FCLT and invariance principle are immediate offshoots of (b), whose proof here does not require the use of the central limit theorem, neither for the classical case for i.i.d. sequences nor for the martingale CLT derived above (Theorem 15.1). Indeed, the latter CLTs just follow as simple consequences!

We now provide infinitesimal conditions for a process to be a Gaussian process with time-dependent mean and variance function. The original idea goes back to Khinchin (1933) and was made use of by Rosén (1967) and Billingsley (1968).

Let $\{X(t) : 0 \le t \le T\}$ be a real-valued stochastic process on $(\Omega, \mathcal{F}, P)$ having continuous sample paths. Assume that $\beta(t)$ and $\sigma(t)$ are continuous on $[0, T]$, and for all $0 \le t_1 < t_2 < \cdots < t_k \le t \le T$, and real numbers $u_1, u_2, \ldots, u_k$, one has

$$\lim_{h\downarrow 0}(1/h)\mathbb{E}\left\{\exp\left(\sum_{1\le j\le k} iu_jX(t_j)\right)[X(t_k+h)-X(t_k)-h\beta(t_k)X(t_k)]\right\}=0, \tag{15.13}$$

$$\lim_{h\downarrow 0}(1/h)\mathbb{E}\left\{\exp\left(\sum_{1\le j\le k} iu_jX(t_j)\right)[(X(t_k+h)-X(t_k))^2-h\sigma^2(t_k)]\right\}=0. \tag{15.14}$$

Further, suppose

$$\sup_{0\le t\le T}\mathbb{E}X^2(t)<\infty. \tag{15.15}$$

Also either assume: (i) there exists K such that, for all $t_1 < t_2 < t_3$,

$$(i)\ \mathbb{E}(X(t_2)-X(t_1))^2(X(t_3)-X(t_2))^2\le K(t_3-t_1)^2, \tag{15.16}$$

or, for every t.

$$(ii)\ \lim_{\lambda\to\infty}\limsup_{h\downarrow 0}\frac{1}{h}\int_{[(X(t+h)-X(t))^2>\lambda h]}(X(t+h)-X(t))^2dP=0. \tag{15.17}$$

For fixed $0 \le t_1 < t_2 < \cdots < t_k$ and $(u_1, u_2, \ldots u_k)$, consider the characteristic function of $(X(t_1), \ldots, X(t_k), X(t))$, as a function of $t \ge t_k$ and u given by

[2] See Rosén (1967), Billingsley (1968) for original versions.

$$\begin{aligned} y(t,u) &= \mathbb{E}\exp\{iu_1X(t_1)+\cdots+iu_kX(t_k)+iuX(t)\} \\ &= \mathbb{E}\exp\{iZ+iuX(t)\}, \end{aligned} \tag{15.18}$$

say. One has

$$\frac{\partial}{\partial u}y(t,u)=\mathbb{E}(iX(t)\exp\{iZ+iuX(t)\}). \tag{15.19}$$

Theorem 15.2. Under the above hypotheses (15.13)–(15.17), and notation,

$$y(t,u)=\mathbb{E}\exp\{iZ+iua(t)X(t_k)\}\exp\{-(1/2)u^2b^2\}, \tag{15.20}$$

where

$$a(t)=\exp\left\{\int_{[t_k,t]}\beta(s)ds\right\},\quad b^2(t)=\int_{[t_k,t]}\sigma^2(r)\exp\left\{2\int_{[r,t]}\beta(s)ds\right\}dr, \tag{15.21}$$

which implies that $Y(t)\equiv X(t)-a(t)X(t_k)$ is normally distributed with mean zero and variance $b^2(t)$ and independent of $(X(t_1),\ldots X(t_k))$.

Proof. The main idea of the proof is to derive the equation

$$\frac{\partial}{\partial t}y(t,u)=u\beta(t)\frac{\partial}{\partial u}y(t,u)-(1/2)u^2\sigma^2(t)y(t,u)\quad (t\geq t_k) \tag{15.22}$$

and solve it. To derive (15.22), note that its left side is the limit, as $h\downarrow 0$, of

$$\begin{aligned} &(1/h)\mathbb{E}\exp\{iZ+iuX(t)\}[\exp\{iuX(t+h)-iuX(t)\}-1] \\ &\quad=(1/h)\mathbb{E}\exp\{iZ+iuX(t)\}[iu(X(t+h)-X(t)] \\ &\qquad+(1/2)(iu(X(t+h)-X(t)))^2)+R(t,h)], \end{aligned} \tag{15.23}$$

say, where the remainder is estimated, using $|\exp(iux)-1-iux+u^2x^2/2|\leq\min\{u^2x^2,|u^3x^3|\}$, as

$$\begin{aligned} |R(t,h)| &\leq (1/h)\mathbb{E}(\min\{|u(X(t+h)-X(t))|^2,|u(X(t+h)-X(t))|^3\}) \\ &\leq (1/h)\mathbb{E}\{|u(X(t+h)-X(t))|^3\mathbf{1}[u(X(t+h)-X(t))|^2\leq\lambda h]\} \\ &\quad+(1/h)\mathbb{E}|u(X(t+h)-X(t))|^2\mathbf{1}[u(X(t+h)-X(t))|^2>\lambda h] \\ &\leq |u^3\lambda^{3/2}h^{1/2}|+(1/h)u^2\mathbb{E}(X(t+h)-X(t))^2\mathbf{1} \\ &\qquad[|X(t+h)-X(t)|^2>\lambda h]. \end{aligned}$$

By (15.17ii), $R(t, h)$ goes to zero as $h \downarrow 0$ and then $\lambda \to \infty$. By assumptions (15.13), (15.14), one now has the limit of (15.23) given by (15.22). It remains to solve (15.22), written as

$$\frac{\partial}{\partial t} y(t, u) - u\beta(t)\frac{\partial}{\partial u} y(t, u) = -(1/2)u^2\sigma^2(t)y(t, u). \tag{15.24}$$

The left side is the directional derivative of $y(t, u)$ in the direction $(1, -u\beta(t))$ in the (t, u)-plane. Letting $\alpha(t : v) = v \exp\{-\int_{[t_k,t]} \beta(s)ds\}$, $v \in \mathbb{R}$. The directional derivative of y along the characteristic curve $(t, \alpha(t : v))$, along with (15.24), yields

$$\frac{d}{dt} y(t, \alpha(t : v)) = (\partial/\partial t)y(t, w)|_{w=\alpha(t:v)} - \frac{\partial}{\partial w} y(t, w)|_{w=\alpha(t:v)}$$

$$\beta(t)\alpha(t : v) = -(1/2))\alpha(t : v)^2\sigma^2(t)y(t, \alpha(t : v)),$$

$$\frac{d}{dt} \log y(t, \alpha(t : v)) = -(1/2)\alpha(t : v)^2\sigma^2(t), \tag{15.25}$$

which on integration over $[t_k, t]$ yields, with $y(t_k, v) = \mathbb{E} \exp\{iZ + ivX(t_k)\}$ (see (15.18)),

$$y(t, \alpha(t : v)) = y(t_k, v) \exp\left\{ - (1/2) \int_{[t_k,t]} \alpha(s : v)^2\sigma^2(s)ds \right\}. \tag{15.26}$$

Now choose $v = u \exp\{\int_{[t_k,t]} \beta(s)ds\} = ua(t)$, so that $\alpha(t : v) = ua(t)$. Then (15.26) reduces to

$$\begin{aligned} y(t, u) &= y(t_k, ua(t)) \exp\left\{ - (u^2/2) \int_{[t_k,t]} a^2(s)\sigma^2(s)ds \right\} \\ &= \mathbb{E} \exp\{iZ + iua(t)X(t_k)\} \exp\{-u^2b^2/2\}. \end{aligned} \tag{15.27}$$

This completes the proof. ■

Corollary 15.3. Under the hypothesis of Theorem 15.2, $\{X(t) : 0 \le t \le T\}$, is a Gaussian process.

Proof. First take $k = 0$, i.e., $t_1 = \cdots = t_k = 0 < t$. This proves $X(t)$ is normal for every $t > 0$. Next let $k = 1, 0 < t_1 < t$. Then $X(t_1)$ and $X(t) - a(t)X(t_1)$ are independent with $X(t) - a(t)X(t_1)$ normal. Together with the case $k = 0$, this implies $X(t_1)$ and $X(t)$ are jointly normal. Continuing in this way (or, by induction), one shows that $X(t_1), \ldots X(t_k), X(t)$ are jointly normal. ■

Our next task is to derive an asymptotic version (limit theorem) of Theorem 15.2. Let $\{X_n(t) : 0 \le t \le T\}$ be a sequence of processes with continuous sample paths satisfying the following asymptotic versions of (15.13)–(15.17).

$$\lim_{h\downarrow 0}\limsup_{n\to\infty}(1/h)|\mathbb{E}\exp\left(\sum_{1\le j\le k} iu_jX_n(t_j)\right)[X_n(t_k+h)-X_n(t_k)-h\beta(t_k)X_n(t_k)]|=0, \tag{15.28}$$

$$\lim_{h\downarrow 0}\limsup_{n\to\infty}(1/h)|\mathbb{E}\exp\left(\sum_{1\le j\le k} iu_jX_n(t_j)\right)[(X_n(t_k+h)-X_n(t_k))^2-h\sigma^2(t_k)]|=0. \tag{15.29}$$

Further, suppose

$$\sup_{0\le t\le T}\limsup_{n\to\infty}\mathbb{E}X_n^2(t)<\infty. \tag{15.30}$$

Also assume either: (i) there exists K such that for all $t_1<t_2<t_3$, and all n,

$$(i)\ \mathbb{E}(X_n(t_2)-X_n(t_1))^2(X_n(t_3)-X_n(t_2))^2\le K(t_3-t_1)^2, \tag{15.31}$$

or, for every t,

$$(ii)\ \limsup_{h\downarrow 0}\limsup_{n\to\infty}\mathbb{E}_\lambda[(X_n(t+h)-X_n(t))^2/h]\to 0 \text{ as } \lambda\to\infty, \tag{15.32}$$

where $\mathbb{E}_\lambda U=\mathbb{E}(U\mathbf{1}[U>\lambda])$.

Theorem 15.4. Let $\{X_n(t):0\le t\le T\}$ be a sequence of processes with continuous sample paths satisfying (15.28)–(15.32). If, in addition, their distributions on $C[0,T]$ are a tight sequence, then they converge in distribution to a continuous process $\{X(t):0\le t\le T\}$, satisfying the conclusions of Theorem 15.2.

Proof. Because of tightness, there exists a process X in $C[0,T]$, such that a subsequence $\{X_{n'}\}$, say, of $\{X_n\}$, converges in distribution to X. One may now check that the conditions (15.28)–(15.32) imply that X satisfies all the assumptions (15.13)–(15.17) of Theorem 15.2 (Exercise 2). Therefore, X_n converges in distribution to a process X characterized by Theorem 15.2. ■

We are now ready to prove the Billingsley–Ibragimov FCLT for martingales.

Theorem 15.5 *(FCLT for Martingales with Stationary Increments).* Let $\{Z_n : n=1,2,\dots\}$ be a sequence of stationary ergodic square-integrable martingale differences, i.e., denoting by $\mathcal{F}_n$ the sigma-field generated by $\{Z_1,\dots,Z_n\}$, one has

$$\mathbb{E}(Z_n|\mathcal{F}_{n-1})=0,\quad \mathbb{E}(Z_n^2)=\sigma^2>0 \text{ for all } n=1,2,\dots \tag{15.33}$$

Then, writing $S_n=Z_1+\cdots+Z_n(n\ge 1)$, $S_0=0$, the polygonal process $X_n(t)=S_{[nt]}/\sqrt{n}+(nt-[nt])Z_{[nt]+1}/\sqrt{n}$ $(t\ge 0)$ converges to a Brownian motion with mean zero and variance parameter σ^2.

Proof. First, let us construct a doubly infinite sequence $\{\tilde{Z}_n : -\infty < n < \infty\}$ such that $\{\tilde{Z}_n : j+1 \leq n \leq j+m\}$ has the same distribution as $\{Z_1, Z_2, \ldots, Z_m\}$, $(-\infty < j < \infty, m = 1, 2, \ldots)$. This being a consistent specification, Kolmogorov's existence theorem provides a doubly stationary sequence whose consecutive m terms have the same joint distribution as that of $(Z_1, Z_2, \ldots, Z_m)$, for every $m \geq 1$. With a minor abuse of notation, let us denote this doubly infinite sequence also as $\{Z_n\}$. Let $\mathcal{G}_n$ be the sigma-field generated by $\{Z_j : -\infty < j \leq n\}$ for all integers n, positive, negative, or zero. One may now check that

$$\mathbb{E}(Z_{n+1}|\mathcal{G}_n) = 0, \quad \text{for all } n. \tag{15.34}$$

By stationarity, $\mathbb{E}(Z_{n+1}|\sigma\{Z_{n+1-j} : j = 1, \ldots, m\}) = \mathbb{E}(Z_{m+1}|\sigma\{Z_1, \ldots, Z_m\}) = 0$. This being true for all $m = 1, 2, \ldots,$ one arrives at (15.34). We will now verify the hypotheses of Theorem 15.4 with $\beta(t) = 0$ and $\sigma^2(t) = \sigma^2$ for all t. In order to verify the hypothesis of Theorem 15.4, note that

$$\begin{aligned}
&X_n(t_k + h) - X_n(t_k) \\
&\quad = \sum_{[nt_k] < j < [n(t_k+h)]} Z_j/\sqrt{n} + (n(t_k + h) - [n(t_k + h)])Z_{[n(t_k+h)]+1}/\sqrt{n} \\
&\quad -(nt_k - [nt_k])Z_{[n(t_k+h)]+1}/\sqrt{n} - (nt_k - [nt_k])Z_{[nt_k]+1}/\sqrt{n} \\
&\quad = \sum_{[nt_k] < j < [n(t_k+h)]+1} Z_j/\sqrt{n} + \mathcal{E}(n, h),
\end{aligned} \tag{15.35}$$

where $\mathbb{E}\mathcal{E}^2(n, h)$ goes to zero as $n \to \infty$, uniformly for all h. Then, recalling $\mathcal{G}_{[nt]} = \sigma\{Z_j : j \leq [nt]\}$, one has $\mathbb{E}(\sum_{[nt_k]<j<[nt_k+h]} Z_j/\sqrt{n}|\mathcal{G}_{[nt_k]}) = 0$, so that (15.28) holds. To verify (15.29), fix an $h > 0$. In view of (15.35), it is enough to prove

$$\mathbb{E}(1/h)\Bigg(\sum_{[nt_k] < j < [n(t_k+h]} Z_j/\sqrt{n})^2|\mathcal{G}_{[nt_k]}\Bigg) \to \sigma^2, \tag{15.36}$$

in L^1 as $n \to \infty$. Using (15.34), the left side of (15.36) equals

$$\mathbb{E}\Bigg(\sum_{[nt_k] < j < [n(t_k+h)]} Z_j^2/nh|\mathcal{G}_{[nt_k]}\Bigg), \tag{15.37}$$

which has the same distribution as the sequence $\mathbb{E}(\sum_{0<j<m(n)} Z_j^2/[nh]|\mathcal{G}_0)$, where $m(n) = [n(t_k + h)] - [nt_k]$ differs from $[nh]$ by at most 2. Hence, (15.37) differs in distribution from $\mathbb{E}(\sum_{0<j<[nh]} Z_j^2/[nh]|\mathcal{G}_0)$ by a quantity that goes to zero in L^1. But by Birkhoff's mean ergodic theorem (see Theorem 4.1), the latter sum divided by nh converges in L^1 to $\mathbb{E}(Z_1^2) = \sigma^2$. Thus, (15.29) follows, with $\sigma^2 = \sigma^2(t_k)$.

As to (15.30), one has $\sup_n \mathbb{E}X_n(t)^2 \leq (T+1)\sigma^2$. We now proceed to the proof of (15.32) and tightness of the sequence $\{X_n(t) : 0 \leq t \leq 1\}$. First, consider the estimate of $\mathbb{E}S_n^4$ assuming $|Z_j|$ is bounded by a constant C. Of the n^4 terms in the expansion, the expectation vanishes of every term in which the term with the highest index k, say Z_k, occurs once as one of the four factors. The remaining terms yield $\mathbb{E}S_n^4 = \sum_k \mathbb{E}Z_k^4 + 4\sum_{j<k} \mathbb{E}Z_j Z_k^3 + 6\sum_{i,j<k} \mathbb{E}Z_i Z_j Z_k^2$. The number of summands in the first two terms together is no more than $n + 4n(n-1)/2 < 2n^2$. The third term is bounded by

$$6C^2 \sum_{k>1} \mathbb{E}(S_{k-1}^2) = 6C^2\sigma^2 \sum_{k>1} (k-1) = 3C^4 n(n-1) < 3C^4 n^2.$$

Hence,

$$\mathbb{E}(S_n^4) < 5C^4 n^2. \tag{15.38}$$

One may use this estimate to prove tightness[3] and also the condition (15.32) for bounded Z_k (Exercise 3). To prove them under the second moment condition, we will use the following truncation argument. For every nonnegative random variable U and every $\lambda > 0$, let $\mathbb{E}_\lambda U = \mathbb{E}(U\mathbf{1}[U > \lambda])$. Define the truncated variable $Z_{k,u} = Z_k \mathbf{1}[Z_k \leq u]$, and $Y_{k,u} = Z_{k,u} - \mathbb{E}(Z_{k,u}|\mathcal{G}_{k-1})$, and the remainder $W_{k,u} = Z_k - Y_{k,u} = Z_k - Z_{k,u} - \mathbb{E}(Z_k - Z_{k,u}|\mathcal{G}_{k-1})$, since $\mathbb{E}(Z_k|\mathcal{G}_{k-1}) = 0$. Both sequences $\{Y_{k,u}\}$, $\{W_{k,u}\}$ are martingale differences. Their respective partial sums are denoted by $S_{k,u} = \sum_{1\leq j\leq k} Y_{j,u}$ and $R_{k,u} = \sum_{1\leq j\leq k} W_{j,u}$. We will prove that

$$\lim_{\lambda\to\infty} \mathbb{E}_\lambda (1/n) \max_{1\leq k\leq n} S_k^2 = 0. \tag{15.39}$$

The tightness of the sequence of processes $\{X_n\}$ will then follow from Lemma 3 below. Observe the following inequalities, for some $\Delta > 0$,

$$\begin{aligned}
&(1/n) \max_{1\leq k\leq n} S_k^2 \leq (2/n) \max_{1\leq k\leq n} S_{k,u}^2 + (2/n) \max_{1\leq k\leq n} R_{k,u}^2;\\
&\mathbb{E}_\Delta (1/n) \max_{1\leq k\leq n} S_{k,u}^2 \leq (1/\Delta)\mathbb{E}(1/n^2) \max_{1\leq k\leq n} S_{k,u}^4 \leq (1/\Delta)(4/3)^4 5(2u)^4.
\end{aligned} \tag{15.40}$$

The second relation above follows from a Chebyshev type inequality, while the last inequality is just Doob's maximal inequality for 4th moments using (15.38) for the martingale $\{S_{k,u}\}$, recalling that $\mathbb{E}|Y_{k,u}| \leq 2u$. Also, using Doob's maximal inequality for second moments for the martingale $\{R_{k,u}\}$

[3] See BCPT Lemma 3, p. 150.

$$\begin{aligned}\mathbb{E}\frac{1}{n}\max_{1\le k\le n} R_{k,u}^2 &\le \frac{4}{n}\mathbb{E}R_{n,u}^2\\ &= 4\mathbb{E}(Z_1 - Z_{1,u} - \mathbb{E}(Z_1 - Z_{1,u}|\mathcal{G}_{k-1}))^2 \le 4\mathbb{E}(Z_1 - Z_{1,u})^2\\ &= 4\mathbb{E}_u Z_1^2. \end{aligned}\tag{15.41}$$

From (15.40) and (15.41), and Billingsley's inequality (Exercise 4) $\mathbb{E}_\Delta(U + V) \le 2(\mathbb{E}_{\Delta/2}U + \mathbb{E}_{\Delta/2}V)$, one obtains $\mathbb{E}_\Delta(1/n)\max_{1\le k\le n} S_k^2 \le 8\mathbb{E}_{u^2}Z_1^2 + Au^4/\Delta$, where A is an absolute constant. Let $u = \Delta^{1/8}$ to get the desired result (15.39), with $\lambda = \Delta$. Theorem 15.5 now follows from Lemma 3 below, since

$$P(\max_{1\le j\le n} |S_{k+j} - S_j| \ge \lambda\sigma\sqrt{n}) \le \mathbb{E}_{\lambda^2}\left(\frac{1}{n}\max_{1\le j\le n} S_j^2/\sigma^2\right). \tag{15.42}$$

This completes the proof. ■

The following are useful tools for establishing tightness under quite general conditions.[4] For the first two lemmas, let $\{P_n : n \ge 1\}$ be a sequence of probability measures on $(C[0,1], \mathcal{B})$. Then the Arzelá–Ascoli theorem provides a useful identification of compact sets in $C[0,1]$ for purposes of checking tightness as follows.[5]

Lemma 1. The sequence $\{P_n : n \ge 1\}$ is tight if:

a For each $\eta > 0$, there is an a such that $P_n(\omega \in C[0,1] : |\omega(0)| > a) \le \eta, \quad n \ge 1$.

b For each $\epsilon > 0, \eta > 0$, there exist δ, $0 < \delta < 1$, and integer n_0 such that $P_n(\omega \in C[0,1] : w_\omega(\delta) := \sup_{|s-t|<\delta} |\omega(s) - \omega(t)| \ge \epsilon) \le \eta, \quad n \ge n_0$.

Remark 15.1. One may use the Arzelá–Ascoli theorem to show that the conditions in Lemma 1 are also necessary for tightness.

Lemma 2. The sequence $\{P_n : n \ge 1\}$ is tight if:

a For each $\eta > 0$, there is an a such that

$$P_n(\omega \in C[0,1] : |\omega(0)| > a) \le \eta, \quad n \ge 1.$$

b For each $\epsilon > 0, \eta > 0$, there exist a $0 < \delta < 1$, and integer n_0 such that for all $0 \le t \le 1$,

$$P_n(\omega \in C[0,1] : \sup_{t\le s\le (t+\delta)\wedge 1} |\omega(s) - \omega(t)| \ge \epsilon) \le \delta\eta, \quad n \ge n_0.$$

[4] See Billingsley (1968) for a more comprehensive treatment of such conditions.

[5] For a proof, see BCPT (Errata), p. 149, Lemma 2.

Proof. Fix $\delta > 0$, and define $A_t = \{\omega \in C[0,1] : \sup_{t \le s \le t+\delta\wedge 1} |\omega(s) - \omega(t)| \ge \epsilon\}$. Each s, t belong to an interval of the form $[j\delta, (j+1)\delta]$. So, in particular, if $|s-t| < \delta$, then the two such intervals either coincide or abut. Hence, $P_n(\omega \in C[0,1] : w_\omega(\delta) \ge 3\epsilon) \le P_n(\cup_{j \le 1/\delta} A_{j\delta})$. Now, condition (b) implies that $P_n(\omega \in C[0,1] : w_\omega(\delta) \ge 3\epsilon) \le (1 + [\delta^{-1}])\delta\eta < 2\eta$. Lemma 1 applies to complete the proof. ■

For the case that P_n is the distribution of a polygonal process, let $Z_1, Z_2, \ldots$ be an arbitrary sequence of real-valued random variables on a probability space $(\Omega, \mathcal{F}, P)$, $S_n = Z_1 + \cdots + Z_n, n \ge 1$, $S_0 = 0$, $\sigma > 0$ a constant, and define

$$X_n(t,\omega) = \frac{1}{\sigma\sqrt{n}} S_{[nt]}(\omega) + (nt - [nt])\frac{1}{\sigma\sqrt{n}} Z_{[nt]+1}(\omega). \tag{15.43}$$

Lemma 3. The sequence $X_n \in C[0,1]$ defined by (15.43) is tight if for each $\epsilon > 0$ there is a $\lambda > 1$, and integer n_0 such that for $n \ge n_0$

$$P(\max_{j \le n} |S_{k+j} - S_k| \ge \lambda\sigma\sqrt{n}) \le \frac{\epsilon}{\lambda^2},\ k = 0, 1, 2, \ldots$$

Proof. Since $X_n(0) = 0$, the tightness of (distributions of) $\{X_n(0)\}$ is trivial. In view of Lemma 2, it is sufficient to show for $\epsilon > 0, \eta > 0$, there is $0 < \delta < 1$ and integer n_0 such that for $0 \le t \le 1, n \ge n_0$,

$$P(\sup_{t \le s \le (t+\delta)\wedge 1} |X_n(s) - X_n(t)| \ge \epsilon) \le \delta\eta. \tag{15.44}$$

For $t = k/n$ and $t + \delta = j/n$, i.e., t, δ integral multiples of $1/n$, this is the same as requiring

$$P\left(\max_{i \le \delta n} \frac{1}{\sigma\sqrt{n}} |S_{k+i} - S_k| \ge \epsilon\right) \le \delta\eta. \tag{15.45}$$

More generally, if $k/n \le t < (k+1)/n$ and $(j-1)/n \le t + \frac{1}{2}\delta < j/n$, the polygonal path satisfies

$$\sup_{t \le s \le t + \frac{1}{2}\delta} |X_n(s) - X_n(t)| \le 2 \max_{0 \le i \le j-k} \frac{1}{\sigma\sqrt{n}} |S_{k+i} - S_k|. \tag{15.46}$$

Thus, taking $n \ge 4/\delta$, one has $j - k \le n\delta$, so that the maximum on the upper bound is not larger than the maximum over $i \le \delta n$. Thus, for tightness of the sequence (of distributions) of $\{X_n\}$, it suffices to check for $\epsilon, \eta > 0$, there is a $0 < \delta < 1$, and integer n_0, such that (15.45) holds for all k and all $n \ge n_0$. Note that if $\delta n = m$ is an integer, then (15.45) simplifies to

$$P(\max_{i\leq m} |S_{k+i} - S_k| \geq \epsilon\sigma\sqrt{m}/\sqrt{\delta}) \leq \delta\eta. \tag{15.47}$$

Let $\lambda = \epsilon/\sqrt{\delta}$. Then one has the further simplification

$$P(\max_{i\leq m} |S_{k+i} - S_k| \geq \lambda\sigma\sqrt{m}) \leq \frac{\eta\epsilon^2}{\lambda^2}. \tag{15.48}$$

With $\eta\epsilon^2$ in place of ϵ, the condition of the lemma yields $\lambda > 0$ and n_1 such that for $n \geq n_1, k \geq 1$,

$$P(\max_{i\leq n} |S_{k+i} - S_k| \geq \lambda\sigma\sqrt{n}) \leq \frac{\eta\epsilon^2}{\lambda^2}.$$

Write $\delta = \frac{\epsilon^2}{\lambda^2} \in (0, 1)$. Take an integer $n_0 \geq n_1/\delta$, so that $[n\delta] \leq n_1$. Then it now follows for $n \geq n_0$, and hence, $[\delta n] \geq n_1$,

$$P(\max_{i\leq[\delta n]} |S_{k+i} - S_k| \geq \lambda\sigma\sqrt{[\delta n]}) \leq \frac{\eta\epsilon^2}{\lambda^2}.$$

Since $\lambda\sqrt{[\delta n]} \leq \epsilon\sqrt{n}$, and $\eta\epsilon^2/\lambda^2 = \delta\eta$, the desired condition (15.45) is satisfied. ■

Remark 15.2. As mentioned before, the statement of Theorem 15.2, Donsker's Theorem (and, therefore, the classical CLT), for sums of i.i.d. random variables is an immediate consequence of Theorem 15.5, whose proof depends on Prokhorov's theorem on tightness[6] that, in the present context, depends on Doob's maximal inequalities. Paul Lévy's martingale characterization of Brownian motion also follows, but with the additional assumption of stationarity (Exercise 5). The result without stationarity can be established using stochastic differential equations.[7]

Exercises

1. Show how each of the following classical limit theorems for independent random variables follows from the more general martingale central limit theory.

   (a) (Lindeberg CLT) For each n, $X_{1,n}, \ldots, X_{k_n,n}$ are independent random variables such that $\mathbb{E}X_{j,n} = 0, \sigma^2_{j,n} = \mathbb{E}X^2_{j,n} < \infty, \sum_{j=1}^{k_n} \sigma^2_{j,n} = 1$, and $k_n \to \infty$ as $n \to \infty$. Assume the

[6] See the Lemma 3, or BCPT, Theorem 7.11, p. 145.

[7] See Ikeda and Watanabe (1981), Theorem 6.1, p.74.

$$[Lindeberg\ Condition] \qquad \lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}(X_{j,n}^2 \mathbf{1}[|X_{j,n}| > \epsilon]) = 0 \tag{15.49}$$

for each $\epsilon > 0$. Then $\sum_{j=1}^{k_n} X_{j,n}$ converges in distribution to the standard normal as $n \to \infty$.

(b) (Lyapunov CLT) For each n, $X_{1,n}, \ldots, X_{k_n,n}$ are independent random variables such that $\sum_{j=1}^{k_n} X_{j,n} = \mu$, $\sum_{j=1}^{k_n} Var X_{j,n} = \sigma^2 > 0$. Assume that

$$[Lyapunov\ Condition] \qquad \lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}|X_{j,n} - \mathbb{E}X_{j,n}|^{2+\delta} = 0, \tag{15.50}$$

for some $\delta > 0$. Then $\sum_{j=1}^{k_n} X_{j,n}$ converges in distribution to the mean μ and variance σ^2 normal distribution as $n \to \infty$.

(c) (Classical I.I.D. CLT) Assume $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with $\mathbb{E}X_1^2 < \infty$. Let $\mu = \mathbb{E}X_1, \sigma^2 = Var X_1$. Then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$ converges in distribution to the standard normal distribution as $n \to \infty$.

2. In reference to the proof of Theorem 15.4, verify that the conditions (15.28)–(15.32) imply that X satisfies all the assumptions (15.13)–(15.17) of the theorem.
3. Use this estimate (15.38) to prove tightness (see BCPT, Lemma 3, p. 150) and also the condition (15.32) for bounded Z_k in the proof of Theorem 15.5.
4. (Billingsey's Inequality) Show that for nonnegative random variables X, Y, and $t > 0$,

$$\mathbb{E}\{(X+Y)\mathbf{1}_{[X+Y\geq t]}\} \leq 2\mathbb{E}\{X\mathbf{1}_{[X\geq \frac{t}{2}]}\} + 2\mathbb{E}\{Y\mathbf{1}_{[Y\geq \frac{t}{2}]}\}, \quad t > 0.$$

[*Hint*: Use $[X + Y > t] \subset [X > t/2] \cup [Y > t/2]$ and inclusion–exclusion to see that $\mathbf{1}_{[X+Y>t]} \leq \mathbf{1}_{[X>t/2]} + \mathbf{1}_{[Y>t/2]}$. Then consider the cases $[X > t/2, Y > t/2]$, $[X > t/2, Y \leq t/2]$, $[X \leq t/2, Y > t/2]$ individually.]

5. Let $X = \{X_t : t \geq 0\}$ be a continuous parameter martingale with (a) continuous sample paths, (b) finite second moments, and (c) stationary ergodic increments over disjoint intervals of the same length. Prove that X is a Brownian motion. [*Hint*: Use Theorem 15.5.]
6. (Ornstein–Uhlenbeck Process) Let $X = \{X(t) : t \geq 0\}$ be a stochastic process with continuous sample paths, satisfying (15.13)–(15.17), with $\beta(t) = \beta$, $\sigma^2(t) = \sigma^2 > 0$, constants. Show that X is Gaussian as well as Markov.

Chapter 16
Stationary Ergodic Markov Processes: SLLN & FCLT

For discrete parameter Markov processes on a general state space, Birkhoff's ergodic theorem provides a natural approach to the existence of invariant probabilities and the corresponding strong law of large numbers in some generality. In addition, it is shown that the notion of an irreducible positive recurrent Markov chain on a countable state space is equivalent to being irreducible ergodic stationary Markov chain having a unique invariant initial distribution.

The strong law of large numbers (SLLN) and the functional central limit theorem (FCLT) for stochastic processes are of great significance in theory and applications. A complete treatment of these for Markov processes with countable state spaces (Markov chains) was given in Chapter 12 using the renewal method. The present brief chapter is devoted to the SLLN and FCLT for Markov processes on general state spaces via Birkhoff's ergodic theorem and martingale theory, respectively.

Proposition 16.1. Let $\mathbf{X} = \{X_0, X_1, \dots\}$ be a Markov process on a state space $(S, \mathcal{S})$ with transition probability $p(x, dy)$ and invariant probability π. If for every x outside a π-null set, one has

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} p^{(m)}(x, B) = \pi(B), \qquad \text{for all } B \in \mathcal{S}, \tag{16.1}$$

then the stationary process $\mathbf{X} = \{X_0, X_1, \dots\}$ is ergodic for the initial distribution π.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_16

Before we prove this proposition, let us recall that (16.1) is equivalent to the statement that for all x outside a π-null set

$$\lim_{n\to\infty}\int_S f(y)\frac{1}{n}\sum_{m=1}^{n} p^{(m)}(x,dy)=\int_S f(y)\pi(dy) \tag{16.2}$$

for all $f\in\mathbb{B}(S)$. The proof of (16.2) is by the method of approximation by simple functions

Proof. Since $\mathcal{S}^{\otimes\infty}$ is generated by finite dimensional sets of the form $A=C\times S^{\infty}$, $C\in\mathcal{S}^{\otimes m+1}$, $m\geq 0$, by the $\pi-\lambda$ theorem, it is enough to prove the proposition with A of this form. For such A, letting $B\in\mathcal{S}^{\otimes\infty}$ and $r>m$, we have

$$\begin{aligned}P([\mathbf{X}\in A]\cap[T^r\mathbf{X}\in B]) &= P([(X_0,\ldots,X_m)\in C]\cap[(X_r,X_{r+1},\ldots,)\in B])\\ &=\mathbb{E}\{\mathbf{1}_C(X_0,\ldots,X_m)P_{X_r}(B)\}\\ &=\mathbb{E}\{\mathbf{1}_C(X_0,\ldots,X_m)\int_S P_y(B)p^{(r-m)}(X_m,dy)\},\end{aligned}$$

where P_y is the distribution of the Markov process with initial state y and transition probability $p(x,dy)$ and $P,\mathbb{E}$ denote the distribution of X, and expected value under the initial distribution π. Summing over $r=m+1,\ldots,n$ and letting $n\to\infty$, one gets

$$\begin{aligned}&\frac{1}{n}\sum_{r=0}^{n-1}P([\mathbf{X}\in A]\cap[T^r\mathbf{X}\in B])\\ &\quad=\frac{1}{n}\sum_{r=0}^{m}P([\mathbf{X}\in A]\cap[T^r\mathbf{X}\in B])+\frac{1}{n}\sum_{r=m+1}^{n-1}P([\mathbf{X}\in A]\cap[T^r\mathbf{X}\in B])\\ &\quad\simeq\frac{n-m-1}{n}\mathbb{E}\left\{\mathbf{1}_A(\mathbf{X})\frac{1}{n-m-1}\sum_{r=1}^{n-m-1}P_y(B)p^{(r)}(X_m,dy)\right\}\\ &\quad\to\mathbb{E}\left\{\mathbf{1}_A(\mathbf{X})\int_S P_y(B)\pi(dy)\right\}=P(\mathbf{X}\in A)P(\mathbf{X}\in B),\end{aligned} \tag{16.3}$$

where $\simeq$ indicates that the difference between the two sides goes to zero as $n\to\infty$. The convergence in (16.3) derives from (16.2) with $f(y)=P_y(B)$. By Proposition 4.5, $\mathbf{X}$ is ergodic. ■

Combining Proposition 16.1 and Theorem 13.3, one may conclude that positive recurrence and ergodicity are equivalent notions for irreducible Markov processes having a countable state space. Namely, one has the following.

Corollary 16.2. Let $\{X_n\}$ be an irreducible Markov chain on a denumerable state space. Then $\{X_n\}$ is positive recurrent with invariant probability π if and only if $\{X_n\}$ is a stationary ergodic Markov process with invariant initial distribution π.

Proof. In view of Corollary 8.5, see (13.19), positive recurrence implies ergodicity of $\{X_n\}$. Conversely, if $\{X_n\}$ is an irreducible ergodic Markov chain, then positive recurrence follows from Birkhoff's ergodic theorem by taking $f(\mathbf{X}) = \mathbf{1}_{[X_0=j]}(\mathbf{X})$ for fixed but arbitrary $j \in S$, for then $f(T^m\mathbf{X}) = \mathbf{1}_{[X_m=j]}(\mathbf{X})$. ■

Remark 16.1. Out of this theory, one has an alternative characterization of ergodicity of stationary Markov chains on a finite state space as the property that 1 is a simple eigenvalue of the transition probability matrix $\mathbf{p}$ (see Exercise 16).

The following result is now a consequence of Birkhoff's ergodic theorem (Theorem 4.1).

Corollary 16.3 *(SLLN for Markov Processes on General State Spaces).* Suppose the transition probability $p(x, dy)$ has an invariant probability π and that (16.1) holds for every x outside a π-null set. Let $f \in L^1(S, \pi)$, and define

$$B := \left\{ \mathbf{x} = (x_0, x_1, \dots) \in S^\infty : \lim_n \frac{1}{n} \sum_{m=0}^{n-1} f(x_m) = \int_S f d\pi \right\}. \tag{16.4}$$

Then $P_y(B) = 1$ for every y outside a π-null set.

Proof. By Birkhoff's ergodic theorem (Theorem 4.1(b))

$$1 = P(\mathbf{X} \in B) = EP(\mathbf{X} \in B | X_0) = EP_{X_0}(B) = \int_S P_y(B)\pi(dy). \tag{16.5}$$

Hence, $P_y(B) = 1$ outside a π-null set. ■

In the next few chapters, we will find broad classes of Markov processes that have invariant probabilities π for which the hypothesis of Proposition 16.1 holds. Ergodicity of the Markov process equivalently refers to the existence of an invariant probability π under which the stationary process with initial distribution π is ergodic in the sense of Definition 4.5. The following Definition 16.1 and Theorem 16.4 capture this essential role of the invariant measure π.

Definition 16.1. An invariant probability π for a transition probability $p(x, dy)$ is said to be *ergodic* if the stationary Markov process with the transition probability $p(x, dy)$ and initial distribution π is ergodic.

Remark 16.2. It should be noted that Birkhoff's ergodic theorem (Theorem 4.1(b)), applied to a stationary ergodic Markov process, implies the hypothesis of Proposition 16.1 (Exercise 16).

Example 1. This example shows that, for a given transition probability, the existence of an ergodic invariant probability does not imply uniqueness nor preclude the existence of an infinite σ-finite invariant measure. Let $S = [-2, 2]$ and define $X_{n+1} = f(X_n) + \epsilon_{n+1}, n = 0, 1, 2, \dots$, where $f(x) = (x + 1)\mathbf{1}_{[-2,0]}(x) + (x - 1)\mathbf{1}_{(0,2]}(x), x \in S$, and $\epsilon_n, n \geq 1$ is an i.i.d. Bernoulli ±1 sequence with equal probabilities. Then, for each fixed $x \in (0, 2]$, if one starts in state $X_0 = x$, the two-point state $X_1(x)$ is identically distributed as $X_1(x - 2)$. In particular, $X_1(x)$ and $X_2(x)$ are independent with the same two-point distribution given by the ergodic invariant probability (for $x \in (0, 2]$)

$$\pi_x^+ = \frac{1}{2}\delta_{\{x\}} + \frac{1}{2}\delta_{\{x-2\}},$$

and similarly, for $x \in [-2, 0]$,

$$\pi_x^- = \frac{1}{2}\delta_{\{x\}} + \frac{1}{2}\delta_{\{x+2\}},$$

are mutually singular ergodic invariant probabilities. In addition, for fixed but arbitrary $0 < a_1 < a_2 < \cdots < 2$, the infinite σ-finite measure

$$m = \sum_{n=1}^{\infty}(\delta_{\{a_n\}} + \delta_{\{a_n-2\}})$$

is a σ-finite invariant measure. (See Exercise 10 for the case $\epsilon_n, n \geq 1$, are i.i.d. uniform on [−1,1].)

We next present an elegant alternative approach to the SLLN for Markov processes whose proof shows, in particular, that the invariant sigma-field may be identified with a sub-sigma-field of $\mathcal{S}$.

Theorem 16.4 *(Ergodicity, SLLN, and the Uniqueness of Invariant Probabilities).* Suppose $\mathbf{X} = \{X_n : n \geq 0\}$ is a stationary Markov process on a state space $(S, \mathcal{S})$ having a transition probability $p(x, dy)$ and invariant initial distribution π. The process is ergodic if and only if there does not exist an invariant distribution $\pi' \neq \pi$ such that $\pi' << \pi$.

Proof. The crucial step in the proof is to first identify the shift-invariant events. For this, let us show that every shift-invariant bounded measurable $h(\mathbf{X})$ is a.s. equal to a random variable $g(X_0)$ where g is a bounded measurable function on $(S, \mathcal{S})$. Let T denote the shift transformation, and let $\mathcal{I}$ denote the shift-invariant sigma-field. If $h(\mathbf{X})$ is invariant, then $h(\mathbf{X}) = h(T^n\mathbf{X})$ a.s. for all $n \geq 1$. Then, by the Markov property,

$$\begin{aligned}\mathbb{E}(h(\mathbf{X})|\sigma(X_0, X_1, \dots X_n)) &= \mathbb{E}(h(T^n\mathbf{X})|\sigma(X_0, X_1, \dots X_n))\\ &= \mathbb{E}(h(T^n\mathbf{X})|\sigma(X_n)) = g(X_n),\end{aligned}$$

where $g(x) = \mathbb{E}(h(X_0, X_1, \dots)|X_0 = x)$. By the martingale convergence theorem applied to the martingale $g(X_n) = \mathbb{E}(h(\mathbf{X})|\sigma(X_0, X_1, \dots X_n)$, it follows that $g(X_n)$

converges, a.s. and in L^1, to $\mathbb{E}(h(\mathbf{X})|\sigma(X_0, X_1, \dots)) = h(\mathbf{X})$. But $g(X_n) - h(\mathbf{X}) = g(X_n) - h(T^n\mathbf{X})$ has the same distribution as $g(X_0) - h(\mathbf{X})$ for each $n \geq 1$. Thus, letting $n \to \infty$, it follows that $g(X_0) - h(\mathbf{X}) = 0$ a.s. In particular, for $G \in \mathcal{I}$, there is $B \in \mathcal{S}$ such that $G = [X_0 \in B]$ a.s. This implies $\pi(B) = P(X_0 \in B) = P(G)$. If $\mathbf{X}$ is not ergodic, then there exists $G \in \mathcal{I}$ such that $0 < P(G) < 1$ and, therefore, $0 < \pi(B) < 1$ for a corresponding set $B \in \mathcal{S}$ as above. But the probability π_B defined by $\pi_B(A) = \pi(A \cap B)/\pi(B)$, $A \in \mathcal{I}$, is invariant. To see this, observe that

$$\begin{aligned}\int_S p(x, A)\pi_B(dx) &= \frac{1}{\pi(B)}\int_B p(x, A)\pi(dx)\\ &= P(X_0 \in B, X_1 \in A)/\pi(B)\\ &= P(X_1 \in B, X_1 \in A)/\pi(B),\end{aligned}$$

by invariance of the event $[X_0 \in B]$. In particular, by stationarity,

$$\begin{aligned}\int_S p(x, A)\pi_B(dx) &= P(X_0 \in A \cap B)/\pi(B)\\ &= \pi(A \cap B)/\pi(B) = \pi_B(A).\end{aligned}$$

Since $\pi_B(B) = 1 > \pi(B)$, and $\pi_B << \pi$, the contrapositive is proven.

To prove the other half, suppose that $\mathbf{X}$ is ergodic and π' is also invariant and absolutely continuous with respect to π. Fix $A \in \mathcal{S}$. By Birkhoff's ergodic theorem, and conditioning on X_0, one has as $n \to \infty$, $\frac{1}{n}\sum_{j=0}^{n-1} p^{(j)}(x; A) \to \pi(A)$ for all x outside a π-null set, and hence outside a π'-null set. The invariance of π' implies

$$\int_S \frac{1}{n}\sum_{j=0}^{n-1} p^{(j)}(x; A)\pi'(dx) = \pi'(A)$$

for all $n \geq 1$. Thus $\pi(A) = \pi'(A)$. Since $A \in \mathcal{S}$ is arbitrary, the proof is complete. ■

We saw above that the strong law of large numbers extends to Markov processes with general state spaces, even if *point recurrence* does not hold. That is, even for Markov processes having a unique invariant distribution, there may not be any point in the state space to which the process returns (infinitely often) with probability one. Thus a more general approach than the renewal decomposition will also be required in order to obtain a central limit theorem. One such more general approach that applies to all ergodic Markov processes is via martingales. So let us now see how to apply the martingale central limit theorem (Theorem 15.5) to an ergodic Markov process $\{X_n : n \geq 0\}$ on a state space S (with σ-field $\mathcal{S}$), having a transition probability $p(x, dy)$ with invariant probability π. As usual, write T for the *transition operator*, $Tg(x) := \int g(y)p(x, dy)$. Then for $f \in L^1(S, \pi)$, with $\mu = \int_S f d\pi$, $T^m\bar{f}(x) = T^m f(x) - \mu$ for all $m \geq 0$, where $\overline{f} = f - \mu$.

As will be clear from the proof below, the key to this application is the following representation of sums of the form $\sum_{m=0}^{n-1} f(X_m)$ for suitable functions $f \in L^2(S, \pi)$. Namely, assume (centering) $\int_S f d\pi = 0$, i.e., $f \in 1^\perp := \{f \in L^2(S, \pi) : \langle f, 1\rangle_\pi \equiv \int_S f d\pi = 0\}$, and that f belongs to the range of $-A := I - T$ as an operator on $L^2(S, \pi)$, i.e., $f = (I - T)g$ for some $g \in L^2(S, \pi)$. Then

$$\sum_{m=0}^{n-1} f(X_m) = \sum_{m=0}^{n-1}[g(X_m)-Tg(X_m)] = \sum_{m=1}^{n}[g(X_m)-Tg(X_{m-1})]+g(X_0)-g(X_n) \tag{16.6}$$

for which $Z_n := \sum_{m=1}^{n}[g(X_m) - Tg(X_{m-1}]$ defines a martingale, i.e., $g(X_m) - Tg(X_{m-1}), m = 1, 2, \ldots,$ is a martingale difference sequence and, by Chebyshev inequality, $(g(X_0) - g(X_n)) = o(\sqrt{n})$ in probability as $n \to \infty$ (Exercise 16).

To obtain a sufficient condition for the range requirement, suppose for $f \in 1^\perp$, the series $g_n(x) := \sum_0^n (T^m f)(x)$ converges in $L^2(S, \pi)$ to g, i.e.,

$$\int (g_n - g)^2 d\pi \to 0 \qquad \text{as } n \to \infty. \tag{16.7}$$

In this case, g satisfies the "Poisson equation"

$$Tg(x) - g(x) = -f(x), \quad \text{or} \quad (T - I)g = -f, \tag{16.8}$$

i.e., f belongs to the *range* of $I - T$ regarded as an operator on $L^2(S, \pi)$.

Theorem 16.5 *(Gordin–Lifsic FCLT for Discrete Parameter Markov Processes*[1]*).* Assume that $p(x, dy)$ admits an invariant probability π and that under this initial invariant distribution the stationary process $\{X_n : n \geq 0\}$ is ergodic. Assume X_0 has distribution π, and let f be a real-valued function on S such that $Ef^2(X_0) < \infty$. Also let $\mu = \int f d\pi$, and assume that $\bar{f} := f - \mu$ is in the range of $I - T$, as an operator on $L^2(S, \pi)$, with $Ag = -\bar{f}$, $g \in L^2(S, \pi)$, where $A = T - I$. Let

$$S_n = \sum_{m=0}^{n-1}(f(X_m) - \mu).$$

Then the polygonal process $\{Z_n(t) = \frac{S_{[nt]}}{\sqrt{n}} + (nt - [nt])\frac{X_{[nt]+1}}{\sqrt{n}} : t \geq 0\}$, with values in $C[0, \infty)$, converges in distribution to σB where $B := \{B_t : t \geq 0\}$ denotes the standard Brownian motion starting at 0 and $\sigma^2 = ||g||_\pi^2 - ||Tg||_\pi^2 = 2\langle g, \bar{f}\rangle_\pi - \langle \bar{f}, \bar{f}\rangle_\pi$.

[1] Gordin and Lifsic (1978). Also see Bhattacharya (1982).

Proof. Suppose that $g \in L^2(S, \pi)$ satisfies

$$g(x) - Tg(x) = \bar{f}(x), \quad \text{or} \quad (I - T)g = \bar{f}, \tag{16.9}$$

i.e., $\bar{f}$ belongs to the *range* of $I - T$ regarded as an operator on $L^2(S, \pi)$. Now it is simple to check that

$$g(X_n) - Tg(X_{n-1}) \qquad (n \geq 1) \tag{16.10}$$

is a *martingale difference sequence*. It is also stationary and ergodic (Exercise 16). Write

$$Z_n := \sum_{m=1}^{n} (g(X_m) - Tg(X_{m-1})). \tag{16.11}$$

Then, by the Billingsley–Ibragimov FCLT (Theorem 15.5), one obtains the asserted convergence in distribution to σB where

$$\sigma^2 = \mathbb{E}(g(X_1) - Tg(X_0))^2 = \mathbb{E}g^2(X_1) + \mathbb{E}(Tg)^2(X_0) - 2\mathbb{E}[Tg(X_0)g(X_1)]. \tag{16.12}$$

Since $\mathbb{E}[g(X_1) \mid \{X_0\}] = Tg(X_0)$, we have $\mathbb{E}[Tg(X_0)g(X_1)] = \mathbb{E}(Tg)^2(X_0)$, so that (16.12) reduces to

$$\sigma^2 = \mathbb{E}\mathbb{E}(g^2(X_1) - \mathbb{E}(Tg)^2(X_0) = \int g^2 d\pi - \int (Tg)^2 d\pi = 2\langle g, \bar{f}\rangle_\pi - \langle \bar{f}, \bar{f}\rangle_\pi. \tag{16.13}$$

The last equality is obtained by writing $Tg = g - \bar{f}$. Also, by (16.9),

$$\begin{aligned} Z_n = \sum_{m=1}^{n} (g(X_m) - Tg(X_{m-1})) &= \sum_{m=0}^{n-1} (g(X_m) - Tg(X_m)) + g(X_n) - g(X_0) \\ &= \sum_{m=0}^{n-1} \bar{f}(X_m) + g(X_n) - g(X_0). \end{aligned} \tag{16.14}$$

Since

$$[\mathbb{E}g(X_n) - g(X_0))/\sqrt{n}]^2 \leq \frac{2}{n}\,(\mathbb{E}g^2(X_n) + \mathbb{E}g^2(X_0)) = \frac{4}{n}\int g^2 d\pi \to 0.$$

■

Example 2. Recall that the (deterministic) cyclic motion on the two-state set $S = \{0, 1\}$ defined by the transition probabilities $p_{01} = p_{10} = 1$ is irreducible with unique invariant probability $\pi_0 = \pi_1 = 1/2$. However, the average of any function over a cycle and the return time are both a.s. constants. Thus the asymptotic normal distribution obtained above is degenerate, i.e., the asymptotic variance is $\sigma^2 = 0$. Also see Exercise 14.

A special case of Theorem 16.5 with countable state space S stated below is of much interest.

Proposition 16.6. Consider an irreducible positive recurrent Markov chain $\{X_n : n = 0, 1, 2, \dots\}$ with countable state space S. It has a unique invariant probability π. If f belongs to the range of $T - I$, then the conclusion of Theorem 16.5 holds. Moreover, if $\sum_{k=1}^{\infty} T^k \bar{f}$ converges in $L^2(S, \pi)$, then $\bar{f}$ belongs to the range of $T - I$ (Exercise 16).

Remark 16.3. Knowing when $0 < \sigma^2 < \infty$ is tantamount to knowing that $\sqrt{n}$ is the correct scaling for the fluctuations. The central limit theorem provides finiteness of σ^2, but not necessarily its positivity. In particular, note that for the above example of deterministic cyclic motion on two states $S = \{0, 1\}$, a time-reversible Markov process for which T is self-adjoint with eigenvalues ± 1, $\bar{f}$ belongs to the range if and only if $\bar{f}(0) = -\bar{f}(1)$. But this condition is equivalent to centering. This example is ruled out by the convergence of the numerical series defining γ, but not a range condition on $\bar{f}$.

Corollary 16.7. In addition to the hypothesis of Theorem 16.5, assume that π is time-reversible and $\lambda = -1$ does not belong to the spectrum of T. Then one has $\sigma^2 > 0$ if and only if f is not constant.

Proof. Consider the contrapositive statement that $\sigma^2 = 0$ if and only if f is constant. Clearly, if f is constant, then $\bar{f} \equiv 0$ and one already obtains $\sigma^2 = 0$ from the variance in Theorem 16.5. Conversely, suppose that $\sigma^2 = 0$. Then one has $\langle g, g\rangle_\pi = \langle Tg, Tg\rangle_\pi$. We have $\bar{f}$ belonging to the range of $A = T - I$ with $Ag = \bar{f}$.

Since T is a self-adjoint contraction on the Hilbert space $L^2(S, \pi)$ with $T1 = 1$, we may apply the functional calculus associated with the spectral measure for self-adjoint bounded operators, see Appendix B. Let $\nu_g(d\lambda) = d\langle E_\lambda g, g\rangle_\pi$, where $T = \int_{[-1,1]} \lambda dE_\lambda$ is the spectral decomposition of T. Then

$$\int_{[-1,1]} \nu_g(d\lambda) = \langle g, g\rangle_\pi = \langle Tg, Tg\rangle_\pi = \langle T^2 g, g\rangle_\pi = \int_{[-1,1]} \lambda^2 \nu_g(d\lambda), \tag{16.15}$$

so that

$$\int_{[-1,1]} (1 - \lambda^2)\nu_g(d\lambda) = 0. \tag{16.16}$$

Since -1 does not belong to $\sigma(T)$, $1 - \lambda^2 > 0$ for all $\lambda \in \sigma(T)\backslash\{1\}$. It follows from (16.16) that g is an eigenvector of T with $\lambda = 1$. Hence, $T^2 g = Tg = g$, and

$$\begin{aligned}\langle \bar{f}, \bar{f}\rangle_\pi &= \langle Tg - g, Tg - g\rangle_\pi \\ &= \langle T^2 g, g\rangle_\pi - 2\langle Tg, g\rangle_\pi + \langle g, g\rangle_\pi \\ &= \langle T^2 g - 2Tg + g, g\rangle_\pi = 0,\end{aligned} \tag{16.17}$$

and hence, $\bar{f} = 0$, and f is a constant π-a.s. ■

Remark 16.4. An important case for positivity of σ^2 is provided in Exercise 16, for the case when $p^{(n)}(x, dy)$, is mutually absolutely continuous with respect to the invariant measure π.

Remark 16.5. Note that

$$\langle \bar{f}, T^k \bar{f} \rangle_\pi = \mathrm{Cov}_\pi\{f(X_0), f(X_k)\}, \quad \sum_{k=1}^{m} \langle \bar{f}, T^k \bar{f} \rangle_\pi = \sum_{k=1}^{m} \mathrm{Cov}_\pi\{f(X_0), f(X_k)\}.$$

The convergence condition of Proposition 16.6 is the condition that the correlation decays to zero at a sufficiently rapid rate for time points k units apart as $k \to \infty$.

Exercises

1. (Simple Eigenvalues) Let $\{X_n\}$ be a Markov chain on a finite state space $S = \{1, 2, \ldots, k\}$ with positive invariant measure $\pi = (\pi_1, \ldots, \pi_k)$ and transition probability matrix $\mathbf{p} = ((p_{ij}))_{i,j \in S}$. Show that ergodicity of $\{X_n\}$ is equivalent to the property that 1 is a simple eigenvalue of $\mathbf{p}$. [*Hint*: Simple implies that the invariant initial distribution π for the stationary process is unique as an eigenvector (with eigenvalue one) and hence as an invariant probability. On the other hand, an ergodic process implies $\lim_n \frac{1}{n} \sum_{m=0}^{n-1} p^m = q$, a matrix with identical rows π. An eigenvector v of p (with eigenvalue one) is an eigenvector of q. Constant rows imply v is a multiple of π, i.e., 1 is simple.
2. (a) In the context of Proposition 16.1, prove that if $g = \sum_{k=0}^{\infty} T^k \bar{f}$ converges in $L^2(S, \pi)$, then $\bar{f}$ belongs to the range of $I - T$ and $(I - T)g = \bar{f}$.
   (b) Suppose S is finite and the Markov chain on S is aperiodic, positive recurrent. Prove that the conclusion of Theorem 16.5 holds for all $f \in L^2(S, \pi)$, with $\bar{f} = f - \int_S f d\pi$, and that $\sigma^2 > 0$ if f is not a constant. [*Hint*: $I - T$ has a bounded inverse on $1^\perp$.]
3. (a) Check that T is self-adjoint on $L^2(\{0, 1\}, \frac{1}{2}\delta_{\{0\}} + \frac{1}{2}\delta_{\{1\}})$ in Example 2, but $\lambda = -1$ is an eigenvalue.
   (b) Extend the example to a finite state space $S = \{0, 1, \ldots, k\}(k \geq 1)$, with a cyclic motion $p(i, (i+1)\mathrm{mod}k) = 1$ $(i = 0, 1, \ldots, k)$, with unique invariant probability uniform on S.
4. Suppose π is an invariant probability for a Markov process with transition probability $p(x, dy)$.

   (a) Assume that for some $n \geq 1$, $p^{(n)}(x, dy)$ is mutually absolutely continuous with respect to π, for each $x \in S$. Prove that π is the unique invariant probability, and
   (b) The variance parameter σ^2 in Theorem 16.5 is strictly positive if $f \in L^2(S, \pi)$ is non-constant π-a.s. [*Hint*: Say $\bar{f} = (T - I)g$. Consider contrapositive with $\sigma^2 = 0$. For $n = 1$, $||Tg||_\pi^2 = ||g||_\pi^2$ if and only if

$(\int_S g(y)p(x,dy))^2 = \int_S g^2(y)p(x,dy)$, for π-a.e. $x \in S$. Argue that this holds iff $g(y)$ is a constant, say $c(x)$, on a set $A(x)$, $p(x,\cdot)$-a.s. Next fix $x_0 \in S$, and explain that $p(x, A(x_0)) = 1$ for all $x \in S$. Deduce that $\pi(g = c(x_0)) = 1$, $\int_S g d\pi = 0$, and therefore, $g = 0$, π a.s. Conclude that $\bar{f} = 0$ π-a.s. For $n > 1$, use the (contraction) property $T^n g||_\pi^2 \leq ||Tg||_\pi^2$.]

5. Consider the simple symmetric random walk on $\{0, 1, 2\}$ with reflecting boundaries at 0 and 2.
   (a) Show $\frac{1}{n}\sum_{r=1}^{n} p^{(r)}$ converges to the matrix whose rows are identically $(1/4, 1/2, 1/4)$.
   (b) Show that $\pi_0 = \pi_2 = 1/4, \pi_2 = 1/2$ is the unique invariant probability.
   (c) Start $\{X_n : n \geq 0\}$ with the invariant initial distribution. Show that $\{X_n : n \geq 0\}$ is ergodic, but $\{X_{2n} : n \geq 0\}$ is not ergodic.
6. Consider the Markov chain $\{X_n : n \geq 0\}$ on $S = \{1, 2, 3\}$ with transition probabilities $p_{11} = 1, p_{23} = p_{32} = 1$, and $p_{ij} = 0$ otherwise. Determine the extremal ergodic invariant probabilities and the collection of all invariant probabilities. Calculate $\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} X_m$ for the initial distribution $\pi = (1/4, 3/8, 3/8)$.
7. Suppose that a stationary Markov process satisfies the SLLN, i.e., assume that $\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} f(X_m) = \int_S f d\pi$ almost surely for every $f \in \mathbb{B}(S)$. Show that the hypothesis of Proposition 16.1 holds.
8. Let $\mathbf{X} = (X_0, X_1, \ldots)$ be a Markov process satisfying the hypothesis of Proposition 16.1. Let $f \in L^1(S^\infty, \mathcal{S}^{\otimes\infty}, P_\pi)$. Prove that $\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} f(T^m\mathbf{X}) = \int_{S^\infty} f(x)P_\pi(d\mathbf{x})$ P_y-almost surely for every $y \in S$ outside a π-null set.
9. For $g \in L^2(S, \pi)$, show that $Z_m = g(X_m) - Tg(X_{m-1}), m = 1, 2, \ldots$, is a stationary and ergodic martingale difference sequence and $(g(X_0) - g(X_n)) = o(\sqrt{n})$ in probability as $n \to \infty$.
10. Let a Markov process on $S = [-2, 2]$ be defined by the equation $X_{n+1} = f(X_n) + \epsilon_{n+1}, n \geq 0$, with f as in Example 1, but $\epsilon_{n+1}, n \geq 0$, i.i.d. uniform on $[-1, 1]$. Show that the triangular distribution π with density $\pi(y) = (2-|y|)/4$, is the unique invariant probability on S, and that the stationary process with initial distribution π is ergodic.

Chapter 17
Linear Markov Processes

The linear Markov processes are most readily described in terms of evolutions obtained by i.i.d. iterated affine linear maps. This chapter addresses the ergodic theory for such processes.

The canonical construction of Markov chains on the space of trajectories is based on Kolmogorov's or Tulcea's existence theorem[1]. In the present and next few sections, another widely used general method of construction of Markov processes by i.i.d. iterated random maps on arbitrary state spaces is illustrated. Markovian models in this form arise naturally in many fields, and they are often easier to analyze in this noncanonical representation.

Example 1 *(The Linear Autoregressive Model of Order One, or the* **AR(1)** *Model).* Let b be a real number and $\{\epsilon_n : n \geq 1\}$ an i.i.d. sequence of real-valued random variables defined on some probability space $(\Omega, \mathcal{F}, P)$. Given an initial random variable X_0 independent of $\{\epsilon_n\}$, define recursively the sequence of random variables $\{X_n : n \geq 0\}$ as follows:

$$X_0, X_1 := bX_0 + \epsilon_1, \qquad X_{n+1} := bX_n + \epsilon_{n+1} \quad (n \geq 0). \tag{17.1}$$

Equivalently, this may be viewed as a composition of "random maps"

$$X_n = \boldsymbol{\alpha}_n \cdots \boldsymbol{\alpha}_1(X_0), \tag{17.2}$$

[1] See BCPT pp. 167–170.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_17

where $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots$ is an i.i.d. sequence of random maps $\boldsymbol{\alpha}_i : \mathbb{R} \to \mathbb{R}$ defined by $\boldsymbol{\alpha}_i(x) := bx + \epsilon_i, x \in \mathbb{R}, i = 1, 2, \ldots$. For now, also observe that $X_0, X_1, \ldots, X_n$ are determined by $\{X_0, \epsilon_1, \ldots, \epsilon_n\}$, and ϵ_{n+1} is independent of the latter. One has, for all Borel sets C,

$$\begin{aligned} P(X_{n+1} \in C \mid \sigma\{X_0, X_1, \ldots, X_n\}) &= P(bx + \epsilon_{n+1} \in C)|_{x=X_n} \\ &= P(\epsilon_{n+1} \in C - bx)|_{x=X_n} \\ &= Q(C - bX_n), \end{aligned} \tag{17.3}$$

where Q is the common distribution of the random variables ϵ_n. Thus, $\{X_n : n \geq 0\}$ is a Markov process on the state space $S = \mathbb{R}$, having the *transition probability* (of going from x to C in one step)

$$p(x, C) := Q(C - bx), \tag{17.4}$$

and the *initial distribution* given by the distribution of X_0. The analysis of this Markov process is, however, facilitated more by its representation (17.1) than by an analytical study of the asymptotics of n-step transition probabilities as evidenced by the proof of the following proposition.

Proposition 17.1. Let b be a real number with $|b| < 1$. Then

$$\mathbb{E} \log^+ |\epsilon_1| < \infty \tag{17.5}$$

is necessary and sufficient that the Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ defined by the random iteration (17.1) converges weakly (in distribution) to a unique invariant probability π given by the distribution of the a.s. limit of the random series $\sum_{n=0}^{\infty} b^n \epsilon_{n+1}$.

Proof. (*Sufficiency*) First note that successive iteration in (17.1) yields

$$\begin{aligned} X_1 &= bX_0 + \epsilon_1, \qquad X_2 = bX_1 + \epsilon_2 = b^2 X_0 + b\epsilon_1 + \epsilon_2 \qquad \cdots \\ X_n &= b^n X_0 + b^{n-1}\epsilon_1 + b^{n-2}\epsilon_2 + \cdots + b\epsilon_{n-1} + \epsilon_n \qquad (n \geq 1). \end{aligned} \tag{17.6}$$

The distribution of X_n is, therefore, the same as that of

$$Y_n := b^n X_0 + \epsilon_1 + b\epsilon_2 + b^2\epsilon_3 + \cdots + b^{n-1}\epsilon_n \qquad (n \geq 1). \tag{17.7}$$

Equivalently, $Y_1, Y_2, \ldots$ is defined by the backward iteration

$$Y_n = \boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_n(X_0), \quad n \geq 1. \tag{17.8}$$

We assume that

$$|b| < 1. \tag{17.9}$$

First consider the case $|\epsilon_n| \leq c$ with probability 1 for some constant c. Then it follows from (17.7) that

$$Y_n \to \sum_{n=0}^{\infty} b^n \epsilon_{n+1} \text{ a.s.,} \tag{17.10}$$

regardless of X_0. Let π denote the distribution of the random variable on the right side in (17.10). Then Y_n converges in distribution to π as $n \to \infty$ (Exercise). Because the distribution of X_n is the same as that of Y_n, it follows that X_n converges in distribution to π. Therefore, π is the unique invariant distribution for the Markov process $\{X_n : n \geq 0\}$, i.e., for $p(x, dy)$. Next consider the more general hypothesis (17.5). Note that this is equivalent to assuming $\sum P(|\epsilon_{n+1}| > c\delta^n) < \infty$ for some $c > 0$, $\delta > 1$ (see Exercise 17) so that, by the Borel–Cantelli lemma,

$$P(|\epsilon_{n+1}| \leq c\delta^n \text{ for all but finitely many } n) = 1.$$

Now choose δ such that $1 < \delta < 1/|b|$. Then, with probability 1, $|b^n \epsilon_{n+1}| \leq c(|b|\delta)^n$ for all but finitely many n. Since $|b|\delta < 1$, the series on the right side of (17.10) is convergent and is the limit of Y_n.

(*Necessity*) Assume $\mathbb{E} \log^+ |\epsilon_1| = \infty$. Let $\delta > (1/|b|) \vee 1$. Then, writing $Z_n = \log^+ |\epsilon_n| / \log \delta$, one has $P(|\epsilon_n| > \delta^n) = P(Z_n > n) = P(Z_1 > n)$, and $1 + \sum_{n=1}^{\infty} P(Z_1 > n) \geq \mathbb{E} Z_1 = \infty$. Since the $Z_n, n \geq 1$, is an i.i.d. sequence, it follows by Borel–Cantelli II[2] that is $P(Z_n > n \text{ i.o.}) = 1$. Thus $P(b^n|\epsilon_n| > (b\delta)^n \text{ i.o.}) = 1$, and therefore, the series $\sum_{n=1}^{\infty} b^n \epsilon_{n+1}$ diverges almost surely. ■

Remark 17.1. The role of the moment condition (17.5) cannot be overemphasized. If this moment is infinite, then no matter how small $|b|$ may be, barring the case $b = 0$, the process does not have an invariant probability[3] (Exercise 6(c) below).

Next, Example 1 has an extension to multidimensional state space.

Example 2 *(General Linear Time Series Model).* Let $\{\epsilon_n : n \geq 1\}$ be a sequence of i.i.d. random vectors with values in $\mathbb{R}^m$ and common distribution Q, and let $\mathbf{B}$ be an $m \times m$ matrix with real entries b_{ij}. Suppose $\mathbf{X}_0$ is an m-dimensional random vector independent of $\{\epsilon_n\}$. Define recursively the sequence of random vectors

$$\mathbf{X}_0, \ \mathbf{X}_{n+1} := \mathbf{B}\mathbf{X}_n + \epsilon_{n+1} \qquad (n = 0, 1, 2, \ldots). \tag{17.11}$$

As in (17.3), (17.4), $\{X_n : n \geq 0\}$ is a Markov process with state space $\mathbb{R}^m$ and transition probability

$$p(\mathbf{x}, C) := Q(C - \mathbf{B}\mathbf{x}) \qquad (\text{for all Borel sets } C \subset \mathbb{R}^m). \tag{17.12}$$

[2] BCPT p. 34.

[3] See Bhattacharya and Majumdar (2017), Theorem 2.1, pp. 290–302

Recall that the *norm of a matrix* $\mathbf{H}$ can be defined by

$$\|\mathbf{H}\| := \sup_{|\mathbf{x}|=1} |\mathbf{H}\mathbf{x}|, \tag{17.13}$$

where $|\mathbf{x}|$ denotes the Euclidean length of $\mathbf{x}$ in $\mathbb{R}^m$. For a positive integer $n > n_0$, write $n = jn_0 + j'$, where $0 \leq j' \leq n_0$. To state some sufficient conditions for the hypothesis of this general result to apply, let us recall the definition of the *spectral radius* $r(\mathbf{B})$ of a matrix $\mathbf{B}$ as the maximum modulus of the eigenvalues of $\mathbf{B}$. The following lemma from linear algebra shows that if $r(B) < 1$, then

$$\|B^{n_0}\| < 1 \qquad \text{for some positive integer } n_0. \tag{17.14}$$

Lemma 1. Let $\mathbf{B}$ be an $m \times m$ matrix. Then the spectral radius $r(\mathbf{B})$ satisfies

$$r(\mathbf{B}) \geq \overline{\lim_{n\to\infty}} \|\mathbf{B}^n\|^{1/n}. \tag{17.15}$$

Proof. Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $\mathbf{B}$. This means $\det(\mathbf{B} - \lambda\mathbf{I}) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_m - \lambda)$, where det is shorthand for determinant and $\mathbf{I}$ is the identity matrix. Let λ_m have the maximum modulus among the λ_i, i.e., $|\lambda_m| = r(\mathbf{B})$. If $|\lambda| > |\lambda_m|$, then $\mathbf{B}-\lambda\mathbf{I}$ is invertible, since $\det(\mathbf{B}-\lambda\mathbf{I}) \neq 0$. Indeed, by the definition of the inverse, each element of the inverse of $\mathbf{B} - \lambda\mathbf{I}$ is a polynomial in λ (of degree $m - 1$ or $m - 2$) divided by $\det(\mathbf{B} - \lambda\mathbf{I})$. Therefore, one may write

$$(\mathbf{B} - \lambda\mathbf{I})^{-1} = (\lambda_1 - \lambda)^{-1}\cdots(\lambda_m - \lambda)^{-1}(\mathbf{B}_0 + \lambda\mathbf{B}_2 + \cdots + \lambda^{m-1}\mathbf{B}_{m-1}) \qquad (|\lambda| > |\lambda_m|), \tag{17.16}$$

where $\mathbf{B}_j$ $(0 \leq j \leq m - 1)$ are $m \times m$ matrices that do not involve λ. Writing $z = 1/\lambda$, one may express (17.16) as

$$\begin{aligned}(B - \lambda\mathbf{I})^{-1} &= (-\lambda)^{-m}(1 - \lambda_1/\lambda)^{-1}\cdots(1 - \lambda_m/\lambda)^{-1}\lambda^{m-1}\sum_{j=0}^{m-1}(1/\lambda)^{m-1-j}\mathbf{B}_j \\ &= (-1)^m z(1 - \lambda_1 z)^{-1}\cdots(1 - \lambda_m z)^{-1}\sum_{j=0}^{m-1} z^{m-1-j}\mathbf{B}_j \\ &= \left(z\sum_{n=0}^{\infty} a_n z^n\right)\sum_{j=0}^{m-1} z^{m-1-j}\mathbf{B}_j, \qquad (|z| < |\lambda_m|^{-1}),\end{aligned} \tag{17.17}$$

for appropriate constants a_n. On the other hand,

$$(\mathbf{B} - \lambda\mathbf{I})^{-1} = -z(\mathbf{I} - z\mathbf{B})^{-1} = -z\sum_{k=0}^{\infty} z^k\mathbf{B}^k \qquad \left(|z| < \frac{1}{\|\mathbf{B}\|}\right). \tag{17.18}$$

To see this, first note that the series on the right is convergent in norm for $|z| < 1/\|B\|$, and then check that term-by-term multiplication of the series $\sum_{k=0}^{\infty} z^k \mathbf{B}^k$ by $\mathbf{I} - z\mathbf{B}$ yields the identity $\mathbf{I}$ after all the cancelations. In particular, writing $b_{ij}^{(k)}$ for the (i, j) element of $\mathbf{B}^k$, the series

$$-z \sum_{k=0}^{\infty} z^k b_{ij}^{(k)} \tag{17.19}$$

converges absolutely for $|z| < 1/\|\mathbf{B}\|$. Since (17.19) is the same as the (i, j) element of the series (17.17), at least for $|z| < 1/\|\mathbf{B}\|$, their coefficients coincide (Exercise 17) and, therefore, the series in (17.19) is absolutely convergent for $|z| < |\lambda_m|^{-1}$ (as (17.17) is).

This implies that, for each $\epsilon > 0$,

$$|b_{ij}^{(k)}| < (|\lambda_m| + \epsilon)^k \qquad \text{for all sufficiently large } k. \tag{17.20}$$

For if (17.20) is violated, one may choose $|z|$ sufficiently close to (but less than) $1/|\lambda_m|$ such that $|z^{k'} b_{ij}^{(k')}| \to \infty$ for a subsequence $\{k'\}$, contradicting the requirement that the terms of the convergent series (17.19) must go to zero for $|z| < 1/|\lambda_m|$.

Now $\|\mathbf{B}^k\| \le m^{1/2} \max\{|b_{ij}^{(k)}| : 1 \le i, j \le m\}$ (Exercise 17). Since $m^{1/2k} \to 1$ as $k \to \infty$, (17.20) implies (17.15). ■

Remark 17.2. The indicated lim sup in (17.15) is actually a limit, with equality[4], referred to as *Gelfand's formula* (see Exercise 17). A version is proven for bounded self-adjoint operators on a Hilbert space in Appendix B.

Proposition 17.2. Assume that $r(B) < 1$. Also assume

$$\mathbb{E} \log^+ |\epsilon_1| < \infty. \tag{17.21}$$

Then the Markov process $\{\mathbf{X}_n : n \ge 0\}$ defined by the random iteration (17.11) converges weakly (in distribution) to a unique invariant probability π given by the distribution of the a.s. limit of the random series

$$\mathbf{Y} := \sum_{n=0}^{\infty} \mathbf{B}^n \epsilon_{n+1}. \tag{17.22}$$

Proof. Using the fact $\|\mathbf{B}_1\mathbf{B}_2\| \le \|\mathbf{B}_1\|\|\mathbf{B}_2\|$ for arbitrary $m \times m$ matrices $\mathbf{B}_1, \mathbf{B}_2$ (Exercise 17), one gets

[4] See Halmos (2017), p. 182.

$$\|\mathbf{B}^n\| = \|\mathbf{B}^{jn_0}\mathbf{B}^{j'}\| \leq \|\mathbf{B}^{no}\|^j \, \|\mathbf{B}^{j'}\| \leq c\|\mathbf{B}^{n_0}\|^j, \; c := \max\{\|\mathbf{B}^r\| : 0 \leq r < n_0\}. \tag{17.23}$$

From (17.14) and (17.23), it follows, as in Example 1, that the series $\sum \mathbf{B}^n \epsilon_{n+1}$ converges a.s. in Euclidean norm if (17.21) holds. Write, in this case,

$$\mathbf{Y} := \sum_{n=0}^{\infty} \mathbf{B}^n \epsilon_{n+1}. \tag{17.24}$$

It also follows, as in Example 1, that no matter what the initial distribution (i.e., the distribution of $\mathbf{X}_0$) is, $\mathbf{X}_n$ converges in distribution to the distribution π of $\mathbf{Y}$. Therefore, π is the unique invariant distribution for $p(\mathbf{x}, d\mathbf{y})$. ■

Two well-known time series models will now be treated as special cases of Example 2, special due to the respective structures of the coefficient matrix $\mathbf{B}$. These are the *pth order autoregressive* (or AR(p)) *model* and the *autoregressive moving average model* ARMA(p, q). These models typically arise in the statistical time series analysis of highly fluctuating data in a wide variety of fields. Questions pertaining to the stationarity of the model are fundamental to such analysis.

Example 3 *(*AR(p) *Model).* Let $p > 1$ be an integer, $\beta_0, \beta_1, \ldots, \beta_{p-1}$ real constants. Given a sequence of i.i.d. real-valued random variables $\{\eta_n : n \geq p\}$, and p other random variables $U_0, U_1, \ldots, U_{p-1}$ independent of $\{\eta_n\}$, define recursively

$$U_{n+p} := \sum_{i=0}^{p-1} \beta_i U_{n+i} + \eta_{n+p} \qquad (n \geq 0). \tag{17.25}$$

The sequence $\{U_n : n \geq 0\}$ is not in general a Markov process, but the sequence of p-dimensional random vectors

$$\mathbf{X}_n := (U_n, U_{n+1}, \ldots, U_{n+p-1})' \qquad (n \geq 0) \tag{17.26}$$

is Markovian. Here, the prime ($'$) denotes transposition, so $\mathbf{X}_n$ is to be regarded as a column vector in matrix operations. To prove the Markov property, consider the sequence of p-dimensional i.i.d. random vectors

$$\boldsymbol{\epsilon}_n := (0, 0, \ldots, 0, \eta_{n+p-1})' \qquad (n \geq 1), \tag{17.27}$$

and note that

$$\mathbf{X}_{n+1} = \mathbf{B}\mathbf{X}_n + \boldsymbol{\epsilon}_{n+1}, \tag{17.28}$$

where $\mathbf{B}$ is the $p \times p$ matrix

$$\mathbf{B} := \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ \beta_0 & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{p-2} & \beta_{p-1} \end{bmatrix}. \tag{17.29}$$

Hence, arguing as in (17.3), (17.4), or (17.12), $\{\mathbf{X}_n : n \geq 0\}$ is a Markov process on the state space $\mathbb{R}^P$. Write

$$\mathbf{B} - \lambda \mathbf{I} = \begin{bmatrix} -\lambda & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\lambda & 1 \\ \beta_0 & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{p-2} & \beta_{p-1} - \lambda \end{bmatrix}.$$

Expanding $\det(\mathbf{B} - \lambda \mathbf{I})$ by its last row, and using the fact that the determinant of a matrix in *triangular form* (i.e., with all zero off-diagonal elements on one side of the diagonal) is the product of its diagonal elements (Exercise 17), one gets

$$\det(\mathbf{B} - \lambda \mathbf{I}) = (-1)^{p+1}(\beta_0 + \beta_1 \lambda + \cdots + \beta_{p-1}\lambda^{p-1} - \lambda^p). \tag{17.30}$$

Therefore, the eigenvalues of $\mathbf{B}$ are the roots of the equation

$$\beta_0 + \beta_1 \lambda + \cdots + \beta_{p-1}\lambda^{p-1} - \lambda^p = 0. \tag{17.31}$$

Finally, in view of (17.15), the following proposition holds (see (17.21) and Exercise 17).

Proposition 17.3. Suppose that the roots of the polynomial equation (17.31) are all strictly inside the unit circle in the complex plane and that the common distribution G of $\{\eta_n : n \geq 1\}$ satisfies

$$\mathbb{E} \log^+ |\eta_n| < \infty. \tag{17.32}$$

Then, (i) there exists a unique invariant distribution π for the Markov process $\{\mathbf{X}_n : n \geq 0\}$ and (ii) no matter what the initial distribution, $\mathbf{X}_n$, converges in distribution to π. In particular, the time series $\{U_n : n \geq 0\}$ converges in distribution to a steady state π_U given, for all Borel sets $C \subset \mathbb{R}$, by

$$\pi_U(C) := \pi(\{\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p : x_1 \in C\}). \tag{17.33}$$

Proof. To see that the last statement follows, simply note that U_n is the first coordinate of $\mathbf{X}_n$, so that $\mathbf{X}_n$ converges to π in distribution implies U_n converges to π_U in distribution. ■

Corollary 17.4. Assume that the roots of the polynomial equation (17.31) are all strictly inside the unit circle in the complex plane and that $E|\eta_n|^r < \infty$ some order $r > 0$. Then the conclusion of Proposition 17.3 holds.

Proof. Simply observe that this implies (17.32) holds by Jensen's inequality.[5] ■

Example 4 *(*ARMA(p, q) *Model).* The *autoregressive moving average model of order* (p, q), in short ARMA(p, q), is defined by

$$U_{n+p} := \sum_{i=0}^{p-1} \beta_i U_{n+i} + \sum_{j=1}^{q} \delta_j \eta_{n+p-j} + \eta_{n+p} \qquad (n \geq 0), \tag{17.34}$$

where p, q are positive integers, $\beta_i (0 \leq i \leq p-1)$ and $\delta_j (1 \leq j \leq q)$ are real constants, $\{\eta_n : n \geq p-q\}$ is an i.i.d. sequence of real-valued random variables, and U_i $(0 \leq i \leq p-1)$ are arbitrary initial random variables independent of $\{\eta_n : n \geq 1\}$. Consider the sequence $\{\mathbf{X}_n : n \geq 0\}$, $\{\boldsymbol{\epsilon}_n : n \geq p-q\}$ of $(p+q)$-dimensional vectors

$$\begin{aligned} \mathbf{X}_n &:= (U_n, \ldots, U_{n+p-1}, \eta_{n+p-q}, \ldots, \eta_{n+p-1})', \\ \boldsymbol{\epsilon}_n &:= (0, 0, \ldots, 0, \eta_{n+p-1}, 0, \ldots, 0, \eta_{n+p-1})' \qquad (n \geq 0), \end{aligned} \tag{17.35}$$

where η_{n+p-1} occurs as the pth and $(p+q)$th elements of $\boldsymbol{\epsilon}_n$.

$$\mathbf{X}_{n+1} = \mathbf{H}\mathbf{X}_n + \boldsymbol{\epsilon}_{n+1} \qquad (n \geq 0), \tag{17.36}$$

where $\mathbf{H}$ is the $(p+q) \times (p+q)$ matrix

$$\mathbf{H} := \left\{ \begin{array}{ccccccccc} b_{11} & \cdot & \cdots & b_{1p} & 0 & \cdot & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ b_{p1} & \cdot & \cdots & b_{pp} & \delta_q & \delta_{q-1} & \cdots & \delta_2 & \delta_1 \\ 0 & \cdot & \cdots & 0 & 0 & 1 & 0 \cdots & 0 & 0 \\ 0 & \cdot & \cdots & 0 & 0 & 0 & 1 \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 \cdots & & \cdot & \cdot & 0 & 0 \cdots & 0 & 1 \\ 0 & 0 \cdots & & \cdot & \cdot & 0 & 0 \cdots & 0 & 0 \end{array} \right\},$$

the first p rows and p columns of $\mathbf{H}$ being the matrix $\mathbf{B}$ in (17.29).

[5] BCPT p. 13.

Note that $U_0, \ldots, U_{p-1}, \eta_{p-q}, \ldots, \eta_{p-1}$ determine $\mathbf{X}_0$, so that $\mathbf{X}_0$ is independent of η_p and, therefore, of ϵ_1. It follows by induction that $\mathbf{X}_n$ and ϵ_{n+1} are independent. Hence, $\{\mathbf{X}_n : n \geq 0\}$ is a Markov process on the state space $\mathbb{R}^{p+q}$.

In order to apply the lemma above, expand $\det(\mathbf{H} - \lambda\mathbf{I})$ in terms of the elements of its pth row to get (Exercise 17)

$$\det(\mathbf{H} - \lambda\mathbf{I}) = \det(\mathbf{B} - \lambda\mathbf{I})(-\lambda)^q. \tag{17.37}$$

Therefore, the eigenvalues of $\mathbf{H}$ are q zeros together with the roots of (17.31). Thus, one has the following proposition.

Proposition 17.5. Under the hypothesis of Proposition 17.3, the ARMA(p, q) process $\{\mathbf{X}_n : n \geq 0\}$ has a unique invariant distribution π, and $\mathbf{X}_n$ converges in distribution to π no matter what the initial distribution is.

Corollary 17.6. Under the hypothesis of Proposition 17.5, the time series $\{U_n : n \geq 0\}$ converges in distribution to π_U given for all Borel sets $C \subset \mathbb{R}$ by

$$\pi_U(C) := \pi(\{\mathbf{x} = (x_1, \ldots, x_{p+q}) \in \mathbb{R}^{p+q} : x_1 \in C\}), \tag{17.38}$$

regardless of the distribution of $(U_0, U_1, \ldots, U_{p-1})$.

In the case that ϵ_n is Gaussian, it is simple to check that under the hypothesis (17.14) in Example 2 the random vector $\mathbf{Y}$ in (17.22) is Gaussian. Therefore, π is Gaussian, so that the stationary vector-valued process $\{\mathbf{X}_n : n \geq 0\}$ with initial distribution π is Gaussian. In particular, if η_n are Gaussian in Example 3, and the roots of the polynomial equation (17.31) lie inside the unit circle in the complex plane, then the stationary process $\{U_n : n \geq 0\}$, obtained when $(U_0, U_1, \ldots, U_{p-1})$ have distribution π in Example 3, is Gaussian. A similar assertion holds for Example 4.

Exercises

1. (a) Show that the a.s. convergence of the backward iteration implies the convergence in distribution of the corresponding Markov process $X_n, n \geq 0$.
   (b) Let $\{\epsilon_n : n = 1, 2, \ldots\}$ be an i.i.d. sequence of symmetric ± 1-valued Bernoulli random variables. Define a Markov process by $X_{n+1} = .5X_n + \epsilon_{n+1}, n = 0, 1, 2, \ldots$. (i) Show that the uniform distribution π on $[-2, 2]$ is the unique invariant probability. (ii) Calculate the asymptotic variance in the CLT for $\frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} X_m$.
2. (a) Let $\mathbf{B}_1, \mathbf{B}_2$ be $m \times m$ matrices (with real or complex coefficients). Define $\|\mathbf{B}\|$ as in (17.13), with the supremum over unit vectors in $\mathbb{R}^m$ or C^m. Show that

$$\|\mathbf{B}_1\mathbf{B}_2\| \leq \|\mathbf{B}_1\|\|\mathbf{B}_2\|.$$

(b) Prove that if $\mathbf{B}$ is an $m \times m$ matrix, then

$$\|\mathbf{B}\| \leq m^{1/2} \max\{|b_{ij}| : 1 \leq i, j \leq m\}.$$

(c) (Gelfand Formula) If $\mathbf{B}$ is an $m \times m$ matrix and $\|\mathbf{B}\|$ is defined to be the supremum over unit vectors in $\mathbb{C}^m$, show that $\|\mathbf{B}^n\| \geq r^n(\mathbf{B})$. Use this together with (17.15) to prove that $\lim \|\mathbf{B}^n\|^{1/n}$ exists and equals $r(\mathbf{B})$. [*Hint*: Let λ_m be an eigenvalue such that $|\lambda_m| = r(\mathbf{B})$. Then there exists $\mathbf{x} \in \mathbb{C}^m$, $\|\mathbf{x}\| = 1$, such that $\mathbf{B}\mathbf{x} = \lambda_m \mathbf{x}$.]

3. Suppose $\sum a_n z^n$ and $\sum b_n z^n$ are absolutely convergent and are equal for $|z| < r$, where r is some positive number. Show that $a_n = b_n$ for all n. [*Hint*: Within its radius of convergence, a power series is infinitely differentiable and may be repeatedly differentiated term by term.]
4. Suppose ϵ_1 is a random vector with values in $\mathbb{R}^k$. Prove that if $\delta > 1$ and $c > 0$, then $\sum_{n=1}^{\infty} P(|\epsilon_1| > c\delta^n) \leq \mathbb{E}|Z|$, where $Z = \frac{\log|\epsilon_1| - \log c}{\log \delta}$.
5. Assume that $\epsilon_n, n \geq 1$, of Example 2 are i.i.d. with common mean vector $\boldsymbol{\mu}$ and finite covariance matrix $\mathbf{D}$.

   (a) Calculate the mean and the covariance matrix of the limiting distribution of $\mathbf{X}$.[*Hint*: Use (17.22).]

   (b) If each ϵ_n is Gaussian $N(\boldsymbol{\mu}, \mathbf{D})$, determine the limiting distribution of X_n.

6. (a) In Example 1, show that $|b| < 1$ is necessary for the existence of a unique invariant probability. [*Hint*: Consider separately the cases $|b| > 1$ and $|b| = 1$.]

   (b) Show by example that $|b| < 1$ is not sufficient for the existence of a unique invariant probability. [*Hint*: Find a distribution Q of the noise ϵ_n with an appropriately heavy tail.]

   (c) Suppose (17.5) does not hold, but $0 \neq |b| < 1$. Show that $\{X_n\}$ does not converge in distribution and does not have an invariant probability. [*Hint*: Show that $\sum_{j=0}^{\infty} P(b^j \epsilon_{j+1}| > \epsilon) = \sum_{j=1}^{\infty} P(\log|\epsilon_1| > -j \log|b|) + P(\log|\epsilon_1| > 0) = \mathbb{E}\frac{\log^+|\epsilon_1|}{-\log|b|} + P(\log|\epsilon_1| > 0) = \infty$. Apply the Borel–Cantelli lemma.]

7. In Example 1, assume $\mathbb{E}\epsilon_n^2 < \infty$, and write $a = \mathbb{E}\epsilon_n$, $X_{n+1} = a + bX_n + \theta_{n+1}$, where $\theta_n = \epsilon_n - a$ $(n \geq 1)$. The *least squares estimates* of a, b are $\hat{a}_N, \hat{b}_N$, which minimize $\sum_{n=0}^{N-1}(X_{n+1} - a - bX_n)^2$ with respect to a, b.

   (a) Show that $\hat{a}_N = \bar{Y} - \hat{b}_N \bar{X}$, $\hat{b}_N = \sum_0^{N-1}(X_{n+1} - \mathbf{a}Y)(X_n - \bar{X}) / \sum_1^N (X_n - \bar{X})^2$, where $\overline{X} := N^{-1}\sum_0^{N-1} X_n$, $\overline{Y} := N^{-1}\sum_1^N X_n$.

   (b) In the case $|b| < 1$, prove that $\hat{a}_N \to a$ and $\hat{b}_N \to b$ a.s. as $N \to \infty$.

   (c) Suppose (17.5) does not hold, but $|b| < 1$, $b \neq 0$. Show that $\{X_n\}$ does not converge in distribution and does not have an invariant probability.

[*Hint*: $\sum_{j=0}^{\infty} P(|b^j \epsilon_{j+1}| > 1) = \sum_{j=1}^{\infty} P(\log|\epsilon_1| > -j \log|b|) = \sum_{j=1}^{\infty} P(\log^+|\epsilon_1| > -j\log|b|) + P(\log|\epsilon_1| > 0) = \mathbb{E}\frac{\log^+|\epsilon_1|}{-\log|b|} + P(\log|\epsilon_1| > 0) = \infty$. Apply the Borel–Cantelli lemma, Part 2, from here.]

8. In Example 2, let $m = 2$, $b_{11} = -4$, $b_{12} = 5$, $b_{21} = -10$, $b_{22} = 3$. Assume ϵ_1 has a finite absolute second moment. Show that $||B|| > 1$. Does there exist a unique invariant probability?
9. (a) Prove that the determinant of an $m \times m$ matrix in triangular form equals the product of its diagonal elements.

   (b) Check (17.30) and (17.37).
10. (Yule–Walker equations) Consider a (non-degenerate) stationary mean-zero (centered) AR(p) process $U_{n+p} = \sum_{j=0}^{p-1} \beta_j U_{n+j} + \eta_{n+p}, n = 0, 1, \ldots$. Show that the first p autocorrelations defined by $\rho_k = \mathbb{E}U_0U_k/\mathbb{E}U_0^2, k = 0, \ldots, p-1$, satisfy the so-called Yule–Walker equations $\boldsymbol{\rho} = \mathbf{R}\boldsymbol{\beta}$, where $\boldsymbol{\rho} = (\rho_0, \ldots, \rho_{p-1})'$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$, and $\mathbf{R} = ((\rho_{|i-j|}))_{0 \le i,j \le p-1}$. In particular, $\mathbf{R}$ is of full rank and symmetric, hence invertible, and the model coefficients are determined from the autocorrelations via $\boldsymbol{\beta} = \mathbf{R}^{-1}\boldsymbol{\rho}$.
11. Show that Propositions 17.1, 17.2 extend to the case of affine linear random maps: (i) $X_{n+1} = c + bX_n + \epsilon_n, n \ge 0$, (ii) $\mathbf{X}_{n+1} = \mathbf{c} + \mathbf{B}\mathbf{X}_n + \epsilon_{n+1}, n \ge 0$, for constants c, $\mathbf{c}$. [Hint: Absorb the constants into ϵ_n.]

Chapter 18
Markov Processes Generated by Iterations of I.I.D. Maps

While all discrete parameter Markov processes on a Polish state space can be represented as i.i.d. iterations of random maps, the properties of the maps obviously play a significant role in their long-run behavior. Non-decreasing monotonicity is one such property for which definitive results can be obtained, as illustrated in this chapter.

The method of construction of Markov processes by i.i.d. iterated random maps, illustrated for linear time series models in Chapter 17, extends to more general Markov processes. The present chapter is devoted to the construction and analysis of some nonlinear models. Before turning to these models, recall that one may regard the process $\{X_n : n \geq 0\}$ defined in the $AR(1)$ example of the previous chapter to be generated by *successive iterations* of an i.i.d. sequence of *random maps* $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\ldots, \boldsymbol{\alpha}_n, \ldots$ defined by

$$x \to \boldsymbol{\alpha}_n x = bx + \epsilon_n \qquad (n \geq 1),$$

$\{\epsilon_n : n \geq 1\}$ being a sequence of i.i.d. real-valued random variables. Each $\boldsymbol{\alpha}_n$ is a random (affine linear) map on the state space $\mathbb{R}^1$ into itself. The Markov sequence $\{X_n : n \geq 0\}$ is defined by successive compositions, or iterations,

$$X_n = \boldsymbol{\alpha}_n \cdots \boldsymbol{\alpha}_1 X_0 \qquad (n \geq 1), \tag{18.1}$$

where the initial X_0 is a real-valued random variable independent of the sequence of random maps $\{\boldsymbol{\alpha}_n : n \geq 1\}$. A similar interpretation holds for the other examples of Chapter 17. Indeed, Theorem 18.1 below says that, under a mild condition on

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_18

the state space, every Markov process in discrete time may be represented as (18.1). Thus the method of the last chapter and the present one is truly a general device for constructing and analyzing Markov processes on general state spaces. We begin with a precise definition of a random map.

Let $(S, \mathcal{S})$ be a measurable (state) space, and $(\Omega, \mathcal{F}, P)$ be a probability space.

Definition 18.1. A *random map* $\boldsymbol{\alpha}$ on S is a measurable function on Ω such that, for each $\omega \in \Omega$, $\boldsymbol{\alpha}(\omega)$ is a map on S into itself and

$$(\omega, x) \longrightarrow \boldsymbol{\alpha}(\omega)x \quad \text{ is measurable} \tag{18.2}$$

on $(\Omega \times S, \mathcal{F} \otimes \mathcal{S})$ into $(S, \mathcal{S})$.

The measurability condition (18.2) guarantees that

$$p(x, B) := P(\boldsymbol{\alpha}x \in B), \quad (x \in S, B \in \mathcal{S}) \tag{18.3}$$

is a transition probability. For: (i) given x, $\omega \to \boldsymbol{\alpha}(\omega)x$ defines a measurable map on Ω into S, so that $\{\omega : \boldsymbol{\alpha}(\omega)x \in B\} \in \mathcal{F}$ and the right side of (18.4) is well defined and (ii) given any $B \in \mathcal{S}$, $x \to p(x, B)$ is measurable due to the measurability of $\boldsymbol{\alpha}(\omega)x$ on the product space $(\Omega \times S, \mathcal{F} \otimes \mathcal{S})$, by a Fubini type argument (Exercise 18).

A *canonical model* of a random map may be given as the identity map $\gamma \to \gamma$ on a probability space $(\Gamma, \mathcal{G}, Q)$, where Γ is a set of maps on S into itself, $\mathcal{G}$ is a σ-field on Γ such that the map

$$(\gamma, x) \longrightarrow \gamma x \quad \text{ is measurable} \tag{18.4}$$

on $(\Gamma \times S, \mathcal{G} \otimes \mathcal{S})$ into $(S, \mathcal{S})$, and Q is an arbitrary probability on $(\Gamma, \mathcal{G})$. On the other hand, given an arbitrary random map $\boldsymbol{\alpha}$, let $\Gamma = \{\boldsymbol{\alpha}(\omega) : \omega \in \Omega\}$, and let $\mathcal{G}$ be the σ-field on Γ generated by $\boldsymbol{\alpha}$. Then (18.4) follows from (18.2), and denoting by Q the distribution of $\boldsymbol{\alpha}$ on Γ, a canonical model for $\boldsymbol{\alpha}$ is furnished by the identity map $\gamma \to \gamma$ on $(\Gamma, \mathcal{G}, Q)$.

Theorem 18.1. Let S be a Borel subset of a complete separable metric space and $\mathcal{S}$ its Borel σ-field. Given a transition probability $p(x, dy)$ on $(S, \mathcal{S})$, there exists a random map $\boldsymbol{\alpha}$ on some probability space $(\Omega, \mathcal{F}, P)$ such that (18.3) holds.

Proof. We will prove the theorem for the case S is a Borel subset of the real line. The general case follows from the fact that there exists a one-to-one map h on S onto a Borel subset C of $\mathbb{R}$ such that h and h^{-1} are both measurable; in other words, $(S, \mathcal{S})$ and $(C, \mathcal{B}^1 \cap C)$ are isomorphic, where $\mathcal{B}^1 \cap C = \{B \cap C : B$ Borel subset of $\mathbb{R}\}$, see Appendix C.

Assume S is a Borel subset of the real line, and let $p(x, dy)$ be a transition probability on $(S, \mathcal{S})$. Let F_x denote the distribution function of $p(x, dy)$ (for each given x), i.e., $F_x(y) := p(x; (-\infty, y])$. Define the *inverse* of F_x by

$$F_x^{-1}(u) := \inf\{t : F_x(t) > u\} \quad u \in (0,1). \tag{18.5}$$

Let U be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ such that U is uniformly distributed on $(0,1)$. As is well known, the random variable $F_x^{-1}(U)$ has the distribution function F_x (Exercise 18). Now define the random map $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha} x := F_x^{-1}(U) \quad (x \in \mathbb{R}). \tag{18.6}$$

The measurability condition (18.2) is easily checked (Exercise 18 4(ii)). ■

The following result is an immediate consequence of Theorem 18.1.

Corollary 18.2. Under the hypothesis of Theorem 18.1, one can construct, on an appropriate probability space $(\Omega, \mathcal{F}, P)$, an i.i.d. sequence of random maps $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$ and a random variable X_0 independent of this sequence such that $X_{n+1} := \boldsymbol{\alpha}_{n+1} X_n$ $(n \geq 0)$, X_0, is a Markov process having the given transition probability $p(x, dy)$ and a given initial distribution of X_0.

There are in general many different random maps $\boldsymbol{\alpha}$ such that (18.3) holds for a given transition probability $p(x, dy)$. This is illustrated by Example 1 below. In many applications, however, the specific representation $X_{n+1} = \boldsymbol{\alpha}_{n+1} X_n$ $(n \geq 0)$ arises from statistical, dynamical, or physical considerations. The ARMA models are also of this kind.

Example 1 *(Two-State Markov Chain).* Let $S = \{0,1\}$ with transition probabilities $p_{ij} := p(i, \{j\})$ $(i, j = 0, 1)$. Assume, for simplicity, $0 < p_{00} < p_{10} < 1$. Let $\Gamma = \{\gamma_1, \gamma_2, \gamma_3\}$, where

$$\gamma_1(0) = \gamma_1(1) = 0; \quad \gamma_2(0) = \gamma_2(1) = 1; \quad \gamma_3(0) = 1,\ \gamma_3(1) = 0. \tag{18.7}$$

Let $Q(\{\gamma_1\}) = p_{00}$, $Q(\{\gamma_2\}) = p_{11}$, $Q(\{\gamma_3\}) = p_{10} - p_{00}$. Then $p_{ij} = P(\boldsymbol{\alpha} i = j)$, where $\boldsymbol{\alpha}$ is a random map on S with distribution Q. For example, take $\boldsymbol{\alpha}$ to be the identity map $\gamma \to \gamma$ on $(\Gamma, \mathcal{G}, Q)$ with $\mathcal{G} =$ class of all subsets of $\{0,1\}$. One may check that the (noncanonical) $\boldsymbol{\alpha}$ given by (18.6) has this distribution Q. A different representation is given by taking $\Gamma = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ with

$$\gamma_1(0) = \gamma_1(1) = 0; \gamma_2(0) = \gamma_2(1) = 1; \gamma_3(0) = 0, \gamma_3(1) = 1; \gamma_4(0) = 1, \gamma_4(1) = 0. \tag{18.8}$$

Take $Q(\{\gamma_i\}) = q_i$ $(i = 1,2,3,4)$, where $q_1 + q_2 + q_3 + q_4 = 1$ and $q_1 + q_3 = p_{00}$, $q_1 + q_4 = p_{10}$. Note that the preceding representation is a special case of this, with $q_4 = 0$; but one may obtain a one-parameter family of distributions Q of the present form, with q_1 arbitrarily chosen from $[0, \min\{p_{00}, p_{10}\}]$.

As mentioned earlier, the representation of Markov processes by iterated maps is often an effective means for analyzing them. In Chapter 17, this representation is made use of to provide geometric rates of convergence to equilibrium for several general classes of Markov processes. We conclude this chapter by showing how the so-called *method of backward iterations* may sometimes be used to prove existence of and convergence to an invariant probability.

Let $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$ be a sequence of i.i.d. random maps on S, and X_0 independent of $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$. By the process obtained by *backward iteration*, we mean

$$Y_0 = X_0, \quad Y_n = \boldsymbol{\alpha}_1\boldsymbol{\alpha}_2 \ldots \boldsymbol{\alpha}_n X_0 \quad (n \geq 1). \tag{18.9}$$

Unlike the forward iterated process $\{X_n : n \geq 0\}$ given by (18.1), the process $\{Y_n\}_{n\geq 0}$ is not in general Markov; in particular, the (joint) distributions of $(X_0, X_1, \ldots, X_n)$ and $(Y_0, Y_1, \ldots, Y_n)$ are generally not the same if $n > 1$ (Exercise). On the other hand, for each n, the (marginal) distribution of Y_n is clearly the same as that of X_n. Thus if one can show that Y_n converges in distribution to some probability, then so does X_n. For certain classes of Markov processes, the sequence $\{Y_n\}_{n\geq 0}$ actually converges almost surely to some random variable Y, so that Y_n converges in distribution to (the law of) Y. This is indeed true of the linear processes considered in Chapter 17 (see (17.7)–(17.10), (17.22)).

Example 2 *(I.I.D. Monotone Increasing Maps).* Let $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$ be a sequence of i.i.d. random maps on an interval $S = I$, with $\boldsymbol{\alpha}_n(\omega)$ monotone increasing for every ω outside a P-null set N. That is, for all $x \leq y$ in I, one has

$$\boldsymbol{\alpha}_n(\omega)x \leq \boldsymbol{\alpha}_n(\omega)y \quad \text{if} \quad \omega \notin N \quad (n \geq 1). \tag{18.10}$$

Assume that *I has a smallest element* $\mathbf{a}$. For each $x \in I$, consider the *backward iterated sequence* $\{Y_n(x)\}_{n\geq 0}$ starting at x,

$$Y_0(x) = x \quad Y_n(x) = \boldsymbol{\alpha}_1\boldsymbol{\alpha}_2 \ldots \boldsymbol{\alpha}_n x \quad (n \geq 1). \tag{18.11}$$

The sequence $\{Y_n(\mathbf{a})\}_{n\geq 1}$ is *increasing* almost surely. That is, if $\omega \notin N$, then

$$\begin{aligned} Y_1(\mathbf{a}) &= \boldsymbol{\alpha}_1(\mathbf{a}) \geq \mathbf{a} = Y_0(\mathbf{a}), \\ Y_{n+1}(\mathbf{a}) &\equiv \boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_n\boldsymbol{\alpha}_{n+1}(\mathbf{a}) \geq \boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_n(\mathbf{a}) = Y_n(\mathbf{a}) \quad (n \geq 1). \end{aligned} \tag{18.12}$$

Let $\underline{Y}$ denote the (a.s.) limit of $Y_n(a)$. If $\underline{Y} < \infty$ a.s., and $\boldsymbol{\alpha}_n$ is continuous a.s., then the distribution π, say, of $\underline{Y}$ is an invariant probability for the Markov process $\{X_n\}_{n\geq 0}$ (Exercise 18). If, in addition, $Y_n(x)$ converges a.s. to the same limit $\underline{Y}$ for every $x \in I$, then π is the *unique* invariant probability of $\{X_n\}_{n\geq 0}$, since in this case the distribution $p^{(n)}(x, dy)$ of $X_n(x)$ converges weakly to π for every $x \in I$ (see Corollary 8.5). We next consider an important class of examples of this.

Example 3 *(A Class of Markov Processes on $S = [0, \infty)$ Generated by IID Monotone Maps).* Let the Markov process on $S = [0, \infty)$ be defined by

$$X_{n+1} = \max\{0, X_n + \epsilon_{n+1}\}, \quad n = 0, 1, \ldots, \tag{18.13}$$

where $\{\epsilon_n : n \geq 1\}$ are i.i.d. random variables with values in $(-\infty, \infty)$, independent of $X_0 \in [0, \infty)$. Consider the family of maps

$$f_\theta(x) = \max\{0, x + \theta\} = (x + \theta)^+, \theta \in (-\infty, \infty). \tag{18.14}$$

One may represent (18.13) as

$$X_n = \alpha_n \circ \alpha_{n-1} \circ \cdots \circ \alpha_1 X_0, \quad \alpha_n x := f_{\epsilon_n}(x) = (x + \epsilon_n)^+,\ n \geq 1. \tag{18.15}$$

For each $\theta \in (-\infty, \infty)$, f_θ is an increasing map (i.e., if $x \leq y$, then $f_\theta(x) \leq f_\theta(y)$), so that $\{\alpha_n : n \geq 1\}$ is a sequence of i.i.d. monotone maps on $S = [0, \infty)$, whose iterations generate the Markov process $\{X_n : n = 0, 1, 2, \dots\}$. Once again, let us consider the nth backward iteration

$$Y_n(z) = \alpha_1 \circ \alpha_2 \circ \cdots \circ \alpha_n X_0 = f_{\epsilon_1} \circ f_{\epsilon_2} \circ \cdots \circ f_{\epsilon_n}(z), n = 0, 1, \dots, \tag{18.16}$$

noting that Y_n has the same distribution as X_n, starting with $X_0 = z$. Also note that $Y_n(0) \uparrow$ as $n \uparrow \infty$, and if the limit, $\underline{Y}$, say, of $Y_n(0)$ is finite (almost surely), and π is the distribution of $\underline{Y}$, then π is an invariant probability, since the distribution $p^{(n)}(0, dy)$ (of $X_n(0)$) converges weakly, i.e., in distribution, to π as $n \to \infty$ (see Proposition 8.4). We now explore the following important result.

Theorem 18.3. For the Markov process defined by (18.13), assume $\mathbb{E}\epsilon_1 < 0$. Then the Markov process has a unique invariant probability π, and $p^{(n)}(z, dy)$ converges to π weakly as $n \to \infty$, for any $z \in [0, \infty)$.

Before proving the theorem, we consider a number of important applications.

1. (*G/G/1 Queue*) In this popular queuing model with a single server, customers arrive one after another at times T_n, $(n = 1, 2, \dots)$, $T_0 = 0$, such that the inter-arrival times $U_n = T_n - T_{n-1}$ are i.i.d. and independent of the service times V_n (for the nth customer), $n \geq 1$, which are also i.i.d. Then the waiting times W_n for the nth customer, $n \geq 1$, satisfy the recursive relation

   $$W_{n+1} = \max\{W_n + V_n - U_n, 0\} = \max\{0, W_n + \epsilon_n\}, \epsilon_{n+1} := V_n - U_{n+1}. \tag{18.17}$$

   Thus the waiting time approaches a steady state if the expected service time of a customer is less than the expected inter-arrival time.
2. (*Lindley–Spitzer Process of Resource Management*) Consider a resource, such as ground water, fish, etc., whose amount at time n (say in year n) is $X_n \geq 0$. At time $n + 1$, a random input R_{n+1} (rainfall, hatchery, etc.) arrives, $R_n, n \geq 1$, i.i.d. nonnegative random variables. A desired consumption level is $c > 0$. If $X_n + R_{n+1} \geq c$, then an amount c is consumed, and if $X_n + R_{n+1} < c$, then $X_n + R_{n+1}$ is consumed. In either case, the remaining amount of the resource at time $n + 1$ is $X_{n+1} = \max\{0, X_n + R_{n+1} - c\}$. Writing $\epsilon_{n+1} = R_{n+1} - c$, one

has the Markov model (18.13). If $\mathbb{E}\epsilon_1 < 0$, that is, $\mathbb{E}R_{n+1} < c$, then the Markov process $\{X_n : n \geq 0\}$ has a unique invariant distribution.

3. (*Problem of Ruin in Insurance*) In the general renewal model of insurance, also known as the Sparre–Andersen model, claims of strictly positive sizes $Z_1, Z_2, \ldots$ arrive at random times $T_1, T_2, \ldots,$ and a constant premium $c > 0$ per unit of time is collected. The sequences $\{Z_n : n \geq 1\}$ and $\{T_n : n \geq 1\}$ are assumed to be independent. It is also assumed that the inter-arrival times $A_i = T_i - T_{i-1} (i = 1, 2, \ldots)$ are i.i.d. with $T_0 = 0$, and $\mathbb{E}A_i = 1/\lambda$ is finite. For an insurance company with an initial cash reserve $u > 0$, the probability of ruin is

$$\psi(u) = P(\sum_{i-1}^{n} Z_i > u + c\sum_{i=1}^{n} A_i \text{ for some } n) = P(S_n > u \text{ for some } n), \tag{18.18}$$

where $S_n = \sum_{i=1}^{n} Z_i$, $\epsilon_i = Z_i - cA_i$, $(n \geq 1)$, $S_0 = 0$. The insurance company requires the following *net profit condition (npc)*:

$$\mathbb{E}\epsilon_1 < 0. \tag{18.19}$$

The (npc) implies that $M := \sup\{S_n : n \geq 0\}$ is finite almost surely, and one may express the ruin probability (18.18) as

$$\psi(u) = P(M > u). \tag{18.20}$$

Note that M is the same as $\underline{Y}$ in the proof of Theorem 18.3, so that

$$\psi(u) = \pi(u, \infty), \tag{18.21}$$

where π is the unique invariant distribution of the Lindley–Spitzer process. If one defines a Markov process $\{X_n : n = 0, 1, \ldots\}$ on $S = [0, \infty)$ as in Theorem 18.3, with $X_0 = x \geq 0$, then of course π is its unique invariant probability, and (18.21) holds.

In the problem of this application to insurance, that Markov process is rather contrived. In the present application, the objective is to find the probability of ruin as a function of the initial asset u. However, this formal link allows one to find many details about the invariant probability of the Markov process in the general context of Theorem 18.3, and in the special contexts of the queue and resource applications, simply because of the huge existing literature on ruin probabilities in insurance. Also see Exercise 18.

In order to prove Theorem 18.3, we first prove a simple lemma.

Lemma 1. With the notation and (18.14) as above, one has for all $n \geq 1$ and all $\theta_i, 1 \leq i \leq n$,

$$f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_n}(0) = \max\left\{0, \sum_{i=1}^{j} \theta_i, 1 \le j \le n\right\}. \tag{18.22}$$

Proof. We use induction. The result is obvious for $n = 1$, since $f_{\theta_1}(0) = \max\{0, \theta_1 + 0\} = \max\{0, \theta_1\}$. Assume that (18.22) holds for some n. Then

$$\begin{aligned} f_{\theta_1} \circ f_{\theta_2} \circ \cdots \circ f_{\theta_n} \circ f_{\theta_{n+1}}(0) &= f_{\theta_1}(\max\{0, \textstyle\sum_{i=2}^{j} \theta_i, 2 \le j \le n+1\}) \\ &= \max\{0, \max\{0, \textstyle\sum_{i=2}^{j} \theta_i, 2 \le j \le n+1\} + \theta_1\} \\ &= \max\{0, \textstyle\sum_{i=1}^{j} \theta_i, 1 \le j \le n+1\}. \end{aligned}$$

■

Proof. Consider the backward iteration $Y_n(z)$ in (18.16) that has, for each n, the same distribution as $X_n(z)$. For $z = 0$, one has by the above lemma,

$$Y_n(0) = \max\{0, \sum_{i=1}^{j} \epsilon_i, 1 \le j \le n\}. \tag{18.23}$$

It is easy to check directly using monotonicity and $f_{\epsilon_{n+1}}(0) \ge 0$ that $Y_n(0) \uparrow \underline{Y}$ for some $\underline{Y}$ as $n \to \infty$. This is also clear from (18.23). We will show that $\underline{Y} < \infty$ almost surely under the hypothesis $\mathbb{E}\epsilon_1 < 0$. It follows by the strong law of large numbers that $\sum_{i=1}^{n} \epsilon_i \to -\infty$ almost surely as $n \to \infty$. Hence, there exists $N = N(\omega)$ such that $\sum_{i=1}^{n} \epsilon_i < 0$ for all $n > N(\omega)$ and, therefore,

$$\underline{Y}(\omega) = \max\{0, \sup(\sum_{i=1}^{n} \epsilon_i, n \ge 1)\} = \max\{0, \sum_{i=1}^{j} \epsilon_i, 1 \le j \le N(?)\} < \infty. \tag{18.24}$$

Letting π denote the distribution of $\underline{Y}$, one then has $Y_n(0)$ converges in distribution to π. Therefore, $X_n(0)$ converges in distribution to π as $n \to \infty$, i.e., $p^{(n)}(0, dy)$ converges weakly to $\pi(dy)$. Since the Markov process has the Feller property, it follows that π is an invariant probability of the process $\{X_n : n \ge 0\}$. To prove that it is the unique invariant probability, consider that, for an arbitrary $z \ge 0$, $Y_n(z)$ that may be represented by the lemma, noting that $f_{\epsilon_n}(z) = \max\{0, z + \epsilon_n\} = f_{\epsilon_n+z}(0)$, as

$$\begin{aligned} Y_n(z) &= f_{\epsilon_1} \circ f_{\epsilon_2} \circ \cdots \circ f_{\epsilon_n}(z) \\ &= f_{\epsilon_1} \circ f_{\epsilon_2} \circ \cdots \circ f_{\epsilon_{n-1}} \circ f_{\epsilon_n+z}(0) \\ &= \max\{0, \sum_{i=1}^{n-1} \epsilon_i + \epsilon_n + z\}. \end{aligned} \tag{18.25}$$

Once again, there exists $N' = N'(\omega)$ such that $\sum_{i=1}^{n-1} \epsilon_i + \epsilon_n + z < 0$ for all $n > N'$, so that (18.25) reduces to $Y_n(z) = \max\{0, \sum_{i=1}^{n-1} \epsilon_i\}$ for all $n > N'$. Thus the limit of $Y_n(z)$ is $\underline{Y}$, the same as that for $Y_n(0)$, as $n \to \infty$. Thus $Y_n(z)$ converges in distribution to π for every z, and so does $X_n(z)$. In particular, $p^{(n)}(z, dy) \Rightarrow \pi(dy)$, by Corollary 8.5. ■

Remark 18.1. It is noteworthy that the limit distribution π of the Markov chain $\{X_n\}$ is reached in a finite number of backward iterations. In a Monte Carlo simulation context, this phenomenon is referred to as a *perfect simulation.*

To conclude, let us consider the nature of the invariant probability π of the Lindley–Spitzer process.

Proposition 18.4. Let $(S, \mathcal{S})$ be a measurable space and $p(x, dy)$ a transition probability on S. Let ρ be a σ-finite measure on $(S, \mathcal{S})$. Suppose that for some n, the n-step transition probability measure $p^{(n)}(x, dy)$ is absolutely continuous with respect to $\rho(dy)$, for every $x \in S$. (i). If π is an invariant probability for $p(x, dy)$, then π is absolutely continuous with respect to ρ. (ii). Let x_0 be a recurrent point and $p^{(n)}(x_0, dy)$, absolutely continuous with respect to ρ for every $n > 0$. Then π is absolutely continuous with respect to ρ as well.

Proof. Suppose $\rho(B) = 0$ for some $B \in \mathcal{S}$. (i) Since $p^{(n)}(x, B) = 0$ for all $x \in S$, by hypothesis, $\pi(B) = \int_S p^{(n)}(x, B)\pi(dx) = 0$. (ii) In this case, $p^{(n)}(x_0, B) = 0$ for every $n = 1, 2, \dots$. Starting from any given $x \in S$, the process $\{X_n : n \geq 0\}$ reaches x_0 with probability one, but the probability is zero that from x_0 the process ever enters B. Since the expected amount of time the process spends in B in a single cycle starting at x_0 and returning to x_0 is zero, $\pi(B) = 0$. ■

Turning to the Lindley–Spitzer process satisfying $\mathbb{E}Z_i < 0$, note that 0 is a positive recurrent state. Denote by G the distribution of Z_1. Let ρ be a sigma-finite measure on $(S, \mathcal{S})$ such that (i) $\rho(0) > 0$ and (ii) for every $B \in \mathcal{B}(0, \infty)$ for which $\rho(B) = 0$, one has $G^{*n}(B) = 0$ for every $n = 1, 2, \dots$. We will show that π is absolutely continuous with respect to ρ. To see this, let $\rho(B) = 0$ for some $B \in \mathcal{B}(0, \infty)$. Then, for every $n = 1, 2, \dots,$

$$p^{(n)}(0, B) = P(\max S_j : j = 1, \dots, n \in B) \leq \sum_{1 \leq j \leq n} P(S_j \in B) = 0.$$

Therefore, the hypothesis of Proposition 18.4 is satisfied. In particular, if G is absolutely continuous with respect to Lebesgue measure, then one may take ρ to be the measure that assigns a unit mass to $\{0\}$ and Lebesgue measure on $(0, \infty)$. Hence, if G is absolutely continuous on $\mathbb{R}$, then π has a point mass at zero and a density on $(0, \infty)$.

Remark 18.2. It is known from Spitzer (1956) that a necessary and sufficient condition for the conclusion of Theorem 18.3 to hold is $\sum_{n=1}^{\infty} \frac{1}{n} P(\epsilon_1 + \epsilon_2 + \cdots + \epsilon_n > 0) < \infty$ (see Exercise 18).

Remark 18.3. In Bhattacharya and Waymire (2021), the ruin problem was introduced, and a treatment of ruin probabilities was given for the general renewal model when the claim sizes are light tailed, i.e., having a finite moment generating function in some neighborhood of zero, and in the so-called heavy-tail case. In the former case, Feller's approach using Blackwell's ladder heights, as well as Blackwell's renewal theorem, plays crucial roles.[1] Recent work has focused on the latter case where Blackwell's ladder heights are not applicable.

The results on linear Markov processes in Chapter 17 may be extended to processes obtained by iterations of general i.i.d. contracting maps on a Polish space (S, ρ). A map $f : S \to S$ is *Lipschitz* (with coefficient L) if there exists $L > 0$ such that $\rho(f(x), f(y)) \leq L\rho(x, y)$, for all x, y in S. The map f is a contraction if it is Lipschitz with the coefficient $L = 1$, and it is a strict contraction if $\rho(f(x), f(y)) < \rho(x, y)$, for all x, y in S. Let α_n, $n \geq 1$, be an i.i.d. sequence of random contractions defined on a probability space $(\Omega, \mathcal{F}, P)$.

We begin with a result on compact metric spaces (S, ρ) due to Dubins and Freedman (1966).

Theorem 18.5. Let (S, ρ) be a compact metric space and Γ the set of all contractions on S, endowed with the supremum norm $|| \cdot ||$. Let Q be a probability on the Borel sigma-field of Γ. If a strict contraction belongs to the support of Q, then the Markov process $X_n(x) := \alpha_n \ldots \alpha_2\alpha_1 x$ $(n \geq 1)$, $X_0(x) = x$ $(x \in S)$, converges in distribution to its unique invariant probability, whatever be the initial state.

Proof. Let γ be a strict contraction in the support of Q. Then, writing γ^j for the jth iterate of γ, one has:

(i) $\text{diam}\bigcap_{0\leq j<\infty} \gamma^j(S) =$ a singleton, say, $\{x_0\}$.
(ii) $\text{diam}(\gamma^j(S)) \downarrow 0$ as $j \uparrow \infty$.

To prove (i), recall that by the finite intersection property,[2] $S_0 := \bigcap_{0\leq j\leq\infty} \gamma^j(S))$ is nonempty. If this set has more than one point say, x_0 and y_0, that would contradict the fact that $\gamma(S_0) = S_0$. For, diam of S_0 is $\rho(x_0, y_0)$, while that of $\gamma(S_0)$ is $\rho(\gamma x_0, \gamma y_0) < \rho(x_0, y_0)$. Note that x_0 is the unique fixed point of γ. To prove (ii), suppose, if possible, there exist $\delta > 0$ and a positive integer j_0 such that for all $j \geq j_0$, $\text{diam}(\gamma^j(S)) \geq \delta$. Then $\text{diam}\bigcap_{j_0\leq j<\infty} \gamma^j(S) = \text{diam}\bigcap_{j_0\leq j<\infty} \gamma^j(S) \geq \delta$, which contradicts (i). We next prove that, given x,

$$P(\sup\{\rho(X_n(x), X_n(y)) : y \in S\}) \to 0 \text{ as } n \to \infty. \tag{18.26}$$

For this, fix $\epsilon > 0$. By (1)(ii), there exists $j(\epsilon)$ such that $\text{diam}\gamma^{j(\epsilon)}(S) < \epsilon$. By the support property, $\delta(\epsilon) := Q(\gamma' \in \Gamma : ||\gamma' - \gamma|| < \epsilon/j(\epsilon)) > 0$. Consider the sequence of independent events

[1] See the specific subject matter texts by Ramasubramanian (2009) and by Rolski et al. (2010) for comprehensive treatments of the ruin problem and much more.

[2] BCPT, p. 242.

$$A_m = \{||\alpha_{j(\epsilon)(m-1)+k} - \gamma|| < \epsilon/j(\epsilon) \text{ for all } k = 1, \ldots, j(\epsilon)\}(m \geq 1).$$

Then $Q(A_m) = \delta(\epsilon)^{j(\epsilon)} > 0$. By Borel–Cantelli Lemma II,[3]

$$P(A_m \text{occurs for infinitely many} m) = 1.$$

But on A_m,

$$||\alpha_{j(\epsilon)(m-1)+1}\alpha_{j(\epsilon)(m-1)+2} \ldots \alpha_{j(\epsilon)(m-1)+j(\epsilon)} - \gamma^{j(\epsilon)}|| < j(\epsilon)\epsilon/j(\epsilon) = \epsilon,$$

as is shown by the bound $\epsilon/j(\epsilon)$ obtained by replacing the α's successively by γ. Hence, on A_m, $\sup\{\rho(X_m(x), X_m(y)) : x, y \in S\} < \epsilon$. The proof of (18.26) is now complete, recalling that all the α's are contractions. Finally, in view of compactness of the space $\mathcal{P}(S)$ of probabilities with the topology of weak convergence, and Feller continuity (Proposition 8.6), there exists an invariant probability π. If one takes the initial state to be X_0 having distribution π, then it follows from (8.6) that, whatever be the initial state y, $X_n(y)$ converges in distribution to π as $n \to \infty$. In particular, π is the unique invariant probability (Corollary 8.5). ■

For the next result, due to Silverstrov and Stenflo (1998), define the bounded Lipschitzian distance d_{BL} on the space $\mathcal{P}(S)$ of probability measures on $(S, \mathcal{S})$ by

$$d_{BL}(\mu, \nu) = \sup\left\{\left|\int_S f d\mu - \int_S f d\nu\right| : f \in BL\right\}, \tag{18.27}$$

where

$$BL := \{f : S \to \mathbb{R}, |f(x) - f(y)| \leq \min(\rho(x, y), 1)\}.^{4} \tag{18.28}$$

The space BL of bounded, Lipschitz functions is endowed with the topology of uniform convergence on compact subsets. A sequence of i.i.d. Lipschitz maps α_n, $n \geq 1$, is defined on a probability space $(\Omega, \mathcal{F}, P)$ as random maps into BL. We consider the Markov process $X_n = \alpha_n\alpha_{n-1} \ldots \alpha_1 X_0$, $n \geq 0$, where X_0 is independent of $\{\alpha_n, n \geq 1\}$. In particular, denote it by $X_n(x)$ if X_0 is the constant $x \in S$. As before, $Y_n(x) = \alpha_1\alpha_2 \ldots \alpha_n x$, $n \geq 1$, denotes the backward iteration. The random Lipschitz coefficient of $\alpha_k\alpha_{k-1} \ldots \alpha_j x$ is defined, for $j \leq k$, by

$$L_j^k = \sup\{\rho(\alpha_k\alpha_{k-1} \ldots \alpha_j x, \alpha_k\alpha_{k-1} \ldots \alpha_j y)/\rho(x, y) : x, y \in S, x \neq y\}. \tag{18.29}$$

Theorem 18.6. Let (S, ρ) be a Polish space and $\{\alpha_n, n \geq 1\}$, $\{X_n, n \geq 0\}$, $\{L_j^k\}$, as above. If for some $r \geq 1$

[3] BCPT, p. 34.

[4] BCPT p.242, p. 34.

$$\mathbb{E}(\log L_1^r) < 0, \tag{18.30}$$

and, for some point x_0 in S,

$$\mathbb{E}\log^+ \rho(\alpha_r\alpha_{r-1}\ldots\alpha_1 x_0, x_0) < \infty, \tag{18.31}$$

then the Markov process X_n, $n \geq 0$, has a unique invariant probability π, and X_n converges in distribution to π as $n \to \infty$, no matter what the initial state is.

Proof. First assume (18.30) holds for $r = 1$. We will establish the following two assertions (a),(b) and then show that the desired result follows from them: (a) $\sup\{\rho(Y_n(x), Y_n(y)) : \rho(x, y) \leq M\} \to 0$ in probability as $n \to \infty$, for all $M > 0$. (b) For some x_0 in S, the sequence of distributions of $\rho(X_n(x_0), x_0)$, $n \geq 1$, is relatively weakly compact. Assume (a), (b) hold. Let $f \in BL$ and $M > 0$. Then

$$\begin{aligned}\sup_{\rho(x,y)\leq M} &|\mathbb{E}f(X_n(x)) - \mathbb{E}f(X_n(y))| \\ &= \sup \rho(x, y) \leq M|\mathbb{E}f(Y_n(x)) - \mathbb{E}f(Y_n(y))| \\ &\leq \sup_{\rho(x,y)\leq M} \mathbb{E}(\rho(Y_n(x), Y_n(y)) \wedge 1) \\ &\to 0 \text{ as } n \to \infty\end{aligned} \tag{18.32}$$

by (a). Also,

$$\begin{aligned}&|\mathbb{E}f(X_n + m(x_0)) - \mathbb{E}f(X_n(x_0))| \\ &= |\mathbb{E}f(Y_n + m(x_0)) - \mathbb{E}f(Y_n(x_0))| \\ &= |\mathbb{E}f(\alpha_1\alpha_2\ldots\alpha_n\alpha_{n+1}\alpha_{n+2}\ldots\alpha_{n+m}x_0) - \mathbb{E}f(\alpha_1\alpha_2..\alpha_n x_0)| \\ &\leq \mathbb{E}[\rho(\alpha_1\alpha_2\ldots\alpha_n\alpha_{n+1}\alpha_{n+2}\ldots\alpha_{n+m}x_0, \alpha_1\alpha_2\ldots\alpha_n x_0) \wedge 1) \\ &\leq P(B_1) + \mathbb{E}(1_{B_2}.\rho(\alpha_1\alpha_2\ldots\alpha_n\alpha_{n+1}\alpha_{n+2}\ldots\alpha_{n+m}x_0, \alpha_1\alpha_2\ldots\alpha_n x_0) \wedge 1),\end{aligned} \tag{18.33}$$

where

$$B_1 = \{\rho(\alpha_{n+1}\alpha_{n+2}\ldots\alpha_{n+m}x_0, x_0) > M\}$$

and

$$B_2 = \{\rho(\alpha_{n+1}\alpha_{n+2}\ldots\alpha_{n+m}x_0, x_0) \leq M\}.$$

Note that $P(B_1) = P(\rho(\alpha_1\alpha_2\ldots\alpha_m x_0, x_0) > M) \to 0$ uniformly in $m = 1, 2, \ldots$, as $M \to \infty$, in view of (b). Given $\epsilon > 0$, and $m \geq 1$, choose $M = M(\epsilon)$ such that $P(B_1) < \epsilon$. Writing $X = \alpha_{n+1}\alpha_{n+2}\ldots\alpha_{n+m}x_0$, the second summand in the last line of (18.33) may be expressed as $\mathbb{E}\mathbf{1}_{B_2}(\rho(Y_n(X), Y_n(x_0)) \wedge 1) \leq \mathbb{E}Z \wedge 1$, where $Z = \sup\{\rho(Y_n(x), Y_n(y) : \rho(x, y) \leq M(\epsilon)\}$. Now, given $\delta > 0$, $\mathbb{E}(Z \wedge 1) = \mathbb{E}(\mathbf{1}_{\{Z>\delta\}} Z \wedge 1) + \mathbb{E}(\mathbf{1}_{\{Z\leq\delta\}} Z \wedge 1) \leq P(Z > \delta) + \delta$. By (a) $P(Z > \delta) \to 0$ as $n \to \infty$. Thus for all sufficiently large n, the first line of (18.33) can be made as small as one likes. In other words,

$$\sup\{|\mathbb{E}f(X_{n+m}(x_0)) - \mathbb{E}f(X_n(x0))| : f \in BL\} \to 0 \text{ as } n \to \infty \quad (m = 1, 2, \ldots). \tag{18.34}$$

Thus $\{X_n(x_0) : n \geq 0\}$ is Cauchy in the d_{BL} distance, and the transition probability is Feller continuous. Therefore, $p^{(n)}(x, dy)$ converges weakly to a unique invariant probability π (See Lemma 2 and Theorem 18.7).

It remains to prove (a) and (b). To prove (a), let L_n $(n \geq 1)$ be the i.i.d. Lipschitz coefficients of α_n $(n \geq 1)$, with $L_1 \equiv L_{1^1}$,

$$L_n = \sup\{\rho(\alpha_n(x), \alpha_n(y))/\rho(x, y) : x \neq y\}\,(n = 1, 2, \ldots),\ L_n(x, x) = 0. \tag{18.35}$$

Then

$$\begin{aligned}\rho(Y_n(x), Y_n(y)) &= \rho(\alpha_1\alpha_2 \ldots \alpha_n x, \alpha_1\alpha_2 \ldots \alpha_n y)\\ &\leq L_1\rho(\alpha_2 \ldots \alpha_n x, \alpha_2 \ldots \alpha_n y)\\ &\leq \cdots \leq L_1 L_2 \cdots L_n \rho(x, y),\end{aligned}$$

so that

$$\sup\{\rho(Y_n(x), Y_n(y)) : \rho(x, y) \leq M\} \leq L_1 L_2 \ldots L_n M, \tag{18.36}$$

and, by (18.30) and the strong law of large numbers, with probability one,

$$\begin{aligned}&(\log L_1 + \log L_2 + \ldots + \log L_n + \log M)/n \to \mathbb{E}\log L_1 < 0,\ \text{ as } n \to \infty; \qquad (18.37)\\ &\log L_1 + \log L_2 + \ldots + \log L_n \to -\infty.\end{aligned}$$

Therefore, the right side of (18.36) goes to zero almost surely, as $n \to \infty$. This proves (a). We next turn to the verification of (b). Note that, by the triangle inequality,

$$\begin{aligned}&|\rho(\alpha_1\alpha_2 \ldots \alpha_n\alpha_{n+1}\alpha_{n+2} \ldots \alpha_{n+m}x_0, x0) - \rho(\alpha_1\alpha_2 \ldots \alpha_n x_0, x_0)| \qquad (18.38)\\ &\quad\leq \rho(\alpha_1\alpha_2 \ldots \alpha_n\alpha_{n+1}\alpha_{n+2} \ldots \alpha_{n+m}x_0, \alpha_1\alpha_2 \ldots \alpha_n x_0)\\ &\quad\leq \sum_{1\leq j\leq m} \rho(\alpha_1 \ldots \alpha_{n+j}x_0, \alpha_1 \ldots \alpha_{n+j-1}x_0)\\ &\quad\leq \sum_{1\leq j\leq m} L_1 L_2 \ldots L_{n+j-1}\rho(\alpha_{n+j-1}x_0, x_0).\end{aligned}$$

First assume $0 < c := -\mathbb{E}\log L_1 < \infty$. Let $0 < \epsilon < c$. By (18.30) and the strong law of large numbers, outside a P-null set N_1, there exists $n_1 = n_1(\omega)$ such that $\log(L_1 L_2 \dots L_n)^{1/n} = (1/n)\sum_{1\leq k\leq n}\log L_k < -(c-\epsilon/2)$, i.e., $(L_1 L_2 \dots L_n)^{1/n} < \exp\{-(c-\epsilon/2)\}$, and

$$L_1 L_2 \dots L_n < \exp\{-n(c-\epsilon/2)\} \text{ for all } n \geq n1(\cdot). \tag{18.39}$$

In view of (18.31), one now has

$$\begin{aligned}\sum_{1\leq k<\infty} & P((\log^+\rho(\alpha_k x_0, x_0))/(\epsilon/2)) > k)\\ &= \sum_{1\leq k<\infty} P((\log^+\rho(\alpha_1 x_0, x_0))/(\epsilon/2)) > k)\\ &\leq \mathbb{E}V < \infty,\end{aligned} \tag{18.40}$$

where $V = \log^+\rho(\alpha_1 x_0, x_0))/(\epsilon/2)$. By the first Borel–Cantelli lemma, it now follows that outside a P-null set N_2, there exists $n_2 = n_2(\omega)$ such that

$$\rho(\alpha_k x_0, x_0) \leq \exp k\epsilon/2 \text{ for all } k \geq n_2(\cdot). \tag{18.41}$$

Applying (18.39) and (18.41) to (18.38), it now follows that outside a P-null set $N = N_1 \cup N_2$, one has

$$\begin{aligned}&\sup_{m\geq 1}|\rho(\alpha_1\alpha_2\dots\alpha_n\alpha_{n+1}\cdots\alpha_{n+m}x_0, x_0) - \rho(\alpha_1\alpha_2\cdots\alpha_n x_0, x_0)|\\ &\quad\leq \sum_{1\leq j<\infty}\exp\{-(n+j-1)(c-\epsilon/2)+(n+j)\epsilon/2\}\\ &\quad= \sum_{1\leq j<\infty}\exp\{-(n+j)(c-\epsilon)+(c-\epsilon/2)\} \text{ for all } n \geq n_3(\cdot)\\ &\quad= \max(n_1(\cdot), n_2(\cdot)).\end{aligned} \tag{18.42}$$

Since the last sum goes to zero as $n \to \infty$, it follows that $\{\rho(\alpha_1\alpha_2\dots\alpha_n x_0, x_0), n \geq 1\}$, is a Cauchy sequence and, therefore, converges outside a P-null set N. This clearly implies (b) in case c is finite, i.e., $-\mathbb{E}\log L_1 \equiv -\mathbb{E}\log L_{1^1} < \infty$. If c equals ∞, then (18.39) and (18.42) hold for any $c > 0$.

Finally, suppose (18.30) holds for some $r > 1$. The argument above may now be applied to the transition probability $p^{(r)}(x, dy)$. It follows that, for every $x \in S$, $p^{(kr)}(x, dy)$ converges to an invariant probability $\tilde{\pi}$, say, in the distance d_{BL} as $k \to \infty$ and, therefore, weakly. This implies that the distribution of X_{kr} with initial distribution μ, namely, $T^{*kr}\mu$, converges weakly to $\tilde{\pi}$, for every probability measure μ, as $k \to \infty$. Letting $\mu = p^{(j)}(x, dy)$, $j > 0$, one then gets $p^{(kr+j)}(x, dy)$ converges weakly to $\tilde{\pi}$, for every j=1,2,.., as $k \to \infty$. On a Polish space (S, ρ),

$(\mathcal{P}(S), d_P)$ is a complete separable metric space. In particular, a Cauchy sequence in the Prokhorov distance d_P converges. A proof of this and of the completeness of the d_{BL} distance is given in Theorems 18.7 and Proposition 18.8 below. This completes the proof of Theorem 18.6. ■

The distance function d_{BL} may be obtained as a special case of *Kantorovich–Rubenstein–Wasserstein metrics* (Rachev (1991)) and is independent due to Dudley (1968). The following lemma implies that $\mathcal{P}(S)$ is Cauchy under the Prokhorov distance d_P, since it is so under the d_{BL} distance.

Lemma 2. Let (S, ρ) be a metric space. On $\mathcal{P}(S)$, the following relation holds: $d_P(\mu, \nu) \le (d_{BL}(\mu, \nu))^{1/2}$.

Proof. Fix $0 < \epsilon \le 1$. For a Borel set B, the function $f(x) = \max\{0, 1 - \epsilon^{-1}\rho(x, B)\}$ satisfies $|f(x) - f(y)| \le \epsilon^{-1}\rho(x, y)$, so that $\epsilon f \in BL : |\epsilon f(x) - \epsilon f(y)| \le 1, |\epsilon f(x) - \epsilon f(y)| \le \rho(x, y)$. Now

$$\begin{aligned}\nu(B) &\le \int_S f d\nu = \epsilon^{-1} \int_S \epsilon f d\nu \\ &= \epsilon^{-1}\Big[\int_S \epsilon f d(\nu - \mu) + \int_S \epsilon f d\mu\Big] \\ &\le \epsilon^{-1} d_{BL}(\mu, \nu) + \mu(B^\epsilon), \end{aligned} \tag{18.43}$$

where $B^\epsilon = \{x : \rho(x, B) < \epsilon\} \ge 0$ on S. The first inequality in (18.43) holds because $f = 1$ on B and $f \ge 0$ on S. The last inequality follows from the facts that $\epsilon f \in BL$, and $f = 0$ outside B^ϵ, $0 \le f \le 1$ on S. Interchanging the roles of ν and μ in (18.43), it now follows that $d_P(\mu, \nu) \le \epsilon^{-1} d_{BL}(\mu, \nu)$. Letting $d_{BL}(\mu, \nu) = \epsilon^2$, one gets $d_P(\mu, \nu) \le \epsilon$, whatever be ϵ, $0 < \epsilon \le 1$. ■

Remark 18.4. One importance of Theorem 18.6 is that it truly extends the linear theory presented in Chapter 17. In other words, all the main results of Chapter 17 follow from Theorem 18.6 almost as immediate corollaries (Exercise 9).

Remark 18.5. Theorem 18.5, due to Dubins and Freedman (1966), cannot be obtained as a special case of Theorem 18.6. One can easily construct contractions with Lipschitz constant 1, perhaps with a strict contraction in the support, but not an atom, for which Theorem 18.5 holds, but Theorem 18.6 does not. On the other hand, Theorem 18.6 only requires $\mathbb{E} \log L < 0$, allowing larger Lipschitz constants than 1 in the support.

Let (S, ρ) be a complete separable metric space. We will show that the space $\mathcal{P}(S)$ of all probability measures on the Borel sigma-field $\mathcal{S}$ of S is complete under the Prokhorov distance d_P and also under the bounded Lipschitz distance d_{BL}. That d_P metrizes the weak topology is a standard fact.[5] It is simple to check that d_{BL} is a

[5] BCPT, Proposition 7.14, pp. 146,147.

metric on $\mathcal{P}(S)$. From Lemma 2, it follows the metric topology under d_{BL} $\mathcal{P}(S)$ is at least as strong as the weak topology; the arguments below will show that the two topologies are the same.

Theorem 18.7. On a complete separable metric space (S, ρ), the space $(\mathcal{P}(S), d_P)$ is a complete separable metric space.

Proof. The proof will involve two steps.

Step 1. Let us begin by proving that, under d_P, a sequence P_n $(n \geq 1)$ in $\mathcal{P}(S)$ is *tight* if, for each $\epsilon > 0$, $\delta > 0$, *there exist a finite set* $\{B_j : j = 1, \ldots, m\}$ *of spheres of radius* δ *such that* $P_n(\bigcup_{1\leq j\leq m} Bj) > 1 - \epsilon$ *for all* n. For this, fix $\epsilon > 0$. By the italicized condition, for each k, there exist a set of m_k spheres $\{B_{k1}, \ldots, B_{kmk}\}$ of radius $1/k$ such that $P_n(\bigcup_{1\leq j\leq mk} B_j) > 1 - \epsilon/2k$ $(k = 1, 2, \ldots)$. The set $D = \bigcap_{k\geq 1}(\bigcup_{1\leq j\leq mk} B_j)$ is *totally bounded*, since given $\eta > 0$, letting k be such that $1/k < \eta$, the mk balls of radius $1/k$ cover D: $D \subset \bigcup_{1\leq j\leq mk} Bj$. On the other hand, $P_n(D^c) < \sum_{1\leq k<\infty} \epsilon/2k = \epsilon$, so that $P_n(D) > 1 - \epsilon$ for all n. Let K be the closure of D. Then K is compact[6] and $P_n(K) > 1 - \epsilon$ for all n. This proves that $\{P_n : n \geq 1\}$ is tight.

Step 2. We now show that if $\{P_n : n \geq 1\}$ is a Cauchy sequence under d_P, then it is tight. We need to check the italicized condition in Step 1. Let $\epsilon > 0$ and $\delta > 0$. Consider $\eta < \min\{\epsilon, \delta\}/2$, and find $n(\eta)$ such that $d_P(P_n, P_{n'}) < \eta$ for all $n, n' \geq n(\eta)$. Because (S, ρ) is a separable metric space, there exist a finite number of spheres B_j $(j = 1, \ldots, m)$ of radius η such that $P_{n(\eta)}(\bigcup B_j : 1 \leq j \leq m) > 1 - \eta$. Let $G_1, \ldots, G_m$ be the spheres with the same centers as $B_1, \ldots, B_m$, respectively, but with radius 2η. Then $(\bigcup B_j : 1 \leq j \leq m)^\eta \subset (\bigcup G_j : j = 1, .., m)$, and, by definition of d_P, $P_n((\bigcup B_j : 1 \leq j \leq m)^\eta + \eta \geq P_{n(\eta)}(\bigcup B_j : 1 \leq j \leq m) \geq 1 - \eta$ for all $n \geq n(\eta)$. That is, $P_n(\bigcup G_j : j = 1, \ldots, m) \geq P_n((\bigcup B_j : 1 \leq j \leq m)^\eta) \geq 1 - 2\eta \geq 1 - \epsilon$, for all $n \geq n(\eta)$. Since G_j are spheres of radius $2\eta < \delta$, it follows from Step 1 that $\{P_n : n \geq n(\eta)\}$ is tight. One may include the other finitely many P_n by using the separability argument to find a finite number of δ-spheres satisfying the italicized requirement of Step 1 by enlarging the family G_j, still with a finite union. ■

Proposition 18.8. On a complete separable metric space (S, ρ), the distance d_{BL} metrizes the weak topology as a complete metric.

Proof. From Theorem 18.7 and Lemma 2, it follows that if P_n is Cauchy in the distance d_{BL}, then it converges weakly to a probability P. However, it does not immediately follow that P_n converges to P in d_{BL}, i.e., $d_{BL}(P_n, P) \to 0$. A natural proof of this comes from the fact[7] that BL is a *uniformity class for weak convergence*; that is, whatever be P and a sequence P_n converging weakly to P, $\sup\{|\int_S f dP_n - \int_S f dP| : f \in BL\} \to 0$ as $n \to \infty$. To see this, fix $\epsilon > 0$.

[6] BCPT, Lemma 4, p. 244.

[7] See Bhattacharya and Ranga Rao (2010), p. 17.

Let P_n converge weakly to P and K a compact set such that $P_n(K) > 1 - \epsilon$ for all n, $P(K) > 1 - \epsilon$. The set BL restricted to K is a compact metric space under the supnorm $||\cdot||_K$ (by the Arzela–Ascoli Theorem).[8] Hence, there exist functions $f_1, \ldots, f_m$, in BL with support contained in K such that $||f - f_j||_K < \epsilon$ for all $f \in BL$ and for all f_j ($j = 1, \ldots, m$). Then, writing the restriction of f to K as f^K (i.e., $f^K = f$ on K, and zero outside), one has

$$\left|\int_S f^K dP_n - \int_S f^K dP\right| \leq \max\left\{\left|\int_S f_j dP_n - \int_S f_j dP\right| : j = 1, .., m\right\} + 2\epsilon \leq 3\epsilon,$$

for all $n \geq n_1$, where n_1 is such that $\int_S |f_j dP_n - \int_S f_j dP| < \epsilon$ for all $j = 1, \ldots, m$ for all $n \geq n_1$. Next,

$$\left|\int_S f^K dP_n - \int_S f dP_n\right| \leq 2||f||P_n(K^c) \leq 2\epsilon$$

for all $f \in BL$, and $|\int_S f^K dP - \int_S f dP| \leq 2\epsilon$. Finally, for all $n \geq n_1$,

$$\begin{aligned}\left|\int_S f dP_n - \int_S f dP\right| &\leq \left|\int_S f^K dP_n - \int_S f^K dP\right| \\ &\quad + \left|\int_S f^K dP_n - \int_S f dP_n\right| + \left|\int_S f^K dP - \int_S f dP\right| \\ &\leq 3\epsilon + 2\epsilon + 2\epsilon = 7\epsilon.\end{aligned}$$

Since these estimates hold for all f in BL,

$$\sup\left\{\left|\int_S f dP_n - \int_S f dP\right| : f \in BL\right\} \to 0,$$

as $n \to \infty$. ■

Exercises

1. (A One-Step Cut-off Map) Let $U_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, $0 \leq \theta < 2\pi$, denote the group of counterclockwise rotation matrices on the unit circle S in $\mathbb{R}^2$. Define $\mathbf{X}_n = \prod_{j=1}^n U_{\Theta_j}\mathbf{X}_0$, $n = 1, 2, \ldots$, where $\mathbf{X}_0 \in S$, and $\Theta_1, \Theta_2, \ldots$ are i.i.d. uniform on $[0, 2\pi)$ under addition modulo 2π.

[8] BCPT, pp. 244–245, or Folland (1984), p. 131.

(a) Show that $\mathbf{X}_1$ is uniformly distributed on S after one iteration (see Example 4 for another cut-off phenomena).

(b) Show that the uniform distribution on S is the unique invariant distribution. [*Hint*: Reformulate the problem in the complex plane using Euler's formula, and check that the distribution of the sum modulo 2π of two independent uniform random variables is uniform on $[0, 2\pi)$.]

2. Prove that (18.3) is a well-defined transition probability on $(S, \mathcal{S})$ if $\boldsymbol{\alpha}$ is a random map according to Definition 18.1.
3. For the function F_x^{-1} defined by (18.5), check that $\boldsymbol{\alpha}x := F_x^{-1}(U)$ satisfies (18.2) and has the distribution function $y \to F_x(y) := p(x; (-\infty, y])$.
4. Prove that the process $\{X_n(x) := \boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 x : n \geq 0\}$, where $\{\boldsymbol{\alpha}_n : n \geq 1\}$ are i.i.d. random maps on $(S, \mathcal{S})$, is a Markov process having the transition probability (18.3). Show that this remains true if the initial state x is replaced by a random variable X_0 independent of $\{\boldsymbol{\alpha}_n : n \geq 1\}$.
5. Show that the hypothesis $\mathbb{E}\epsilon_1 < 0$ in Theorem 18.3 implies that Spitzer's condition holds: $\sum_{n=1}^{\infty} \frac{1}{n} P(\epsilon_1 + \cdots + \epsilon_n > 0) < \infty$.
6. Suppose ϵ_i in Theorem 18.3 has the shifted exponential distribution with density $h(x) = \beta^{-1} \exp\{-\frac{x+c}{\beta}\}\mathbf{1}_{(c,\infty)}(x)$, $(c > \beta > 0)$. Show that the invariant distribution π has a point mass at 0 and a density on $(0, \infty)$ given by $\pi(0) = \beta/\theta, \pi(x) = \theta^{-1} \exp\{-\frac{x+c}{\theta}\}, x > 0$, where $\theta > \beta$ is the solution of the equation $1 - \frac{\beta}{\theta} = e^{-\frac{c}{\theta}}$.
7. Apply Exercise 18 to each of the following models:

   (a) The Lindley–Spitzer resource management model in the case when the distribution of the random input R_n has the exponential distribution with mean $\beta < c$.

   (b) The G/G/1 queuing model when the distributions of the inter-arrival time U and the service time V are both exponential with means β, θ, respectively, with $\beta < \theta$. Show that $\pi(\{0\}) = 1 - \beta/\theta, \pi(x) = \frac{\beta}{\theta}(\beta^{-1} - \theta^{-1}) \exp\{-(\beta^{-1} - \theta^{-1})x\}\mathbf{1}_{(0,\infty)}(x)$.

   (c) The insurance model with exponential with mean β claim size distribution, exponential with mean $1/\lambda$ inter-arrival times of claims, and the premium c per unit time satisfying $\theta = c/\lambda > \beta$. Show that the probability of ruin with initial asset u is $\psi(u) = \int_{(u,\infty)} \pi(x)dx = \frac{\beta}{\theta} \exp\{-(\beta^{-1} - \theta^{-1})u\}, u > 0$.

8. In reference to (18.12), show that if $\underline{\mathrm{Y}} < \infty$ a.s., and α_n is continuous a.s., then the distribution π, say, of $\underline{\mathrm{Y}}$ is an invariant probability for $\{X_n : n \geq 0\}$.
9. Adapt the Lindley–Spitzer model to a non-profit organization (NP0) with initial capital u and i.i.d. random donations (in dollars) each year. Assume a commitment by the NPO to spend an annual amount of c dollars/year in support of its cause. How do such models compare to the insurance models?
10. (a) Show that the results of convergence to a unique invariant probability of the linear models $AR(p)$ and $ARMA(p, q)$ of Chapter 17 follow from Theorem 18.6.

(b) Show that the convergence results extend to affine linear maps as well, i.e., if one adds a constant vector to the deterministic part Hx.

11. Consider a Markov process generated by iterations of an i.i.d. sequence α_n ($n \geq 1$) with common distribution Q (on a space Γ of functions on a metric space). Give an example where the process converges to a unique invariant probability irrespective of the initial distribution, with Q having a finite support and only one element a strict contraction, while the others have Lipschitz coefficients larger than 1.

Chapter 19
A Splitting Condition and Geometric Rates of Convergence to Equilibrium

This chapter builds on the representation of Markov processes in terms of i.i.d. iterated maps by developing the so-called “splitting techniques”that capture the recurrence structure of certain iterated maps in a novel way.

The questions of whether a Markov process has a unique invariant probability and, if so, how fast the process converges to this invariant probability starting from an arbitrary initial state are of basic interest. In this chapter, a “splitting” criterion is shown to imply the existence of a unique invariant probability and to yield geometric, or exponentially, fast rates of convergence to equilibrium in appropriate distances. The main result is first derived for Markov processes generated by the iteration of i.i.d. maps satisfying such a criterion, and then it is applied to different classes of Markov processes.

The results of this chapter may be used to derive important limit theorems such as the central limit theorem in a broad range of contexts.

Let $(S, \mathcal{S})$ be a measurable state space and $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$ an i.i.d. sequence of random maps on S. Thus each $\boldsymbol{\alpha}_n$ is a measurable map on a probability space $(\Omega, \mathcal{F}, P)$ into a measurable space of functions $(\Gamma, \mathcal{G})$, such that $(\gamma, x) \to \gamma x$ is measurable on $(\Gamma \times S, \mathcal{G} \otimes \mathcal{S})$ into $(S, \mathcal{S})$, as in (18.4). Let the distribution of $\boldsymbol{\alpha}_n$ be Q. Then $(\Gamma, \mathcal{G}, Q)$ is a probability space, and, for each n, the distribution of $(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n)$ is the product probability measure $Q \times Q \times \cdots \times Q = Q^n$ on $(\Gamma^n, \mathcal{G}^{\otimes n})$. In this setting, the transition operator T of the Markov process and its iterates T^n on the space $\mathbb{B}(S)$ of real-valued bounded measurable functions on S may be expressed as

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_19

$$Tf(x) = \mathbb{E}f(\boldsymbol{\alpha}_1 x) = \int_{\Gamma} f(\gamma x) Q(d\gamma), \tag{19.1}$$

$$T^n f(x) = \mathbb{E}f(\boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 x) = \int_{\Gamma^n} f(\boldsymbol{\gamma} x) Q^n(d\boldsymbol{\gamma}), \quad x \in S, f \in \mathbb{B}(S), n \geq 1.$$

As usual, the adjoint operator T^* and its iterates T^{*n} are defined, on the space $\mathcal{P}(S)$ of all probability measures on S, by letting $T^*\mu$ be the distribution of $\boldsymbol{\alpha}_1 X_0$ when X_0 has distribution μ (X_0 is independent of $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$). For $n \geq 1$, $T^{*n}\mu$ is the distribution of $\boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 X_0$. Hence, denoting by $p^{(n)}(x, dy)$ the n-step transition probability of the Markov process $X_n = \boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 X_0$, $n \geq 1$, one has

$$\begin{aligned}(T^{*n}\mu)(A) = P(\boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 X_0 \in A) &= \int_S Q^n(\{\boldsymbol{\gamma} \in \Gamma^n : \gamma_n \dots \gamma_1 x \in A\})\mu(dx) \\ &= \int_S p^{(n)}(x, A)\mu(dx), \quad \mu \in \mathcal{P}(S). \end{aligned} \tag{19.2}$$

Note that a probability measure π *is invariant* for the Markov process generated by the iterated maps *if* π *is a fixed point of* $T^* : T^*\pi = \pi$. That is, π is an *invariant initial distribution* if $X_1 \equiv \boldsymbol{\alpha}_1 X_0$ has distribution π when X_0 has distribution π.

The main result of this chapter, Theorem 19.1 below, estimates the distance between $p^{(n)}(x, dy)$ and an invariant probability π in suitable metrics on $\mathcal{P}(S)$ of the form

$$d(\mu, \nu) := \sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|, \quad (\mu, \nu \in \mathcal{P}(S)), \tag{19.3}$$

where $\mathcal{A} \subset \mathcal{S}$ must be selected such that under d, $\mathcal{P}(S)$ is a *complete metric space,* that is, if $d(\mu_n, \mu_m) \to 0$ as $n, m \to \infty$ for a sequence $\{\mu_n\}_{n\geq 1} \subset \mathcal{P}(S)$, then there exists $\nu \in \mathcal{P}(S)$ such that $d(\mu_n, \nu) \to 0$. Two typical selections are (i): $\mathcal{A} = \mathcal{S}$, in which case d is called the *total variation distance* and $(\mathcal{P}(S), d)$ is a complete metric space (Exercise 19), and (ii) in the case S is an interval of $\mathbb{R}$, with $\mathcal{S}$ the Borel σ-field, the *Kolmogorov metric* is defined by the supremum distance between distribution functions and is obtained here by taking $\mathcal{A}$ to be the class of all sets $S \cap (-\infty, x]$, $x \in \mathbb{R}$. Again one may check that this defines a complete metric on an interval $S \subset \mathbb{R}$ (Exercise 19).

For the statement below, recall that $\mu \circ \gamma^{-1}$ is the image (measure) of μ under the map γ (on S into S), i.e., $(\mu \circ \gamma^{-1})(B) = \mu(\gamma^{-1}(B))$, $B \in \mathcal{S}$. Write $\mathcal{R}(\gamma)$ for the *range of* γ, i.e.,

$$\mathcal{R}(\gamma) := \{\gamma(x) : x \in S\}, \tag{19.4}$$

and also write γ_{1n} for the composition

$$\gamma_{1n} := \gamma_n \gamma_{n-1} \dots \gamma_1 \quad \text{for } \boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_n) \in \Gamma^n \quad (n \geq 1). \tag{19.5}$$

The splitting theorem will feature three basic hypotheses: (1) the complete metric space hypothesis described above, (2) a contractive hypothesis, and (3) a splitting condition. For the two metrics noted above, the complete metric space condition is always implied. The second condition always holds for the total variation metric and, in the case of monotone, maps on an interval. That the second condition holds is also implied for the Kolmogorov metric. To gain some insight into the splitting condition (3), let us first consider the "most ergodic" Markov process conceivable, namely that of a sequence of i.i.d. random variables.

Example 1. Suppose that $X_1, X_2, \dots$ is an i.i.d. sequence with state space $(S, \mathcal{S})$ and defined on a probability space $(\Omega, \mathcal{F}, P)$ with a common distribution $P \circ X_n^{-1} = \mu$. Then one may represent this Markov process by iterations of the i.i.d. *constant* random maps defined for $\omega \in \Omega, n \geq 1$, by $\alpha_n(\omega) : S \to S$ via $\alpha_n(\omega)(x) = X_n(\omega)$, for all $x \in S$. So $\Gamma = \{\eta_z : \eta_z(x) = z \text{ for all } x, z \in S\}$, with $\mathcal{G} = \sigma\{\{\eta_z : z \in B\} : B \in \mathcal{S}\}$. Now observe that the range of the maps η_z is the singleton set $\{z\}$, so for any $A \in \mathcal{S}$, the range of η_z is either a subset of A or a subset of A^c. It is in such a sense that we say the maps "split" the class $\mathcal{A} := \mathcal{S}$.

Theorem 19.1. *Let* $\boldsymbol{\alpha}_n$ $(n \geq 1)$ *be i.i.d. random maps on* S *with common distribution* Q*. Suppose there exists* $\mathcal{A} \subset \mathcal{S}$ *with the following properties:*

1. $(\mathcal{P}(S), d)$ is a complete metric space, where d is defined by (19.3).
2. $d(\mu \circ \gamma^{-1}, \nu \circ \gamma^{-1}) \leq d(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}(S)$ and for Q-almost all $\gamma \in \Gamma$.
3. *(Splitting)* There exist $\delta > 0$, $N \in \mathbb{N}$, and for every $A \in \mathcal{A}$ a set $\Gamma_A \subset \Gamma^N$ belonging to $\mathcal{G}^{\otimes N}$, such that: (i) $Q^N(\Gamma_A) \geq \delta$ and (ii) either $\mathcal{R}(\gamma_{1N}) \subset A$ or $\mathcal{R}(\gamma_{1N}) \subset A^c$, for all $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_N) \in \Gamma_A$.

Then the Markov process generated by the iteration of $\{\alpha_n\}_{n \geq 1}$ has a unique invariant probability π, and the distribution $T^{*n}\mu$ of $X_n = \boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 X_0$, with initial distribution μ, satisfies

$$d(T^{*n}\mu, \pi) \leq (1 - \delta)^{[n/N]} d(\mu, \pi) \leq (1 - \delta)^{[n/N]}, \quad (n \geq 1, \mu \in \mathcal{P}(S)), \tag{19.6}$$

where $[x]$ denotes the integer part of $x \in \mathbb{R}$.

Proof. Let $A \in \mathcal{A}$. By assumption (3), if $\boldsymbol{\gamma} \in \Gamma_A$, then either $\mu(\gamma_{1N}^{-1}(A)) = 1$ for all μ or $\mu(\gamma_{1N}^{-1}(A)) = 0$ for all μ $(\mu \in \mathcal{P}(S))$. Therefore, for all $\mu, \nu \in \mathcal{P}(S)$,

$$\begin{aligned}\left|(T^{*N}\mu)(A) - (T^{*N}\nu)(A)\right| &= \left|\int_{\Gamma_A} (\mu \circ \gamma_{1N}^{-1})(A) - (\nu \circ \gamma_{1N}^{-1})(A) Q^N(d\boldsymbol{\gamma})\right| \\ &\quad + \left|\int_{\Gamma_A^c} (\mu \circ \gamma_{1N}^{-1})(A)) - (\nu \circ \gamma_{1N}^{-1})(A)) Q^N(d\boldsymbol{\gamma})\right| \\ &= \left|\int_{\Gamma_A^c} (\mu \circ \gamma_{1N}^{-1})(A) - (\nu \circ \gamma_{1N}^{-1})(A)) Q^N(d\boldsymbol{\gamma})\right| \\ &\leq (1 - \delta) d(\mu, \nu). \end{aligned} \tag{19.7}$$

We have used here the facts: (a) $Q^N(\Gamma_A^c) \leq 1-\delta$ and (b) $d(\mu \circ \gamma_{1N}^{-1}, \nu \circ \gamma_{1N}^{-1}) \leq d(\mu, \nu)$ (a.e. Q^N). Note that (b) follows from assumption (2) by induction on N. The inequality (19.7) implies that

$$d(T^{*N}\mu, T^{*N}\nu) \leq (1-\delta)d(\mu,\nu), \quad (\mu, \nu \in \mathcal{P}(S)), \tag{19.8}$$

that is, the map T^{*N} on the (complete) metric space $(\mathcal{P}(S), d)$ is a *strict contraction.* It now follows from the contraction mapping principle (see lemma below) that T^{*N} has a unique fixed point $\pi \in \mathcal{P}(S)$ and that, with $k = n-[n/N]N, n = k+[n/N]N$,

$$\begin{aligned} d(T^{*n}\mu, \pi) = d(T^{*n}\mu, T^{*n}\pi) &= d\left(T^{*k}\left(T^{*[\frac{n}{N}]N}\mu\right), T^{*k}\left(T^{*[\frac{n}{N}]N}\pi\right)\right) \\ &\leq d\left(T^{*[\frac{n}{N}]N}\mu, T^{*[\frac{n}{N}]N}\pi\right) \\ &\leq (1-\delta)^{[\frac{n}{N}]}d(\mu, \pi), \ (\mu \in \mathcal{P}(S)). \end{aligned} \tag{19.9}$$

The first inequality in (19.9) makes use of the assumption (2), implying $d(T^*\mu, T^*\nu) \leq d(\mu,\nu)$, since for all $A \in \mathcal{A}$,

$$\begin{aligned} |(T^*\mu)(A) - (T^*\nu)(A)| &= \left|\int_\Gamma (\mu \circ \gamma^{-1})(A)Q(d\gamma) - \int_\Gamma (\nu \circ \gamma^{-1})(A)Q(d\gamma)\right| \\ &\leq d(\mu, \nu). \end{aligned} \tag{19.10}$$

Finally, if one takes $\mu = T^*\pi$ in place of μ in (19.9), one gets $d(T^*\pi, \pi) = d(T^{*(n+1)}\pi, \pi) \to 0$ as $n \to \infty$. Hence, $T^*\pi = \pi$. That is, π is a fixed point of T^*. To prove that π is the unique fixed point of T^*, simply note that every fixed point of T^* is a fixed point of T^{*N}, and T^{*N} has a unique fixed point.

Before deriving some important corollaries of Theorem 19.1, let us restate the "splitting" condition 3 as the following:
3$'$ *There exist* $\delta > 0$ *and* N *such that, for each* $A \in \mathcal{A}$, *there is a set* $F_A \subset \Omega$ *with the properties*:

(a) $P(F_A) \geq \delta$.
(b) *For every* $\omega \in F_A$, *the range of* $\alpha_{1,N} \equiv \alpha_N \cdots \alpha_1$ *is contained either in* A *or in* A^c.

The statement (3) in the theorem is the canonical version of this with $(\Omega, \mathcal{F}, P) = (\Gamma^\infty, \mathcal{G}^{\otimes\infty}, Q^\infty)$.

Corollary 19.2 *(Doeblin Minorization Theorem*[1]*).* Let $p(x, dy)$ be a transition probability on $(S, \mathcal{S})$. Assume there exists a nonzero measure λ on $(S, \mathcal{S})$ such that

[1] See Bhattacharya and Majumdar (2007), Bhattacharya and Waymire (2002) for this and related refinements.

$$p^{(m)}(x,B) \geq \lambda(B) \quad \text{for all } B \in \mathcal{S}, x \in S, \tag{19.11}$$

for some $m \geq 1$. Then there exists a unique invariant probability π for $p(x,dy)$, and one has, for every $\mu \in \mathcal{P}(S)$,

$$\sup_{B \in \mathcal{S}} \left|(T^{*n}\mu)(B) - \pi(B)\right| \leq (1-\delta)^{[n/m]} \quad (n \geq 1), \tag{19.12}$$

where $\delta := \lambda(S) > 0$.

Proof. We will prove the assertion under the additional assumption that S is a Borel subset of a Polish space and $\mathcal{S}$ is its Borel σ-field. A different proof applicable without this extra assumption is sketched in Exercise 19. The result is simple to check for the case $\delta = 1$, in which case $\pi = \lambda$. Assume $0 < \delta < 1$. Then $p^{(m)}$ can be represented as the sum

$$p^{(m)}(x,B) = \delta\mu(B) + (1-\delta)q(x,B) \quad (x \in S, B \in \mathcal{S}), \tag{19.13}$$

where $\mu = \lambda/\lambda(S)$ is a probability measure, and q is the transition probability given by $q(x,B) = (p^{(m)}(x,B) - \lambda(B))/(1-\lambda(S))$. Now a Markov process with transition probability $p^{(m)}$, thought of as a one-step transition probability, may be generated by iterations of i.i.d. random maps $\{\boldsymbol{\alpha}_n\}_{n\geq1}$ as follows. Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined three independent i.i.d. sequences $\{\boldsymbol{\beta}_n\}_{n\geq1}$, $\{Z_n\}_{n\geq1}$, and $\{\varepsilon_n\}_{n\geq1}$ with the following properties: (i) $P(\boldsymbol{\beta}_n x \in B) = q(x,B)$ for all $x \in S, B \in \mathcal{S}$, (ii) Z_n $(n \geq 1)$ have the common distribution μ, and (3) ε_n $(n \geq 1)$ have the common Bernoulli distribution, $P(\varepsilon_n = 1) = \delta$, and $(\varepsilon_n = 0) = 1-\delta$. Define $\boldsymbol{\alpha}_n(\omega)$ to be the (constant) map $\boldsymbol{\alpha}_n(\omega)x = Z_n(\omega)$ $(x \in S)$ if $\varepsilon_n(\omega) = 1$ and $\boldsymbol{\alpha}_n(\omega) = \boldsymbol{\beta}_n(\omega)$ if $\varepsilon_n(\omega) = 0$. Then $P(\boldsymbol{\alpha}_n x \in B) = \delta P(Z_n \in B) + (1-\delta)P(\boldsymbol{\beta}_n x \in B) = \delta\mu(B) + (1-\delta)q(x,B) = p^{(m)}(x,B)$.

We claim that $\{\boldsymbol{\alpha}_n\}_{n\geq1}$ so defined satisfies the hypothesis of Theorem 19.1 with $\mathcal{A} = \mathcal{S}$, $N = 1$. To see this, first note that d is the *total variation distance* $d(\mu,\nu) = \|\mu - \nu\|_{TV} := \sup\{|\mu(B) - \nu(B)| : B \in \mathcal{S}\}$, under which $\mathcal{P}(S)$ is a complete metric space. Also, $d(\mu \circ \gamma^{-1}, \nu \circ \gamma^{-1}) = \sup\{|\mu(\gamma^{-1}(B)) - \nu(\gamma^{-1}(B))| : B \in \mathcal{S}\} \leq d(\mu,\nu)$. Finally, we will prove "splitting" in the form of the italicized statement $3'$ preceding the statement of the present theorem. For every $A \in \mathcal{S}$, choose $F_A = \{\omega : \varepsilon_1(\omega) = 1\}$. Then $P(F_A) = \delta$, and for every $\omega \in F_A$, the *range* of $\boldsymbol{\alpha}_1(\omega)$ *is the singleton* $\{Z_1(\omega)\}$, which is contained either in A or in A^c. If one writes the adjoint operator T_1^* corresponding to the (one-step) transition probability $p^{(m)}$ (i.e., $(T_1^*\mu)(B) = \int p^{(m)}(x,B)\mu(dx) = (T^{*m}\mu)(B)$), then, by Theorem 19.1 with $N = 1$, there exists a unique fixed point π of T_1^* and $d(T_1^{*k}\mu, \pi) \leq (1-\delta)^k$ for every $\mu \in \mathcal{P}(S)$. That is, $d(T^{*mk}, \pi) \leq (1-\delta)^k$. Also, $d(T^{*(mk+r)}, \pi) = d(T^{*mk}\mu', \pi) \leq (1-\delta)^k$, with $\mu' = T^{*r}\mu$.

A simple application of Corollary 19.2 is to possibly non-irreducible Markov chains on countable state spaces.

Corollary 19.3. Let $\mathbf{p} = ((p_{ij}))_{i,j\in S}$ be a transition probability on a countable state space S. Suppose there exist $m \geq 1$ and $j \in S$ such that $p_{ij}^{(m)} > \epsilon > 0$ for all i. Then the Markov process has a unique invariant probability π, and

$$\sup_i \sum_{j\in S} \left| p_{ij}^{(n)} - \pi_j \right| \leq 2(1-\delta)^{[n/m]} \quad (n \geq 1), \tag{19.14}$$

where $\delta := \sum_{j\in S} \inf\{p_{ij}^{(m)} : i \in S\}$, $\pi_j := \pi(\{j\})$. Thus if $\delta > 0$, one has exponential convergence to equilibrium in total variation distance.

Proof. Use Corollary 19.2 with $\lambda(\{j\}) = \delta_j := \inf\{p_{ij}^{(m)} : i \in S\}$, $j \in S$ to prove the existence of a unique invariant probability π and to get the estimate

$$\sup_{B\subset S} \left| p^{(n)}(i, B) - \pi(B) \right| \leq (1-\delta)^{[n/m]}, \tag{19.15}$$

since $\lambda(S) = \sum_{j\in S} \delta_j = \delta$. It is simple to check that $\sum_{j\in S} |p_{ij}^{(n)} - \pi_j|$ is twice the left side of (19.15) (Exercise 19).

Corollary 19.4 *(Convergence to Equilibrium of Finite State Irreducible Chains).* Let $\mathbf{p}$ be an irreducible transition probability matrix on a finite state space S. Then (a) $\mathbf{p}$ has a unique invariant probability $\boldsymbol{\pi} = \{\pi_j : j \in S\}$, and (b) if d is the period of the chain, there exist $\delta > 0$, and a positive integer ν such that

$$\sum_{j\in S} \left| \frac{1}{d} \sum_{u=0}^{d-1} p_{ij}^{(nd+u)} - \pi_j \right| \leq (1-\delta)^{[n/\nu]} \quad (n \geq 1). \tag{19.16}$$

In particular,

$$\sup_{i\in S} \sum_{j\in S} \left| \frac{1}{N} \sum_{m=1}^{N} p_{ij}^{(m)} - \pi_j \right| \leq \frac{1}{[N/d]} \sum_{n=0}^{[N/d]} (1-\delta)^{[n/\nu]} = O\left(\frac{1}{N}\right), \quad (N \geq 1). \tag{19.17}$$

Proof. If $\mathbf{p}$ is aperiodic, then, by the lemma below, there exists an integer $\nu \geq 1$ such that $p_{ij}^{(\nu)} > 0$ for all i, j, and Corollary 19.3 applies. Assume now that $\mathbf{p}$ is periodic with period $d > 1$ and with the cyclical sets C_r, $0 \leq r \leq d-1$ (see Proposition 10.3(b)). Then, for each r, $\mathbf{p}^d$ is an aperiodic and irreducible transition probability matrix on C_r. Let ν_r be the smallest integer such that $p_{ij}^{(\nu_r d)} > 0$ for all $i, j \in C_r$. Then, by Corollary 19.3, $\mathbf{p}_d$ has a unique invariant probability $\boldsymbol{\pi}_r = \{\pi_{r,j} : j \in C_r\}$ on C_r, and one has

$$\sup_{i\in C_r} \sum_{j\in C_r} \left| p_{ij}^{(nd)} - \pi_{r,j} \right| \leq (1-\delta_r)^{[\frac{n}{\nu_r}]} \quad (n \geq 1), \tag{19.18}$$

where $\delta_r := \min\{p_{ij}^{(\nu_r d)} : i, j \in C_r\} > 0$. It follows that if $i \in C_s$, $j \in C_r$ with $p_{ij}^{(nd+u)} = 0$ if $u \neq r - s (\mathrm{mod}\, d)$,

$$\begin{aligned}\sum_{j\in C_r}\left|p_{ij}^{(nd+u)} - d\pi_j\right| &= \sum_{j\in C_r}\left|\sum_{k\in C_r} p_{ik}^{(u)}(p_{kj}^{(nd)} - d\pi_j)\right| \\ &\leq \sum_{k\in C_r} p_{ik}^{(u)} \sum_{j\in C_r}\left|p_{kj}^{(nd)} - d\pi_j\right| \\ &\leq (1-\delta_r)^{[n/\nu_r]} \quad \text{if } u = r - s (\mathrm{mod}\, d).\end{aligned}$$

Let $\delta = \min\{\delta_r : 0 \leq r \leq d-1\}$.

Define the probability measure $\boldsymbol{\pi} = \{\pi_j : j \in S\}$ as

$$\boldsymbol{\pi} = (1/d)\sum_{r=0}^{d-1}\boldsymbol{\pi}_r, \tag{19.19}$$

where $\boldsymbol{\pi}_r$ is extended to S by setting $\boldsymbol{\pi}_r(S\backslash C_r) = 0$.

Summing (19.19) over $u = 0, 1, \ldots, d-1$, one then obtains

$$\frac{1}{d}\sum_{u=0}^{d-1} p_{ij}^{(nd+u)} = \frac{1}{d}\, p_{ij}^{(nd+r)} \quad (r := j - i\ (\mathrm{mod}\, d)), \tag{19.20}$$

so that for all i, j,

$$\begin{aligned}\sum_{j\in S}\left|\frac{1}{d}\sum_{u=0}^{d-1} p_{ij}^{(nd+u)} - \pi_j\right| &\leq \sum_{r=0}^{d-1}\sum_{j\in C_r}\left|\frac{1}{d}\, p_{ij}^{(nd+r)} - \frac{1}{d}\pi_{r,j}\right| \\ &\leq \sum_{r=0}^{d-1}\frac{1}{d}(1-\delta)^{[n/\nu_r]} \leq (1-\delta)^{[n/\nu]},\end{aligned} \tag{19.21}$$

where $\nu = \max\{\nu_r : 0 \leq r \leq d-1\}$. This proves (19.18). The inequality (19.17) follows by breaking up the sum (over m) into consecutive blocks of d summands and applying (19.16) to each block.

The invariance of $\boldsymbol{\pi}$ and its uniqueness follow from (19.17), using Corollary 8.5.

Part (b) of the following lemma was used in the above proof.

Lemma 1. Let $\mathbf{p}$ be an irreducible aperiodic transition probability matrix on a countable state space S.

(a) Then, for each pair (i, j), there exists an integer $\nu(i, j)$ such that $p_{ij}^{(n)} > 0$ for all $n \geq \nu(i, j)$.

(b) If S is finite, there exists ν_0 such that $p_{ij}^{(n)} > 0$ for all i, j, if $n \geq \nu_0$.

Proof.

(a) Let $B_{ij} = \{\nu \geq 1 : p_{ij}^{(\nu)} > 0\}$. For each j, B_{jj} is closed under addition, since $p_{jj}^{(\nu_1+\nu_2)} \geq p_{jj}^{(\nu_1)} p_{jj}^{(\nu_2)}$. By hypothesis, the greatest common divisor (g.c.d.) of B_{jj} is 1. We now argue that, if $B = B_{jj}$ is a set of positive integers closed under addition, then the smallest subgroup G of $\mathbb{Z}$ (which is a *group* under addition) is $\mathbb{Z}$. Note that G equals $\{u - v : u, v \in B\}$. If B does not equal $\mathbb{Z}$, then $1 \notin G$, so that $G = \{rn : n \in \mathbb{Z}\}$ for some $r > 1$. But, since $B \subset G$, this would imply that the g.c.d. of $B \geq r$, a contradiction.

We have shown that $1 \in G$, i.e., there exists an integer $b \geq 1$ such that $b + 1, b$ both belong to B_{jj}. Let $\nu_j = (2b + 1)^2$. If $n \geq \nu_j$, one may write $n = q(2b+1)+r$, where r and q are integers, $0 \leq r < 2b+1$, and $q \geq 2b+1$. Then $n = q\{b + b + 1\} + r\{b + 1 - b\} = (q - r)b + (q + r)(b + 1) \in B$. Thus $b_{jj}^{(n)} > 0$ for all $n \geq \nu_j$. Find $k \equiv k_{ij}$ such that $p_{ij}^{(k)} > 0$ and then $p_{ij}^{(n+k)} \geq p_{ij}^{(k)} p_{jj}^{(n)} > 0$ for all $n \geq \nu_j$. Now take $\nu(i, j) = k_{ij} + \nu_j$.

(b) If S is finite, let $\nu_0 = \max\{\nu_j + k_{ij} : i, j \in S\}$. Then, for all i, j, one has $p_{ij}^{(n)} > 0$ provided $\nu \geq \nu_0$.

Example 2 *(A Non-irreducible Markov Chain with Unique Invariant Probability).* Consider the following transition probability matrix $\mathbf{p}$ on the state space $S = \{1, 2, 3\}$ in the context and notation of Corollary 19.3.

$$\mathbf{p} = \begin{bmatrix} q & p & 0 \\ 0 & q & p \\ 0 & p & q \end{bmatrix} \qquad (0 < p < 1, q = 1 - p). \tag{19.22}$$

Then, with $m = 1$, $j = 2$, $\delta_1 = \delta_3 = 0$ and $\delta_2 = \delta = \min\{p, q\}$. Also, it is easy to solve for the invariant probability π to get $\pi_1 = 0, \pi_2 = \pi_3 = \frac{1}{2}$. Then (19.14) yields

$$\sum_{j=1}^{3} \left| p_{ij}^{(n)} - \pi_j \right| \leq 2(1 - \min\{p, q\})^n \quad (n \geq 1). \tag{19.23}$$

Note that the state 1 is inessential or transient. That is, if the initial state is 2 or 3, then the Markov process will remain in $\{2, 3\}$ and never visit 1; on the other hand, if the initial state is 1, then, with probability one, the process will enter the set $\{2, 3\}$ and never return to 1 again.

The next example is that of a finite *irreducible aperiodic* Markov chain, namely, the case $p_{ij}^{(m)} > 0$ for all $i, j \in S$, for some $m \geq 1$.

Example 3 *(A Birth–Death Chain with Two Reflecting Boundaries).* Let $S = \{1, 2, \ldots, L\}$, and $\mathbf{p} = ((p_{ij})) \equiv ((p_{i,j}))$ is given by

$$p_{i,i-1} = \delta_i > 0,\ p_{i,i} = \gamma_i > 0,\ p_{i,i+1} = \beta_i > 0,\ \delta_i + \gamma_i + \beta_i = 1\ (2 \le i \le L-1),$$
$$p_{1,1} = \gamma_1 > 0,\ p_{12} = \beta_1 = 1-\gamma_1 > 0,\ p_{1,L-1} = \delta_L > 0,\ p_{L,L} = \gamma_L = 1-\delta_L > 0.$$

It is simple to check that $p_{ij}^{(m)} > 0$ for all i, j if $m = L - 1$, so that (19.14) holds.

The study of the existence of unique invariant probabilities and stability is often relatively simpler for those cases in which the transition probabilities $p(x, dy)$ have a density $p(x, y)$, say, with respect to some reference measure $\mu(dy)$ on the state space. In the case of Markov chains on a countable state space, this measure may be taken to be the counting measure, assigning mass 1 to each singleton in the state space.

Example 4 *(Maps on the Real Number Line).* For a class of simple examples with an uncountable state space, let $S = \mathbb{R}$ and f a bounded measurable function on $\mathbb{R}$, $a \le f(x) \le b$. Let $\{\epsilon_n\}$ be an i.i.d. sequence of real-valued random variables whose common distribution has a strictly positive continuous density φ with respect to Lebesgue measure on $\mathbb{R}$. Consider the Markov process

$$X_{n+1} := f(X_n) + \epsilon_{n+1} \qquad (n \ge 0), \tag{19.24}$$

with X_0 arbitrary (independent of $\{\epsilon_n\}$). Then the transition probability $p(x, dy)$ has the density

$$p(x, y) := \varphi(y - f(x)). \tag{19.25}$$

Note that

$$\varphi(y - f(x)) \ge \psi(y) \text{ for all } x \in \mathbb{R}, \tag{19.26}$$

where

$$\psi(y) := \min\{\varphi(y - z) : a \le z \le b\} > 0.$$

Now Corollary 19.2 applies with λ as the measure with density ψ and $m = 1$.

In contrast to this class, when $p(x, dy)$ are mutually singular for all or most x, one may have infinitely many mutually singular invariant probabilities. This may happen, e.g., if in (19.24), the ε_n are discrete. The example below and Exercises 2–19 demonstrate the dramatic difference in behavior of X_n, such as governed by (19.24), that may arise with the same f but with ε_n having a positive density in one case and being discrete in the other case. Assumptions such as monotonicity, or contraction, of f are therefore invoked to guarantee stability in distribution, i.e., the convergence in distribution to a unique invariant probability, irrespective of X_0.

Example 5 *(An Erdös Problem).* Let $\{\epsilon_n : n = 1, 2, \ldots\}$ be an i.i.d. sequence of symmetric Bernoulli ±1-valued random variables, and let $0 < b < 1$. According to Proposition 17.1, the distribution π_b of the random series $\sum_{n=0}^{\infty} \pm b^n := \sum_{n=0}^{\infty} \epsilon_{n+1} b^n$ is the unique invariant probability for the Markov chain $X_{n+1} = bX_n + \epsilon_{n+1}, n = 0, 1, 2 \ldots$ In the case $b = 1/2$, define $T : [0, 1) \to [-2, 2)$ by $T(x) := \sum_{n=0}^{\infty}(2\theta_{n+1} - 1)2^{-n}$, where $x := \sum_{n=1}^{\infty} \theta_n 2^{-n}$ is the unique binary expansion of $x \in [0, 1)$ having $\theta_n = 0$ for infinitely many n. Then $\pi_{\frac{1}{2}} = \lambda \circ T^{-1}$, where λ is Lebesgue measure on $[0, 1)$. But $Tx = 4\sum_{n=1}^{\infty} \theta_n 2^{-n} - \sum_{n=0}^{\infty} 2^{-n} = 4x - 2, x \in [0, 1)$. Thus $\pi_{\frac{1}{2}}$ is the normalized Lebesgue measure on $[-2, 2)$. Next consider the case $b = \frac{1}{3}$. Let K denote the standard Cantor subset of $[0, 1]$ obtained by successively removing middle one-third intervals, or equivalently, $K := \{x = \sum_{n=1}^{\infty} \alpha_n 3^{-n} : \alpha_n \in \{0, 2\}$ for all $n\}$. Define $T : K \to [-3/2, 3/2]$ by $T(x) := \sum_{n=0}^{\infty}(\alpha_{n+1} - 1)3^{-n}, x \in K$. Then $\pi_{\frac{1}{3}} = \mu \circ T^{-1}$, where μ is the continuous singular Cantor distribution supported on $K \subset [0, 1]$. But $T(x) = 3x - 3/2, x \in K$. Thus $\pi_{\frac{1}{3}}$ is supported on a Cantor subset.[2] A famous conjecture by Paul Erdös that π_b is absolutely continuous remained unresolved for nearly sixty years until Solomyak (1995) proved it for $1/2 \leq b < 1$.

Example 6 *(Markov Chain Monte Carlo/MCMC).* In the *Metropolis–Hastings algorithm*, one considers estimating an unknown strictly positive probability density f on a state space $(S, \mathcal{S}, \mu)$, with respect to a sigma finite measure μ, specified up to a normalizing constant that is not computable, but the ratio $f(x)/f(y)$ is computable. For this, the algorithm constructs a Markov chain on S whose unique invariant probability, and asymptotic distribution, is f. For this purpose, consider a distribution given by a positive Markov kernel (*proposal distribution*), i.e., strictly positive transition probability density $q(x, y) \equiv q(y|x)$ from which it is easy to select an observation y, given x. Beginning with an initial state $X_0 = x_0$, one draws an observation Y_0 with distribution $q(\cdot|x)\mu(dy)$. One lets $X_1 = Y_0$ (as the state of the Markov chain at time 1) with probability given by the acceptance ratio $a(X_0, Y_0)$, where

$$a(x, y) = \min\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\}, \tag{19.27}$$

and let $X_1 = X_0$ with probability $1 - a(X_0, Y_0)$. The transition probability of this Markov chain is given by

$$p(x, dy) = q(y|x)a(x, y)\mu(dy) + [\int_S (1 - a(x, z))q(z|x)\mu(dz)]\delta_x(dy). \tag{19.28}$$

Letting

[2] For further results on the problem of delineating the structure of π_b beginning with early results of Erdos (1937), see Peres et al. (1999).

$$R(x) := \int_S (1 - a(x, y)) q(y|x) \mu(dy), \tag{19.29}$$

one has

$$p(x, B) = \int_B p(x, y)\mu(dy) + R(x)\delta_x(B), \quad 1 = p(x, S) = \int_S p(x, y)\mu(dy) + R(x). \tag{19.30}$$

That is, the chain has a density component $p(x, y) \equiv p(y|x) = q(y|x)a(x, y)$ on $\{y \in S : y \neq x\}$ and a point mass at $\{x\}$ with probability $R(x)$. Setting $p(x, x) = 0$ for all x, the transition density component $p(x, y)$ satisfies the detailed balance condition

$$f(y)p(y, x) = f(x)p(x, y) \quad \text{for all } x, y \in S. \tag{19.31}$$

To see this, note that if $f(y)q(x|y) > f(x)q(y|x)$, then $a(x, y) = 1$, and the right side in (19.31) equals $f(x)q(y|x)$. But in this case, $a(y, x) = f(x)q(y|x)/[f(y)q(x|y)]$, and the left side of (19.31) is given by the ratio $f(y)q(x|y)f(x)q(y|x)/[f(y)q(x|y)] = f(x)p(x, y)$. The case $f(y)q(x|y) < f(x)q(y|x)$ is similar.

Theorem 19.5. The pdf f is an invariant probability density with respect to μ for the Markov chain:

$$\int_S p(x, B) f(x)\mu(dx) = \int_B f(y)\mu(dy) \quad \text{for all } B \in \mathcal{S}. \tag{19.32}$$

Proof. First use (19.30) and then (19.31) to obtain that

$$\begin{aligned}
\int_S p(x, B) f(x)\mu(dx) &= \int_S \left[\int_B (p(x, y)\mu(dy)\right] f(x)\mu(dx) + \int_S R(x) f(x)\delta_x(B)\mu(dx) \\
&= \int_S \left[\int_B p(y, x) f(y)\mu(dy)\right] \mu(dx) + \int_B R(x) f(x)\mu(dx) \\
&= \int_B \left[\int_S p(y, x)\mu(dx)\right] f(y)\mu(dy) + \int_B R(y) f(y)\mu(dy) \\
&= \int_B \left[\int_S p(y, x)\mu(dx) + R(y)\right] f(y)\mu(dy) \\
&= \int_B f(y)\mu(dy).
\end{aligned}$$

For the last equality, we have used the last relation in (19.30).

Proposition 19.6. In the Metropolis–Hastings algorithm, assume $f(y) > 0$ for all y and $q(x, y) = q(y|x) > 0$ (μ-a.e. in y, for every x). Then, if h is a bounded measurable function, $\frac{1}{n}\sum_{m=0}^{n-1} h(X_m)$ converges almost surely to

$\int_S h(y)\pi(dy)$ for μ-almost every initial state $X_0 = x$. Here $\pi(dy) = f(y)\mu(dy)$ is the invariant measure of the Markov chain $\{X_n : n = 0, 1, \dots\}$ constructed above. (That is, π has the density f with respect to μ).

Proof. Consider the stationary Markov process starting with initial distribution π. We denote by P_π the distribution of the Markov process with initial distribution π and by P_x its distribution starting at x ($X_0 = x$). According to Birkhoff's ergodic theorem, we need to show that the shift-invariant σ-field $\mathcal{I}$ is trivial. First note that, since f is strictly positive (μ-a.e.), μ and π are absolutely continuous with respect to each other (i.e., $\pi(B) = 0$ if and only if $\mu(B) = 0$). Suppose now, if possible, that $\mathcal{I}$ is not trivial so that there exists $G \in \mathcal{I}$ such that $0 < P_\pi(G) < 1$. From the proof of Theorem 16.4, it follows that there exists a set $B \in \mathcal{S}$ such that $[X_0 \in B] = G$ almost surely (with respect to P_π) and that the probability π_B defined by $\pi_B(A) = \pi(A \cap B)/\pi(B)$, $A \in \mathcal{S}$, is an invariant probability for the Markov process. In particular $\pi_B(B^c) = 0$. However, since $\pi(B^c) > 0$, $\mu(B^c) > 0$. This leads to the contradiction

$$0 = \pi_B(B^c) = \int_S p(x, B^c)\pi_B(dx) \geq \int_S \left[\int_{B^c} p(x, y)\mu(dy)\right]\pi_B(dx) > 0. \tag{19.33}$$

Here $p(x, y)$ is the strictly positive density component of the transition probability $p(x, dy)$, see (19.30). The second equality in (19.33) follows from the invariance of π_B, while the last (strict) inequality follows from the positivity of the integral within brackets [ ], for every x. Thus $\mathcal{I}$ is trivial, and the stationary Markov process with distribution P_π is ergodic. Therefore, by Birkhoff's ergodic theorem, $\frac{1}{n}\sum_{j=1}^{n} h(X_j)$ converges π-almost surely to $\int_S h(y)\pi(dy)$. Conditioning on X_0, it then follows that for π-almost all x, $\frac{1}{n}\sum_{j=1}^{n} h(X_j)$ converges to $\int_S h(y)\pi(dy)$ almost surely, with respect to P_x. ■

Remark 19.1. A simple illustration is obtained by taking $q(x, y) = g(y)$, $y \in S$, independent of x, i.e., *independent sampling*, and then $\gamma = \pi/g$ is bounded. Note that in this case, the acceptance ratio is $a(x, y) = \min\{\frac{f(y)}{f(x)}\frac{g(x)}{g(y)} \wedge 1\}$ (Exercise 19).

We now turn to the so-called *Gibbs sampler*[3] as an alternative to the Metropolis–Hastings algorithm. The latter is difficult to apply directly to state spaces S of dimension $d > 1$ partly because of the problem with directly generating random vectors and partly because of the slow rate convergence to stationarity of multidimensional chains. The Gibbs sampler alleviates these problems by using several one-dimensional problems to manage a multidimensional problem. To illustrate this, consider the two-dimensional problem with (X, Y) having density $f(x, y)$, on $S \subset \mathbb{R}^2$, with respect to the product $\mu \times \nu$ of two sigma-finite measures on $(S, \mathcal{S})$. Let $f_{Y|X}(y|x)$ denote the conditional density of Y (at y) given

[3] See Gelman et al. (1995) for an early exposition. Also see Chib and Greenberg (1995).

$[X = x]$. Similarly, define $f_{X|Y}(x|y)$. Assume that it is possible to sample from the conditional distributions.

Algorithm:

(i) Generate X_1 with density $f_{X|Y}(\cdot|y_0)$.
(ii) Generate Y_1 with density $f_{Y|X}(\cdot|X_1)$.
(iii) Given (X_n, Y_n), generate X_{n+1} with density $f_{X|Y}(\cdot|Y_n)$, and generate Y_{n+1} with density $f_{Y|X}(\cdot|X_{n+1}), n = 1, 2, \ldots$

Theorem 19.7. Assume S is an open rectangle in $\mathbb{R}^2$ and $f(x, y) > 0$ for all $(x, y) \in S$. Then (a) the Markov chain $\{(X_n, Y_n) : n \geq 0\}$ has the invariant density $f(x, y)$, and (b) for every bounded, measurable real-valued function h on S,

$$\frac{1}{n}\sum_{m=0}^{n-1} h(X_m, Y_m) \to \int_S h(x, y)\mu(dx)\nu(dy), \tag{19.34}$$

almost surely for all initial states (x_0, y_0) outside a $\mu \times \nu$-null set as $n \to \infty$.

Proof. (a) The transition probability density of the Markov chain $(X_n, Y_n), n \geq 0$, is given by

$$q(x_1, y_1|x_1, y_0) = f_{X|Y}(x_1|y_0) f_{Y|X}(y_0|x_1) = \frac{f(x_1, y_0)}{f_Y(y_0)}\frac{f(x_1, y_1)}{f_X(x_1)}, \tag{19.35}$$

so that

$$\begin{aligned}
&\int_S q(x_1, y_1|x_1, y_0) f(x_0, y_0)\mu(dx_0)\nu(dy_0) \\
&\quad = \int_S \left(\int_S f(x_0, y_0)\mu(dx_0)\right) \frac{f(x_1, y_0) f(x_1, y_1)}{f_Y(y_0) f_X(x_1)}\nu(dy_0) \\
&\quad = \int_S \frac{f_Y(y_0) f(x_1, y_0) f(x_1, y_1)}{f_Y(y_0) f_X(x_1)}\nu(dy_0) \\
&\quad = f(x_1, y_1).
\end{aligned} \tag{19.36}$$

This establishes part (a). For (b), the proof of convergence is similar to that of Proposition 19.6 and left as Exercise 19.

The next application of Theorem 19.1 is to Markov processes generated by iterations of i.i.d. monotone maps already introduced in Chapter 18. Indeed, the basic notion of "splitting" in this chapter may be thought of as a generalization of that appearing on iterations of i.i.d. continuous monotone maps on a compact interval,[4] originally due to Dubins and Freedman (1966). We will say γ is *monotone increasing* if $\gamma x \leq \gamma y$ whenever $x \leq y$. Similarly, γ is monotone *decreasing* if $\gamma x \geq \gamma y$ whenever $x \leq y$.

Corollary 19.8 *(Convergence of Iterations of I.I.D. Monotone Maps).* Let $S \subset \mathbb{R}$ be an interval, finite or infinite, and $\mathcal{S}$ its Borel σ-field. Suppose $\{\boldsymbol{\alpha}_n\}_{n\geq 1}$ is an i.i.d. sequence of monotone random maps on S with the following property:

(Splitting) There exist $m \geq 1$, $x_0 \in S$, and $\delta > 0$, such that

$$P(\boldsymbol{\alpha}_m \dots \boldsymbol{\alpha}_1 x \geq x_0 \text{ for all } x, \quad \text{or} \quad \boldsymbol{\alpha}_m \dots \boldsymbol{\alpha}_1 x \leq x_0 \text{ for all } x) \geq \delta. \tag{19.37}$$

Then the Markov process $\{X_n(x) := \boldsymbol{\alpha}_n \dots \boldsymbol{\alpha}_1 x\}_{n\geq 0}$, $x \in S$, has a unique invariant probability π and

$$\sup_{x\in S} |P(X_n(x) \leq z) - \pi((-\infty, z] \cap S)| \leq (1-\delta)^{[n/m]} \quad (n \geq 1). \tag{19.38}$$

Proof. One applies Theorem 19.1 with $\mathcal{A} = \{(-\infty, z] \cap S : z \in \mathbb{R}\}$, $N = m$. In this case, $d(\mu, \nu)$ is the supremum (or, uniform) distance between the distribution functions of μ and ν and $(\mathcal{P}(S), d)$ is a complete metric space. Also, if γ is a monotone increasing map on S, then $d(\mu \circ \gamma^{-1}, \nu \circ \gamma^{-1}) = \sup\{\mu(\gamma^{-1}((-\infty, z] \cap S)) - \nu(\gamma^{-1}((-\infty, z] \cap S))| : z \in \mathbb{R}\}$. If γ is monotone increasing, then $\gamma^{-1}((-\infty, z] \cap S)$ is an interval of the type $(-\infty, z'] \cap S$ or $(-\infty, z') \cap S$ (the latter may arise if γ is not continuous). Clearly, $|\mu((-\infty, z'] \cap S) - \nu((-\infty, z'] \cap S)| \leq d(\mu, \nu)$. Since the distribution function is right continuous, and d is the uniform distance, it follows that $|\mu((-\infty, z') \cap S) - \nu((-\infty, z') \cap S)| \leq d(\mu, \nu)$. If γ is monotone decreasing, then the same argument works on taking the μ and ν-measures of the complement of $\gamma^{-1}((-\infty, z] \cap S)$.

It remains to check the "splitting" condition 3. We will check the version of 3′ appearing in italics preceding the statement of Corollary 19.2. For each $A = (-\infty, z] \cap S$, let F_A be the set appearing in parentheses in (19.37). Fix $\omega \in F_A$, and write $\gamma := \boldsymbol{\alpha}_m(\omega) \dots \boldsymbol{\alpha}_1(\omega)$. If $\gamma x \leq x_0$ for all x, then $\gamma x \in A \equiv (-\infty, z] \cap S$ for all x if $z \geq x_0$, and $\gamma x \in A^c \equiv (z, \infty) \cap S$ for all x if $z < x_0$. Similar argument applies if $\gamma x \geq x_0$ for all x.

An important generalization of Corollary 19.8 to multidimension is provided by Theorem 19.9 below. For its statement, define two metrics on the space $\mathcal{P}(S)$ of all probability measures on the Borel sigma-field $\mathcal{S}$ of a Borel measurable subset S of $\mathbb{R}^k$ (S is nonempty and not a singleton). Consider $\mathcal{A} \subset \mathcal{S}$ comprising all sets A of the form

$$A = \{y \in S : \varphi(y) \leq x\}, \tag{19.39}$$

[4] The Corollary 19.8 also extends to closed sets $S \subset \mathbb{R}^k$, with a coordinatewise partial order. For this, see Bhattacharya and Lee (1988). Also see Bhattacharya and Majumdar (2010a), and Chakroborty and Rao (1998).

where φ a non-decreasing continuous function on S into $\mathbb{R}^k$, $x \in \mathbb{R}^k$, and define

$$d_{\mathcal{A}}(\mu, \nu) = \{\sup |\mu(A) - \nu(A)| : A \in \mathcal{A}\},\ \mu, \nu \in \mathcal{P}(S). \tag{19.40}$$

Also define for $a > 0$, $\mu, \nu \in \mathcal{P}(S)$,

$$d_a(\mu, \nu) = \sup\{|\int_S hd\mu - \int_S hd\nu| : h \in \mathcal{H}_a\}, \tag{19.41}$$

where $\mathcal{H}_a$ is the class of all real-valued non-decreasing Borel measurable functions h on $S, 0 \leq h \leq a$. The (partial) ordering $\leq$ on $\mathbb{R}^k$ used here is the coordinatewise order: $x \leq y$ if $x(i) \leq y(i)$ for all $1 \leq i \leq k, (x = (x(1), \ldots, x(k)), y = (y(1), \ldots, y(k)))$. Note that $d_a(\mu, \nu) = ad_1(\mu, \nu)$. Let Γ denote the class of all non-decreasing Borel measurable functions on S into S, and $\mathcal{G}$ a sigma-field on Γ such that $(\gamma, x) \to \gamma(x)$ is measurable on $(\Gamma \times S, \mathcal{G} \otimes \mathcal{S})$ into $(S, \mathcal{S})$.

Theorem 19.9 *(Convergence to Equilibrium of Monotone Markov Processes in Multidimension).* Assume S is a closed subset of $\mathbb{R}^k$. Let Q be a probability measure on $(\Gamma, \mathcal{G})$ with the following splitting property: (H): There exist a positive integer N, $\delta_i > 0$, and measurable subsets F_i of $(\Gamma^N, \mathcal{G}^{\otimes N})(i = 1, 2)$, and $x_0 \in S$ such that:

1. $Q^N(F_1) = \delta_1$, where $F_1 := \{\gamma \in \Gamma^N : \tilde{\gamma}x \leq x_0 \text{ for all } x \in S\}$
2. $Q^N(F_2) = \delta_2$, where $F_2 := \{\gamma \in \Gamma^N : \tilde{\gamma}x \geq x_0 \text{ for all } x \in S\}$

where $\tilde{\gamma}$ is the composition

$$\tilde{\gamma} = \gamma_N \gamma_{N-1} \cdots \gamma_1 \text{ for } \gamma = (\gamma_1, \ldots, \gamma_N) \in \Gamma^N. \tag{19.42}$$

Then, letting $\delta = \min\{\delta_1, \delta_2\}$, one has

$$d_1(T^{*n}\mu, T^{*n}\nu) \leq (1 - \delta)^{[n/N]} \quad \text{for all } \mu, \nu \in \mathcal{P}(S), n \geq 1. \tag{19.43}$$

Also, there exists a unique invariant probability π of the Markov process generated by i.i.d. iterations with common distribution Q, and the following holds:

$$d_1(T^{*n}\mu, \pi) \leq (1 - \delta)^{[n/N]} \quad \text{for all } \mu \in \mathcal{P}(S), n \geq 1. \tag{19.44}$$

Proof. For $h \in \mathcal{H}_1$, one has

$$\begin{aligned}\int_S hd(T^{*N}\mu) - \int_S hd(T^{*N}\nu) = &\int_S \left\{\int_{\Gamma^N} h(\tilde{\gamma}x) Q^N(d\gamma)\right\} \mu(dx) \\ &- \int_S \left\{\int_{\Gamma^N} h(\tilde{\gamma}x) Q^N(d\gamma)\right\} \nu(dx)\end{aligned}$$

$$= \sum_{1\le i\le 4} \left\{ \int_S h_i(x)\mu(dx) - \int_S h_i(x)\nu(dx) \right\}. \tag{19.45}$$

Here

$$\begin{aligned} h_1(x) &= \int_{F_1\setminus(F_1\cap F_2)} h(\tilde{\gamma}x)Q^N(d\gamma), \\ h_2(x) &= \int_{F2\setminus(F1\cap F2)} h(\tilde{\gamma}x)Q^N(d\gamma), \\ h_3(x) &= \int_{(F_1\cup F2)^c} h(\tilde{\gamma}x)Q^N(d\gamma), \\ h_4(x) &= \int_{F_1\cap F_2} h(\tilde{\gamma}x)Q^N(d\gamma). \end{aligned} \tag{19.46}$$

Since $\tilde{\gamma}x = \tilde{\gamma}x_0$ on $F_1 \cap F_2$, the difference between the two integrals in (19.45) for $i = 4$ vanishes. Now h_1 and h_3 are non-decreasing and $0 \le h_1(x) \le h(x_0)(Q^N(F_1) - Q^N(F_1 \cap F_2)) := a_1, h_1(x) \in \mathcal{H}_{a_1}$. Hence,

$$|\int_S h_1(x)\mu(dx) - \int_S h_1(x)\nu(dx)| \le a_1 d_1(\mu,\nu). \tag{19.47}$$

Also, $0 \le h_3(x) \le 1 - Q^N(F_1 \cup F_2) := a_3, h_3(x) \in \mathcal{H}_{a_3}$, so that

$$|\int_S h_3(x)\mu(dx) - \int_S h_3(x)\nu(dx)| \le a_3 d_1(\mu,\nu). \tag{19.48}$$

Next, consider the nonincreasing function

$$h_2'(x) = \int_S \int_{F_2\setminus(F_1\cap F_2)} (1 - h(\tilde{\gamma}x))Q^N(d\gamma), \tag{19.49}$$

$0 \le h_2'(x) \le (1 - h(x_0))(Q^N(F_2) - Q^N(F_1 \cap F_2)) := a_2$, say, $a_2 - h_2'(x)$ being a non-decreasing function, $0 \le a_2 - h_2'(x) \le a_2$, belonging to $\mathcal{H}_{a_2}$. Hence, $|\int_S h_2'(x)\mu(dx) - \int_S h_2'(x))\nu(dx)| \le a_2 d_1(\mu,\nu)$. But the left side equals $|\int_S h_2(x)\nu(dx) - \int_S h_2(x)\mu(dx)|$, so that

$$|\int_S h_2(x)\mu(dx) - \int_S h_2(x)\nu(dx)| \le a_2 d_1(\mu,\nu). \tag{19.50}$$

The relations (19.47)–(19.50) yield

$$\begin{aligned}
d_1(T^{*N}\mu, T^{*N}\nu) &\leq \sup_{h\in\mathcal{H}_1} (a_1 + a_2 + a_3) d_1(\mu, \nu) \\
&= \sup_{h\in\mathcal{H}_1} \{h(x_0)(Q^N(F_1) - Q^N(F_1 \cap F_2) \\
&\quad + (1 - h(x_0))(Q^N(F_2) - Q^N(F_1 \cap F_2)) \\
&\quad + 1 - (1 - Q^N(F_1 \cup F_2))d_1(\mu, \nu) \\
&\leq \max\{Q^N(F_1) - Q^N(F_1 \cap F_2), Q^N(F_2) - Q^N(F_1 \cap F_2)\} \\
&\quad + Q^N(F_1 \cup F_2) d_1(\mu, \nu) = \delta d_1(\mu, \nu). \qquad (19.51)
\end{aligned}$$

Iterating the inequality $d_1(T^{*N}\mu, T^{*N}\nu) \leq \delta d_1(\mu\nu)$ with μ, ν replaced by $T^{*N}\mu, T^{*N}\nu$, and so on, and using

$$\begin{aligned}
d_1(T^*\mu, T^*\nu) &= \sup_{h\in\mathcal{H}_1} \{|\int_S [\int_S h(y)p(x, dy)]\mu(dx) - \int_S [\int_S h(y)p(x, dy)]\nu(dx)|\} \\
&= \sup_{h\in\mathcal{H}_1} |\int_S \int_\Gamma h(\tilde{\gamma}x)Q(d\gamma)\mu(dx) - \int_S \int_\Gamma h(\tilde{\gamma}x)Q(d\gamma)\nu(dx)| \\
&\leq d_1(\mu, \nu),
\end{aligned}$$

we arrive at (19.43). Note that the inner integral $\int_\Gamma h(\tilde{\gamma}x)Q(d\gamma)$ above is a non-decreasing function of x bounded between 0 and 1. If one could prove that $(\mathcal{P}(S), d_1)$ is a complete metric space, then the proof of the theorem would be complete, letting π be the limit of the Cauchy sequence $\{T^{*n}\nu\}$ with $\nu = T^{*m}\mu$ in (19.43). Unfortunately, that is a rather complex issue (see Remark 19.2). Instead, we consider the metric $d_{\mathcal{A}}$, and note that $d_1 \geq d_{\mathcal{A}}$. $h \equiv 1 - \mathbf{1}_A \in \mathcal{H}_1$ for every $A \in \mathcal{A}$. Hence, (19.43) holds with $d_{\mathcal{A}}$ replacing d_1. Lemma 2 below proves that $(\mathcal{P}(S), d_{\mathcal{A}})$ is a complete metric space. Hence, the Cauchy sequence $\{T^{*n}\mu\}$ under $d_{\mathcal{A}}$ converges to a limit π, say, which is easily seen to be the unique invariant probability of the Markov process. The relation (19.44) follows.

Lemma 2 *(Completeness of* $(\mathcal{P}(S), d_{\mathcal{A}})$*).* If S is a closed subset of $\mathbb{R}^k$, then $(\mathcal{P}(S), d_{\mathcal{A}})$ is a complete metric space.

Proof. Let $\{P_n\}$ be a Cauchy sequence in $(\mathcal{P}(S), d_{\mathcal{A}})$. Due to the completeness of $\mathbb{R}$, $P_n(A)$ converges uniformly on $\mathcal{A}$ to some function $P_\infty(A)$, $0 \leq P_\infty(A) \leq 1$. Let $\hat{P}_n$ be the extension of P_n to $\mathbb{R}^k$, i.e., $\hat{P}_n(B) = P_n(B \cap S)$ for all Borel sets B in $\mathbb{R}^k$. Taking φ to be the identity map on S, $\varphi(y) = y$, as one of the functions in the definition of $\mathcal{A}$, the sequence $\{F_n(x) : x \in \mathbb{R}^k\}$ of distribution functions of $\{\hat{P}_n\}$ is a Cauchy sequence converging uniformly on $\mathbb{R}^k$ to a distribution function F of a probability measure $\hat{P}$ on $\mathbb{R}^k$. Since S is closed and $P_n(S) = 1$ for all n, it follows from Alexandrov's Theorem[5] that $1 \geq \hat{P}(S) \geq \limsup_n P_n(S) = 1$ so that $\hat{P}(S) = 1$. Let P be the restriction of $\hat{P}$ to S. Then $P \in \mathcal{P}(S)$. We will show that $P_\infty = P$. Let O be an open subset of S. Then $O = G \cap S$, where G is an open subset of $\mathbb{R}^k$. Now $\liminf_n P_n(O) = \liminf_n \hat{P}_n(G) \geq \hat{P}(G) = P(O)$, proving the desired result that P_n converges weakly to P. Note that the two equalities here follow from the definition of $\hat{P}$ and P, while the inequality follows from Alexandrov's Theorem, using the weak convergence of $\hat{P}_n$ to $\hat{P}$. We have proved that P_n converges weakly to P. This implies that for every non-decreasing continuous function φ on S (into $\mathbb{R}^k$), $P_n \circ \varphi^{-1}$ (on $\mathbb{R}^k$) converges weakly to $P \circ \varphi^{-1}$, so that the distribution function $F_n(x)$ of $P_n \circ \varphi^{-1}$ converges to that of $P \circ \varphi^{-1}$, say $F(x)$, uniformly (as argued above). With A as defined in (19.39), this says that $P_n(A) \equiv F_n(x) \to F(x) \equiv P(A)$, as $n \to \infty$. Hence, $P(A) = P_\infty(A)$ for all A of the form (19.39).

The following example[6] shows that the hypothesis that S is closed, or some such restriction is necessary for the result of Theorem 19.9.

Example 7. Let $C \subset [0, 1]$ be the Cantor "middle-third" set, P be the Cantor distribution with support C, and X be a random variable with distribution P. The distributions P_n of $X + 3^{-n}$ satisfy $P_n([0, 1]\backslash C) = 1$, with distribution functions converging uniformly to the distribution function of P, but $P([0, 1]\backslash C) = 0$. Hence, $(\mathcal{P}([0, 1]\backslash C), d_{\mathcal{A}})$ is not complete.

Remark 19.2. One can extend Theorem 19.9 to state spaces that are open or semi-closed rectangles, finite, or infinite.[7] But completeness of $\mathcal{P}(S)$ under the metric d_1 seems to be a delicate issue.[8]

Definition 19.1. A *Markov process* on a Borel subset S of $\mathbb{R}^k$ having the transition probability $p(x, dy)$ is said to be *monotone increasing* if $p(x, dy)$ is *stochastically larger* than $p(x', dy)$ whenever $x' \leq x$.

This definition means that if X and X' have distributions $p(x, dy)$ and $p(x', dy)$, respectively, then X is stochastically larger than X': $P(X > y) \geq P(X' > y)$ for all $y \in \mathbb{R}^k$. Equivalently, writing F_z as the distribution function of a random variable with distribution $p(z, dy)$, a monotone increasing Markov process satisfies $F_x(y) \leq F_{x'}(y)$ for all y, if $x' \leq x$. In this sense, the Markov processes generated by iterations of i.i.d. monotone non-decreasing maps are monotone increasing. Further results pertaining to monotone Markov processes are presented in Chapter 24 in the context of *coupling*.

[5] See BCPT Theorem 7.1, pp. 137–139; Billingsley (1968): pp. 11–14.

[6] [Personal communication (1994)] This example was kindly furnished by Professor B.V. Rao, ISI Kolkata, India.

[7] See Bhattacharya and Lee (1997).

[8] Some general insights and detailed analysis for subsets S of $\mathbb{R}$ and $\mathbb{R}^2$ may be found in Chakroborty and Rao (1998).

We leave the proof of the following corollary to Exercise 19. Note that there is a certain symmetry in the splitting hypothesis (H).

Corollary 19.10 *(Convergence to Equilibrium of Monotone Decreasing Markov Processes).* Theorem 19.9 holds for iterations of monotone nonincreasing i.i.d. maps.

Remark 19.3. When the state space is an interval, the above corollary follows as a special case of Corollary 19.8. However, a dramatic difference between nondecreasing and nonincreasing cases is that the necessity condition for the existence of a unique equilibrium in Proposition 19.7 may fail in the nonincreasing case. We cite an example of randomly iterated quadratic maps $\gamma_{\theta_i}(x) = \theta_i x(1-x)(i = 1, 2)$ on $[0, 4]$, with $\theta_1 = 3.18, \theta_2 = 3.20$. The maps leave the interval $[u, v]$ invariant, where $u = \min\{1 - 1/\theta_1\}, \gamma_{\theta_1}(\theta_2/4)\} = 0.5088$, $v = \theta_2/4 = 0.80$. The maps $\gamma_{\theta_i}(i = 1, 2)$ are decreasing on $[u, v]$. They are also periodic, each with a two-period stable orbit. An even number of random iterates are increasing satisfying the splitting condition on a subinterval $[u, q_1]$ of $[u, v]$, and also on $[q_2, v]$, while an odd number of iterations take the first subinterval to the second and vice versa. The Markov process is periodic, or cyclical, and the transition probability $p^{(n)}(x, dy)$ converges only in Cesaro mean to a unique invariant probability, although no splitting condition holds.[9]

The final result of this chapter is an FCLT for monotone Markov processes.

Theorem 19.11 *(FCLT for Monotone Markov Processes).* Let the hypothesis of Theorem 19.9 hold for a Markov process $\{X_n : n \geq 0\}$ generated by i.i.d. monotone non-decreasing maps. (a) (*One-dimensional Case*). Let S be an interval. If f may be expressed as the difference between two monotone non-decreasing functions in $L^2(S, \pi)$, then the FCLT holds for the polygonal process $Y_n(t) = (1/\sqrt{n})S_n(t) = (1/\sqrt{n})[f(X_1) + \cdots + f(X_{[nt]}) + (nt - [nt])f(X_{[nt]+1})]$ $(0 \leq t \leq 1)$. (b) (Multidimensional Case). For $k > 1$, let the state space be a closed subset S of $\mathbb{R}^k$. The FCLT holds for f that may be expressed as the difference between two bounded measurable monotone increasing functions on S. (c) The variance parameter σ^2 for the limiting Wiener measure in (a), (b) is given by $\sigma^2 = \int_S g^2 d\pi - \int_S (Tg)^2 d\pi$, where $(I - T)g = f - \int_S f d\pi$, i.e., $f - \int_S f d\pi$ belongs to the range of $I - T$ in $L^2(S, \pi)$.

Proof. (a) First let us state a simple result: for all probability measures μ on $\mathbb{R}$ such that $\int_{\mathbb{R}} x^2 \mu(dx) < \infty$, the following equality holds

$$\int_{\mathbb{R}} x^2 \mu(dx) - (\int_{\mathbb{R}} x\mu(dx))^2 = (1/2) \int_{\mathbb{R}} \int_{\mathbb{R}} (x - y)^2 \mu(dx)\mu(dy), \qquad (19.52)$$

[9] For the details, we refer to Bhattacharya and Majumdar (2007), p. 315.

which is easily proved by expanding the square on the right side and integrating. Next, let f be monotone non-decreasing on S, $f \in L^2(S, \pi)$. By (19.52), $\mu(dy) = p^{(N)}(x, dy)$, where $p(x, dy)$ is the transition probability of the Markov process. Writing $\overline{f} = \int_S f d\pi$ and $|| \cdot ||_2$ for the L_2 norm in $L_2(S, \pi)$, one has using (19.52)

$$\begin{aligned}
&||T^N(f - \overline{f})||_2^2\\
&\quad = \int_S \left\{ \int_S (f(y) - \overline{f}) p^{(N)}(x, dy) \right\}^2 \pi(dx)\\
&\quad = \int_S \left[\int_S (f(y) - \overline{f})^2 p^{(N)}(x, dy) \right.\\
&\qquad \left. -(1/2) \int_S \int_S (f(y) - f(z))^2 p^{(N)}(x, dy) p^{(N)}(x, dz) \right] \pi(dx)\\
&\quad = ||f - \overline{f}||_2^2 - (1/2) \int_S \left[\int_S (f(y) - f(z))^2 p^{(N)}(x, dy) p^{(N)}(x, dz) \right] \pi(dx).
\end{aligned} \tag{19.53}$$

The splitting condition now yields (see (19.42))

$$\begin{aligned}
&\int_S \int_S (f(y) - f(z))^2 p^{(N)}(x, dy) p^{(N)}(x, dz)]\\
&\quad \geq \int_{z \geq x_0} \int_{y \leq x_0} (f(y) - f(x_0))^2 p^{(N)}(x, dy) p^{(N)}(x, dz)\\
&\qquad + \int_{z \leq x_0} \int_{y > x_0} (f(y) - f(x_0))^2 p^{(N)}(x, dy) p^{(N)}(x, dz)\\
&\quad \geq Q^N(F_2) \int_{y \leq x_0} (f(y) - f(x_0))^2 p^{(N)}(x, dy)\\
&\qquad + Q^N(F_1) \int_{y > x_0} (f(y) - f(x_0))^2 p^{(N)}(x, dy)\\
&\quad \geq \min\{Q^N(F_1), Q^N(F_2)\} \int_S (f(y) - f(x_0))^2 p^{(N)}(x, dy).
\end{aligned} \tag{19.54}$$

Therefore,

$$\begin{aligned}
&\int_S \left[\int_S \int_S (f(y) - f(z))^2 p^{(N)}(x, dy) p^{(N)}(x, dz) \right] \pi(dx)\\
&\quad \geq \min \left\{ Q^N(F_1), Q^N(F_2) \right\} \int \int (f(y) - f(x_0))^2 p^{(N)}(x, dy) \pi(dx)
\end{aligned}$$

$$
\begin{aligned}
&= \min\left\{Q^N(F_1), Q^N(F_2)\right\} \int (f(y) - f(x_0))^2 \pi(dy) \\
&\geq \min\left\{Q^N(F_1), Q^N(F_2)\right\} ||f - \overline{f}||_2^2 \geq \delta ||f - \overline{f}||_2^2. \qquad (19.55)
\end{aligned}
$$

The relations (19.53) and (19.55) imply that

$$
||T^N(f - \overline{f})||_2^2 \leq (1 - \frac{1}{2}\delta)||f - \overline{f}||_2^2. \qquad (19.56)
$$

So,

$$
||T^N(f - \overline{f})||_2 \leq \theta ||f - \overline{f}||_2, \qquad (19.57)
$$

where $\theta := (1 - \frac{1}{2}\delta)^{1/2} < 1$. If f is non-decreasing, then Tf is non-decreasing, and $T^n f$ is non-decreasing for all n; it follows on iteration that $||T^n(f - \overline{f})||_2 \leq \theta^{[n/N]}$ for all n, so that $g := \sum_{0 \leq n < \infty} T^n(f - \overline{f})$ is well defined and belongs to $L^2(S, \pi)$, and one has $(I - T)g = f - \overline{f}$. That is, $f - \overline{f}$ belongs to the range of $I - T$. The proof of the convergence in distribution to Brownian motion of the process $\{Y_n(t) : 0 \leq t \leq 1\}$ now follows from the Billingsley–Ibragimov FCLT (Theorem 15.5), provided the initial distribution is π. If $f = f_1 - f_2$ with f_1, f_2, both monotone non-decreasing functions in $L^2(S, \pi)$, then both functions, minus their means, are in the range of $I - T$, and so is their difference. Hence, the Billingsley–Ibragimov FCLT applies for this f as well. It remains to prove (a) under an arbitrary initial distribution. Assume for now that $f \in L^2(S, \pi)$ is monotone non-decreasing. Let $X_j(x)$, $j = 0, 1, \ldots$ denote the Markov process starting at x, while, as above, X_j, $j = 0, 1, \ldots$, is the process under initial distribution π. For simplicity, write $\tilde{f} = f - \overline{f}$ and

$$
\begin{aligned}
S_{m,m'}(x) &= (1/\sqrt{n}) \sum_{m \leq j \leq m'} \tilde{f}(X_j(x)), \\
S_{m,m'} &= (1/\sqrt{n}) \sum_{m \leq j \leq m'} \tilde{f}(X_j). \qquad (19.58)
\end{aligned}
$$

Thus $S_{0,n}(x) = S_{0,n_0-1}(x) + S_{n_0,n}(x)$, and similarly for $S_{0,n}(n_0 < n)$. Note that $S_{0,n_0}(x), S_{0,n_0} \to 0$ as $n \to \infty$ (for every n_0). Also, for every $r \in \mathbb{R}$, $y \to h_n(y) := P(S_{0,n-n_0}(y) > r)$ is monotone non-decreasing, and $P(S_{n_0,n}(x) > r) = \mathbb{E}h_n(X_{n_0}(x)) = \int_S h_n(y)p^{(n_0)}(x, dy)$. Therefore, by Theorem 19.9,

$$
\begin{aligned}
&\sup_{n > n_0} |\int_S h_n(y)p^{(n_0)}(x, dy) - \int_S h_n(y)\pi(dy)| \\
&\quad = \sup_{n > n_0} |P(S_{0,n}(x) > r) - P(S_{0,n-n_0} > r)| \to 0, \qquad (19.59)
\end{aligned}
$$

as $n_0 \to \infty$. Hence, given $\epsilon > 0$, one may choose $n_0 = n_0(\epsilon)$ such that the left side of (19.59) is less than $\epsilon/3$, and then choose $n(\epsilon)$ such that for all $n \geq n(\epsilon)$, one has

$$\begin{aligned}&|P(S_{0,n}(x) > r) - P(S_{0,n} > r)|\\&\quad \leq |P(S_{0,n}(x) > r) - P(S_{0,n-n_0} > r)| + |P(S_{0,n} > r) - P(S_{0,n-n_0} > r)|\\&\qquad < \epsilon, \quad \text{for all } n \geq n(\epsilon).\end{aligned} \tag{19.60}$$

Note that $S_{0,n}$ and $S_{0,n-n_0}$ differ by a sum of n_0 consecutive terms whose distribution is that of S_{0,n_0-1}, which goes to zero as $n \to \infty$. By (19.60), for every x, $S_{0,n}(x)$ converges in distribution to the corresponding limiting normal distribution of $S_{0,n}$. An entirely analogous argument shows that all finite dimensional distributions of the process $\{Y_n\}$, starting at x, say $\{Y_n(x)\}$, converge to those of the limiting Brownian motion (Exercise 19). For the tightness of $\{Y_n(x) : n = 0, 1, \ldots\}$, define, for each $r > 0$, and positive integers $n_0 < n_1 = n_0 + 1 < n_2 = n_0 + 2 < \cdots < n_{N+1} = n$, the sets

$$A(y) = \left[\max_{0 \leq i \leq N} S_{n_0+i,n_0+i+1}(y) > r\right], \quad B(y) = \left[\max_{0 \leq i \leq N} S_{n_0+i,n_0+i+1}(y) \geq -r\right]. \tag{19.61}$$

Let A, B be the corresponding events for the stationary sequence $\{X_j\}$. Since $P(A(y))$, $P(B(y))$ are non-decreasing in y, Theorem 19.9 yields as $n \to \infty$,

$$\begin{aligned}P(A(y)) - P(A) &= \int_S P(A(y))p^{(n_0)}(x, dy) - \int_S P(A)p^{(n_0)}(x, dy) \to 0\\P(B(y)) - P(B) &= \int_S P(B(y))p^{(n_0)}(x, dy) - \int_S P(B)p^{(n_0)}(x, dy) \to 0,\end{aligned} \tag{19.62}$$

uniformly for all $r > 0$. Because the process $\{Y_n\}$, corresponding to the sequence $\{X_j\}$, i.e., with X_0 having the distribution π, converges in distribution to Brownian motion, (19.62) implies that $\{Y_n(x) : n = 0, 1, \ldots\}$ is tight (Exercise 19) and, therefore, converges in distribution to Brownian motion with variance parameter σ^2 (see Theorem 15.5 for the computation of σ^2). On integration of the distribution of $Y_n(x)$ with respect to x, it now follows that $\{Y_n\}$ converges in distribution to Brownian motion under every initial distribution. If $f = f_1 - f_2$, both f_1, f_2 are monotone non-decreasing elements of $L^2(S, \pi)$, and let $Y_n^{(1)}, Y_n^{(2)}$ be the two processes as above corresponding to f_1, f_2, respectively. Letting $S(i)_{m,m'}(y)$, $S(i)_{m,m'}(i = 1, 2)$, be the quantities corresponding to those in (19.58). In place of $A(y)$, consider the set $A^{(1)}(y) = [\max_{0 \leq i \leq N} S(1)_{n_0+i,n0+i+1}(y) > r_1]$, $A^{(2)}(y) = [\max_{0 \leq i \leq N} S(2)_{n_0+i,n_0+i+1}(y) > r_2]$, $r_1 > 0, r_2 > 0$, and similarly define $B^{(i)}(y)(i = 1, 2)$. Apply the above results separately to these quantities for $i = 1, 2$ (Exercise 19).

(b) Unfortunately, for $k > 1$, the relation (19.55) does not hold.[10]

But the boundedness assumption simplifies much. It is enough to consider the case f bounded, $c \leq f(x) \leq d, c < d$, by translation, if necessary, one may assume $c \geq 0$. Then, writing $\tilde{f} = f - \int_S f d\pi$, $D(u) = \{y : f(y) > u\}$,

$$\begin{aligned} |T_n \tilde{f}(x)| &= |\int_{[c,d]} P(f(X_n(x)) > u)du - \int_{[c,d]} P(f(X_n) > u)du| \\ &= |\int_{[c,d]} p^{(n)}(x, D(u))du - \int_{[c,d]} \pi(D(u))du| \leq (1-\delta)^{[n/N]}, \end{aligned}$$

by Theorem 19.9. Note that $\mathbf{1}_{D(u)}$ belongs to $\mathcal{H}_1$. It follows that the infinite series $\sum_{0 \leq n < \infty} T^n \tilde{f}(x)$ converges uniformly to define a bounded non-decreasing function, say $g(x)$, on S, and $(I - T)g = \tilde{f}$. The rest of the proof of part (b) is the same as that of the proof of part (a).

The following result of Dubins and Freedman (1966) indicates the important role of splitting for monotone Markov processes.

Proposition 19.7. Let $\{\boldsymbol{\alpha}_n\}_{n \geq 1}$ be an i.i.d. sequence of increasing continuous maps on a closed bounded interval $[a, b]$. Then the condition (19.37) is *necessary* as well as sufficient for the existence of a unique non-degenerate invariant probability π.

Proof. Define a backward iteration by

$$Y_0(x) \equiv x, \qquad Y_n(x) := \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2 \cdots \boldsymbol{\alpha}_n x \qquad (n \geq 1). \tag{19.63}$$

Then $Y_n(x)$ and $X_n(x)$ have the same distribution. Also,

$$\begin{gathered} Y_1(a) \geq a, \qquad Y_2(a) = Y_1(\boldsymbol{\alpha}_2 a) \geq Y_1(a), \ldots \\ Y_{n+1}(a) = Y_n(\boldsymbol{\alpha}_{n+1} a) \geq Y_n(a), \ldots \end{gathered}$$

i.e., the sequence of random variables $\{Y_n(a) : n \geq 0\}$ is increasing. Similarly, $\{Y_n(b) : n \geq 0\}$ is decreasing. Let the limits of these two sequences be $\underline{Y}$, $\bar{Y}$, respectively. As $Y_n(a) \leq Y_n(b)$ for all n, $\underline{Y} \leq \bar{Y}$. If $P(\underline{Y} < \bar{Y}) > 0$, then $\underline{Y}$ and $\bar{Y}$ cannot have the same distribution. In other words, $Y_n(a)$ (and, therefore, $X_n(a)$) and $Y_n(b)$ (and, therefore, $X_n(b)$) converge in distribution to different limits π_1, π_2, say, both invariant probabilities by Proposition 8.4(a). On the other hand, if $\underline{Y} = \bar{Y}$ a.s., then these limiting distributions are the same, say π. Also, $Y_n(a) \leq Y_n(x) \leq Y_n(b)$ for all x, so that in this case $Y_n(x)$ converges in distribution to the same limit π, whatever x may be. Therefore, π is the unique invariant probability. The assumption of non-degeneracy of π, i.e., π, does not assign all its mass at a single point, and

[10] [Personal Communication (2019)] This was kindly pointed out by Dr. Eduardo A. Silva of the Universidade de Brasilia, Brazil.

rules out the case that with probability 1 all $\boldsymbol{\alpha}(\omega)$ have a common fixed point. Then there exist $c < d$ such that $P(\overline{Y} < c) > 0$ and $P(\underline{Y} > d) > 0$. There exists m such that $P(Y_m(b) < c) > 0$ and $P(Y_m(a) > d) > 0$. Now any $x_0 \in [c, d]$ satisfies (19.37).

Remark 19.4. Example 3, Chapter 18, and Theorem 18.3 provide instances of Markov processes on $[0, \infty)$ generated by i.i.d. continuous, monotone maps that have unique invariant probabilities[11] although, in general, splitting does not hold.

Example 8 *(Iterates of Quadratic Maps[12]).* Consider the random quadratic (logistic) maps on [0, 1] of the form

$$\gamma_\theta(x) = \theta x(1-x), 0 \le x \le 1, \tag{19.64}$$

for parameters $\theta \in (0, 4)$. One may check (Exercise 19) that if $1 < \theta_1 < \theta_2 < 4$, then for any $\theta \in [\theta_1, \theta_2]$, the interval $[a, b] = [(1 - \frac{1}{\theta_1}) \wedge \gamma_{\theta_1}(\frac{\theta_2}{4}), \frac{\theta_2}{4}]$ is invariant under γ_θ. Moreover, for $1 < \theta_1 < \theta_2 \le 3$, γ_{θ_i} have attractive fixed points located at $p_i = 1 - \frac{1}{\theta_i}, i = 1, 2$, respectively. Consider the case $1 < \theta_1 < \theta_2 < 2$, where γ_{θ_i} is selected with probability $u_i, 0 < u_i < 1, u_1 + u_2 = 1, i = 1, 2$. Let $S = [p_1, p_2] \subset [0, 1/2]$ and choose $x_0 \in (p_1, p_2)$ such that for m sufficiently large $\gamma_{\theta_1}^m(p_2) < x_0$ and $\gamma_{\theta_2}^m(p_1) > x_0$. In particular, the maps $\gamma_{\theta_i}, i = 1, 2$, are both increasing on $S = [p_1, p_2]$, and splitting holds in Corollary 19.8 with $\delta = \min\{u_1^m, u_2^m\}$. Moreover, since for any $x \in (0, 1)\backslash[p_1, p_2]$, the successive iterates will enter $[p_1, p_2]$ in a finite number $n(x)$ of steps. Thus one obtains the existence of a unique invariant probability π on (0, 1) with an exponential rate of convergence starting from any $x \in (0, 1)$. Although the maps γ_{θ_i} are decreasing in the case $2 < \theta_1 < \theta_2 \le 3$, the composites $\gamma_{\theta_i}\gamma_{\theta_j}$ are increasing. Thus the conclusions are similar with the details left as Exercise[13] 19.

The following example[14] from economics provides an application of Corollary 19.8.

Example 9 *(A Descriptive Model of Capital Accumulation).* Consider an economy that has a single producible good. The economy starts with an initial stock $X_0 = x > 0$ of this good that is used to produce an output Y_1 in period 1. The *output* Y_1 is not a deterministic function of the *input* x. In view of the randomness of the state of nature, Y_1 takes one of the values $f_r(x)$ with probability $p_r > 0$ $(1 \le r \le N)$. Here, f_r are *production functions* having the following properties:

[11] For the speed of convergence to the invariant probability in these examples, see Lund and Tweedie (1996) and Bhattacharya and Majumdar (2010a) (with polynomial rates).

[12] Markov processes generated by i.i.d. iterations of quadratic maps were considered by Bhattacharya and Rao (1993). Also see Athreya and Dai (2000), Bhattacharya and Majumdar (2004), (2007).

[13] These examples and much more are presented in Bhattacharya and Majumdar (2007).

[14] This example may be found in Mirman (1980).

(a) f_r is twice continuously differentiable, $f_r'(x) > 0$, and $f_r''(x) < 0$ for all $x > 0$.
(b) $\lim_{x\downarrow 0} f_r(x) = 0$, $\lim_{x\downarrow 0} f_r'(x) > 1$, $\lim_{x\uparrow\infty} f_r'(x) = 0$.
(c) If $r > r'$, then $f_r(x) > f_{r'}(x)$ for all $x > 0$.

The strict concavity of f_r in (i) reflects a *law of diminishing returns*, while (iii) assumes an ordering of the technologies or production functions f_r, from the least productive f_1 to the most productive f_N.

A fraction β $(0 \le \beta < 1)$ of the output Y_1 is *consumed*, while the rest $(1-\beta)Y_1$ is invested for the production in the next period. The total stock X_1 at hand for investment in period 1 is $\theta X_0 + (1-\beta)Y_1$. Here, $\theta < 1$ is the rate of *depreciation* of capital used in production. This process continues indefinitely, each time with an independent choice of the production function f_r with probability p_r, $1 \le r \le N$. Thus, the capital X_{n+1} at hand in period $n+1$ satisfies

$$X_{n+1} = \theta X_n + (1-\beta)\varphi_{n+1}(X_n) \qquad (n \ge 0), \tag{19.65}$$

where φ_n is the random production function in period n,

$$P(\varphi_n = f_r) = p_r > 0 \qquad (1 \le r \le N),$$

and the φ_n $(n \ge 1)$ are independent. Thus the Markov process $\{X_n(x) : n \ge 0\}$ on the *state space* $(0, \infty)$ may be represented as

$$X_n(x) = \boldsymbol{\alpha}_n \cdots \boldsymbol{\alpha}_1 x,$$

where, writing

$$g_r(x) := \theta x + (1-\beta) f_r(x), \qquad 1 \le r \le N, \tag{19.66}$$

one has

$$P(\alpha_n = g_r) = p_r \qquad (1 \le r \le N). \tag{19.67}$$

Suppose, in addition to the assumptions already made, that

$$\theta + (1-\beta)\lim_{x\downarrow 0} f_r'(x) > 1 \qquad (1 \le r \le N), \tag{19.68}$$

i.e., $\lim_{x\downarrow 0} g_r^{\hat{\prime}}(x) > 1$ for all r. As $\lim_{x\to\infty} g_r'(x) = \theta + (1-\beta)\lim_{x\to\infty} f_r'(x) = \theta < 1$, it follows from the strictly increasing and strict concavity properties of g_r that each g_r has a *unique fixed point* a_r on the state space $S = (0, \infty)$

$$g_r(a_r) = a_r \qquad (1 \le r \le N), \tag{19.69}$$

namely, a_r is the point where g_r crosses the line $y = x$. Note that by property (iii) of $f_r, a_1 < a_2 < \cdots < a_N$. If $y \ge a_1$, then $g_r(y) \ge g_r(a_1) \ge g_1(a_1) = a_1$, so that

$X_n(x) \geq a_1$ for all $n \geq 0$ if $x \geq a_1$. Similarly, if $y \leq a_N$, then $g_r(y) \leq g_r(a_N) \leq g_N(a_N) = a_N$, so that $X_n(x) \leq a_N$ for all $n \geq 0$ if $x < a_N$. As a consequence, if the initial state x is in $[a_1, a_N]$, then the process $\{X_n(x) : n \geq 0\}$ remains in $[a_1, a_N]$ forever. In this case, one may take $S = [a_1, a_N]$ to be the effective state space. Also, if $x \geq a_1$, then the nth iterate of g_1, namely $g_1^{(n)}(x)$, decreases as n increases. For if $x \geq a_1$, then $g_1(x) \leq x$, $g_1^{(2)}(x) = g_1(g_1(x)) \leq g_1(x)$, etc. The limit of this decreasing sequence is a fixed point of g_1 (Exercise 19) and, therefore, must be a_1. Similarly, if $x \leq a_N$, then $g_N^{(n)}(x)$ increases, as n increases, to a_N. In particular,

$$\lim_{n\to\infty} g_1^{(n)}(a_N) = a_1, \qquad \lim_{n\to\infty} g_N^{(n)}(a_1) = a_N.$$

Thus, there exists an integer n_0 such that

$$g_1^{(n_0)}(a_N) < g_N^{(n_0)}(a_1). \tag{19.70}$$

This means that if $x_0 \in [g_1^{(n_0)}(a_n), g_N^{(n_0)} a_1)]$, then

$$P(X_{n_0}(x) \leq x_0 \text{ for all } x \in [a_1, a_N]) \geq P(\boldsymbol{\alpha}_n = g_1 \text{ for } 1 \leq n \leq n_0) = p_1^{n_0} > 0,$$
$$P(X_{n_0}(x) \geq x_0 \text{ for all } x \in [a_1, a_N]) \geq P(\boldsymbol{\alpha}_n = g_N \text{ for } 1 \leq n \leq n_0) = p_N^{n_0} > 0.$$

Hence, the condition (19.37) of Corollary 19.8 holds, with $m = n_0$, and there exists a unique invariant probability π, if the state space is taken to be $[a_1, a_N]$.

Next fix the initial state x in $(0, a_1)$. Then $g_1^{(n)}(x)$ increases, as n increases. The limit must be a fixed point and, therefore, a_1. Since $g_r(a_1) > a_1$ for $r = 2, \ldots N$, there exists $\epsilon > 0$ such that $g_r(y) > a_1$ $(2 \leq r \leq N)$ if $y \in [a_1 - \epsilon, a_1]$. Now find n_ϵ such that $g_1^{(n_\epsilon)}(x) \geq a_1 - \epsilon$. If $\tau_1 := \inf\{n \geq 1 : X_n(x) \geq a_1\}$, then it follows from the above that

$$P(\tau_1 > n_\epsilon + k) \geq p_1^k \qquad (k \geq 1),$$

because $\tau_1 > n_\epsilon + k$ implies that the last k among the first $n_\epsilon + k$ functions $\boldsymbol{\alpha}_n$ are g_1. Since p_1^k goes to zero as $k \to \infty$, it follows from this that τ_1 is a.s. finite. Also $X_{\tau_1}(x) < a_N$ as $g_r(y) \leq g_r(a_N) \leq g_N(a_N) = a_N$ $(1 \leq r \leq N)$ for $y \leq a_1$, so that in a single step it is not possible to go from a state less than a_1 to a state larger than a_N. By the strong Markov property, and the result in the preceding paragraph on the existence of a unique invariant distribution and stability on $[a_1, a_N]$, it follows that $X_{\tau_1+r}(x)$ converges in distribution to π, as $r \to \infty$ (Exercise 19). From this, one may show that $p^{(n)}(x, dy)$ converges weakly to $\pi(dy)$ for all x, as $n \to \infty$, so that π is the unique invariant probability on $(0, \infty)$ (Exercise 19).

In the same manner, it may be checked that $X_n(x)$ converges in distribution to π if $x > a_N$. Thus, no matter what the initial state x is, $X_n(x)$ converges in distribution to π. Therefore, on the state space $(0, \infty)$, there exists a unique invariant distribution π (assigning probability 1 to $[a_1, a_N]$), and stability holds. In analogy with the case of Markov chains, one may call the set of states $\{x; 0 < x < a_1 \text{ or } x > a_N\}$ *inessential.*

The theory presented here, and in the previous Chapter 19 on iterations of monotone maps, extends to monotone maps on general partial ordered spaces S. We will illustrate this by an important example.

Example 10 *(Propp–Wilson Algorithm and the Gibbs Sampler for an Ising Model: Coupling from the Past).* Consider the problem of random sampling from a distribution π on a finite but enormously large state space S, where π is given only up to a normalizing constant that is extremely difficult to calculate. A basic approach here is *Gibbs sampling,* in which an appropriate Markov chain X_n is constructed having π as an invariant distribution and such that the distribution of X_n converges to π (say, in total variation distance) as $n \to \infty$. Then an observation of X_n for a large enough n may be viewed, approximately, as an observation from π. Having a large number of independent observations of this kind enables one to obtain an empirical distribution that is a good estimate of π. Example 6 above also deals with this problem, but under somewhat different assumptions. In the present example, S is the space of *configurations* on the finite two-dimensional lattice $L = \{(i, j) : 1 \le i, j \le M\}$, i.e., $S = \{-1, +1\}^L$. For all considerations below, one may alternatively choose S to be a subset $\{-1, +1\}^{L_1}$ of $\{-1, +1\}^L$ by fixing the signs at all sites in $L \backslash L_1$, referred to as the *boundary* of L_1. Thus an element $s = \{s_{ij} : (i, j) \in L\}$ of S is an assignment of -1 or $+1$ to each of the M^2 lattice sites of L. The probability measure[15] π of interest is given by

$$\pi(\{s\}) = c_\beta e^{\beta H(s)}, \quad s \in S \qquad (\beta > 0), \tag{19.71}$$

where β is a real-valued parameter and

$$H(s) = \sum_{(i,j),(i',j')} s_{ij} s_{i'j'}, \tag{19.72}$$

the sum being over *neighboring sites* $(i, j), (i', j')$, i.e., either $|i - i'| = 1$ and $j = j'$, or $i = i'$ and $|j - j'| = 1$. The Markov chain we construct has the following transition probabilities $p(s, \{s'\})$: given a configuration $s \in S$ and a site $(k, \ell) \in L$, let $s^{k\ell+}$ denote the element of S that is the same as s at all sites (i, j) except perhaps at (k, ℓ), and at (k, ℓ) the value of $s^{k\ell+}$ is $+1$. Similarly, $s^{k\ell-}$ is obtained by setting -1 at the site (k, ℓ), but keeping it the same as s at all other sites. Define

[15] In the physics literature, both in the definition of $H(s)$ and the exponent, defining π, each, has a minus sign. The signs are included there to make alignment of spins the least energetic (ground states), as well as to make them the most likely in the case $\beta > 0$. However, since they cancel, we omit them for convenience.

$$p\left(s,\{s^{k\ell+}\}\right)=\frac{1}{M^2}\pi_+(k\ell|s),\quad p\left(s,\{s^{k\ell-}\}\right)=\frac{1}{M^2}\pi_-(k\ell|s),\quad (k,\ell)\in L, s\in S, \tag{19.73}$$

where

$$\pi_+(k\ell|s)=\frac{\pi(\{s^{k\ell+}\})}{\pi(\{s^{k\ell+}\})+\pi(\{s^{k\ell-}\})},\quad \pi_-(k\ell|s)=1-\pi_+(k\ell|s). \tag{19.74}$$

All other transitions have zero probabilities. Note that the normalizing constant c_β cancels out from the numerator and denominator in (19.74). Indeed, since the signs of only the sites neighboring to (k,ℓ) are affected by switching from s to $s^{k\ell\pm}$, one has

$$p\left(s,\{s^{k\ell+}\}\right)=\frac{1}{M^2}\left(\frac{\exp\{\beta\sum^{*k\ell}s_{ij}\}}{\exp\{\beta\sum^{*k\ell}s_{ij}\}+\exp\{-\beta\sum^{*k\ell}s_{ij}\}}\right), \tag{19.75}$$

where $\sum^{*k\ell}$ is the sum over all sites (i,j) in L neighboring to (k,ℓ). Similarly, $p(s,\{s^{k\ell-}\})$ is given by (19.75) with β replaced by $-\beta$ in the numerator. One may now check that

$$\begin{aligned}&p\left(s,\{s^{k\ell+}\}\right)\pi(\{s\})=p(s^{k\ell+},\{s\})\pi\left(\{s^{k\ell+}\}\right),\\&p\left(s,\{s^{k\ell-}\}\right)\pi(\{s\})=p\left(s^{k\ell-},\{s\}\right)\pi\left(\{s^{k\ell-}\}\right),\quad\text{for all }(k,\ell)\in L, s\in S.\end{aligned} \tag{19.76}$$

Observe that $s=s^{k\ell+}$ if $s_{k\ell}=+1$ and $s=s^{k\ell-}$ if $s_{k\ell}=-1$. Thus if $s_{k\ell}=+1$, the first relation in (19.76) is trivially true. If $s_{k\ell}=+1$, then the numerator on the left side of the second relation equals $(c_\beta/M^2)\exp\{\beta a_{k\ell}\}$, say, where $a_{k\ell}=H(s)-\sum^{*k\ell}s_{ij}$, and the same is true for the right side; the denominators of both sides are obviously the same. An entirely analogous argument applies to the case $s_{k\ell}=-1$. From Proposition 8.8, it now follows that π is invariant and the chain is time-reversible.

Assume henceforth that $\beta>0$, referred to as the *ferromagnetic model*. We now construct monotone increasing i.i.d. maps $\boldsymbol{\alpha}_n$ $(n\geq 1)$ whose iterations give rise to a Markov process with the above transition probability. For $(k,\ell)\in L$ and $u\in(0,1)$, define the map $f_{k\ell,u}$ on S by setting $f_{k\ell,u}(s)=s'$, where s' equals s at all sites $(i,j)\neq(k,\ell)$ and with $s'_{k\ell}=+1$ if $0<u<\pi_+(k\ell|s)$, $s'_{ij}=-1$ if $\pi_+(k\ell|s)\leq u<1$. Let us show that $f_{k\ell,u}$ is monotone increasing on S with respect to the usual *partial order:* $s\leq s'$ if $s_{ij}\leq s'_{ij}$ for all $(i,j)\in L$. If $s\leq t$, then, for every $(i,j)\neq(k,\ell)$, one has $(f_{k\ell,u}(s))_{ij}=s_{ij}\leq t_{ij}=(f_{k\ell,u}(t))_{ij}$. At the site (k,ℓ), if $u\geq\pi_+(k\ell|s)$, then $(f_{k\ell,u}(s))_{k\ell}=-1\leq(f_{k\ell,u}(t))_{k\ell}$. Thus it remains to show that $\pi_+(k\ell|s)\leq\pi_+(k\ell|t)$, i.e., if $f_{k\ell,u}(s)$ is $+1$ at (k,ℓ), so is $f_{k\ell,u}(t)$. Now, $\pi_+(k\ell|s)$ is an increasing function of $\sum^{*k\ell}s_{ij}$ since $\beta>0$ and $u\to e^u/(e^u+e^{-u})=e^{2u}/(e^{2u}+1)$ is increasing on $(0,\infty)$. Using the fact that $\sum^{*k\ell}s_{ij}\leq\sum^{*k\ell}t_{ij}$, it now follows that $\pi_+(k\ell|s)\leq\pi_+(k\ell|t)$. Thus $f_{k\ell,u}$ is increasing on S for all $(k,\ell),u$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined two independent i.i.d. sequences $\{U_n\}_{n\geq 1}$ and $\{\varepsilon_n\}_{n\geq 1}$, where U_n are uniform on $(0, 1)$ and ε_n is uniform on L. Define $\boldsymbol{\alpha}_n = f_{\varepsilon_n, U_n}$ $(n \geq 1)$. Then (1) $\boldsymbol{\alpha}_n$ $(n \geq 1)$ generate, by iteration, the Markov process whose transition probability is given by (19.73), or (19.75), and (2) $\boldsymbol{\alpha}_n$ are monotone increasing on S. This is true whether S is unrestricted, i.e., $S = \{-1, +1\}^L$, or $S = \{-1, +1\}^{L_1}$, $L_1 \subset L$.

Consider the backward iterations starting at the smallest element a and the largest element b of S. Thus, a is the constant function assigning -1 to all sites (except at boundary points, if there is a boundary). Similarly, b is the constant function assigning $+1$ to all sites (excluding the boundary). Let

$$Y_n(a) = \boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_n a, \quad Y_n(b) = \boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_n b \quad (n \geq 1). \tag{19.77}$$

Now $Y_n(a)$ increases as n increases, since $Y_{n+1}(a) = Y_n(\alpha_{n+1}a) \geq Y_n(a)$; similarly, $Y_n(b)$ decreases as n increases, and $Y_n(a) \leq Y_n(b)$ for all n. Since there are only finitely many sites, and the probability that $Y_n(a)$ remains constant for infinitely many n is zero (Exercise 19), it follows that there is a *finite (random) time* T such that[16] $Y_T(a) = Y_T(b)$. Also, as shown in the proof of Proposition 19.7, the distribution of $Y_T(a)$ $(= Y_T(b))$ is the unique invariant probability π of the Markov chain. Thus one has achieved an *exact random sampling* of one observation from the distribution π, by recording $Y_T(a) = Y_T(b)$. By repeating this procedure, each time independently of all the preceding, one may take a random sample of any size.

Remark 19.5. While the ferromagnetic case rests on monotonicity of the maps for the Propp–Wilson algorithm, an extension has been developed for the so-called *anti-monotone models* with some success as well.[17]

Exercises

1. (*Doeblin Minorization Theorem*) Use the following steps to give an alternate derivation of Corollary 19.2:

   (i) Letting d denote the total variation metric, show that $d_1(\mu, \nu) := \sup\{|\int_S f d\mu - \int_S f d\nu| : f \in \mathbb{B}(S), |f| \leq 1\} = 2d(\mu, \nu)$, for all $\mu, \nu \in \mathcal{P}(S)$. [*Hint*: Here is an outline of an argument. Let $f = \sum_{i=1}^{k} c_i \mathbf{1}_{[A_i]}$, where $|c_i| \leq 1$, and the A_i's are disjoint. Then $|\int_S f d\mu - \int_S f d\nu| = \sum_{i\in I^+} c_i(\mu(A_i) - \nu(A_i)) + \sum_{i\in I^-} c_i(\mu(A_i) - \nu(A_i))$, where $I^+ := \{i \leq k : \mu(A_i) > \nu(A_i)\}$ and $I^- := \{1, 2, \ldots, k\}\backslash I^+$. Thus,

[16] A version of this so-called coupling from the past method already appears in the proof of Proposition 19.7. The application to Monte Carlo simulations was effectively developed by Propp and Wilson (1998).

[17] See Haggstrom and Nelander (1998) for an extension exploiting earlier ideas of Wilfrid Kendall.

$|\int_S f d\mu - \int_S f d\nu| \leq \mu(A^+) - \nu(A^+) + \mu(A^-) - \nu(A^-) \leq 2d(\mu, \nu)$, where $A^+ := \cup_{i \in I^+} A_i$, $A^- := \cup_{i \in I^-} A_i$. Thus $d_1(\mu, \nu) \leq 2d(\mu, \nu)$. Conversely, for $A \in \mathcal{S}$, take $f = \mathbf{1}_A - \mathbf{1}_{S \backslash A}$, so that $|\int_S f d\mu - \int_S f d\nu| = 2|\mu(A) - \nu(A)|$.

(ii) Use (19.13) to show $d_1(T^{*m}\mu, T^{*m}\nu) \leq (1-\delta) d_1(\mu, \nu)$.

(iii) Iterate (ii) (by induction) to prove

$$d_1(T^{*mk}\mu, T^{*mk}\nu) \leq (1-\delta)^k d_1(\mu, \nu), k \geq 1.$$

(iv) Show that for all $\mu \in \mathcal{P}(S)$, the sequence $\{T^{*mk}\mu : k \geq 1\}$ is Cauchy for the metric d_1 and therefore has a limit π that is the unique invariant probability.

(v) Use (iii) and (iv) with $\nu = \pi$ to complete the proof of (19.12).

2. Consider the mixture $p(x, dy) = \alpha\lambda(dy) + \beta q(x, dy), x, y \in S, \alpha > 0, \beta \geq 0$, $\alpha + \beta = 1$, where $\lambda(dy)$ and $q(x, dy)$ are probability measures on $(S, \mathcal{S})$ for each $x \in S$. Show that $p(x, dy)$ satisfies Doeblin minorization.
3. (i) Show that $\mathcal{P}(S)$ is a complete metric space under the total variation metric. (ii) Show that the Komogorov metric makes $\mathcal{P}(S)$ a complete metric space when S is an interval.
4. In the context of Corollary 19.3, show that

$$\sum_{j \in S} |p_{ij}^{(n)} - \pi_j| = 2 \sup_{B \subset S} \left| p^{(n)}(i, B) - \pi(B). \right|$$

5. Calculate the invariant probability and acceptance ratio in the case of *independence sampling,* where $q(x, y) = g(y), y \in S$ is independent of $x \in S$.
6. Verify that the probability that $Y_n(a)$ remains constant for infinitely many n is zero where $Y_n(a)$ is given by (19.77).
7. Let $S = [-2, 2]$, $X_{n+1} = f(X_n) + \varepsilon_{n+1}$ $(n \geq 0)$, X_0 independent of the i.i.d. sequence $\{\varepsilon_n : n \geq 1\}$ on $[-1, 1]$, and $f(x) = (x+1)\mathbf{1}_{[-2,0]}(x) + (x-1)\mathbf{1}_{(0,2]}$. Suppose $P(\varepsilon_n = 1) = P(\varepsilon_n = -1) = \frac{1}{2}$.

   (i) Show that with $X_0 \equiv x \in (0, 2]$, $\{X_n(x) : n \geq 1\}$ is i.i.d. with common distribution $\pi_x = \frac{1}{2}\delta_x + \frac{1}{2}\delta_{x-2}$ (i.e., $\pi_x(\{x\}) = \pi_x(\{x-2\}) = \frac{1}{2}$), so that π_x is invariant.

   (ii) If $X_0 \equiv x \in [-2, 0]$, show that $\{X_n(x) : n \geq 1\}$ is i.i.d. with common distribution $\gamma_x = \frac{1}{2}\delta_x + \frac{1}{2}\delta_{x+2}$, which is, therefore, invariant.

   (iii) Show that the Markov process has uncountably many mutually singular invariant probabilities.

   (iv) Show that the uniform distribution π (with p.d.f. $\frac{1}{4}$) on $[-2, 2]$ is invariant, and prove that if X_0 has distribution π, $(X_0 + \cdots + X_n)/(n+1)$ converges a.s. to a nonconstant random variable Z as $n \to \infty$. Find the distribution of Z.

8. In Exercise 19, assume ϵ_n are i.i.d. uniform on $[-1, 1]$.

(i) Prove that $\{X_{2n}(x) : n \geq 1\}$ is i.i.d. with common p.d.f. given by $\pi(y) = (1/4)(2 - |y|)\mathbf{1}_{[-2,2]}(y)$ if $x \in [-2, 2]$.

(ii) Show that π is the unique invariant probability, and $(X_0 + X_1 + \cdots + X_n)/(n+1)$ converges a.s. to a constant as $n \to \infty$, no matter what the initial distribution is.

(iii) Show that the sequence $\{X_n : n \geq 1\}$ is 2-dependent. [A sequence $\{X_n : n \geq 1\}$ is *m-dependent* if, for every $k \geq 0$, $\{X_j : 1 \leq j \leq k\}$ and $\{X_j : j \geq k + m\}$ are independent.]

9. In Exercise 19, modify f as follows. Let $0 < \delta < \frac{1}{2}$. Define $f_\delta(x) := f(x)$ for $-2 \leq x < -\delta$, and $\delta \leq x \leq 2$, and linearly interpolate between $(-\delta, \delta)$, so that f_δ is continuous.

   (i) Show that, for $x \in [\delta, 1]$ (or, $x \in [-1, -\delta]$), $\{X_n(x) : n \geq 1\}$ is i.i.d. with common distribution π_x (or, γ_x).

   (ii) For $x \in (1, 2]$ (or, $[-2, -1)$)$\{X_n(x) : n \geq 2\}$ is i.i.d. with common distribution π_x (or, γ_x).

   (iii) For $x \in (-\delta, \delta)\{X_n(x) : n \geq 1\}$ is i.i.d. with common distribution π_{-x+1}.

10. In reference to the capital accumulation example 9, complete the argument that X_{τ_1+r} converges in distribution to π and, consequently, $p^{(n)}(x, dy)$ converges weakly to π as well.

11. Let $\{Z_n : n \geq 1\}$ be an i.i.d. sequence with values in a measurable space $(A, \mathcal{A})$. Let $X_n = g(Z_n, Z_{n+1}, \ldots, Z_{n+m-1})$, $n \geq 1$, where g is a measurable function on $(A^n, \mathcal{A}^{\otimes s})$ into a measurable state space $(S, \mathcal{S})$. Show that $\{X_n : n \geq 1\}$ is *m-dependent.*

12. Let $\{\boldsymbol{\alpha}_n : n \geq 1\}$ be *i.i.d. random contractions* on a compact metric space $(S, \rho) : \rho(\boldsymbol{\alpha}_n x, \boldsymbol{\alpha}_n y) \leq \rho(x, y)$ for all x, y (a.s.), defined on a probability space $(\Omega, \mathcal{F}, P)$. Suppose there exists $F \in \mathcal{F}$ such that: (i) $P(F) > 0$ and (ii) $\alpha_1(\omega)$ is a *strict contraction* for $\omega \in F$, i.e., $\rho(\alpha_1(\omega)x, \alpha_1(\omega)y) < \rho(x, y)$ for all $x, y \in S$. Prove that there exists a unique invariant probability π for $X_n := \boldsymbol{\alpha}_n \cdots \boldsymbol{\alpha}_1 X_0$ $(n \geq 0)$ and that X_n converges in distribution to π, whatever be X_0.

13. In reference to Example 8, assume $1 < \theta_1 < \theta_2 < 4$. Show that: (a) For any $\theta \in [\theta_1, \theta_2]$, the interval $[a, b] = [(1 - \frac{1}{\theta_1}) \wedge \gamma_{\theta_1}(\frac{\theta_2}{4}), \frac{\theta_2}{4}]$ is invariant under γ_θ. (b) For $1 < \theta_1 < \theta_2 < 3$, γ_{θ_i} have attractive fixed points located at $p_i = 1 - \frac{1}{\theta_i}$, $i = 1, 2$, respectively. (c) For $1 < \theta_1 < \theta_2 < 2$, both maps are increasing. (d) For $2 < \theta_1 < \theta_2 < 3$, both maps are decreasing, but $\theta_i\theta_j$ is increasing for $i, j = 1, 2$. Complete the proof of the existence of a unique invariant probability for i.i.d. random iterates of the maps in this case.

14. Suppose that S comprises a single essential class of aperiodic states. Show that there is an integer ν such that $p_{ij}^{(\nu)} > 0$ for all $i, j \in S$ by filling in the details of the following steps:

   (a) For a fixed (i, j), let $B_{ij} = \{\nu \geq 1 : p_{ij}^{(\nu)} > 0\}$. Then for each state j, B_{jj} is closed under addition.

(b) (Basic Number Theory Lemma) If B is a set of positive integers having greatest common divisor 1 and if B is closed under addition, then there is an integer b such that $n \in B$ for all $n \geq b$. [*Hints*:

(i) Let G be the smallest additive subgroup of $\mathbb{Z}$ that contains B. Then argue that $G = \mathbb{Z}$ since if d is the smallest positive integer in G, it will follow that if $n \in B$, then, since $n = qd + r, 0 \leq r \leq d$, one obtains $r = n - qd \in G$ and hence $r = 0$, i.e., d divides each $n \in B$ and thus $d = 1$.

(ii) If $1 \in B$, then each $n = 1 + 1 + \cdots + 1 \in B$. If $1 \notin B$, then by (a), $1 = \alpha - \beta$ for $\alpha, \beta \in B$. Check $b = (\alpha + \beta)^2 + 1$ suffices; for if $n > (\alpha + \beta)^2$, then writing $n = q(\alpha + \beta) + r, 0 \leq r < \alpha + \beta$, $n = q(\alpha + \beta) + r(\alpha - \beta) = (q + r)\alpha + (q - r)\beta$, and in particular $n \in B$ since $q + r > 0$ and $q - r > 0$ by virtue of $n > (r + 1)(\alpha + \beta)$.]

(c) For each (i, j), there is an integer b_{ij} such that $\nu \geq b_{ij}$ implies $\nu \in B_{ij}$. [*Hint*: Obtain b_{jj} from (ii) applied to (i) and then choose k such that $p_{ij}^{(k)} > 0$. Check that $b_{ij} = k + b_{jj}$ suffices.

(d) Check that $\nu = \max\{b_{ij} : i, j \in S\}$ suffices for the statement of the exercise.

15. Complete the proof of part (b) of Theorem 19.7. [*Hint*: Follow similar argument as that for Proposition 19.6.
16. Provide a proof for Corollary 19.10.
17. Show that all finite dimensional distributions of the process $\{Y_n\}$, starting at x, say $\{Y_n(x)\}$, converge to those of the limiting Brownian motion as asserted in the proof of Theorem 19.11.
18. In the context of Theorem 19.11, show that convergence in distribution (to Brownian motion) of the process $\{Y_n\}$ corresponding to the sequence $\{X_j\}$, when X_0 has the distribution π, implies that the sequence $\{Y_n(x) : n = 0, 1, \dots\}$ is tight in the case $X_0 = x$.
19. Complete the proof of part (a) for Theorem 19.11 as indicated.
20. Let Γ be a finite set, say $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_k\}$, and assume the support of Q is Γ. Show that if $\cup_{j=1}^{k} \gamma_j(S)$ is a finite set $\{x_1, x_2, \dots, x_m\}$, then d defined by (19.3) is not a metric on $\mathcal{P}(S)$ if $\mathcal{P}(S)$ includes a μ for which $\mu(\{x_1, x_2, \dots, x_m\}) = 0$.

Chapter 20
Irreducibility and Harris Recurrent Markov Processes

Unlike discrete parameter Markov processes on a finite state space where notions of irreducibility and recurrence properties may be defined pointwise on the state space, for Markov processes on a general state space this may not be possible. Various notions of "neighborhood recurrence" become possible either through topological considerations or measure theoretic considerations relative to some reference measure. The notion of Harris recurrence considered in this chapter is of the latter type.

Suppose that $\{X_n\}_{n=0}^{\infty}$ is a Markov process taking values in a measurable state space $(S, \mathcal{S})$ and having transition probabilities

$$P(X_{n+1} \in A | X_0, X_1, \ldots, X_n) = p(X_n, A), \quad A \in \mathcal{S}.$$

When S is a finite or countably infinite set with power set $\mathcal{S}$ then *irreducibility* and *recurrence* refer to the properties that for each $x, y \in S$

$$p_{x,y}^{(n)} \equiv p^{(n)}(x, \{y\}) = P_x(X_n = y) > 0 \quad \text{for some } n \geq 1, \tag{20.1}$$

and

$$P_x(X_n = y \quad \text{for some } \ n \geq 1) = 1, \tag{20.2}$$

respectively. Such "point–recurrence" is illustrated by the simple symmetric random walk on the integers. However, considerations of a general random walk on $S = \mathbb{R}$, say, having continuously distributed increments makes it clear that this is not always

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_20

an appropriate way to describe *regenerative structures*. As it would be for these examples, without any significant loss of generality we will assume throughout this chapter that S is Borel subset of a Polish space without explicitly repeating it.

A quite natural generalization[1] is given as follows.

Definition 20.1. Let φ be a nonzero σ-finite measure on $(S, \mathcal{S})$. A Markov process $\{X_n\}_{n=0}^{\infty}$ is said to be *φ-irreducible* if for each $x \in S$, $A \in \mathcal{S}$ with $\varphi(A) > 0$,

$$p^{(n)}(x, A) = P(X_n \in A | X_0 = x) > 0 \quad \text{for some } n = n(x) \geq 1.$$

If $\{X_n\}_{n=0}^{\infty}$ is φ-irreducible and for each $x \in S$, $A \in \mathcal{S}$ with $\varphi(A) > 0$, it is also true that

$$P(X_n \in A \quad \text{for some } \ n \geq 1 | X_0 = x) = 1,$$

then $\{X_n\}_{n=0}^{\infty}$ is said to be *φ-recurrent,* or *Harris recurrent.*

In the case when S is finite or countably infinite, the notions of φ-irreducibility/recurrence coincide with the pointwise notions with *counting measure* $\varphi(A) =$ cardinality of $A \subset S$.

Another form of irreducibility and recurrence is given by the following definition.

Definition 20.2. Suppose that $\{X_n\}_{n=0}^{\infty}$ is a Markov process with values in a measurable state space $(S, \mathcal{S})$ having transition probabilities $p(x, dy)$. If there is a set $A_0 \in \mathcal{S}$ such that

$$P(X_n \in A_0 \text{ for some } n \geq 1 | X_0 = x) = 1 \quad \text{for all } x \in S,$$

then A_0 is said to be a *recurrent set,* and $\{X_n : n = 0, 1, 2, \ldots\}$ is said to be *A_0−recurrent.*

Definition 20.3. If there exists a set $A_0 \in \mathcal{S}$, a probability ν on $(S, \mathcal{S})$ with $\nu(A_0) = 1$, a positive integer N, and a positive number λ such that

$$p^{(N)}(x, A) \geq \lambda \cdot \nu(A), \quad \text{for all } A \in \mathcal{S}, x \in A_0, \tag{20.3}$$

then we say $\{X_n\}_{n=0}^{\infty}$ is *locally minorized*, and the set A_0 is called a *small set* (with respect to ν).

Notice that in the case $A_0 = S$ Definition 20.3 is simply Doeblin's minorization (See Corollary 19.2).

[1] The basic ideas in this generality are due to Harris (1956).

Example 1. Consider the Markov process[2] $X = \{X_n : n \geq 0\}$ on $[0, \infty)$ defined by iterated maps of the form $\gamma_{u,t}(x) = ux + t, x \geq 0$, where $u \in [0, 1]$ and $t \geq 0$. Specifically, suppose that $(U_n, T_n), n = 1, 2, \ldots,$ is an i.i.d. sequence independent of X_0, for which U_n is uniform on $[0, 1]$ and T_n is, independently of U_n, exponentially distributed with mean $1/2$. Then $X_n = \gamma_n \cdots \gamma_1(X_0), n \geq 1$, i.e., $X_{n+1} = U_{n+1}X_n + T_{n+1}, n \geq 0$. One may check that X has an absolutely continuous transition probability with continuous density $p(x, y) = 2e^{-2y}\frac{e^{2(x\wedge y)}-1}{2x}, x, y \geq 0$. Moreover the probability π, with pdf $\pi(x) = 4xe^{-2x}, x \geq 0$, is easily checked to be a time-reversible invariant probability (Exercise 5). For sufficiently large $k > 1, 0 \leq \delta < 1$, the interval $A_0 \equiv A_0(k, \delta) = [\delta k, k]$ is a small set with respect to the probability measure $\pi_{k,\delta}(dy) = \{(2\delta k + 1)e^{-2\delta k} - (2k + 1)e^{-2k}\}^{-1}4ye^{-2y}\mathbf{1}_{[k\delta,k]}(y)dy$ with $N = 1$ since, for $x \in A_0$, $A \in \mathcal{B}[0, \infty)$, one has upon splitting the integral over $[k\delta, k] = [k\delta, x] \cup (x, k]$ for $k\delta < x < k$,

$$\begin{aligned} p(x, A) &\geq \int_{A\cap[\delta k,k]} p(x, y)dy \\ &\geq \frac{1}{2k}\int_{A\cap[k\delta,k]} 4ye^{-2y}dy \\ &= \lambda_{k,\delta}\pi_{k,\delta}(A), \end{aligned} \tag{20.4}$$

where $\pi_{k,\delta}(A) = \frac{\pi(A\cap A_0)}{\pi(A_0)}$, and

$$\begin{aligned} \lambda_{k,\delta} &= \frac{1}{2k}\pi(A_0) \\ &= \left\{\frac{2k\delta + 1}{2k}e^{-2\delta k} - \frac{2k + 1}{2k}e^{-2k}\right\} \\ &\leq \frac{2k + 1}{2k}\left(e^{-2\delta k} - e^{-k}\right) < 1, \end{aligned} \tag{20.5}$$

for k sufficiently large such that $e^{-2\delta k} < \frac{2k}{2k+1}$. Also see Exercise 5(c).

Remark 20.1. Clearly, an A_0-recurrent locally minorized Markov process is ν-irreducible and Harris recurrent. It may be shown[3] that the converse is also true if $\mathcal{S}$ is countably generated, e.g., if S is a Borel subset of a Polish space with $\mathcal{S} = \mathcal{B}(S)$. Indeed, if the Markov process on S is φ-irreducible, then for every B with $\varphi(B) > 0$ there is an $A_0 \subset B$ such that $\varphi(A_0) > 0$ and A_0 is a small set.[4]

[2] This example arises naturally in the analysis of the Le Jan-Sznitman cascade associated with the Navier–Stokes equations in the case of the Bessel majorizing kernel; see Bhattacharya and Waymire (2021),Chapter 28, and Dascaliuc et al. (2022a,b).

[3] See Orey (1971).

[4] See Meyn and Tweedie (1993) page 109. Also see Orey (1971), Chapter 6.

Let us see how to use the local minorization and recurrence of A_0 by observing the motion at times when it reaches A_0. We first consider the case $N = 1$. Specifically, starting with $X_0 \in A_0$, let

$$\tau^{(0)} \equiv \tau_{A_0}^{(0)} = 0, \quad \tau^{(j)} \equiv \tau_{A_0}^{(j)} = \inf\left\{n > \tau_{A_0}^{(j-1)} : X_n \in A_0\right\}, \quad j = 1, 2, \ldots, \tag{20.6}$$

denote the successive times at which the process $\{X_n\}_{n=0}^{\infty}$ visits A_0. Then, by the strong Markov property the process $X_{\tau^{(0)}} = x \in A_0, X_{\tau^{(1)}}, X_{\tau^{(2)}}, \ldots$, viewed only on its visits to A_0 is a Markov process with state space A_0 having transition probabilities (Exercise 12) given, for $x \in A_0$, $B \in A_0 \cap \mathcal{S}$, by

$$\begin{aligned} p_{A_0}(x, B) &= P(X_{\tau_{A_0}^{(1)}} \in B | X_0 = x) \\ &= \sum_{n=1}^{\infty} P\left(X_n \in B, \tau_{A_0}^{(1)} = n | X_0 = x\right). \end{aligned} \tag{20.7}$$

First notice that if $\{X_n\}_{n=0}^{\infty}$ is A_0-recurrent and locally minorized on A_0 with $N = 1$, say, then

$$p_{A_0}(x, B) \geq P(X_1 \in B | X_0 = x) \geq \lambda \nu(B) \quad \text{for all } B \in A_0 \cap \mathcal{S}, x \in A_0. \tag{20.8}$$

That is, the process viewed on its visits to A_0 satisfies Doeblin's minorization and therefore there is a unique invariant probability measure π_0 on $(A_0, A_0 \cap \mathcal{S})$ such that (see Corollary 19.2)

$$\sup_{\substack{x \in A_0 \\ B \in A_0 \cap \mathcal{S}}} \left| p_{A_0}^{(n)}(x, B) - \pi_0(B) \right| \leq (1 - \lambda)^n. \tag{20.9}$$

Similar considerations apply for $N > 1$ by observing the process on a time scale of every N steps and as it visits A_0. In the case $N = 1$ the locally minorized Harris recurrent process is called *strongly aperiodic*.

Consider the extension of $p_{A_0}(x, \cdot)$ in (20.7) to a nonnegative set function on $(S, \mathcal{S})$ by defining, for each $x \in S$, $B \in \mathcal{S}$, $(B \neq \emptyset)$,

$$\begin{aligned} p_{A_0}(x, B) &:= \sum_{n=1}^{\infty} P(X_n \in B, X_j \in A_0^c, 1 \leq j < n | X_0 = x) \\ &= \sum_{n=1}^{\infty} {}_{A_0}p^{(n)}(x, B), \end{aligned} \tag{20.10}$$

where ${}_{A_0}p^{(n)}(x, B) := P(X_n \in B, X_j \in A_0^c, 1 \leq j < n | X_0 = x)$. Note that $p_{A_0}(x, B)$ is the expected number of visits to B by $\{X_n : n \geq 0\}$ during the cycle $[1, \tau_{A_0}](= [1, \infty)$ if $\tau_{A_0} = \infty)$, that is, it equals $\mathbb{E}(N(B)|X_0 = x)$, where

$$N(B) := \sum_{n=1}^{\infty} \mathbf{1}_{[X_n \in B, X_j \in A_0^c \text{ for } 1 \le j < n]}. \tag{20.11}$$

One might guess that averaging $p(x, \cdot)$ over A_0 with respect to an invariant probability π_{A_0}, say, for the process viewed on its returns to A_0 might produce an invariant measure for the process. That this is the case and, moreover, that it is a unique (up to constant multiples) σ-finite invariant measure is the objective from here.

The following lemma estimates $\mathbb{E}N(B)$ for a particular class of sets, without the requirements of irreducibility or recurrence.

Lemma 1. Define $B(m, \delta) := \{x \in A_0^c : p^{(m)}(x, A_0) > \delta\}, m \ge 1, \delta > 0$. Then for $x \in A_0$,

$$p_{A_0}(x, B(m, \delta)) \le m\left(1 + \frac{1}{\delta}\right) \quad (m \ge 1, \delta > 0). \tag{20.12}$$

Proof. First consider $B \equiv B(1, \delta), x \in A_0$. Let $\tau^{(k)}$ denote the time of the k-th visit to $B = B(1, \delta)$ before returning to A_0. Then

$$p_{A_0}(x, B) = \mathbb{E}_x N(B) \le 1 + \sum_{k=1}^{\infty} P_x(N(B) \ge k+1) = 1 + \sum_{k=1}^{\infty} P_x\left(\tau^{(k+1)} < \infty\right). \tag{20.13}$$

Note that $[N(B) \ge k + 1] = [\tau^{(k+1)} < \infty], k \ge 0$, and, noting $B \subset A_0^c$, so that $[\tau^{(k+1)} < \infty] \subset [\tau^{(k)} < \infty, X_{\tau^{(k)}+1} \in A_0^c]$. Therefore, by the strong Markov property with stopping time $\tau^{(k)}$, one has

$$\begin{aligned} P_x(\tau^{(k+1)} < \infty) &\le P_x(\tau^{(k)} < \infty, X_{\tau^{(k)}+1} \in A_0^c | \mathcal{F}_{\tau^{(k)}}) \\ &\le P_x(\tau^{(k)} < \infty)(1 - \delta) \le (1 - \delta)^k P_x(\tau^{(1)} < \infty) \\ &\le (1 - \delta)^k. \end{aligned} \tag{20.14}$$

Hence,

$$\mathbb{E}_x N(B(1, \delta)) \le 1 + \sum_{k=1}^{\infty} (1 - \delta)^{k-1} = 1 + \frac{1}{\delta}. \tag{20.15}$$

Next let $m > 1$ and apply the same argument to the m-step time scale process $X_{mj}, j = 0, 1, 2, \ldots$.

Harris recurrent Markov processes admit a nontrivial σ-finite invariant measure which is unique up to scalar multiples, though not necessarily a probability. We prove this under the alternative characterization of Harris recurrence indicated in Remark 20.1.

Before embarking on the main tasks of this chapter, the following simple but useful result may be noted.

Lemma 2. Let φ be a nonzero σ-finite measure. Consider a φ-irreducible and φ-recurrent Markov process with transition probability $p(x, dy)$. If π is a nonzero σ-finite invariant measure for $p(x, dy)$, then φ is absolutely continuous with respect to π.

Proof. If $\pi(B) = 0$, then $\int_S p^{(n)}(x, B)\pi(dx) = 0$. Thus, for each $n \geq 1$, $p^{(n)}(x, B) = 0$ π-a.s. This implies that B is not a φ-recurrent set, and therefore $\varphi(B) = 0$.

Theorem 20.1. Assume that there exists a recurrent set $A_0 \in \mathcal{S}$ which is also a small set. Then the following hold: (a) The transition probability $p_{A_0}(x, \cdot)$ on $(A_0, A_0 \cap \mathcal{S})$ admits a unique invariant probability π_{A_0} and

$$\sup_{x \in A_0, B \in A_0 \cap \mathcal{S}} |p_{A_0}^{(Nk)}(x, B) - \pi_{A_0}(B)| \leq (1 - \lambda)^k, \quad k = 1, 2, \ldots, \tag{20.16}$$

where $N \geq 1$ and $\lambda > 0$ are as in Definition 20.3. (b) The measure π defined by

$$\pi(B) := \int_{A_0} p_{A_0}(x, B)\pi_{A_0}(dx), \quad B \in \mathcal{S} \tag{20.17}$$

is σ-finite and it is, up to scalar multiples, the unique σ-finite invariant measure for $p(x, dy)$.

Proof. We first derive two general identities for every nonempty $A \in \mathcal{S}$. Note that the quantities ${}_A p^{(n+1)}(x, dy)$ satisfy (see (20.10)), upon conditioning with respect to $\mathcal{F}_n = \sigma\{X_0, X_1, \ldots X_n\}$,

$$\begin{aligned} {}_A p^{(n+1)}(x, B) &\equiv P(X_{n+1} \in B, X_j \in A^c \text{ for } 1 \leq j \leq n | X_0 = x) \\ &= \mathbb{E}(p(X_n, B)\mathbf{1}[X_n \in A^c]\mathbf{1}[X_j \in A^c, 1 \leq j \leq n-1] | X_0 = x) \\ &= \int_{A^c} p(y, B) {}_A p^{(n)}(x, dy), \quad n \geq 1. \end{aligned} \tag{20.18}$$

From this one gets (see (20.10))

$$p_A(x, B) = p(x, B) + \int_{A^c} p(y, B) p_A(x, dy), \quad x \in S, \ B \in \mathcal{S}. \tag{20.19}$$

Hence, under the given assumption for $x \in A_0$, applying (20.14) to the N-step transition probability, one has

$$p_{A_0}(x, B) \geq p^{(N)}(x, B) \geq \lambda\nu(B), \quad B \in A_0 \cap \mathcal{S}, \tag{20.20}$$

where $N, \lambda > 0$ and the probability measure ν are as in Definition 20.3. Part (a) now follows from the Doeblin minorization theorem (Corollary 19.2). To prove (b), observe that due to recurrence of A_0, one has $S = \cup_{m=1}^{\infty} \cup_{n=1}^{\infty} B(m, \frac{1}{n})$, where $B(m, \frac{1}{n}) = \{x \in S : p^{(m)}(x, A_0) > \frac{1}{n}\}$. By Lemma 1, $\pi(B(m, \frac{1}{n})) \leq m(1+n)$. Hence π is σ-finite. The invariance of π follows from (20.19) and the fact that $\pi = \pi_{A_0}$ on $(A_0, A_0 \cap \mathcal{S})$. Namely,

$$\begin{aligned}\int_S p(y, B)\pi(dy) &= \int_{A_0} p(y, B)\pi_{A_0}(dy) + \int_{A_0^c} p(y, B) \int_{A_0} p_{A_0}(x, dy)\pi_{A_0}(dx)\\ &= \int_{A_0} [p(x, B) + \int_{A_0^c} p(y, B) p_{A_0}(x, dy)]\pi_{A_0}(dx)\\ &= \int_{A_0} p_{A_0}(y, B)\pi_{A_0}(dy) = \pi(B). \qquad (20.21)\end{aligned}$$

For the proof of uniqueness, let π' be a nonzero σ-finite invariant measure. Let us first show that *the restriction of π' to A_0 is invariant for the process on A_0, i.e., with respect to $p_{A_0}(x, dy)$*. For this we derive the following sequence of identities (see (20.10) for notation)

$$\pi'(C) = \sum_{k=1}^{n} \int_{A_0} {}_{A_0}p^{(k)}(x, C)\pi'(dx) + \int_{S\backslash A_0} {}_{A_0}p^{(n)}(x, C)\pi'(dx),\ C \in \mathcal{S}, n = 1, 2, \ldots. \qquad (20.22)$$

For $n = 1, {}_{A_0}p^{(n)}(x, C) \equiv p(x, C)$ so that (20.22) is just the invariance of π'. As induction hypothesis, assume (20.22) holds for some n. Using the invariance of π', and (20.18), one obtains

$$\begin{aligned}\int_{S\backslash A_0} {}_{A_0}p^{(n)}(x, C)\pi'(dx) &= \int_S \int_{S\backslash A_0} {}_{A_0}p^{(n)}(x, C)p(y, dx)\pi'(dy)\\ &= \int_{A_0} \int_{S\backslash A_0} {}_{A_0}p^{(n)}(x, C)p(y, dx)\pi'(dy)\\ &\quad + \int_{S\backslash A_0} \int_{S\backslash A_0} {}_{A_0}p^{(n)}(x, C)p(y, dx)\pi'(dy)\\ &= \int_{A_0} {}_{A_0}p^{(n+1)}(y, C)\pi'(dy)\\ &\quad + \int_{S\backslash A_0} {}_{A_0}p^{(n+1)}(y, C)\pi'(dy). \qquad (20.23)\end{aligned}$$

Replace the second integral in (20.22) by this to derive the desired relation for $n+1$, completing the induction argument. Letting $n \to \infty$ in (20.22), using (20.10), and using the fact that the second term in (20.22) goes to zero as $n \to \infty$, because of recurrence of A_0, we have

$$\pi'(C) = \int_{A_0} p_{A_0}(x, C)\pi'(dx), \quad C \in \mathcal{S}. \tag{20.24}$$

In particular,

$$\pi'(C) = \int_{A_0} p_{A_0}(x, C)\pi'(dx), \quad C \in A_0 \cap \mathcal{S}. \tag{20.25}$$

This proves the invariance (with respect to p_{A_0}) of the restriction of π' to A_0. Moreover, this makes $\frac{\pi'(\cdot)}{\pi'(A_0)}$ an invariant *probability* on A_0, so that by uniqueness of the invariant probability π on $A_0 \cap \mathcal{S}$, one has $\pi'(\cdot) = \pi'(A_0)\pi(\cdot)$ on $A_0 \cap \mathcal{S}$. From (20.25), and with $\pi' = \pi'(A_0)\pi$, the uniqueness of π' up to a constant multiple ($\pi'(A_0)$) follows.

As an important consequence we note some conditions under which the invariant measure π obtained under Harris recurrence is normalizable to an invariant probability.

Corollary 20.2. Assume that there exists a recurrent set A_0 which is also a small set. If $\sup_{x \in A_0} \mathbb{E}_x \tau_{A_0} < \infty$, then the Markov process $\{X_n : n \geq 0\}$ with transition probability $p(x, dy)$ has a unique invariant probability π.

The corollary follows using (20.10), (20.11), and the statement following them.

Due to the significance of this last result, we present another proof, which will further show that $\frac{1}{n} \sum_{m=1}^{n} p^{(m)}(x, dy)$ converges to $\pi(dy)$ in the total variation metric as $n \to \infty$, for every initial state x, under the assumption of Corollary 20.2.

Proposition 20.1 *(Regeneration Lemma).* Let $\{X_n\}_{n=0}^{\infty}$ be a Markov process on $(S, \mathcal{S})$. Assume that the process is A_0–recurrent and locally minorized on A_0 with $N = 1$. *(a)*: $\{X_n\}_{n=0}^{\infty}$ has a representation by i.i.d. random cycles between regeneration times $\rho^{(1)}, \rho^{(2)}, \ldots$, namely

$$U_j := \left(X_{\rho^{(j)}+1}, \ldots, X_{\rho^{(j+1)}}, \rho^{(j+1)} - \rho^{(j)}\right), \quad j = 0, 1, 2, \ldots \tag{20.26}$$

are independent for $j \geq 0$ and identically distributed for $j \geq 1$. *(b)*: If, in addition, $c := \sup_{x \in A_0} \mathbb{E}(\tau^{(1)} | X_0 = x) < \infty$, then $\mathbb{E}(\rho^{(j+1)} - \rho^{(j)}) \leq c/\lambda$.

Proof. The idea of the proof is simple. When the current state is x then in the next step it moves to y following the transition probability $p(x, dy)$. For $x \in A_0$, the transition probability has the representation on $A_0 \times \mathcal{S}$, with $0 < \lambda < 1$ (see Definition 20.3), $x \in A_0$, given by

$$p(x, dy) = \lambda\nu(dy) + (1 - \lambda)q(x, dy), \; q(x, dy) = (1 - \lambda)^{-1}\{p(x, dy) - \lambda\nu(dy)\}. \tag{20.27}$$

That is, when $x \in A_0$, in the next step with probability λ it has distribution $\nu(dy)$, and with probability $1 - \lambda$ it has distribution $q(x, dy)$. So one way to construct a Markov process with transition probabilities $p(x, dy)$ starting at any initial state x_0 is to first let $\theta_n, n \geq 0$, be an iid sequence of Bernoulli random variables with $P(\theta_n = 1) = \lambda$, $P(\theta_n = 0) = 1 - \lambda$. Then the process moves according to $p(x, dy)$ until $X_n \in A_0$, in which case X_{n+1} has distribution $\nu(dy)$ if $\theta_n = 1$, and has distribution $q(X_n, dy)$ if $\theta_n = 0$.Let $\tau^{(0)} = 0$, $\tau^{(1)} = \tau_{A_0} = \inf\{n \geq 1 : X_n \in A_0\}$, $\tau^{(j+1)} = \inf\{n > \tau^{(j)} : X_n \in A_0\}, n = 0, 1, 2 \dots.$ By the recurrence of A_0, $\tau^{(j)} < \infty$ a.s. for all j. The process $(X, \theta) = \{(X_n, \theta_n) : n = 0, 1, 2, \dots\}$ is a Markov process on $S \times \{0, 1\}$, with the obvious product σ-field $\tilde{\mathcal{S}} = \{B \times \{0\} : B \in \mathcal{S}\} \cup \{B \times \{1\} : B \in \mathcal{S}\}$ (Exercise 6). For this process define the stopping times

$$\rho^{(0)} = 0, \rho^{(j)} = \inf\{n > \rho^{(j-1)} : X_n \in A_0, \theta_n = 1\}, \quad j = 1, 2, \dots. \tag{20.28}$$

It is simple to check that $\rho^{(j)} < \infty$ a.s. for all j (Exercise 6). By the strong Markov property, at time $\rho^{(j)}$, the process (X, θ) starts afresh with initial distribution $\nu \times \delta_{\{1\}}$, $X_{\rho^{(j)}}$ has distribution ν and $\theta_{\rho^{(j)}} = 1$, independently of the past, i.e., the pre-$\rho^{(j)}$sigma-field $\mathcal{G}_{\rho^{(j)}}$, say, for $j = 1, 2, \dots$. This proves part (a).

To prove part (b) observe that for $j \geq 1$, $\mathbb{E}(\rho^{(j+1)} - \rho^{(j)}) = \mathbb{E}(\rho^{(1)} - \rho^{(0)} | X_0 \in A_0, \theta_0 = 1) = \mathbb{E}(\rho^{(1)} | X_0 \in A_0, \theta_0 = 1)$, since in this case $\rho^{(0)} = 0$ and

$$\rho^{(1)} - \rho^{(0)} = \rho^{(1)} = \sum_{j=1}^{\infty} \tau^{(j)} \mathbf{1}_{[\theta_{\tau^{(i)}}=0 \text{ for } 1 \leq i < j,\ \theta_{\tau^{(j)}}=1]}, \tag{20.29}$$

the first term on the right being $\tau^{(1)} \mathbf{1}_{[\theta_{\tau^{(1)}}=1]}$. Taking expectations and noting that (i) $\theta_{\tau^{(j)}}$ is independent of $\tilde{\mathcal{F}}_{\tau^{(j-1)}} := \sigma\{(X_n, \theta_n) : n \leq \tau^{(j-1)}\}$, and of $\tau^{(j)}$, and (ii) $\mathbb{E}(\tau^{(j)} - \tau^{(j-1)} | \tilde{\mathcal{F}}_{\tau^{(j-1)}}) \leq c := \sup_{x \in A_0} \mathbb{E}(\tau^{(1)} | X_0 = x)$, one has

$$\begin{aligned}
\mathbb{E}\rho^{(1)} &= \sum_{j=1}^{\infty} \mathbb{E}\left(\mathbf{1}_{[\theta_{\tau^{(i)}}=0, 1 \leq i \leq j-1]} \tau^{(j)} \mathbb{E}(\mathbf{1}_{[\theta_{\tau^{(j)}}=1]})\right) \\
&= \lambda \sum_{j=1}^{\infty} \mathbb{E}\mathbb{E}\left(\mathbf{1}_{[\theta_{\tau^{(i)}}=0, 1 \leq i \leq j-1]} \{\tau^{(j-1)} + \mathbb{E}(\tau^{(j)} - \tau^{(j-1)} | \tilde{\mathcal{F}}_{\tau^{(j-1)}}\right) \\
&\leq \lambda \sum_{j=1}^{\infty} \mathbb{E}\left(\mathbf{1}_{[\theta_{\tau^{(i)}}=0, 1 \leq i \leq j-1]} \{\tau^{(j-1)} + c\}\right) \\
&= \lambda \sum_{j=1}^{\infty} c(1-\lambda)^{j-1} + \lambda \sum_{j=1}^{\infty} \mathbb{E}\left(\mathbf{1}_{[\theta_{\tau^{(i)}}=0, 1 \leq i \leq j-1]} \tau^{(j-1)}\right) \\
&= c + \lambda \sum_{j=1}^{\infty} \mathbb{E}\left(\mathbf{1}_{[\theta_{\tau^{(i)}}=0, 1 \leq i \leq j-2]} \{\tau^{(j-2)} + \mathbb{E}\{\mathbf{1}_{[\theta_{\tau^{(j-1)}}=0]} (\tau^{(j-1)} - \tau^{(j-2)} | \tilde{\mathcal{F}}_{\tau^{(j-2)}})\}\right) \\
&= c + c\lambda \sum_{j=2}^{\infty} (1-\lambda)^{j-1} + \lambda(1-\lambda) \sum_{j=3}^{\infty} \mathbb{E}\left(\mathbf{1}_{[\theta_{\tau^{(i)}}=0,\ 1 \leq i \leq j-2]} \tau^{(j-2)}\right)
\end{aligned}$$

$$\begin{aligned}&\le \cdots \le c + c(1-\lambda) + c(1-\lambda)^2 + \cdots\\ &= c\sum_{j=1}^{\infty}(1-\lambda)^{j-1} = c/\lambda.\end{aligned}$$

This completes the proof of part (b).

Theorem 20.4 *(Ergodicity of Strongly Aperiodic Harris Recurrent Processes).* Let $p(x, dy)$ be a transition probability on $(S, \mathcal{S})$. Assume that a Markov process $\{X_n : n \ge 0\}$ with this transition probability is A_0-recurrent and locally minorized, according to Definition 20.3, and that the process is strongly aperiodic, i.e., $N = 1$. If, in addition,

$$\sup_{x\in A_0} \mathbb{E}_x \tau^{(1)} < \infty,$$

then there exists a unique invariant probability π and, whatever the initial distribution of $\{X_n : n \ge 0\}$,

$$\sup_{A\in\mathcal{S}^{\otimes\infty}} \left| Q^{(n)}(A) - Q_\pi(A) \right| \longrightarrow 0 \quad \text{as } n \to \infty, \tag{20.30}$$

where $Q^{(n)}$ is the distribution of the after-n process $X_n^+ := \{X_{n+m} : m \ge 0\}$, and Q_π is the distribution of the process with initial distribution π. In particular,

$$\sup\{|p^{(n)}(x, B) - \pi(B)| : B \in \mathcal{S}\} \to 0 \ \text{ for all } x \in S. \tag{20.31}$$

Proof. The proof of the existence and uniqueness of an invariant probability follows from Corollary 20.2. The proof of (20.30) follows by a coupling argument. We will first show that the measure π_1 defined by

$$\pi_1(B) := \mathbb{E}_\nu \sum_{n=\rho^{(0)}+1}^{\rho^{(1)}} \mathbf{1}_{[X_n\in B]} \qquad (B \in \mathcal{S}) \tag{20.32}$$

is invariant under $p(x, dy)$. Here $\mathbb{E}_\nu$ denotes expectation of events in $\sigma\{X_n : n \ge 0\}$ when X_0 has distribution ν. Note that the right side in (20.32) also equals the expectation of the sum over $\rho^{(j)} + 1 \le n \le \rho^{(j+1)}$, whatever be the distribution of X_0. Now for nonnegative bounded measurable f on $(S, \mathcal{S})$, define

$$V_j(f) := \sum_{n=\rho^{(j-1)}+1}^{\rho^{(j)}} f(X_n). \tag{20.33}$$

By Proposition 20.1, $V_j(f)$, $j \geq 1$, are i.i.d. with expected value $\int_S f d\pi_1$. By the strong law of large numbers, $N^{-1}\sum_{j=1}^{N} V_j(f) \to \int_S f d\pi_1$ a.s. as $N \to \infty$. In particular, with $f = 1$, a.s. as $N \to \infty$ one has

$$\frac{\rho^{(N)} - \rho^{(0)}}{N} \to \mathbb{E}(\rho^{(1)} - \rho^{(0)}) = \pi_1(S), \qquad \frac{\rho^{(N)}}{N} \to \pi_1(S). \tag{20.34}$$

For each $n = 1, 2, \ldots,$ define

$$N_n := \sup\{j \geq 1 : \rho^{(j)} \leq n\}. \tag{20.35}$$

Then by (20.34) one has a.s. as $n \to \infty$,

$$\frac{\rho^{(N_n)}}{N_n} \to \pi_1(S). \tag{20.36}$$

But $0 \leq n - \rho^{(N_n)} \leq \rho^{(N_n+1)} - \rho^{(N_n)}$, and $\frac{\rho^{(N_n+1)} - \rho^{(N_n)}}{N_n} \to 0$ a.s. since $\frac{\rho^{(k+1)} - \rho^{(k)}}{k} = \frac{\rho^{(k+1)}}{k+1}\frac{k+1}{k} - \frac{\rho^{(k)}}{k} \to 0$ a.s. as $k \to \infty$. Hence (20.36) implies a.s. as $n \to \infty$,

$$\frac{n}{N_n} \to \pi_1(S). \tag{20.37}$$

In the same manner for all $n > \rho^{(1)}$, if f is bounded and measurable, then a.s. as $n \to \infty$,

$$\begin{aligned}
\frac{1}{n}\sum_{m=1}^{n} f(X_m) &= \frac{1}{n}\sum_{m=1}^{\rho^{(1)}} f(X_m) + \frac{1}{n}\sum_{j=1}^{N_n-1} V_j(f) + \frac{1}{n}\sum_{m=N_n}^{n} f(X_m) \\
&= \frac{1}{N_n - 1}\frac{N_n - 1}{n}\sum_{j=1}^{N_n-1} V_j(f) + \frac{1}{n}\sum_{m=1}^{\rho^{(1)}} f(X_m) + \frac{1}{n}\sum_{m=N_n}^{n} f(X_m) \\
&\to \frac{1}{\pi_1(S)}\int_S f d\pi_1.
\end{aligned} \tag{20.38}$$

Hence $\pi(B) := \pi_1(B)/\pi_1(S)$ is the unique invariant probability for $\{X_n : n \geq 0\}$, i.e., for $p(x, dy)$.

We will complete the proof by a coupling argument. One may construct a (common) probability space $(\Omega', \mathcal{F}', P')$, say, on which are defined two independent families $\{(X_n, \theta_n) : n \geq 0\}$ and $\{(\widetilde{X}_n, \widetilde{\theta}_n) : n \geq 0\}$ as above with $\{(\theta_n : n \geq 0\}$ and $\{\widetilde{\theta}_n : n \geq 0\}$ iid Bernoulli $0 - 1$ $(1 - \lambda, \lambda)$ and independent of X_0 and $\widetilde{X}_0$. Let X_0 have (an arbitrary initial) distribution μ and let $\widetilde{X}_0$ have the invariant distribution. Let $\{\rho^{(j)} : j \geq 0\}$ and $\{\widetilde{\rho}^{(j)} : j \geq 0\}$ denote the corresponding independent sequences of regeneration times, and let $\{Y_0 = \rho^{(0)}, Y_k = \rho^{(k)} - \rho^{(k-1)}\ (k \geq 1)\}$,

$\{\widetilde{Y}_0 = \widetilde{\rho}^{(0)}, \widetilde{Y}_k = \widetilde{\rho}^{(k)} - \widetilde{\rho}^{(k-1)}\ (k \geq 1)\}$ be the corresponding sequences of lifetimes of two independent renewal processes. Under the hypothesis, Y_k, or $\widetilde{Y}_k$ $(k \geq 1)$ has a lattice span 1, and $\mathbb{E}Y_k = \mathbb{E}\widetilde{Y}_k = \mathbb{E}(\rho^{(2)} - \rho^{(1)}) < \infty$. Hence, as in the coupling proof of the Renewal Theorem 8.5 in Bhattacharya and Waymire (2021), $\rho := \inf\{n \geq 0 : R_n = \widetilde{R}_n = 0\} < \infty$ a.s., where $\{R_n : n \geq 0\}$ and $\{\widetilde{R}_n : n \geq 0\}$ are the two residual lifetime processes corresponding to $\{Y_k : k \geq 0\}$ and $\{\widetilde{Y}_k : k \geq 0\}$, respectively. Note that ρ is the first common renewal epoch, i.e., there are $m, \widetilde{m}$ such that $\rho^{(m)} = \tilde{\rho}^{(\tilde{m})} = \rho$. Now define

$$X'_n := \begin{cases} \widetilde{X}_n & \text{if } \rho > n, \\ X_n & \text{if } \rho \leq n. \end{cases} \tag{20.39}$$

Since ρ is a stopping time for the Markov process $\{(W_n, \widetilde{W}_n) : n \geq 0\}$, with $W_n := (X_n, \theta_n)$, $\widetilde{W}_n = (\widetilde{X}_n, \widetilde{\theta}_n)$, one may use the strong Markov property to see that $\{X_n : n \geq 0\}$ and $\{X'_n : n \geq 0\}$ have the same distribution. Hence, for all $A \in \mathcal{S}^{\otimes\infty}$,

$$\begin{aligned} |P(X_n^+ \in A) - P(\widetilde{X}_n^+ \in A)| &\leq |P((X')_n^+ \in A, \rho \leq n) - P(\widetilde{X}_n^+ \in A, \rho \leq n)| \\ &\quad + |P((X')_n^+ \in A, \rho > n) - P(\widetilde{X}_n^+ \in A, \rho > n)| \\ &\leq P(\rho > n). \end{aligned} \tag{20.40}$$

Here we have used the fact $(X')_n^+ = X_n^+$ on $[\rho \leq n]$.

Remark 20.2. For the case $N > 1$, one first constructs $X_0, X_1, \ldots, X_{N-1}$ with transition probabilities $p(x, dy)$. Then apply the above proof to the N (N-skeleton) Markov processes given by $\{X_{Nk} : k \geq 0\}$, $\{X_{1+Nk} : k \geq 0\}, \ldots, \{X_{N-1+Nk} : k \geq 0\}$, each having transition probability $p^{(N)}(x, dy)$. Let $\pi_0, \ldots, \pi_{N-1}$, respectively, be the invariant probabilities to which these Markov processes converge in total variation distance. Then the Markov process $\{X_n : n \geq 0\}$ converge in total variation distance to $\pi = \frac{1}{N}\sum_{j=0}^{N-1} \pi_j$.

Remark 20.3. Theorem 20.4 is due to Athreya and Ney (1978), and Nummelin (1978). In the context of Harris recurrence, the bivariate regenerative model (X, θ) introduced here is often referred to as a *splitting model*.[5] An essentially identical construction that exploits local minorization, often referred to as *Nummelin splitting*,[6] provides an approach to results of the type given here based on the construction of a representation of the process having an *atom*; that is a set of states B for which transition probabilities from states $x \in B$ do not otherwise depend on x. For this purpose one defines a bivariate Markov process on $S^* = S \times \{0, 1\}$ as follows. Assume local minorization by a small set A_0 with respect to a probability measure ν on $(S, \mathcal{S})$, concentrated on A_0, i.e., for some $0 < \lambda < 1$, $p(x, A) \geq$

[5] Athreya and Ney (1978).

[6] Nummelin (1978).

$\lambda \mathbf{1}_{A_0}\nu(A), x \in S, A \in \mathcal{S}$. For $x \in S$, $A \in \mathcal{S}$, define $x^{(0)} = (x, 0)$, $x^{(1)} = (x, 1)$, $A^{(0)} = A \times \{0\}$, $A^{(1)} = A \times \{1\}$, $A^* = A \times \{0, 1\}$. $A^{(0)}$ and $A^{(1)}$ are considered to be (coded) copies of A. Let $\mathcal{S}^* = \sigma\{A^{(0)}, A^{(1)} : A \in \mathcal{S}\}$. Give $\mathcal{S}^*$ the product σ-field. If γ is a probability measure on $(S, \mathcal{S})$, then the *splitting* of γ is the probability γ^* on $(S^*, \mathcal{S}^*)$ defined by $\gamma^*(A^{(0)}) = (1-\lambda)\gamma(A \cap A_0) + \lambda\gamma(A \cap A_0^c)$, $\gamma^*(A^{(1)}) = \lambda\gamma(A \cap A_0)$. The transition probabilities of X^* are then defined as above by splitting transition probabilities in such a way that $X^* = (X, \theta) \in S^*$ inherits Harris recurrence from X. Moreover, X is a Markov process with the transition probabilities $p(x, A), x \in S, A \in \mathcal{S}$. The set $B = A_0^{(1)}$ is an atom of X^* since $p(x^{(1)}, \cdot)$ does not depend on $x^{(1)} \in B$. From here techniques from renewal theory developed for countable state Markov chains can be extended and applied.

We conclude this chapter by consideration of conditions for an exponential rate of convergence for the total variation convergence (20.31) provided in the preceding Ergodic Theorem 20.4, referred to as *geometric ergodicity*.

Definition 20.4 *(Geometric Ergodicity).* A Markov process on $(S, \mathcal{S})$ with invariant probability π is *geometrically ergodic* if

$$||p^{(n)}(x,\cdot) - \pi(\cdot)||_{TV} = \sup_{B \in \mathcal{S}} |p^{(n)}(x, B) - \pi(B)| \le C(x)r^n, \quad n = 1, 2, \ldots,$$

for some $r < 1$, where $C(x) < \infty$ for π-a.e. $x \in S$. The process is said to be *uniformly ergodic* in the case that C is a constant.

The obvious first question is to provide some condition under which local minorization of a Harris recurrent Markov process provides geometric ergodicity. A standard approach to this problem borrows a notion of *Lyapunov function* $V : S \to [1, \infty]$, with $V(x) < \infty$ for $x \in A_0$, to control the drift of the Markov process as follows.

Theorem 20.5 *(Foster–Tweedie Drift Condition).* Suppose that X is a φ-irreducible, aperiodic Markov process with invariant probability π on a state space $(S, \mathcal{S})$. Assume the local minorization condition $p(x, A) \ge \lambda\mathbf{1}_{A_0}(x)\nu(A), x \in S$, for some $0 < \lambda < 1$ and probability ν. If there is a Lyapunov function $V \ge 1$, with $V(x) < \infty$ for $x \in A_0$, such that for some $0 < \beta < 1$,

$$\mathbb{E}_x V(X_1) \le \beta V(x) + b\mathbf{1}_{A_0}(x), \qquad \text{for all } x \in S, \tag{20.41}$$

then X is geometrically ergodic.

We will prove an alternate version of this theorem[7] using coupling with the advantage of providing a computable rate of convergence, but under somewhat stronger conditions. The coupling is an independent coupling of two Markov

[7] See Rosenthal (2002), Roberts and Rosenthal (2004), for both this and its extension to the more general Foster–Tweedie Theorem 20.5.

processes X, X' that occurs when the two processes enter the small set A_0 and an independent coin toss is “head”, (an outcome with probability λ).

Theorem 20.6. Let $X = \{X_n : n = 0, 1, \dots\}$ be a Markov process on the state space $(S, \mathcal{S})$ starting at $X_0 = x \in S$ and having transition probabilities $p(x, dy)$ and invariant initial probability π. Assume that there is a small set $A_0 \subset S$, a Lyapunov function $V \geq 1$ satisfying the drift condition (20.41) with parameters $\beta < 1, b \in \mathbb{R}$, a probability measure ν on S, and $\lambda > 0$ such that local minorization (20.3) holds with parameters $N = 1, \lambda$. Assume also that

$$(i) \sup_{x \in A_0} V(x) < \infty, \quad \text{and } (ii)\ d = \inf_{x \in A_0^c} V(x) > \left[\frac{b}{1-\beta}\right] - 1. \tag{20.42}$$

Let $\alpha = \beta + \frac{b}{d+1} < 1$, and $B = \max\{1, \alpha^{-1}(1-\lambda) \sup_{(x,y) \in A_0 \times A_0} \overline{R}v\}$, where, $v(x, y) = \frac{V(x)+V(y)}{2}$, and for $(x, y) \in A_0 \times A_0$,

$$\overline{R}v(x, y) = (1-\lambda)^{-2} \int_S \int_S v(z, w)(p(x, dz) - \lambda\nu(dz))(p(y, dw) - \lambda\nu(dw)). \tag{20.43}$$

Then, for any $1 \leq j \leq n$,

$$||P_x(X_n \in \cdot) - \pi(\cdot)||_{TV} \leq (1-\lambda)^j + \alpha^n B^{j-1} \int_S v(x, y)\pi(dy). \tag{20.44}$$

In particular,

$$||P_x(X_n \in \cdot) - \pi(\cdot)||_{TV} \leq c(x)r^n, \tag{20.45}$$

for $r = \max\{(1-\lambda), B\alpha\}$, $c(x) = 2B^{-1} \int_S v(x, y)\pi(dy)$.

Proof. The proof is by a coupling that encompasses the type of “splitting”of probabilities involved in the use of local minorization. Next let us define the Markov processes (X, X') to be coupled. To start, fix arbitrary $x \in S$. Let $X_0 = x$ and $X'_0 = x'$ be a sampled value from the invariant distribution π. The process is defined recursively as follows: Let $\theta_n, n = 0, 1, 2, \dots$ be an i.i.d. sequence of Bernoulli $1, 0$ random variables with respective probabilities $\lambda, 1-\lambda$, independent of X_0, X'_0. If $x = x'$, then couple $X_1 = X'_1$ with value sampled from the distribution $p(x, dy)$. Suppose $x \neq x'$, but $(x, x') \in A_0 \times A_0$. If $\theta_0 = 1$, then sample X'_1 from ν, but if $\theta_0 = 0$, then sample X'_1 from $q(x', dy) = \frac{p(x',dy)-\lambda\nu(dy)}{1-\lambda}$. If $\theta_0 = 1$, then take $X_1 = X'_1$, but if $\theta_0 = 0$, then sample X_1 from $q(x, dy)$, independently from X'_1. Finally, if $x \neq x'$, $(x, x') \notin A_0 \times A_0$, then independently sample X_1 and X'_1 from $p(x, dy)$ and $p(x', dy)$, respectively. Now repeat this sampling beginning with X_1, X'_1 in place of x, x', and so on. One may readily check that X and X' are Markov processes starting from δ_x and π, respectively, and having the common transition probability $p(x, dy)$.

We first check that the technical condition (20.42) implies that the coupled processes satisfy a (bivariate) drift condition outside $A_0 \times A_0$ with Lyapunov function $v(x, x') = \frac{V(x)+V(x')}{2}$ for the small set $A_0 \times A_0$ and parameter α; by assumption, $v(x, x')$ is bounded on $A_0 \times A_0$. Specifically, if $(x, x') \notin A_0 \times A_0$, then either $x \in A_0^c$ or $x' \in A_0^c$, or both; thus, $v(x, x') \geq \frac{1+d}{2}$ and $\mathbb{E}_x V(X_1) + \mathbb{E}_{x'} V(X_1) \leq \beta V(x) + \beta V(x') + b$. Then,

$$\begin{aligned}
\mathbb{E}_{(x,x')} v(X_1, X_1') &= \frac{1}{2}[\mathbb{E}_x V(X_1) + \mathbb{E}_{x'} V(X_1')] \\
&= \frac{1}{2}[\mathbb{E}_x V(X_1) + \mathbb{E}_{x'} V(X_1)] \\
&\leq \frac{1}{2}[\beta V(x) + \beta V(x') + b] \\
&= \beta v(x, x') + \frac{b}{2} \\
&\leq \beta v(x, x') + \frac{b}{2} \frac{v(x, x')}{\frac{1+d}{2}} \\
&= (\beta + b/(1+d)) v(x, x') \\
&= \alpha v(x, x'), \quad (x, x') \notin A_0 \times A_0. \qquad (20.46)
\end{aligned}$$

In view of (ii) of (20.42), one has $\alpha = \beta + b/(1+d) < 1$ as required for a bivariate Lyapunov function.

Coupling occurs at time

$$T_c = \inf\{n : X_n = X_n'\} = \inf\{n : (X_n, X_n') \in A_0 \times A_0, \theta_n = 1\}, \qquad (20.47)$$

the almost sure finiteness of which will follow from the bound on $P(X_n \neq X_n')$ as follows. Let N_n denote the number of visits to $A_0 \times A_0$ by (X_k, X_k') in time $k \leq n$. Then, for any $1 \leq j \leq n$,

$$\begin{aligned}
P(X_n \neq X_n') &= P(X_n \neq X_n', N_{n-1} \geq j) + P(X_n \neq X_n', N_{n-1} < j) \\
&\leq (1-\lambda)^j + P(X_n \neq X_n', N_{n-1} < j), \qquad (20.48)
\end{aligned}$$

since the event $[X_n \neq X_n', N_{n-1} \geq j]$ implies that $[\theta_i = 0]$ for j distinct values of i corresponding to visits to $A_0 \times A_0$; else $X_n = X_n'$. To bound the second term define of (20.48)

$$M_n = \alpha^{-n} B^{-N_{n-1}} v(X_n, X_n') \mathbf{1}_{[X_n \neq X_n']}, \quad n = 0, 1, 2 \ldots, (N_{-1} = 0). \qquad (20.49)$$

We will first show that $M_n, n \geq 0$, is a supermartingale, and therefore has decreasing expected values with increasing n. This will be accomplished by considering the cases $[(X_n, X'_n) \notin A_0 \times A_0]$ and $[(X_n, X'_n) \in A_0 \times A_0]$, separately. Observe that for $(X_n, X'_n) \notin A_0 \times A_0$, $N_n = N_{n-1}$, so that, recalling that (X, X') is Markov and noting that $[X_{n+1} \neq X'_{n+1}] \subset [X_n \neq X'_n]$,

$$\begin{aligned}
&\mathbb{E}[M_{n+1}|\sigma((X_j, X'_j), 1 \leq j \leq n)] \\
&\quad = \alpha^{-n-1} B^{-N_{n-1}} \mathbb{E}[v(X_{n+1}, X'_{n+1}) \mathbf{1}_{[X_{n+1} \neq X'_{n+1}]} | \sigma(X_n, X'_n)] \\
&\quad \leq \alpha^{-n-1} B^{-N_{n-1}} \mathbf{1}_{[X_n \neq X'_n]} \mathbb{E}[v(X_{n+1}, X'_{n+1}) | \sigma(X_n, X'_n)] \\
&\quad = M_n \mathbb{E}[v(X_{n+1}, X'_{n+1}) | \sigma(X_n, X'_n)] / \alpha v(X_n, X'_n) \\
&\quad \leq M_n, \qquad (20.50)
\end{aligned}$$

where this final inequality uses the bivariate Lyapunov drift bound outside $A_0 \times A_0$. On the other hand, if $(X_n, X'_n) \in A_0 \times A_0$, then $N_n = N_{n-1} + 1$, so that

$$\begin{aligned}
&\mathbb{E}[M_{n+1}|\sigma((X_j, X'_j), 1 \leq j \leq n)] \\
&\quad = \alpha^{-n-1} B^{-N_{n-1}-1} \mathbb{E}[v(X_{n+1}, X'_{n+1}) \mathbf{1}_{[X_{n+1} \neq X'_{n+1}]} | \sigma(X_n, X'_n)] \\
&\quad \leq \alpha^{-n-1} B^{-N_{n-1}-1} \mathbb{E}[v(X_{n+1}, X'_{n+1}) \mathbf{1}_{[X_n \neq X'_n]} | \sigma(X_n, X'_n)] \\
&\quad = \alpha^{-n-1} B^{-N_{n-1}-1} (1-\lambda) \overline{R} v(X_n, X'_n) \mathbf{1}_{[X_n \neq X'_n]} \\
&\quad = M_n \alpha^{-1} B^{-1} (1-\lambda) \overline{R} v(X_n, X'_n) / v(X_n, X'_n) \leq M_n, \qquad (20.51)
\end{aligned}$$

by definition of B. Thus, M_n is a supermartingale. As a consequence, one has the following bound on the second term of (20.48), using $B \geq 1$ and the property $v \geq 1$,

$$\begin{aligned}
P(X_n \neq X'_n, N_{n-1} < j) &= P(X_n \neq X'_n, N_{n-1} \leq j-1) \\
&\leq P(X_n \neq X'_n, B^{-N_{n-1}} \geq B^{-(j-1)}) \\
&\leq B^{j-1} \mathbb{E}[\mathbf{1}_{[X_n \neq X'_n]} B^{-N_{n-1}}] \\
&\leq B^{j-1} \mathbb{E}[\mathbf{1}_{[X_n \neq X'_n]} B^{-N_{n-1}} v(X_n, X'_n)] \\
&= \alpha^n B^{j-1} \mathbb{E} M_n \\
&\leq \alpha^n B^{j-1} \mathbb{E} M_0 = \alpha^n B^{j-1} \mathbb{E} v(X_0, X'_0). \qquad (20.52)
\end{aligned}$$

The theorem now follows from (20.48) using the Borel–Cantelli lemma.

Remark 20.4. If the local minorization and the Foster–Tweedie drift condition hold for $p^{(N)}(x, dy)$ for some $N > 1$, rather than for $p(x, dy)$, then apply the above argument to the N-skeleton Markov processes $\{X_{Nk} : k \geq 0\}, \ldots, \{X_{N-1+Nk} : k \geq 0\}$, each with transition probability $p^{(N)}(x, dy)$ as described in Remark 20.2.

It follows that $\frac{1}{N}\sum_{j=0}^{N-1} p^{(Nk+j)}(x,\cdot)$ converges to π in the total variation norm as $k\to\infty$.

Remark 20.5. The assumption $\sup_{x\in A_0} V(x) < \infty$ is redundant since it can be deduced[8] from the other conditions of the theorem. Also the condition (ii) of (20.42) makes the choice of the bivariate Lyapunov function simpler, but is otherwise not necessary.[9]

Observe that it follows from the exponential bound in the above proof that, choosing $j = [\delta n]$ for sufficiently small $\delta > 0$, $\lim_{n\to\infty} P(X_n \neq X_n') = 0$ at an exponential rate. In particular, the coupling time is almost surely finite (i.e., successful coupling). Moreover, since the time τ_{A_0} to reach A_0 starting at $x \in S$ is smaller than the coupling time one obtains the following corollary.

Corollary 20.7. Let τ_{A_0} be the hitting time of the small set A_0. Then, under the conditions of Theorem 20.6, one has

$$\mathbb{E}_x s^{\tau_{A_0}} < \infty, \quad x \in S, \tag{20.53}$$

where $1 < s < \max\{(1-\lambda)^{-\delta}, \alpha B^{-\delta}\}$, for $0 < \delta < \frac{\ln\alpha}{\ln B}$.

Proof. Simply observe that $\tau_{A_0} \le T_c$, where T_c is the coupling time (20.47). So, for $s > 1$

$$\begin{aligned}
\mathbb{E}_x s^{\tau_{A_0}} &\le \mathbb{E}_x s^{T_c}\\
&\le s + \sum_{n=1}^{\infty} s^n P_x(T_c = n)\\
&= s + \sum_{n=1}^{\infty} s^n \{x(T_c > n-1) - P_x(T_c > n)\}\\
&\le 1 + s + (s-1)\sum_{n=1}^{\infty} s^n P_x(T_c > n)\\
&\le 1 + s + (s-1)\sum_{n=1}^{\infty} s^n c(x) r^n < \infty,
\end{aligned} \tag{20.54}$$

for $r = \max\{(1-\lambda), \alpha B\} < 1$, $c(x) = 2B^{-1}\int_S v(x,y)\pi(dy)$, $1 < s < \frac{1}{r}$.

[8] See Roberts and Rosenthal (2004) for proof.

[9] See Roberts and Rosenthal (2004) for a more general statement.

Remark 20.6. It is actually true[10] that under geometric ergodicity for every A such that $\pi(A) > 0$, there is an $r = r_A > 1$ such that $\mathbb{E}_\pi r^{\tau_A} < \infty$.

We now turn to an important consequence of Theorem 20.6.

Definition 20.5. A real-valued sequence $\{X_n : n \geq 0\}$ is a nonlinear autoregressive process of order $k > 1$, if it is of the form

$$X_{n+1} = h(X_{n+1-k}, \ldots, X_{n-1}, X_n) + \eta_{n+1} \quad n \geq 0, \tag{20.55}$$

where h is a real-valued measurable function on $\mathbb{R}^k$ and $\{\eta_n : n \geq 0\}$ is an i.i.d. sequence.

Theorem 20.8 *(Geometric Ergodicity of a Class of Nonlinear Autoregressive Processes).* Let $\{X_n : n \geq 0\}$ be a nonlinear autoregressive sequence of order k of the form (20.55), with h bounded on compacts, and on the set $\{y = (y_1, \ldots, y_k) : |y| = \sum_{j=1}^k |y_j| \geq R\}$ one has

$$|h(y)| \leq \sum_{1 \leq i \leq k} a_i |y_i| + c, \tag{20.56}$$

where a_i's, c, and R are positive constants, $\sum_{1 \leq i \leq k} a_i < 1$, and the distribution of η_n has an absolutely continuous component g (with respect to Lebesgue measure λ_k), which is positive a.e. Also, assume $\mathbb{E}|\eta_n| < \infty$. Then (a) the Markov process $\{Y_n = (X_{n+1-k}, \ldots, X_n)' : n \geq k - 1\}$ is geometrically ergodic, and (b) the distribution of X_n converges to a steady state distribution exponentially fast in total variation.

Proof. We will show that the k-tuple $(X_n, X_{n+1}, \ldots, X_{n+k-1}), n \geq k - 1$, is a geometrically ergodic Markov process. To simplify notation we take $k = 2$. The general case is entirely analogous with a little messier notation (Exercise 7). Define the Lyapunov function (for $k = 2$)

$$V(y) = \max\{|y_1|, |y_2|\} + 1 \quad y = (y_1, y_2) \in \mathbb{R}^2. \tag{20.57}$$

One has, defining the norm on $\mathbb{R}^2$ by $|y| = |y_1| + |y_2|$, $y = (y_1, y_2)'$, and writing $Y_n = (Y_{n,1}, Y_{n,2})'$ to denote the components of of the random (column) vector Y_n,

$$Y_{n+1} = [X_n, h(Y_n) + \eta_{n+1}]' = [Y_{n,2}, h(Y_n) + \eta_{n+1}]',$$

$$Y_{n+2} = [h(Y_n)+\eta_{n+1}, h(Y_{n+1})+\eta_{n+2}]' = [h(Y_n)+\eta_{n+1}, h([Y_{n,2}), h(Y_n)+\eta_{n+1}])+\eta_{n+2}]',$$

$$|Y_{n+2,1}| \leq |h(Y_n) + \eta_{n+1}| + |Y_{n+2,2}| = |[h(Y_{n,2}, h(Y_n) + \eta_{n+1}) + \eta_{n+2}|.$$

[10] Nummelin and Tuominen (1983).

One then has on the event $[|Y_n| > R]$,

$$|Y_{n+2,1}| \le a_1|Y_{n,1}| + a_2|Y_{n,2}| + c + |\eta_{n+1}|, \tag{20.58}$$

and

$$\begin{aligned}|Y_{n+2,2}| &\le |\eta_{n+2}| + [a_1|Y_{n,2}| + a_2(|h(Y_n) + \eta_{n+1}| + c)\mathbf{1}_{[|Y_{n,2}|+a_2|h(Y_n)+\eta_{n+1}|\ge R]} + a_2 R + c \\ &\le |\eta_{n+1}| + |\eta_{n+2}| + c + a_1|Y_{n,2}| + a_2(a_1|Y_{n,1}| + a_2|Y_{n,2}| + c) + R. \qquad (20.59)\end{aligned}$$

Therefore, on $[Y_n > R]$, one has

$$\begin{aligned}V(Y_{n+2}) &= \max\{|Y_{n+2,1}|, |Y_{n+2,2}|\} + 1 \\ &\le |\eta_{n+1}| + |\eta_{n+2}| + 1 + 2c + R + \max\{a_1|Y_{n,1}| + a_2|Y_{n,2}|, a_1|Y_{n,2}| \\ &\quad + a_2(a_1|Y_{n,1}| + a_2|Y_{n,2}|)\} \\ &\le |\eta_{n+1}| + |\eta_{n+2}| + 2c + R + 1 + (a_1 + a_2)V(Y_n). \qquad (20.60)\end{aligned}$$

Now let $\theta > 1$ be chosen so large that

$$(a_1 + a_2) + 2(2\mathbb{E}|\eta_1| + 2c + 1 + R)/\theta R = 1 - \epsilon < 1.$$

In the part of (20.60) without $V(Y_n)$ as a factor, multiply and divide by $V(Y_n)$. Using the lower bound $V(Y_n) \ge \theta R/2$, for this $\epsilon > 0$ one then obtains,

$$\mathbb{E}[V(Y_{n+2})|Y_n = y] \le (1 - \epsilon)V(y) \text{ for all } |y| > \theta R. \tag{20.61}$$

Thus the relations (20.41) or (20.42) are easily verified with $S = \mathbb{R}^2$, $A_0 = \{y \in \mathbb{R}^2 : |y| \le \theta R\}$, $b = 0$, for the Markov process $\{Y_{2n} : n = 0, 1, \dots\}$, i.e., on the time scale $Y_0, Y_2, Y_4, \dots$. Next, note that this Markov process has an a.e. positive density component no smaller than (Exercise 8)

$$u(x, y) = g(y_1 - h(x)) \prod_{2\le j\le k} g(y_j - h(x_j, \dots, x_k, y_1, \dots y_{j-1}), \quad (k = 2). \tag{20.62}$$

In the case that the distribution of η_1 is absolutely continuous with density g, $u(x, y)$ is the transition probability density of the Markov process $\{Y_{2n} : n \ge 0\}$. Because of the a.e. positivity assumption for g, it is now straightforward to check that the Markov process $\{Y_{2n} : n \ge 0\}$ is (i) aperiodic and (ii) A_0 is a recurrent set with respect to Lebesgue measure λ_k (Exercise 9).

It remains to show that A_0 is a small set. For the proof of this we make the simplifying assumption that h is continuous and g is bounded below by a positive (λ_2-a.e.) function f (See Remark 20.7). Let $G(x) = \int_{A_0} u(x, y)\lambda_2(dy)$. Then $G(x) \ge F(x)$, where F is obtained by replacing in the last integral of g by f in the expression for $u(x, y)$ in (20.61). Here λ_2 is Lebesgue measure on $\mathbb{R}^2$. Then

$$\delta := \inf\{G(x) : x \in A_0\} \geq \inf\{F(x) : x \in A_0\} > 0.$$

Let $\nu(B) = \lambda_2(A_0 \cap B)/\lambda(A_0)$ for all Borel subsets B of $\mathbb{R}^2$. Then ν is a probability measure on $\mathbb{R}^2$, and denoting by q the transition probability of $\{Y_{2n} : n \geq 0\}$, one has $q(x, B) \geq \delta\nu(B)\mathbf{1}_{A_0}(x)$. Hence the Markov process $\{Y_{2n} : n \geq 0\}$ is geometrically ergodic by Theorem 20.5 or 20.6. The same proof applies to the process $\{Y_{2n+1} : n = 0, 1, \ldots\}$. Also, the existence of a a nonzero density component η_n (with full support) implies that the process is aperiodic (see (20.62)). The geometric ergodicity of the process $\{Y_n : n \geq 0\}$ follows easily from this (Exercise 11). This proves part (a) of the theorem. Part (b) is an easy consequence of this. ■

Remark 20.7. Theorem 20.8 is a slightly extended version of Theorem 1 in Bhattacharya and Lee (1995) (Correction, ibid, 1999). It implies several other results, as described in the following example.

***Example 2** (SETAR Model[11]).* The so-called self-exciting threshold autoregressive model (SETAR) is defined by

$$X_{n+1} = h(X_{n-k+1}, \ldots, X_n) + \eta_{n+1}, \tag{20.63}$$

where $\{\eta_n : n > 0\}$ is an i.i.d. sequence with an absolutely continuous common distribution with a positive density (a.e.) and $h : \mathbb{R}^k \to \mathbb{R}$ is specified as one of m different linear autoregressive models depending on which of m regimes (intervals) a random variable Z takes its values. Typically Z is one of the k coordinates, say $y_j (j = 1, \ldots, k)$, or some function of the k coordinates, or a variable independent of $\{\eta_n : n \geq 1\}$. Thus,

$$h(y) = c_i + \sum_{1 \leq j < k} a_{ij} y_j, \quad \text{if } r_{i-1} < Z \leq r_i, i = 1, \ldots, m, y = (y_1, \ldots, y_k), \tag{20.64}$$

where $-\infty = r_0 < r_1 < \cdots < r_m = \infty$. Under the condition that $\max_{1 \leq i \leq m} \sum_{k_{i-1} \leq j < k_i} |a_{ij}| < 1$, the conclusions of Theorem 20.6 hold. One may use the same Lyapunov function $V(y) = \max\{|y_i| : i = 1, \ldots, k\} + 1$, as in the proof of Theorem 20.8, on each regime, and arrive at the desired inequality on the time scale $Y_0, Y_k, Y_{2k}, \ldots, \mathbb{E}[V(Y_{kn}|Y_{(k-1)n} = y] \leq (1-\epsilon)V(y)$; first for $k = 2$ for simplicity, and then generally (Exercise 11). This result holds if η_n has an absolutely continuous component which is a.e. positive, somewhat extending the original result of Chan and Tong (1985).

[11] Chan and Tong (1985).

Exercises

1. Let $Y_j = X_{\tau_j}$, $j \geq 0$, be the process defined on a recurrent set A_0 (see (20.6)). Use the strong Markov property to prove that $\{Y_j : j \geq 1\}$ is a Markov process on $(A_0, A_0 \cap \mathcal{S})$ with the transition probability $p_{A_0}(x, dy)$ given by (20.7).
2. Suppose that $p(x, dy)$ has, for each $x \in \mathbb{R}$, a density with respect to Lebesgue measure on $\mathbb{R}$. Show that $p^{(n)}(x, \{y\}) = 0$ for each singleton $\{y\}$.
3. Prove the regeneration lemma (Proposition 20.1) in the case $\lambda = 0$
4. Assume that A_0 is recurrent and locally minorized (Definitions 20.2, 20.3). Define, as in the proof of Theorem 20.1 (a), $F_j := \{X_n \in B \text{ for some } n \in [\tau_j, \tau_{j+1})\}$, $j \geq 0$, with $Y_0 = X_0 \in A_0$ having distribution π_{A_0}. (i) In the strongly aperiodic case ($N = 1$) show that the tail σ-field of the stationary Markov process $\{Y_j : j \geq 0\}$ is trivial. (ii) Show $P_y(\cap_{n=0}^{\infty} \cup_{j=n}^{\infty} F_j) = 1$ for all $y \in A_0$. [*Hint*: Use (20.16) with $N = 1$. (iii) For general $N \geq 1$, show that $\{Y_j : j \geq 0\}$ is a stationary ergodic Markov process. [An alternative procedure for proving ergodicity of Y_j, $j \geq 0$, is to use Theorem 16.4.]
5. (a) Show that $\pi(dy) = 4ye^{-2y}\mathbf{1}_{[0,\infty)}(y)dy$ is a time-reversible invariant probability.[*Hint*: Consider $A_0 = [\delta k, k]$ for $k \geq 1$ and $V(x) = e^{cx}, 0 < c < 1, x \geq 0$. (b) Recall Example 1 and prove geometric ergodicity of the Markov process on $[0, \infty)$ defined by iterated maps $X_n = \gamma_{(U_n,T_n)} \cdots \gamma_{(U_1,T_1)}(X_0), n \geq 1$, of the form $\gamma_(u, t)(x) = ux + t, t \geq 0$, where, independently of X_0, $U_1, U_2, \ldots$ is an iid sequence of uniformly distributed random variables on $[0, 1]$, and $T_1, T_2, \ldots$ is an i.i.d. sequence, independent of $U_n, n \geq 1$, of exponentially distributed random variables with mean $1/2$. (c) Show that one may take the small set in Example 1 as $A_0 = (0, a]$ for some $a > 0$, and ν having density $2e^{-2y}\mathbf{1}_{[a,\infty)}(y)$.
6. (a) In the context of the proof of Proposition 20.1, (i) show that $\{(X_n, \theta_n) : n \geq 0\}$ is a Markov process, (ii) for the proof of part (b) of the proposition show that $\theta_{\tau^{(j)}}$ is independent of $\tilde{\mathcal{F}}_{\tau^{(j-1)}} = \sigma\{(X_n, \theta_n) : 0 \leq n \leq \tau^{(j-1)}\}$ and of $\tau^{(j)}$. (b) In the context of the proof of Theorem 20.4, show that both X and X' are Markov processes with the transition probability $p(x, dy)$.
7. Extend the proof of Theorem 20.8 to the case of general $k \geq 2$ under the simplifying assumption in the last paragraph on h and on the absolutely continuous component of the distribution of η_n.
8. Show that the density component of the transition probability of the process $\{Y_{kn} : n \geq 0\}$, on the time scale $0, k, 2k, \ldots$, is no smaller than $u(x, y)$ in (20.62).
9. Under the hypothesis of Theorem 20.8 show that the Markov process $\{Y_{kn} : n \geq 0\}$ satisfies the hypothesis of Theorem 20.5 or 20.6. You may assume the simplifying assumption to prove that A_0 is a small set made in the last paragraph of the proof of the theorem.
10. Fix $0 < a < 1$ and define $X_{n+1} = aX_n + T_{n+1}, n = 0, 1, \ldots,$ where $T_n, n \geq 1$, is an i.i.d. sequence of mean one exponentially distributed random variables. Then, $p(x, dy) = e^{ax}e^{-y}\mathbf{1}_{[ax,\infty)}(y)dy$. The objective is to establish geometric

ergodicity for this example. Let $A_0 = [0, k]$. (i) With $N = 1$, determine k such that A_0 is a small set for ν_k given by $\nu_k(B) = \int_{B\cap[ak,k]} e^{-y}dy/\lambda_k$, $\lambda_k = e^{-ak} - e^{-k}$. [*Hint*: For $x \in A_0$, $B \subset A_0$, show that $p(x, B) = e^{ax} \int_{B\cap[ax,k]} e^{-y}dy \geq \lambda_k \nu_k(B)$. Let $V_c(x) = e^{cx}$. (ii) For $x > k$, show $\mathbb{E}_x V_c(X_1) = e^{acx}(1-c)^{-1} \leq \beta_{c,k} V_c(x)$, where $\beta_{c,k} = e^{-(1-a)ck}(1-c)^{-1} < 1$ for k sufficiently large, i.e., $k > \frac{1}{c(1-a)} \ln \frac{1}{1-c}$. (iii) For $x \in [0, k]$, show that $\mathbb{E}_x V_c(X_1) \leq \beta_{c,k} V_c(x)+b$, for b sufficiently large. [*Hint*: Consider $b \geq \max_{0\leq x\leq k} e^{acx}(1-c)^{-1} = e^{ack}(1-c)^{-1}$. (iv) Determine $0 < c < 1$ such that the condition (20.42)(ii) for the bivariate drift holds. [*Hint*: Check $e^{kc} > ([\frac{b}{1-\beta_{c,k}}] - 1)$, or equivalently, refine k so that $e^{kc} > 2e^{ack}/(1-c)$ $k > \frac{1}{c(1-a)} \ln(\frac{2}{1-c})$ is sufficient.

11. (a) Show that Theorem 20.5 applies to the the proof of Theorem 20.8, and so does Theorem 20.6. (b) Write out a proof of the geometric ergodicity of the SETAR model (Example 2).

Chapter 21
An Extended Perron–Frobenius Theorem and Large Deviation Theory for Markov Processes

An extension of the Perron–Frobenius theorem to compact linear operators on $C_b(S)$ for locally compact and σ-compact metric spaces S is presented. This yields some of the basic large deviation theory originating with Cramér, Sanov, Donsker, and Varadhan presented in three parts. In the first part the extended Perron–Frobenius theorem is used to obtain a large deviation theorem of the Cramér–Chernoff type for a class of Markov processes. In the second part the large deviation framework is extended to large deviations for the empirical distributions of a class of Markov processes originally obtained by Donsker and Varadhan. As a corollary the third part involves large deviations for empirical distributions of i.i.d. random variables originating with Sanov.

Throughout this chapter (S, ρ) is a locally compact and σ-compact Polish space, and for each $x \in S$, let $q(x, dy)$ be a finite measure on the Borel σ-field $\mathcal{S}$ of S, and such that $\sup_x q(x, S) < \infty$, and $x \to q(x, B)$ is measurable for every $B \in \mathcal{S}$. Define a bounded linear operator $T_q : C_b(S) \to C_b(S)$ by

$$T_q f(x) = \int_S f(y)q(x, dy),\ x \in S. \tag{21.1}$$

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_21

The notion of the support of a finite measure often appears in this chapter. Here is the definition.

Definition 21.1. The *support of a finite measure* μ on $(S, \mathcal{S})$ is the smallest closed set C_μ such that $\mu(C_\mu) = \mu(S)$, i.e., $C_\mu := \cap_{C \in \mathcal{C}} C$, where $\mathcal{C}$ is the class of all closed subsets of S such that $\mu(C) = \mu(S)$. Equivalently, $C_\mu = \mathcal{S} \setminus \cup_{G \in \mathcal{O}} G$ where $\mathcal{O}$ is the class of all open $G \subset S$ with $\mu(G) = 0$.

Theorem 21.1 *(Extended Perron–Frobenius Theorem for Compact T_q).* Assume that the operator T_q defined by (21.1) is compact, i.e., T_q maps bounded subsets of $C_b(S)$ to relatively compact subsets of $C_b(S)$ for the uniform norm. Let $\Gamma = \overline{T_q(B_1^+)} \cup \{1\}$ be the closure of the set of images of $B_1^+ = \{f \in C_b(S) : f > 0, ||f|| = 1\}$ under T_q, with added constant function $f \equiv 1$.

Define

$$\Lambda^+ = \{\lambda > 0 : T_q f \geq \lambda f \text{ on } S \text{ for some } f \in \Gamma\}. \tag{21.2}$$

Assume $\inf_x q(x, S) > 0$, $\sup_x q(x, S) < \infty$. Also assume the Doeblin minorization condition: there is a finite nonzero measure ψ on $(S, \mathcal{S})$ such that $q(x, dy) \geq \psi(dy)$, for all $x \in S$, and ψ is fully supported. Then

a $\Lambda^+ \neq \emptyset$, and Λ^+ has a maximum element λ^+, which can be paired with a function $f^+ \in \Gamma$ such that $T_q f^+(x) = \lambda^+ f^+(x)$ for all $x \in S$,

b There is a unique probability measure $\widetilde{\pi}$ such that

$$\lim_{n\to\infty} \left\| \frac{q^{(n)}(x, dy) f^+(y)}{f^+(x)\lambda^{+n}} - \widetilde{\pi}(dy) \right\|_{tv} = 0, \tag{21.3}$$

where $|| \cdot ||_{tv}$ denotes the total variation norm. Moreover the convergence in total variation norm is exponentially fast, uniformly in $x \in S$. Also, $\widetilde{\pi}$ is the unique invariant probability of a Markov chain with transition probability $\widetilde{p}(x, dy) = \frac{q(x,dy) f^+(y)}{f^+(x)\lambda^+}$. In particular,

$$\lim_{n\to\infty} \int_S g(y) q^{(n)}(x, dy) \frac{f^+(y)}{(\lambda^+)^n f^+(x)} = \int_S g d\widetilde{\pi}, \tag{21.4}$$

for all bounded measurable g, the convergence being uniform in x.

Proof. (a) Clearly, Λ^+ is nonempty, since $\lambda_m := \inf\{q(x, S) : x \in S\}$ is an element; for one may take $f \equiv 1$ and $\lambda = \lambda_m$ in (21.2). Also Λ^+ is bounded above by $\lambda_M := \sup\{q(x, S) : x \in S\}$. Denote the supremum of Λ^+ by λ^+, and let $\lambda_n \in \Lambda^+$ increase to λ^+ as $n \uparrow \infty$. Also let f_n be corresponding elements of Γ paired with λ_n. There exists a subsequence of $\{f_n : n \geq 1\}$ which converges to $f = f^+ \in \Gamma$, by the Arzela–Ascoli Theorem.[1] To reduce notation, we let this paired subsequence be denoted $\{(\lambda_n, f_n)\}$. Then (21.2) leads to

$$T_q f^+(x) \geq \lambda^+ f^+(x) \quad \text{for all } x \in S. \tag{21.5}$$

We now show that there is equality in (21.5). For otherwise, there is some $x = x_0$ such that $T_q f^+(x_0) > \lambda^+ f^+(x_0)$. In view of continuity, this implies that there exists an open neighborhood U of x_0 such that $T_q f^+(x) > \lambda^+ f^+(x)$, or, $T_q f^+(x) - \lambda^+ f^+(x) > 0$, for all $x \in U$, and $T_q f^+(x) - \lambda^+ f^+(x) \geq 0$ outside U. Applying T_q to the difference $T_q f^+ - \lambda^+ f^+$, and using the fact that $q(x, dy) \geq \psi(dy)$ has full support, it follows that $T_q(T_q f^+ - \lambda^+ f^+)(x) \geq \int_S (T_q f^+(x) - \lambda^+ f^+(x))\psi(dx) > 0$ on S. Writing $\epsilon = \inf_{x \in S} T_q(T_q f^+ - \lambda^+ f^+)(x)$, one has $T_q(T_q f^+) \geq \lambda^+ T_q f^+ + \epsilon \geq (\lambda^+ + \delta) T_q f^+$, where $\delta = \frac{\epsilon}{||T_q f^+||}$. Also, $\frac{f^+}{||f^+||} \in B_1^+$ implies that $g = T_q f^+/||f^+|| \in \Gamma$. So, dividing the extreme sides by $||f^+||$, one gets a contradiction to the maximality of λ^+. Hence $T_q f^+ = \lambda^+ f^+$ on S. This proves (a). For (b), first note that under the hypothesis for (a), $f^+(x) > 0$ for all $x \in S$. Therefore, $\widetilde{p}(x, dy) := q(x, dy) \frac{f^+(y)}{\lambda^+ f^+(x)}$, $x \in S$, defines a transition probability on S. That is, apart from obvious measurability, $\widetilde{p}(x, dy)$ is a probability measure for every $x \in S$. By the additional hypothesis imposed, this transition probability satisfies Doeblin's condition. Hence $\widetilde{p}^{(n)}(x, dy)$ converges in variation norm to a unique invariant probability $\pi(dy)$ for $\widetilde{p}$ (See Corollary 20.2). One also has the relation

$$\widetilde{p}^{(n)}(x, dy) = q^{(n)}(x, dy) \frac{f^+(y)}{\lambda^{+n} f^+(x)}, \quad n = 1, 2 \ldots. \tag{21.6}$$

This is easy to check for $n = 2$:

$$\begin{aligned}
\widetilde{p}^{(2)}(x, dy) &= \int_S \mathbf{1}_S(z) \widetilde{p}(z, dy) \widetilde{p}(x, dz) \\
&= \int_S \mathbf{1}_S(z) q(z, dy) \frac{f^+(y)}{\lambda^+ f^+(z)} q(x, dz) \frac{f^+(z)}{\lambda^+ f^+(x)} \\
&= \int_S \mathbf{1}_S(z) (q(z, dy) q(x, dz)) \frac{f^+(y)}{\lambda^{+2} f^+(x)} \\
&= q^{(2)}(x, dy) \frac{f^+(y)}{\lambda^{+2} f^+(x)}.
\end{aligned} \tag{21.7}$$

The general formula (21.6) now follows by induction on n (Exercise 2). To check that $\widetilde{p}(x, dy)$ satisfies Doeblin's condition, note that

[1] Folland (1984), p. 131.

$$\widetilde{p}(x,dy) \geq \frac{f^+(y)}{\lambda^+ f^+(x)}\psi(dy) \geq \frac{f^+(y)}{\lambda^+ \sup_x f^+(x)}\psi(dy). \tag{21.8}$$

Hence $\widetilde{p}^{(n)}(x,dy)$ converges in total variation norm to the unique invariant probability $\pi(dy)$, say, uniformly for all $x \in S$. Hence part (b) of the theorem holds. Therefore, for all bounded measurable functions g on S,

$$\lim_{n\to\infty}\int_S g(y)q^{(n)}(x,dy)\frac{f^+(y)}{\lambda^{+n} f^+(x)} = \int_S g(y)\widetilde{\pi}(dy), \tag{21.9}$$

uniformly in $x \in S$. ■

Remark 21.1. It appears that part (a) of Theorem 21.1 may also be proven by an application of a general theorem[2] due to Krein and Rutman (1948). However the statement and proof provided here is self-contained and particularly suited to its application in large deviation theory.

Remark 21.2. We call λ^+ in Theorem 21.1 the largest eigenvalue of T_q.

The following corollary is an immediate consequence of Theorem 21.1, and relation (21.6). (Exercise 1)

Corollary 21.2. Let $q(x,dy) = p(x,dy)e^{v(y)}$ satisfy the hypotheses of Theorem 21.1, where $p(\cdot,dy)$ is a transition probability of a Markov process $X_0 = x$, $X_1, X_2, \ldots$, and v is a measurable function. Then (i) $q^{(n)}(x,S) = \mathbb{E}_x \exp\{v(X_1) + \ldots + v(X_n)\}$, and (ii) one has

$$\lim_{n\to\infty}\frac{1}{n}\log q^{(n)}(x,S) = \log\lambda^+, \tag{21.10}$$

where λ^+ denotes the largest eigenvalue of the operator T_q defined by $q(x,dy) = p(x,dy)e^{v(y)}$.

Remark 21.3. It is enough for the theorem to hold if the hypotheses hold for $q^{(n_0)}(x,dy)$ for some $n_0 \geq 1$ (Exercise 3). Also, Theorem 21.1 holds if $q(x,dy)$ has a density $q(x,y)$ with respect to a measure $\mu(dy)$ with full support such that $y \to q(x,y)$ is continuous in y and $g(y) := \inf_x q(x,y) > 0$ for all y, and $\sup_{x,y} q(x,y) < \infty$. In this case one takes $\psi(dy) = g(y)\mu(dy)$.

Remark 21.4. In Theorem 21.1 the assumption that the measure ψ has full support is only used to show that there is equality in $T_q f^+ = \lambda^+ f^+$ for the maximum element of Λ^+. There are cases where this assumption is not needed to prove the equality. For example, in the case $q(x,dy) = e^{v(y)}\mu(dy)$, where $\mu(dy)$ is the common distribution of i.i.d. random variables, it is simple to check that the equality

[2] The authors thank Patrick De Leenheer for this resource.

holds with $\lambda^+ = m(v) = \int_S e^{v(y)}\mu(dy)$, and $f^+ = 1$. Here $\psi(dy)$ may be taken to be $\mu(dy)$.

Remark 21.5. In case the X_j's are i.i.d. with common distribution π, taking hv in place of $v(h \in \mathbb{R})$, (21.10) is just the statement that the moment generating function $m(h)$ of $v(X_1)$ satisfies $\frac{1}{n}\log m(h)^n \to \log m(h)$, i.e., $\lambda^+ = m(h)$.

The so-called *first-level* large deviation problem can be stated as follows. Let $X_0, X_1, \ldots$ be a stationary ergodic Markov process on S having transition probability kernel $p(x, dy)$ and unique invariant probability π. Let v be measurable function on S and assume that $\mathbb{E}_x e^{hv(X_1)} < \infty$ for all $h \in \mathbb{R}$, and $x_0 = x \in S$. Then, just as in the i.i.d. case of the Cramér–Chernoff theorem,[3] one seeks the large deviation rate $I(a) = -\lim_{n\to\infty} \frac{1}{n}\log P(\sum_{j=1}^n v(X_j) > na)$ for a deviation from the sample mean given by $a > m := \int_S v(x)\pi(dx)$. The size-bias approach given here is precisely along the same lines as that of the classical Cramér–Chernoff theorem for i.i.d random variables.

Remark 21.6. The proof that $h \to \lambda^+(h)$ is differentiable and $\frac{d}{dh}\lambda^+(h) \neq 0$ appears to require perturbation theory. A direct proof for finite S is indicated in Exercise 18. For the i.i.d. case of the Cramèr-Chernoff theorem, it is also simple to show, using moment generating functions, that the kernel $q_h(x, dy) = p(x, dy)e^{hv(y)}$ satisfies the hypothesis of Theorem 21.1 for all $h \in \mathbb{R}$.

Lemma 1. Assume that the kernel $q_h(x, dy) = p(x, dy)e^{hv(y)}$ satisfies the hypothesis of Theorem 21.1 for all $h \in \mathbb{R}$. Let $\lambda^+(h)$ denote the largest eigenvalue of the transition operator for the (non-normalized) kernel $q_h(x, dy)$. Let $X_0, X_1, \ldots$ be a Markov process with transition probability kernel $p(x, dy) \geq \psi(dy)$ satisfying the Doeblin minorization with respect to a nonzero measure ψ, and let $\widetilde{X}_0, \widetilde{X}_1, \ldots$ be the Markov process with transition probability kernel $\widetilde{p}(x, dy) = \frac{f^+(y)}{\lambda^+(h^*)f^+(x)}e^{h^*v(y)}p(x, dy)$ for a given $h = h^*$, where, suppressing the dependence in f^+ on h^* for notational convenience, f^+ is the normalized positive eigenfunction of the operator $T_{q_{h^*}}f(x) = \int_S f(y)e^{h^*v(y)}p(x, dy)$ corresponding to $\lambda^+(h^*)$. Then the respective distributions, $P_{x,n}$ and $\widetilde{P}_{x,n}$, of $(X_0, \ldots, X_n)$ and $(\widetilde{X}_0, \ldots, \widetilde{X}_n)$, starting at $X_0 = \widetilde{X}_0 = x$, are mutually absolutely continuous with

$$\frac{d\widetilde{P}_{x,n}}{dP_{x,n}}(x_0, \ldots, x_n) = \frac{f^+(x_n)}{\lambda^{+n}(h^*)f^+(x_0)}e^{h^*\sum_{j=1}^n v(x_j)}, \tag{21.11}$$

and

$$\frac{dP_{x,n}}{d\widetilde{P}_{x,n}}(\widetilde{x}_0, \ldots, \widetilde{x}_n) = \frac{f^+(\widetilde{x}_0)\lambda^{+n}(h^*)}{f^+(\widetilde{x}_n)}e^{-h^*\sum_{j=1}^n v(\widetilde{x}_j)}. \tag{21.12}$$

[3] See BCPT p. 94.

If X is stationary and ergodic with transition probability kernel $p(x, dy)$ and unique invariant probability π with $\mathbb{E}_\pi v(X_0) = m$, then $\widetilde{X}$ has transition probabilities $\widetilde{p}(x, dy)$ with a unique invariant probability $\widetilde{\pi}(dy)$ with $\int_S v(y)\widetilde{\pi}(dy) = a$, say.

Proof. Consider the Markov process $\{\widetilde{X}_n : n \geq 0\}$ on S defined by the consistent specification ($x_0 = x$),

$$\begin{aligned}
\widetilde{P}_x(\widetilde{X}_0 \in dx_0, \widetilde{X}_1 \in dx_1, \dots, \widetilde{X}_n \in dx_n) &= \prod_{j=1}^{n} \widetilde{p}(x_{j-1}, dx_j) \\
&= \frac{f^+(x_n)}{f^+(x_0)\lambda^{+n}(h^*)} e^{\sum_{j=1}^{n} h^* v(x_j)} \prod_{j=1}^{n} p(x_{j-1}, dx_j) \\
&= \frac{f^+(x_n)}{f^+(x_0)\lambda^{+n}(h^*)} e^{\sum_{j=1}^{n} h^* v(x_j)} P_x(X_0 \in dx_0, X_1 \in dx_1, \dots, X_n \in dx_n).
\end{aligned}$$

Similarly, or by inversion,

$$\begin{aligned}
P_x(X_1 \in dx_1, \dots, X_n \in dx_n) &= \prod_{j=1}^{n} p(x_{j-1}, dx_j) \\
&= \frac{p(x_0, dx_1)e^{h^* v(x_1)} f^+(x_1)}{\lambda^+(h^*) f^+(x_0)} \cdot \frac{p(x_1, dx_2)e^{h^* v(x_2)} f^+(x_2)}{\lambda^+(h^*) f^+(x_1)} \cdots \frac{p(x_{n-1}, dx_n)e^{h^* v(x_n)} f^+(x_n)}{\lambda^+(h^*) f^+(x_{n-1})} \\
&= \frac{f^+(x_0)\lambda^{+n}(h^*)}{f^+(x_n)} e^{-\sum_{j=1}^{n} h^* v(x_j)} \prod_{j=1}^{n} p(x_{j-1}, dx_j) \\
&= \frac{f^+(x_0)\lambda^{+n}(h^*)}{f^+(x_n)} e^{-\sum_{j=1}^{n} h^* v(x_j)} P_x(\widetilde{X}_0 \in dx_0, \widetilde{X}_1 \in dx_1, \dots, \widetilde{X}_n \in dx_n).
\end{aligned}$$

This establishes the mutual absolute continuity of the distributions of $(X_0, X_1, \dots, X_n)$ and $(\widetilde{X}_0, \widetilde{X}_1, \dots, \widetilde{X}_n)$, with $X_0 = \widetilde{X}_0 = x$, as asserted.

That $\{\widetilde{X}_0, \widetilde{X}_1, \dots\}$ has an unique invariant probability $\widetilde{\pi}$ follows from Theorem 21.1, or Corollary 21.2. ■

Remark 21.7. In the case that π is a time-reversible invariant probability for $p(x, dy)$, one may readily check that $\widetilde{\pi}(dx) = Z^{-1}(h^*)(f^+(x))^2 e^{h^* v(x)}\pi(dx)$ is the time-reversible invariant probability for $\widetilde{p}(x, dy)$, where $Z(h^*) = \int_S (f^+(x))^2 e^{h^* v(x)}\pi(dx)$ is the normalization constant. (Exercise 4)

Remark 21.8. Under the hypothesis of Lemma 1, $h \to \lambda^+(h)$ is continuous. Let $A = \{y : v(y) > a\}$ for some $a > 0$. If $\inf_x p(x, A) > 0$, then one can show that $\lambda^+(h) \to \infty$ exponentially fast as $h \to \infty$. (Exercise 19)

Theorem 21.3 *(Large Deviations for a Class of Markov Processes on Locally Compact & σ-compact Polish S).* Assume the conditions of the framework leading up to Lemma 1 and that $\mathbb{E}_\pi v(X_1) = 0$. Then, for $a > 0$, the large deviation rate is given by

$$\begin{aligned} I(a) &= -\lim_{n\to\infty}\frac{1}{n}\log P_x\left(\sum_{j=1}^{n} v(X_j) \geq na\right) = -\inf_{h>0}\{\log\lambda^+(h) - ah\} \\ &= ah^* - \log\lambda^+(h^*). \end{aligned}$$

Define $h^* = \infty$ if the infimum is not attained, i.e., the infimum is ∞.

Proof. First consider the case $h^* < \infty$. Observe that for $h > 0$,

$$\begin{aligned} q_h^{(n)}(x, S) &= \mathbb{E}_x e^{h\sum_{j=1}^{n} v(X_j)} \\ &\geq e^{nha} P_x\left(\sum_{j=1}^{n} v(X_j) \geq na\right). \end{aligned} \tag{21.13}$$

Now one has, using (21.6), the general relation

$$\frac{\lambda^{+n} f^+(x)}{\sup_y f^+(y)} \leq q_h^{(n)}(x, S) \leq \frac{\lambda^{+n} f^+(x)}{\inf_y f^+(y)}.$$

In particular,

$$\begin{aligned} \frac{1}{n}\log P_x\left(\sum_{j=1}^{n} v(X_j) \geq na\right) &\leq \frac{1}{n}\left(\log q_h^{(n)}(x, S) - nha\right) \\ &\leq \frac{1}{n}\left\{\log\left(\lambda^{+n}(h)\frac{\sup_x f^+(x)}{\inf_x f^+(x)}\right) - nha\right\} \\ &\leq \log\lambda^+(h) + \frac{1}{n}\log\frac{\sup_x f^+(x)}{\inf_x f^+(x)} - ha. \end{aligned} \tag{21.14}$$

Thus,

$$\begin{aligned} &\frac{1}{n}\log P_x\left(\sum_{j=1}^{n} v(X_j) \geq na\right) \\ &\leq \inf_{h>0}\left\{\log\lambda^+(h) + \frac{1}{n}\log\frac{\max_x f^+(x)}{\min_y f^+(y)} - ha\right\}, \end{aligned} \tag{21.15}$$

and, for each $h > 0$,

$$\limsup_{n\to\infty}\frac{1}{n}\log P_x\left(\sum_{j=1}^{n} v(X_j)\geq na\right)\leq \log\lambda^{+}(h)-ha, \tag{21.16}$$

so that

$$\limsup_{n\to\infty}\frac{1}{n}\log P_x\left(\sum_{j=1}^{n} v(X_j)\geq na\right)\leq \inf_{h>0}\left\{\log\lambda^{+}(h)-ha\right\}=\log\lambda^{+}(\tilde{h})-\tilde{h}a, \tag{21.17}$$

say. For the lower bound we consider the mutually absolutely continuous size-bias change of distribution of $X_0, X_1, \ldots$ defined in Lemma 1, under which the transformed Markov process $\widetilde{X}_0, \widetilde{X}_1, \ldots$ is an ergodic Markov process with unique invariant probability $\widetilde{\pi}$ for the transition probability $\widetilde{p}(x, dy) = f^{+}(y)e^{h^{*}v(y)}p(x, dy)$, where h^{*} is the size-bias parameter such that $\mathbb{E}_{\widetilde{\pi}}v(\widetilde{X}_0) = a$.

Then, for any $\epsilon > 0$, one has by the ergodic theorem that $\widetilde{P}$-a.s.

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} v(\widetilde{X}_j)=\int_S v(y)\widetilde{\pi}(dy)=a. \tag{21.18}$$

Define

$$D_n=\left\{(y_1,\ldots,y_n):\frac{1}{n}\sum_{j=1}^{n} v(y_j)\in(a-\epsilon, a+\epsilon)\right\}. \tag{21.19}$$

Then, writing $A_n = \frac{1}{n}\sum_{j=1}^{n} v(X_j)$, $\widetilde{A}_n = \frac{1}{n}\sum_{j=1}^{n} v(\widetilde{X}_j)$, P_x for the distribution of $(X_0, X_1, \ldots)$, $\widetilde{P}_x$ for the distribution of $(\widetilde{X}_0, \widetilde{X}_1, \ldots)$, and recalling that under the invariant distribution $\widetilde{\pi}$ for $\{\widetilde{X}_n : n \geq 0\}$, $v(\widetilde{X}_n)$ has mean a, one has using Lemma 1,

$$\begin{aligned}
&P_x(A_n > a-\epsilon)\geq P_x(A_n\in(a-\epsilon, a+\epsilon))=\mathbb{E}_{P_{\widetilde{\pi}}}\mathbf{1}[(X_1,\ldots,X_n)\in D_n]\\
&\quad\geq \mathbb{E}_x\mathbf{1}[(X_1,\ldots,X_n)\in D_n]\exp\{-nh^{*}(a+\epsilon)+h^{*}\sum_{j=1}^{n} v(X_j)\}\\
&\quad=\exp\{-nh^{*}(a+\epsilon)+n\log\lambda^{+}(h^{*})\}\mathbb{E}_x\mathbf{1}[(X_1,\ldots,X_n)\in D_n]\frac{1}{\lambda^{+n}(h^{*})}e^{h^{*}\sum_{j=1}^{n} v(X_j)}\\
&\quad=\exp\{-nh^{*}(a+\epsilon)+n\log\lambda^{+}(h^{*})\}\widetilde{\mathbb{E}}_x\frac{f^{+}(\widetilde{X}_0)}{f^{+}(\widetilde{X}_n)}\mathbf{1}[(\widetilde{X}_1,\ldots,\widetilde{X}_n)\in D_n]\\
&\quad=\exp\{-n[h^{*}(a+\epsilon)-\log\lambda^{+}(h^{*})]\}\mathbb{E}_x\frac{f^{+}(\widetilde{X}_0)}{f^{+}(\widetilde{X}_n)}\mathbf{1}[\widetilde{A}_n\in(a-\epsilon, a+\epsilon)].
\end{aligned} \tag{21.20}$$

By the property of Harris positive recurrent Markov processes (see Chapter 20) $\mathbf{1}[\widetilde{A}_n \in (a-\epsilon, a+\epsilon)] \to 1$ $\widetilde{P}_x$-a.s. as $n \to \infty$. Also, it follows from Lebesgue dominated convergence that

$$\lim_{n\to\infty}\left\{\widetilde{\mathbb{E}}_x\frac{f^+(\widetilde{X}_0)}{f^+(X_n)}-\widetilde{\mathbb{E}}_x\frac{f^+(\widetilde{X}_0)}{f^+(X_n)}\mathbf{1}[\widetilde{A}_n\in(a-\epsilon,a+\epsilon)]\right\}$$

$$=\lim_{n\to\infty}\widetilde{\mathbb{E}}_x\frac{f^+(\widetilde{X}_0)}{f^+(X_n)}\mathbf{1}[\widetilde{A}_n\notin(a-\epsilon,a+\epsilon)]=0. \tag{21.21}$$

It follows from the bound (21.20) that $\liminf_{n\to\infty}\frac{1}{n}\log P_{\widetilde{\pi}}(A_n>a-\epsilon)+h^*\epsilon\geq\log\lambda^+(h^*)-h^*a$. Since the left side is an increasing function of ϵ, it follows that

$$\liminf_{n\to\infty}\frac{1}{n}\log P_{\widetilde{\pi}}(A_n\geq a)\geq\log\lambda^+(h^*)-h^*a.$$

Thus,

$$\liminf_{n\to\infty}\frac{1}{n}\log P_{\widetilde{\pi}}(A_n\geq a)\geq\log\lambda^+(h^*)-h^*a.$$

Since the lower bound cannot exceed the upper bound, and the infimum in (21.17) is no more than $\lambda^+(h^*)-h^*a$, one has $\tilde{h}=h^*$.

Now consider the case $h^*=\infty$. The proof for the upper bound holds with the infimum equal to $-\infty$ in (21.17), which implies that the "limit" is $-\infty$. ■

Remark 21.9. From Theorem 21.3, one obtains an indirect proof of the relation

$$\frac{dlog\lambda^+(h)}{dh}=m(h)=\mathbb{E}_{\widetilde{\pi}(h)}v,$$

assuming $h\to\lambda^+(h)$ is differentiable.

Remark 21.10. By changing signs one may derive a large deviation rate for

$$\lim_{n\to\infty}\frac{1}{n}\log P\left(\sum_{j=1}^{n}(v(X_j)-\mathbb{E}_{\widetilde{\pi}}v)\leq -na\right)$$

for all $a>0$. (Exercise 18)

Remark 21.11. If S is compact the boundedness condition on $p(x,y)$ will be satisfied if $p(x,y)$ is jointly continuous in (x,y).

The proof of the following corollary is left as Exercise 5.

Corollary 21.4. The conditions of Theorem 21.3 are satisfied if (i) $x\to\int_S e^{hv(y)}p(x,dy)$ is continuous in total variation norm for every $h>0$, (ii) $\inf_{x\in S}\int_S e^{hv(y)}p(x,dy)>0$, (iii) $p(x,dy)\geq\psi(dy)$ for all $x\in S$, where the minorizing measure $\psi(dy)$ has full support, and (iv) $\sup_{x\in S}\int_S e^{hv(y)}p(x,dy)<\infty$.

Corollary 21.5 *(Cramér–Chernoff Large Deviations).* Suppose that $X_1, \dots$ is an i.i.d. sequence of random variables, and v a measurable function on S such that $m(h) = \mathbb{E}e^{hv(X_1)} < \infty$ for all $h \in \mathbb{R}$. Then, $\lambda^+(h) \equiv m(h)$ and, for $a > \mathbb{E}v(X_1)$,

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\sum_{j=1}^{n} v(X_j) \geq na\right) = \inf_{h>0}\{\log m(h) - ha\}.$$

Proof. Let S be the support of the distribution $p(dy)$ of X_1 on $\mathbb{R}$. The conditions of the theorem are easily checked for the canonical distribution on S (Exercise 21). In particular, $Tf(x) = \int_S f(y)e^{hv(y)}p(dy)$ maps to constants, and $T1(x) = \int_S e^{hv(y)}p(dy) = m(h)1(x), x \in S$, for the constant function $1(x) \equiv 1 \forall x \in S$. Thus $\lambda^+(h) = m(h)$, and $f^+ \equiv 1$ (Exercise 21). ■

Remark 21.12. While we have focused this chapter on large deviations in the presence of Markov dependence, another departure from the classic Cramér–Chernoff theory occurs by considering i.i.d. summands for which the moment generating function may be infinite. By contrast to such "light-tailed"conditions, Nagaev (1969) considered large deviations for a class of distributions[4] referred to as *stretched exponential distributions* $P(X > t) = ce^{-t^r}$, $t \geq 0$, for a parameter $0 < r < 1$. In particular, for example, it is shown that if $X_1, X_2, \dots$ is an i.i.d. sequence of stretched exponentially distributed random variables with mean m, then for $a > m$,

$$\lim_{n\to\infty} \frac{1}{n^r} \log P(S_n > na) = -(a-m)^r, \tag{21.22}$$

where $S_n = \sum_{j=1}^{n} X_j$, i.e., the probabilities of deviations of $\frac{S_n}{n}$ above the mean m decay more slowly than exponentially. Such a deviation can occur when merely one of the independent summands eventually takes a very large value, i.e., for certain heavy-tailed distributions. Other naturally important directions involve nonhomogeneous Markov processes, e.g., Dietz and Sethuraman (2005), and Markov processes in a random environment, e.g., Seppalainen (1994).

The following proposition provides a somewhat more friendly "operator norm" lower bound[5] on the large deviation rate than the precise rate derived from the spectral radius.

Proposition 21.6 *(Spectral Radius Bound).* Suppose that S is compact. For a bounded linear operator T acting on the Banach space $C_b(S)$,

[4] A generalization that includes results of Nagaev (1969) was recently given in Gantert et al. (2014).

[5] In fact, for bounded linear operators on a Banach space one has the Gelfand formula for the spectral radius as $\lim_{n\to\infty} ||T^n||^{\frac{1}{n}}$; see Chapter 5, Exercise 6 as an application of the subadditive ergodic theorem.

$$\lambda^+ \le ||T||_{op} := \sup_{||f||_u=1} ||Tf||_u,$$

where $||f||_u = \sup_{x\in S} |f(x)|$.

Proof.

$$\lambda^+||f^+||_u = ||\lambda^+ f^+||_u = ||Tf^+||_u.$$

Thus, dividing by $||f^+||_u$,

$$\lambda^+ \le \sup_{||f||_u=1} ||Tf||_u = ||T||_{op}.$$

■

Example 1 *(Two-State Markov Chain).* This example illustrates the nature of explicit computations involved in this theory. Consider $S = \{-1, 1\}$ and transition probabilities $p_{-1,-1} = p = p_{1,1}$, $p_{-1,1} = q = 1 - p = p_{1,-1}$, having invariant probability $\pi = (\frac{1}{2}, \frac{1}{2})$ with mean $m = 0$. Let $v(y) = y, y \in \{-1, 1\}$. Then $T_{q_h} = \begin{pmatrix} pe^{-h} & qe^{h} \\ qe^{-h} & pe^{h} \end{pmatrix}$ has eigenvalue of maximum modulus

$$\lambda^+(h) = p\cosh(h) + \sqrt{p^2\cosh^2(h) - p^2 + q^2},\ h > 0. \tag{21.23}$$

Of course if $p = 1/2$, then this is the familiar moment generating function $\cosh(h)$ for the i.i.d. case. The computation of the large deviation rate function

$$I(a) = -\inf_{h>0}\{\log\lambda^+(h) - ah\} = \sup_{h>0}\{ah - \log\lambda^+(h)\},\ 0 < a < 1, \tag{21.24}$$

generally requires numerical approximation methods.[6] For the operator norm bound, consider

$$\begin{aligned} \det(T'_{q_h}T_{q_h} - \rho I) &= \det\begin{pmatrix} (p^2+q^2)e^{-2h} - \rho & 2pq \\ 2pq & (p^2+q^2)e^{2h} - \rho \end{pmatrix} \\ &= \rho^2 - 2(p^2+q^2)\cosh(2h)\rho + (p^2 - q^2). \end{aligned} \tag{21.25}$$

Thus, the eigenvalue of maximal magnitude for $T'_{q_h}T_{q_h}$ is

$$\rho(h) = (p^2+q^2)\cosh(2h) + \sqrt{(p^2+q^2)^2\cosh^2(2h) - (p^2-q^2)^2},\ h > 0. \tag{21.26}$$

So, (Exercise 8)

[6] MATLAB has a routine for numerically computing Legendre transforms.

$$||T_{q_h}||_{op} = \sqrt{(p^2+q^2)\cosh(2h) + \sqrt{(p^2+q^2)^2\cosh^2(2h) - (p^2-q^2)^2}}. \tag{21.27}$$

In particular, the Legendre transform calculation is not computationally simplified using this bound. In the special case $p = 1/2$ these calculations reduce to

$$\lambda^+(h) = \cosh(h) < \sqrt{\cosh(2h)} = ||T_{q_h}||_{op}, \quad h > 0.$$

In particular, the lower bound on the large deviation rate derived from the operator norm is precisely $\frac{1}{2}I(a)$ in this case (Exercise 8).

Example 2. Consider the random dynamical system

$$X_{n+1} = g(X_n) + \sigma Z_{n+1},\ n = 0, 1, 2 \ldots, \tag{21.28}$$

where $Z_1, Z_2, \ldots$ is an i.i.d. standard normal sequence, $\sigma > 0$, and g is a bounded function on $S = \mathbb{R}$. Then $p(x, dy) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(y - g(x))^2\}dy, x \in S$. In this case, after completing the square in the exponent, one has for $f \in C_0(\mathbb{R})$, $v(y) = y, y \in S$,

$$T_{q_h} f(x) = e^{\frac{1}{2}\sigma^2 h^2 + g(x)h}\mathbb{E}f(\sigma Z + \sigma^2 h + g(x)), \tag{21.29}$$

where Z is a standard normal random variable. Thus,

$$||T_{q_h}||_{op} \le e^{h||g||_\infty + \frac{1}{2}\sigma^2 h^2},$$

so that $\log\lambda^+(h) - ah \le \frac{1}{2}\sigma^2 h^2 + h(||g||_\infty - a)$. Thus, for $a > ||g||_\infty$,

$$\inf_{h>0}\{\log\lambda^+(h) - ah\} \le -\frac{(a - ||g||_\infty)^2}{2\sigma^2}.$$

Thus, $I(a) = \sup_{h>0}\{ah - \log\lambda^+(h)\} \ge \frac{(a-||g||_\infty)^2}{2\sigma^2}$, for $a > ||g||_\infty$.

For i.i.d. random variables a generalization of large deviations of sample averages from the mean of the underlying distribution can also be formulated[7] in terms of large deviations of the empirical distribution function from that of the underlying distribution. This leads to the next topic. To focus on the concepts rather than the generality, we first consider i.i.d. random variables taking finitely many values.

Let S be a finite set, say $S = \{1, 2, \ldots, r\}$. Consider the set $\mathcal{P}$ of all probability measures on S with full support. Given a sequence of i.i.d. random variables $X_1, \ldots, X_n$ with values in S having common distribution μ, let $E_n = \frac{1}{n}(\delta_{X_1} +$

[7] Sanov (1957).

$\cdots + \delta_{X_n})$ be the empirical distribution. Denote by $\mathcal{P}_n$ the set of probabilities in $\mathcal{P}$ in the range of E_n (as a map on Ω). Note that an element of $\mathcal{P}_n$ corresponds to n observations partitioned into r groups or less. One may show by a simple combinatorial argument that the number of such partitions is no more than $(n+1)^r$ (Exercise 26). Then one has for $\nu \in \mathcal{P}_n$

$$\begin{aligned} P_\mu(E_n = \nu) &= \left[\prod_{i=1}^{r} \mu(i)^{n\nu(i)}\right] \#J_q(\nu) \\ &= \exp\left\{\sum_{i=1}^{r} n\nu(i)\log\mu(i)\right\} \#J_q(\nu) \\ &= \exp\{-n[D(\nu||\mu) + H(\nu)]\}\,\#J_q(\nu), \ \ (\nu \in \mathcal{P}_n), \end{aligned} \tag{21.30}$$

where (i) $J_q(\nu)$ stands for the *type* of ν, defined by the set of all permutations of balls marked $1, \ldots, n$, distributed in r boxes such that the number of balls in box i is $n\nu(i)$ $(i = 1, \ldots, r)$, (ii) $H(\nu) = -\sum_i \nu(i)\log\nu(i)$ is the *(Shannon) entropy* of ν, and (iii) $D(\nu||\mu) = \sum_i \nu(i)\log[\nu(i)/\mu(i)]$ is the *relative entropy* of ν with respect to μ, also called the *Kullback–Liebler divergence*. We will show that

$$(n+1)^{-r}\exp\{nH(\nu)\} \le \#J_q(\nu) \le \exp\{nH(\nu)\}. \qquad \nu \in \mathcal{P}_n$$

For the right side use (21.30) to get, since $D(\nu||\nu) = 0$,

$$1 \ge P_\nu(E_n = \nu) = \exp\{-nH(\nu)\}\#J_q(\nu). \qquad \nu \in \mathcal{P}_n$$

For the left side,

$$1 = \sum_{\gamma \in \mathcal{P}_n} P_\nu(E_n = \gamma) \le P_\nu(E_n = \nu)(n+1)^r = (n+1)^r \exp\{-nH(\nu)\}\#J_q(\nu), \tag{21.31}$$

using ($\nu \in \mathcal{P}_n$), the fact that $P_\nu(E_n = \nu)$ *is the maximum among* $P_\nu(E_n = \gamma)$ *over all* $\gamma \in \mathcal{P}_n$ (see Lemma below), and that $\#\mathcal{P}_n$ is less than the number of ways n balls may be distributed in r boxes. We then have from (21.30) and ($\nu \in \mathcal{P}_n$),

Theorem 21.7 *(Method of Types).*

$$(n+1)^{-r}\exp\{-nD(\nu||\mu)\} \le P_\mu(E_n = \nu) \le \exp\{-nD(\nu||\mu)\}.$$

Theorem 21.8 *(Sanov's Theorem).* (a) For every closed subset F of $\mathcal{P}$,

$$\limsup_{n\to\infty} \frac{1}{n}\log P_\mu(E_n \in F) \le -\inf_{\nu \in F} \exp\{D(\nu||\mu)\}.$$

(b) For every open subset G of $\mathcal{P}$,

$$\liminf_{n\to\infty}\frac{1}{n}\log P_\mu(E_n\in G)\geq -\inf_{\nu\in G}\exp\{D(\nu||\mu)\}.$$

Proof. For the proof of part (a), one has from Theorem 21.7,

$$\begin{aligned}P_\mu(E_n\in F)&=\sum_{\nu\in F\cap\mathcal{P}_n}P_\mu(E_n=\nu)\\&\leq\sum_{\nu\in F\cap\mathcal{P}_n}\exp\{-nD(\nu||\mu)\}\\&\leq\#\mathcal{P}_n\sup_{\nu\in F}\exp\{-nD(\nu||\mu)\}\\&=\#\mathcal{P}_n\exp\{\sup_{\nu\in F}-nD(\nu||\mu)\}\\&=\#\mathcal{P}_n\exp\{-n\inf_{\nu\in F}D(\nu||\mu)\}\\&\leq(n+1)^r\exp\{-n\inf_{\nu\in F}D(\nu||\mu)\}.\end{aligned}\tag{21.32}$$

By taking logs on both sides and then dividing by n, the result follows. To prove part (b) we will use the facts (i) $\cup_{n\geq r}\mathcal{P}_n$ is dense in $\mathcal{P}$, and (ii) $\nu\to D(\nu||\mu)$ is continuous in the weak topology on $\mathcal{P}$. Therefore, given μ, there exists a sequence $\nu_n\in G\cap\mathcal{P}_n$, such that $D(\nu_n||\mu)\to\inf_{\nu\in G}D(\nu||\mu)$. Hence, by the method of types theorem, one has for all sufficiently large n,

$$P_\mu(E_n\in G)=\sum_{\nu\in G}P_\mu(E_n=\nu)\geq P_\mu(E_n=\nu_n)\geq(n+1)^{-r}\exp\{-nD(\nu_n||\mu)\}.\tag{21.33}$$

So that

$$\liminf_{n\to\infty}\frac{1}{n}\log P_\mu(E_n\in G)\geq-\lim D(\nu_n||\mu)=-\inf_{\nu\in G}D(\nu||\mu).$$

■

It remains to prove the italicized statement that was used after (21.31).

Lemma 2.

$$P_\nu(E_n=\nu)=\max_{\gamma\in\mathcal{P}_n}P_\nu(E_n=\gamma).$$

Proof. Let $\gamma\in\mathcal{P}_n$. Then,

$$P_\nu(E_n=\nu)=\left(n!/\left[\prod_{i=1}^r(n\nu(i))!\right]\right)\prod_{i=1}^r\nu(i)^{n\nu(i)},\tag{21.34}$$

$$P_\nu(E_n = \gamma) = \left(n!/\left[\prod_{i=1}^{r}(n\gamma(i))!\right]\right)\prod_{i=1}^{r}\nu(i)^{n\gamma(i)}. \tag{21.35}$$

Therefore,

$$\begin{aligned}
&P_\nu(E_n = \nu)/P_\nu(E_n = \gamma)\\
&= \left(\prod_{i=1}^{r}[(n\gamma(i))!]\right)/([(n\nu(i))!])\prod_{i=1}^{r}(\nu(i)^{n\nu(i)-n\gamma(i)})\\
&\geq \prod_{i=1}^{r}(n\nu(i))^{n(\gamma(i)-\nu(i))}\prod_{i=1}^{r}(\nu(i)^{n\nu(i)-n\gamma(i)}) \text{ (note } k!/m! \geq m^{k-m})\\
&= n^{n\sum_i(\nu(i)-\gamma(i))} = n^0 = 1.
\end{aligned}$$

■

Remark 21.13. It may be noted that part (a) of Theorem 21.8 holds for all measurable sets F, not just closed F. The proof of part (b), however, requires that G be open.

Next we turn to the problem of obtaining *large deviation rates for empirical measures of Markov chains.* This includes the Donsker–Varadhan extension to Markov chains of Sanov's Theorem 21.8 for large deviations of empirical measures for i.i.d random variables. We begin with a definition. From this point on $\lambda^+(v)$ will denote the maximum eigenvalue of the operator T_q with kernel $e^{v(y)}p(x, dy)$.

Definition 21.2. A sequence $P_n, n \geq 1$, of probability measures on (the Borel sigma-field of) S is said to satisfy the large deviation principle (LDP) with a rate function I if

$$\limsup_{n\to\infty}\frac{1}{n}\log P_n(C) \leq -\inf(I(x) : x \in C) \text{ for all closed } C \subset S, \tag{21.36}$$

and

$$\liminf_{n\to\infty}\frac{1}{n}\log P_n(O) \geq -\inf(I(x) : x \in O) \text{ for all open } O \subset S, \tag{21.37}$$

where $I : S \to [0, \infty]$ is lower semicontinuous, i.e., satisfies: $\{x : I(x) \leq d\}$ is closed for every $d \geq 0$.

We continue to assume that the state space S is a locally compact and σ-compact Polish space, and a Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ with state space S defined on a probability space $(\Omega, \mathcal{F}, P)$ has the transition probability $p(x, dy)$. Let $L(S)$ be the space of lower semicontinuous functions v bounded below such that the transition operator T_q with kernel $q = q(x : v) = p(x, dy)\exp\{v(y)\}$ is compact and the hypothesis of Theorem 21.1 is satisfied. The space $\mathcal{P}$ of probability measures on S, endowed with the weak (weak-star) topology is a Polish space. Let $E_n = n^{-1}\sum_{i=1}^{n}\delta_{X_i}$ denote the empirical measure of the Markov process. We plan to establish an LDP for the distributions P_n of E_n, with initial state x. The probabilities $P_x(\cdot)$ expectations $\mathbb{E}_x(\cdot)$ are computed under the given Markov process, with initial value x.

One has, from Corollary 21.2 with $F(\gamma) = \int_S v d\gamma$,

$$\int_S \exp\{nF\}dP_n = \mathbb{E}_x \exp\{nF(E_n)\} = \mathbb{E}_x \exp\{v(X_1) + \cdots + v(X_n)\},$$

and

$$\limsup \frac{1}{n} \log \int_S \exp\{nF\}dP_n = \log \lambda^+(v).$$

Writing $G = F - \log \lambda^+(v)$, one then has

$$\limsup \frac{1}{n} \log \int_S \exp\{nG\}dP_n = 0. \tag{21.38}$$

For any Borel subset B of $\mathcal{P}$ one then has $P_n(B) \leq \exp\{\sup(-nG(\gamma) : \gamma \in B)\}\int_S \exp\{nG\}dP_n = \exp\{-\inf[nG(\gamma) : \gamma \in B]\}\int_S \exp\{nG\}dP_n$, $\limsup_n \frac{1}{n}\log P_n(B) \leq -\inf\{G(\gamma) : \gamma \in B\} = -\inf[\int_S v d\gamma - \log\lambda^+(v) : \gamma \in B]$. Optimizing over all such $v \in L(S)$, one gets

$$\begin{aligned}\limsup \frac{1}{n} \log P_n(B) &\leq \inf\{-\inf[\int_S v d\gamma - \log \lambda^+(v) : \gamma \in B] : v \in L(S)\}\\ &= -\sup\{\inf[\int_S v d\gamma - \log \lambda^+(v) : \gamma \in B] : v \in L(S)\}.\end{aligned} \tag{21.39}$$

The following lemma shows that the supremum and infimum in the last line may be interchanged for compact B.

Lemma 3. If C is compact, then, writing $I(\gamma) = \sup\{\int_S v(y)\gamma(dy) - \log\lambda^+(v) : v \in L(S)\}$, one has

$$\limsup_{n\to\infty} \frac{1}{n} \log P(E_n \in C) \leq -\inf\{I(\gamma) : \gamma \in C\}.$$

Proof. First note that the inequality (21.39) holds for all measurable sets in $\mathcal{P}$, not just compact C. Given $\gamma \in C$ with $I(v) < \infty$ and $\epsilon > 0$, find $v_\gamma \in L(S)$ such that $\int_S v_\gamma(y)\gamma(dy) - \log\lambda^+(v_\gamma) > I(\gamma) - \epsilon$. Note that I is lower semicontinuous (Exercise 11), and $O_\gamma = \{\gamma' : I(\gamma') > I(\gamma) - \epsilon\}$ is an open neighborhood of

γ. Since C is compact, there now exists a finite set $\{\gamma_1, \gamma_2, \ldots, \gamma_m\}$ such that C is contained in $\cup_{i=1}^m O_{\gamma_i}$.

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log P(E_n \in C) &\leq \limsup_{n\to\infty} \frac{1}{n} \log[m \max\{P(E_n \in O_{\gamma_i}) : i = 1, .., m)\} \\ &= \limsup_{n\to\infty} \frac{1}{n} \log \max\{P(E_n \in O_{\gamma_i}) : i = 1, .., m\} \\ &= \limsup_{n\to\infty} \max[\frac{1}{n} \log P(E_n \in O_{\gamma_i}) : i = 1, .., m].\end{aligned} \tag{21.40}$$

But, by (21.39), which holds for all measurable sets B,

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \\ &\leq \max_{i=1,\ldots,m} \left(-\sup\left\{\inf\left[\int_S v d\gamma - \log\lambda^+(v) : \gamma \in O_{\gamma_i}\right] : v \in L(S)\right\}\right) \\ &\leq \max_{i=1,\ldots,m} \{-I(\gamma_i) + \epsilon : i = 1, \ldots, m\} \\ &= -\min_{i=1,\ldots,m} |\{I(\gamma_i) - \epsilon : i = 1, \ldots, m\} \\ &= -\min\{I(\gamma_i) : i = 1, \ldots, m\} + \epsilon \leq -\inf[I(\gamma) : \gamma \in C] + \epsilon.\end{aligned} \tag{21.41}$$

The desired result follows by letting $\epsilon \downarrow 0$. ■

We have therefore arrived at the following. Although the supremum in the proposition below may be taken over the class $L(S)$, in view of Lemma 3, it turns out it is the same when taken over the smaller class $C_b(S)$. In any case, the upper bound surely holds over this smaller set.

Proposition 21.9. For compact C, the upper bound for the LDP holds for the sequence P_n of distributions of the empirical measures E_n, with rate function

$$I(\gamma) = \sup\{\int_S v(y)\gamma(dy) - \log\lambda^+(v) : v \in C_b(S)\}.$$

For the upper bound for all closed sets C we need an additional condition defined as follows.

Definition 21.3. A sequence of probability measures $P_n, n = 1, \ldots,$ on a metric space S is said to be *e*xponentially tight if for every d, however, large, there exists a compact set K_d such that $\limsup \frac{1}{n} \log P_n(S \backslash K_d) \leq -d$.

Lemma 4. Suppose (21.36) holds, with $S = \mathcal{P}$, for all compact $C \subset \mathcal{P}$. If, in addition, $\{P_n : n = 1, 2, \ldots\}$ is exponentially tight, then the upper bound holds for all closed $C \subset \mathcal{P}$.

Proof. Let C be a closed subset of $\mathcal{P}$. With K_d as in Definition 21.3 for the metric space $\mathcal{P}$. One has,

$$\begin{aligned}
\limsup_{n\uparrow\infty}\frac{1}{n}\log P_n(C) &= \limsup\frac{1}{n}\log[P_n(C\cap K_d)+P_n(C\cap K_d^c)]\\
&\le \limsup_n\frac{1}{n}\log(2\max[P_n(C\cap K_d),\,P_n(K_d^c)])\\
&= \limsup_n\frac{1}{n}\log(\max[P_n(C\cap K_d),\,P_n(K_d^c)])\\
&\le \max(\limsup_n\frac{1}{n}\log P_n(C\cap K_d),\,-d)\\
&\le \max(-\inf\{I(\gamma):\gamma\in C\cap K_d\},\,-d)\\
&\le \max(-\inf\{I(\gamma):\gamma\in C\},\,-d). \qquad (21.42)
\end{aligned}$$

The desired result follows by letting $d\uparrow\infty$. ■

A simple criterion for exponential tightness is the following.

Lemma 5. A sequence $\{P_n\}$ of distributions on S is exponentially tight if there exists a function $F : S \to \mathbb{R}$ with the properties (i) $\{x : F(x)\le d\}$ is compact for all $d>0$, and (ii) $b := \limsup_{n\to\infty}\frac{1}{n}\log\int_S \exp\{nF\}dP_n < \infty$.

Proof. Assume (i),(ii). For a given $d>0$, choose $d' = b+d$. Clearly, $P_n(F > d') \le \exp\{-nd'\}\int_s \exp\{nF\}dP_n$. Therefore, $\limsup_{n\to\infty}\frac{1}{n}\log P_n(F > d') \le -d'+b = -d$. Now let $K_d = \{x\in S : F(x)\le d'\}$. ■

We will assume exponential tightness for $\{P_n\}$ for now, illustrating it for the classical i.i.d. case of Sanov later. To avoid possible confusion, we denote the operator with kernel $q(x, dy)$ by $\tilde{T}_q, whileT_pf(x) = \int_S f(y)p(x, dy)$. In particular, given a transition probability $p(x, dy)$, let $q(x, dy) = e^{v(y)}p(x, dy)$. For the lower bound in (21.37), consider a lower semicontinuous function f on the state space S, $f(x) > 0$ a.e. with respect to a finite nonzero measure μ dominating the transition probability $p(x, dy) = p(x, y)\mu(dy)$. Assume that $T_pf(x) = \int_S f(y)p(x, dy)$ is finite, and the kernel q,with $v(y) = \log\frac{f(y)}{T_pf(y)}$ satisfies the hypothesis of Theorem 21.1. Then,

$$q(x, dy) = p(x, dy)\exp\{v(y)\} = p(x, dy)\frac{f(y)}{T_pf(y)}. \qquad (21.43)$$

$$\int_S q(x, dy)T_pf(y) = \int_S p(x, dy)f(y) = T_pf(x), \qquad (21.44)$$

or

$$\tilde{T}_q T_p f(x) = T_p f(x). \tag{21.45}$$

That is, $T_p f(x)$ is an eigenfunction of the $\tilde{T}_q$, and

$$\begin{aligned}\widetilde{p}(x, dy) &= p(x, dy) \exp\{v(y)\} \frac{T_p f(y)}{T_p f(x)} \\ &= p(x, dy) \frac{f(y)}{T_p f(x)},\ (v(y) := \log \frac{f(y)}{T_p f(y)}).\end{aligned} \tag{21.46}$$

Denote by $\gamma(f)$ the unique invariant probability of the Markov process with transition probability $\widetilde{p}$ in (21.46). Let $\gamma = \gamma(f) \in O$, where O is an open subset of $\mathcal{P}$. Tilting the distribution to the Markov process with transition probability $\widetilde{p}$, one has

$$P_n(O) = P(E_n \in O) = \int_S R_n \mathbf{1}\left[\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \in O\right] d\widetilde{Q}_n, \tag{21.47}$$

where $\widetilde{Q}_n$ is the distribution of the Markov process $(\widetilde{X}_0 = x, \widetilde{X}_1, \ldots, \widetilde{X}_n)$ with transition probability $\widetilde{p}$, and Q_n that of $(X_0 = x, X_1, \ldots, X_n)$ with transition probability p. Then, by Lemma 1, relation(21.12),

$$R_n = \frac{dP_n}{d\widetilde{P}_n} = \frac{f^+(\tilde{x}_n)\lambda^+(0)^n}{f^+(\tilde{x}_0)} \exp\left\{-\sum_{j=1}^{n} v(\tilde{x}_j)\right\}. \tag{21.48}$$

Since O is open and $\gamma \in O$, under $\widetilde{P}_n$ the indicator function in (21.47) converges in probability to 1, as $n \to \infty$. Hence

$$\begin{aligned}\frac{1}{n}\log P_n(O) &= \frac{1}{n}\log \widetilde{\mathbb{E}} \prod_{1 \le i \le n} \frac{T_p f(\widetilde{X}_{i-1})}{f(\widetilde{X}_i)} + o(1) \\ &\ge \widetilde{\mathbb{E}} \frac{1}{n} \sum_{1 \le i \le n} \log \frac{T_p f(\widetilde{X}_{i-1})}{f(\widetilde{X}_i)} + o(1),\end{aligned} \tag{21.49}$$

where $\widetilde{\mathbb{E}}$ denoting expectation under $\widetilde{P}_n$. By ergodicity under $\widetilde{p}$, $\gamma(f)$, (21.49) converges to

$$\begin{aligned}&\int_S \int_S \log \frac{T_p f(x)}{f(y)} \gamma(f)(dx) \widetilde{p}(x, dy) \\ &\quad = \int_S \int_S (\log(T_p f(x))\gamma(f)(dx) - \int_S \int_S (\log f(y))\gamma(f)(dx)\widetilde{p}(x, dy)\end{aligned}$$

$$= \int_S (\log(T_p f(x))\gamma(f)(dx) - \int_S \log f(y)\gamma(f)(dy)$$

$$= \int_S (\log \frac{T_p f(x)}{f(x)})\gamma(f)(dx), \tag{21.50}$$

using (i) $\int_S p(x, dy) = 1$ and (ii) $\int_S \widetilde{p}(x, dy)(\gamma(f)(dx)) = \gamma(f)(dy)$. Thus $\liminf \frac{1}{n} \log P_n(O) \geq \int_S (\log \frac{T_p f(x))}{f(x)})\gamma(f)(dx)$. Since this is true for every $\gamma = \gamma(f) \in O$, one gets

$$\begin{aligned}&\liminf_{n\to\infty} \frac{1}{n} \log P_n(O)\\ &\quad \geq \sup\left\{\int_S \left(\log \frac{T_p f(x)}{f(x)}\right)\gamma(f)(dx) : f > 0, T_p f < \infty, \gamma = \gamma(f) \in O)\right\}\\ &\quad = -\inf\left\{\int_S \log \frac{f(x)}{T_p f(x)}\gamma(f)(dx) : f > 0, T_p f < \infty, \gamma = \gamma(f) \in O)\right\}.\end{aligned} \tag{21.51}$$

We have mostly arrived at the following main result.

Theorem 21.10 *(Donsker–Varadhan Large Deviation Theorem for Markov Processes).* Under the hypothesis of Theorem 21.1, for all kernels $q(x, dy) = p(x, dy) \exp\{v(y)\}$, the following LDP holds (a) For all closed set C,

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \leq -\inf\{I(\gamma) : \gamma \in C\}, \tag{21.52}$$

where $I(\gamma)$ is given in Proposition 21.9
(b) For all open sets O,

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n(O) \geq -\inf\{I(\gamma) : f > 0,\ f \text{ measurable},\ T_p f < \infty,\ \gamma = \gamma(f) \in O\}, \tag{21.53}$$

where $I(\gamma(f))$ is given by

$$I(\gamma) = I(\gamma(f)) = \int_S (\log[f(x)/T_p f(x)])\gamma(f)(dx). \tag{21.54}$$

Proof. Note that (21.53) is the same as (21.51). Part (a) follows from Proposition (21.9) and exponential tightness for which we refer to Rezakhanlou (2017).[8] ■

[8] Rezakhanlou (2017).

Remark 21.14. The rate function for the LDP in Theorem 21.10 is that given in Proposition 21.9. That this rate function holds for open sets O is not proved here. For finite state Markov chains a proof, using variational arguments, may be found in Varadhan (2008). For the general case we refer to Donsker and Varadhan (1975–1983), or the comprehensive presentation of their results by Rezakhanlou (2017). We will derive it for the i.i.d. case in Sanov's theorem.

The following interesting and useful lemma by Donsker and Varadhan shows that the large deviation rate $I(\gamma)$ in Sanov's Theorem is the same as the Kullback–Liebler divergence $D(\nu||\mu)$, where μ is the common distribution of the i.i.d. random variables.

Lemma 6. Let $D(\nu||\mu) = \int_S (\log(\frac{d\nu}{d\mu}))d\nu$ the Kullback–Liebler divergence of ν with respect to μ, μ and ν being probability measures on the Borel σ-field $\mathcal{S}$ of S. The following relations hold:

$$\begin{aligned} D(\nu||\mu) &= \sup_{f\in C_b(S)} \left[\int_S f d\nu - \log\left(\int_S e^f d\mu\right)\right] \\ &= \sup_{f\in B_b(S)} \left[\int_S f d\nu - \log\left(\int_S e^f d\mu\right)\right], \end{aligned} \tag{21.55}$$

where $B_b(S)$ is the set of all bounded Borel measurable functions on S.

Proof. Let[9] us write the first supremum in (21.55) as $I_1(\nu)$, and the second one as $I_2(\nu)$. To show that $D(\nu||\mu) \geq I_2(\nu)$, let $\frac{d\nu}{d\mu} = h$ for some measurable function h. Note that if ν is not absolutely continuous with respect to μ, then $D(\nu||\mu) = \infty$. One has, for every $f \in B_b(S)$,

$$\begin{aligned} \int f d\nu - D(\nu||\mu) &= \int h(f - \log h)d\mu \\ &= \log\exp\{\int (f - \log h)hd\mu\} \\ &\leq \log\left[\int \exp\{f - \log h\}hd\mu\right] \\ &= \log\int \exp\{f\}d\mu, \end{aligned}$$

proving $D(\nu||\mu) \geq I_2(\nu)$. Next, to prove $I_1(\nu) = I_2(\nu)$ one uses, for each $\epsilon > 0$, Lusin's theorem[10] approximating $f \in B_b(S)$ by a continuous function f_ϵ such that $||f_\epsilon|| \leq ||f||$, and $(\mu + \nu)\{x : f_\epsilon(x) \neq f(x)\} \leq \epsilon$. In particular, $|\int(f - f_\epsilon)d\nu| \leq$

[9] The proof follows Rezakhanlou (2017).

[10] Folland (1984), p. 211.

$2\epsilon||f||$,

$$\int f_\epsilon d\nu \le \log\left(\int_S \exp\{f_\epsilon\}d\mu\right) + I_1(\nu),$$

and,

$$\begin{aligned}\int f d\nu &= \int f_\epsilon d\nu + \int (f - f_\epsilon)d\nu \\ &\le \int f_\epsilon d\nu + 2\epsilon||f|| \le 2\epsilon||f|| + I_1(\nu) + \log\{\int \exp\{f_\epsilon\}d\mu\} \\ &\le 2\epsilon||f|| + I_1(\nu) + \log\left[\int \exp\{f - 2\epsilon||f||\}d\mu\right] \\ &= 4\epsilon||f|| + I_1(\nu) + \log\left[\int \exp\{f\}d\mu\right]. \end{aligned} \tag{21.56}$$

Letting $\epsilon \downarrow 0$, one arrives at the desired result $I_2(\nu) \le I_1(\nu)$ and hence $I_1(\nu) = I_2(\nu)$.

Finally, to prove $D(\nu||\mu) \le I_2(\nu)$, first assume that $I_2(\nu)$ is finite. In this case ν is absolutely continuous with respect to μ: $\nu << \mu$. To see this, let $\mu(B) = 0$ for some Borel set. Let $f = L\mathbf{1}_B$ in (21.56). We get (after letting $\epsilon \downarrow 0$), $L\nu(B) \le I_1(\nu)$ ($= I_2(\nu)$), which cannot hold for all L unless $\nu(B) = 0$. Assume that $h = \frac{d\nu}{d\mu}$ is bounded away from zero and infinity. Let $f = \log h$ in (21.56).

If $\log h$ is bounded, then (letting $\epsilon \downarrow 0$ in (21.56)) one has

$$\int \log h d\nu \le I_1(\nu) - \log \int h d\mu = I_2(\nu). \tag{21.57}$$

This proves the desired result if $\log h$ is bounded. In general, let

$$h_{\epsilon,M} = \begin{cases} \epsilon & \text{if } h \le \epsilon \\ h & \text{if } \epsilon < h < M \\ M & \text{if } h \ge M. \end{cases}$$

Then, by (21.56),

$$\int \log h_{\epsilon,M} d\nu \le I_1(\nu) + \log \int h_{\epsilon,M} d\mu.$$

By letting $\epsilon \downarrow 0$, and using the monotone convergence theorem, one obtains, with $h_M = \lim_{\epsilon\downarrow 0} h_{\epsilon,M}$,

$$\int \log h_M d\nu \le I_2(\nu) + \log \int h_M d\mu.$$

Next, let $M \uparrow \infty$, and using the dominated convergence theorem with $h_M \leq h$, one gets

$$\int \log h d\nu \leq I_1(\nu) + \log \int h d\mu.$$

In the case the supremum in $I_2(\nu)$ is not achieved, given $\delta > 0<$ chose $v = g_\delta$ such that $\int g_\delta d\nu - \log \int \exp\{g_\delta\} d\mu > I_2(\nu) - \delta$. ■

We now show that for i.i.d. random variables with common distribution μ the rate function given in Proposition 21.9, namely $D(\nu||\mu)$ applies to open sets in Theorem 21.10, as for closed sets. For this fix $\nu = hd\mu$, i.e., $\frac{d\nu}{d\mu} = h$. Suppose the supremum in

$$\sup\{\int v d\nu - \log \lambda^+(v) : v \in C_b(S)\} \tag{21.58}$$

is attained at v^*. Note that here $\lambda^+(v) = \int e^{v(y)} \mu(dy)$.

Let $g \in C_b(S)$. For all $\epsilon > 0$ one has

$$\int (v^* + \epsilon g) d\nu - \log \lambda^+(v^* + \epsilon g) \leq \int v^* d\nu - \log \lambda^+(v^*),$$

i.e.,

$$\epsilon \int g d\nu - \left[\log \lambda^+(v^* + \epsilon g) - \log \lambda^+(v^*)\right] \leq 0.$$

Dividing by ϵ and letting $\epsilon \downarrow 0$, one obtains (Exercise 22)

$$\int g d\nu \equiv \int g h d\mu \leq \int g e^{v^*} d\mu / \lambda^+(v^*). \tag{21.59}$$

Replacing g by $-g$, one gets

$$\int -g d\nu \equiv \int -g h d\mu \leq \int -g e^{v^*} d\mu / \lambda^+(v^*).$$

Therefore,

$$\int g d\nu \equiv \int g h d\mu = \int g e^{v^*} d\mu / \lambda^+(v^*).$$

Since this is true for all $g \in C_b(S)$, we get

$$h = e^{v^*} / \lambda^+(v^*) \, a.e.(\mu).$$

This says that $v^* = \log h$ attains the supremum in (21.58). (Note that $\lambda^+(\log h) \equiv \int e^{\log h} d\mu = 1$.) A similar but somewhat more elaborate argument holds for Markov processes. We refer to Varadhan (2008) for the case of finite state Markov chains, and Donsker and Varadhan (1975a, 1976, 1983), or Rezakhanlou (2017) for the general case. The proof of Theorem 21.10 (b) provides a hint for the computation of $I(\nu)$ for invariant probabilities ν of tilted chains with transition probabilities of the form $\tilde{p}(x, dy) = p(x, dy)\frac{f(y)}{T_p f(x)}$ for $f > 0$, $T_p f < \infty$.

For the next result for i.i.d. random variables with common distribution π, let

$$c(v) = \log \int_S \exp\{v(y)\}\pi(dy).$$

Corollary 21.11 *(Sanov's Theorem).* Consider a sequence of i.i.d. random variables $X_n (n = 1, 2, \ldots)$ with values in S, with common distribution π, and let P_n denote the distribution of the empirical $E_n = \frac{1}{n}\sum_{1\le i\le n} \delta_{x_i} (n = 1, 2, \ldots)$. Then the sequence $\{P_n\}$ satisfies the LDP with rate function

$$I(\gamma) = \sup\left\{\int_S v d\gamma - c(v) : v \in C_b(S)\right\} = D(\nu||\pi). \tag{21.60}$$

Proof. The result follows immediately from Theorem 21.10, with exponential tightness derived in Proposition 21.12, and Lemma 6. Remark 21.11 shows that the same $I(\gamma)$ applies to open sets as well. ■

Remark 21.15. It may be shown that $I(\gamma)$ in (21.60) equals the *r*elative entropy, or *K*ullback–Liebler divergence $D(\gamma||\pi)$.

Remark 21.16. If one specializes (21.54) (21.53) to the i.i.d. cases, one readily shows that $\gamma(f)$ is the probability $\frac{f d\pi}{\int f d\pi}$, and $I(\gamma(f)) = D(\gamma(f)||\pi)$. Thus only those γ are considered which are absolutely continuous with respect to π. All other γ such that $D(\gamma||\pi) = \infty$, and, therefore, can be omitted.

Proposition 21.12. In the case of i.i.d. random variables the sequence $\{P_n\}$ of distributions of empirical measures is exponentially tight.

Proof. Let $K_j (j = 1, 2, \ldots)$ be compact subsets of S such that $\pi(K_j^c) \le \exp\{-j^2\}$, where π is the common distribution of the i.i.d. sequence. Define

$$D_j = \{\gamma \in \mathcal{P} : \gamma(K_j^c) \le j^{-1}\} \qquad (j = 1, 2, \ldots).$$

Then the sets D_j are closed and the sets $C_d = \bigcap_{j=1}^{\infty} D_j$ are compact, by Prokhorov's theorem (Exercise 12). One then has

$$\begin{aligned}
P_n(C_d) &= P_n\left(\bigcup_{j=d}^{\infty} D_j^c\right) = P\left(E_n \in \bigcup_{j=d}^{\infty} D_j^c\right) \\
&\leq \sum_{j=d}^{\infty} P(E_n \in D_j^c) = \sum_{j=d}^{\infty} P(E_n(K_j^c) \geq j^{-1}) \\
&= \sum_{j=d}^{\infty} P(nj^2 E_n(K_j^c) \geq nj) \\
&= \sum_{j=d}^{\infty} e^{-nj}\mathbb{E}\left(\exp\{njE_n(K_j^c)\}\right) \qquad \text{(by Chebyshev's inequality)} \\
&= \sum_{j=d}^{\infty} e^{-nj}\left[\sum_{r=0}^{n}\exp\{nj^2\frac{r}{n}\}\cdot P(E_n(K_j^c) = \frac{r}{n})\right] \\
&= \sum_{j=d}^{\infty} e^{-nj}\left[\sum_{r=0}^{n}\exp\{rj^2\}\binom{n}{r}(\pi(K_j^c))^r\pi(K_j)^{n-r}\right] \\
&= \sum_{j=d}^{\infty} e^{-nj}\left[\exp\{j^2\}\pi(K_j^c) + \pi(K_j)\right]^n \\
&\leq \sum_{j=d}^{\infty} e^{-nj}[1+\pi(K_j)]^n \leq 2^n\sum_{j=d}^{n} e^{-nj} \leq 2^{n+1}e^{-nd}.
\end{aligned}$$

From this exponential tightness of $\{P_n\}$ follows (Exercise 13). ■

It is interesting to see how Sanov's Theorem 21.8 implies the Cramér–Chernoff theorem. In the case of bounded random variables, this is essentially a consequence of the following (Exercise 23)

Lemma 7 *(Contraction Principle).* Suppose the sequence $\{P_n\}$ of distributions of $E_n = \frac{1}{n}\sum_{m=1}^{n}\log\delta_{X_m}$ $\{P_n\}$ satisfies a LDP with rate function I, on (S,ρ). Let $\varphi : S \to S'$ be a continuous map into a Polish space (S',ρ'). Then the family $\{P_n' = P_n \circ \varphi^{-1}\}$ satisfies the LDP with rate function $I'(x') = \inf\{I(x) : \varphi(x) = x'\}$.

Corollary 21.13 *(Cramér–Chernoff).* Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^d$ having distribution π. Let $\lambda(u) = \log\mathbb{E}e^{u\cdot X_1} = \log\int_{\mathbb{R}^d} e^{u\cdot x}\pi(dx)$. Assume $\int_{\mathbb{R}^d} e^{x\cdot r}\pi(dx) < \infty$. Define

$$P_n(A) = P(\frac{1}{n}\sum_{j=1}^{n} X_j \in A), \quad A \in \mathcal{B}, n = 1, 2, \ldots,$$

where $\mathcal{B}$ is the Borel σ-field of $\mathbb{R}^d$.

Then $\{P_n : n \geq 1\}$ satisfies the LDP with rate function

$$I(x) = \lambda^*(x) = \sup_{u \in \mathbb{R}^d} (x \cdot u - \lambda(u)).$$

Proof. To use the contraction principle to prove Cramér–Chernoff we first restrict the proof to the case of bounded random variables, say $\pi(S) = 1$, $S = \{x \in \mathbb{R}^d : |x| \leq m\}$, and define a linear functional φ on $\mathcal{P} = \mathcal{P}(S)$ by

$$\varphi(\nu) = \int_S x\nu(dx), \quad \nu \in \mathcal{P}.$$

Apply Lemma 7 to $\mathcal{P}(S)$ in place of S.

Then, since $f(x) = x$ is bounded and continuous on S, φ is continuous for the weak*topology on $\mathcal{P}$. Let P_n denote the distribution of $\frac{1}{n}\sum_{j=1}^n X_j$, and define

$$P_n'(A) = P_n(\{\nu : \varphi(\nu) \in A\}, \quad A \in \mathcal{B}.$$

Equivalently,

$$\begin{aligned} P_n'(A) &= P\left(\frac{1}{n}\sum_{j=1}^n \delta_{X_j} \in \{\nu : \varphi(\nu) \in A\}\right) \\ &= P\left(\int_S \frac{1}{n}\sum x\delta_{X_j}(dx) \in A\right) \\ &= P\left(\frac{1}{n}\sum_{j=1}^n X_j \in A\right). \end{aligned} \tag{21.61}$$

It now follows from the contraction principle (Lemma 7) that $\{P_n'\}$ satisfies the LDP with rate

$$I'(a) = \inf_{\nu \in \mathcal{P}} \{D(\nu || \pi) : \varphi(\nu) = a\}. \tag{21.62}$$

So the objective is to show that I' may be equivalently expressed in the familiar form

$$I(a) = \sup_{u \in \mathbb{R}^d} (a \cdot u - \log \int_S e^{u \cdot x} \pi(dx)). \tag{21.63}$$

To show that $I' = I$ we will use the Minimax Lemma 21.14 given below in order to interchange a supremum with an infimum that occurs in the following re-expression of the large deviation rate I'. For this one simply notes the continuity of

$\nu \to D(\nu||\pi)$ and $\nu \to \int_S u \cdot x\nu(dx)$.

$$\begin{aligned} I'(a) &= \inf_{\nu\in\mathcal{P}}\{D(\nu||\pi) : \varphi(\nu) = a\} \\ &= \inf_{\nu\in\mathcal{P}} \sup_{u\in\mathbb{R}^d} \left\{D(\nu||\pi) - u \cdot \left(\int_S x\nu(dx) - a\right)\right\} \\ &= \inf_{\nu\in\mathcal{P}} \sup_{u\in\mathbb{R}^d} \left\{D(\nu||\pi) - \int_S u \cdot x\nu(dx) + u \cdot a\right\} \\ &= \sup_{u\in\mathbb{R}^d} \inf_{\nu\in\mathcal{P}} \left\{D(\nu||\pi) - \int_S u \cdot x\nu(dx) + u \cdot a\right\} \\ &= \sup_{u\in\mathbb{R}^d} \left\{-\log \int_S e^{u\cdot x}\pi(dx) + u \cdot a\right\} = I(a), \end{aligned} \tag{21.64}$$

where the first equality follows from the fact that the supremum is ∞ unless $\varphi(\nu) = a$, and the last equality follows because for any bounded, measurable function f on S one has $\log \int_S e^{f(x)}\pi(dx) = \sup_{\nu\in\mathcal{P}}\{\int_S f d\nu - D(\nu||\pi)\}$ (Exercise 27). To remove the boundedness assumption complete the steps in Exercise 24. ■

The exchange of the supremum with infimum used in the above proof involves a *minimax* formula due to Fan (1953). Another application of the minimax lemma occurs in the Special Topics Chapter 22. The simple proof given here is due to Borwein and Zhuang (1986). The lemma relies on a definition that captures the property of a function of two variables that is convex-like in one variable, and concave like in the other.

Definition 21.4. A function $f : X \times Y \to \mathbb{R}$ is said to be convex-concave like on $X \times Y$ if for $0 \le t \le 1$, (a) for $x_1, x_2 \in X$ there is an $x_3 \in X$ such that $f(x_3, y) \le tf(x_1, y) + (1 - t)f(x_2, y)$ for all $y \in Y$; and (b) for $y_1, y_2 \in Y$ there is a $y_3 \in Y$ such that $f(x, y_3) \ge tf(x, y_1) + (1 - t)f(x, y_2)$ for all $x \in X$.

To simplify notation write $\sup_Y \equiv \sup_{y\in Y}$, and similarly for $\min_X$.

Proposition 21.14 *(Fan's Minimax Formula).* Suppose that X, Y are nonempty sets with f convex-concave like on $X \times Y$. If X is compact and $f(\cdot, y)$ is lower semicontinuous on X for each $y \in Y$, then

$$p := \inf_X \sup_Y f(x, y) = \sup_Y \inf_X f(x, y).$$

Proof. If $p = -\infty$, then the result is trivial since $\inf_X \sup_Y f(x, y) \ge \sup_Y \inf_X f(x, y)$. So assume p is finite. Let $a \in \mathbb{R}$ with $a < p$. Since $K(y) = \{x \in X : f(x, y) \le a\}$ is compact for each y, and $\cap_{y\in Y} K(y) = \emptyset$, there exist $y_1, \ldots, y_n$ in Y such that $a < \min_X \sup_{1\le j\le n} f(x, y_j)$. Let

$$C = \{(z, r) = (z_1, \ldots, z_n, r) \in \mathbb{R}^n \times \mathbb{R} : \exists\, x \in X, f(x, y_j) \le r + z_j, j = 1, ..., n\}.$$

Then C is convex subset of $\mathbb{R}^{n+1}$ since $f(\cdot, y)$ is convex-like. Also, by construction, the point $(z, r) = (0, 1 + \max_{1\le j\le n} f(x, y_j))$ belongs to the interior C^o of C for any $x \in X$. Moreover, the point $(z, r) = (0, a) \notin C$ since $a < \min_X \sup_{1\le j\le n} f(x, y_j)$. By the separation theorem[11] there exists $(\lambda_1, \dots, \lambda_n, \overline{r}) \neq 0$, such that for $(z, r) \in C$, $z = (z_1, \dots, z_n)$, $r \in \mathbb{R}$,

$$\sum_{j=1}^{n} \lambda_j z_j + \overline{r}r \ge \overline{r}a.$$

Since $C + \mathbb{R}_+^{n+1} \subset C$, one has $\lambda_j \ge 0, \overline{r} \ge 0$. Moreover, clearly the point $(z, r) = (0, 1+\max_{1\le j\le n} f(x, y_j)) \in C^0$ so that, in fact, one has $\overline{r} > 0$. Thus, for all $x \in X$, $r \in \mathbb{R}$, writing $z_j = f(x, y_j) + r$, $1 \le j \le n$, $z = (z_1, \dots, z_n)$, then $(z, -r) \in C$, so that one has

$$\sum_{j=1}^{n} \frac{\lambda_j}{\overline{r}} f(x, y_j) + \left(\sum_{j=1}^{n} \frac{\lambda_j}{\overline{r}} - 1\right) r \ge a. \tag{21.65}$$

Thus, considering $r \to -\infty$, one must have

$$\sum_{j=1}^{n} \frac{\lambda_j}{\overline{r}} = 1. \tag{21.66}$$

Since $f(x, \cdot)$ is concave like, it follows from (21.66) and (21.65) that for $j = 1, \dots n$, some $y \in Y$, $f(x, y) \ge a$ for all $x \in X$. In particular, $\sup_Y \inf_X f(x, y) \ge a$. Since $a \le p$ is arbitrary, one has $\sup_Y \inf_X f(x, y) \ge \inf_X \sup_Y f(x, y)$. Now use lower semicontinuity of $f(x, y)$ and $\sup_Y f(x, y)$ on the compact set X to obtain the asserted equality. ■

Corollary 21.15. Suppose that X, Y are nonempty sets with f concave-convex like on $X \times Y$. If X is compact and $f(\cdot, y)$ is upper semicontinuous on X for each $y \in Y$, then

$$p := \sup_X \inf_Y f(x, y) = \inf_Y \sup_X f(x, y).$$

Proof. Note that since

$$\inf_X \sup_Y f(x, y) = -\sup_X(-\sup_Y f(x, y)) = -\sup_X \inf_Y(-f(x, y)),$$

[11] See BCPT, p. 12 for a proof in one-dimension.

the roles of infimum and supremum can be interchanged assuming the corresponding exchange of convex-concave like conditions to concave-convex like, and lower to upper semicontinuity conditions entailed in replacing f by $-f$. ■

The following proposition yields a simple variational formula as a corollary.

Proposition 21.16. Let g be a bounded measurable function, P, Q probability measures such that $P << Q$. Let $\lambda > 0$. Then,

$$\int_S g dP - \lambda D(P||Q) = \lambda \log \int_S e^{g/\lambda} dQ - \lambda D(P||P^*), \tag{21.67}$$

where

$$\frac{dP^*}{dQ} = \frac{e^{g/\lambda}}{\int_S e^{g/\lambda} dQ}. \tag{21.68}$$

Proof. The left side of the asserted equation may be expressed as follows:

$$\begin{aligned}\int_S g dP - \lambda D(P||Q) &= \int_S g dP - \lambda \int_S \log\left(\frac{dP}{dQ}\right) dP \\ &= \int_S g dP - \lambda \int_S \log\left(\frac{dP}{dP^*}\right) dP - \lambda \int_S \log\left(\frac{dP^*}{dQ}\right) dP \\ &= \int_S \left(g - \lambda \log\left(\frac{dP^*}{dQ}\right)\right) dP - \lambda D(P||P^*) \\ &= \int_S \left(g - \lambda \log\left(\frac{e^{\frac{g}{\lambda}}}{\int_S e^{g/\lambda} dQ}\right)\right) dP - \lambda D(P||P^*) \\ &= \int_S \left(g - \lambda \log\left(e^{\frac{g}{\lambda}}\right) + \lambda \log\left(\int_S e^{g/\lambda} dQ\right)\right) dP - \lambda D(P||P^*) \\ &= \int_S \left(g - \lambda g/\lambda + \lambda \log\left(\int_S e^{g/\lambda} dQ\right)\right) dP - \lambda D(P||P^*) \\ &= \int_S \lambda \left(\log \int_S e^{g/\lambda} dQ\right) dP - \lambda D(P||P^*) \\ &= \lambda \log \int_S e^{g/\lambda} dQ - \lambda D(P||P^*). \end{aligned} \tag{21.69}$$

This completes the proof. ■

Corollary 21.17 *(Donsker–Varadhan Variational Formula).* Under conditions of the theorem one has

$$\log \mathbb{E}_Q e^{\lambda g} = \max_{P \in \mathcal{P}} \big(\lambda \int_S g dP - D(P||Q)\big). \tag{21.70}$$

Moreover, the argmax P^* is unique.

Proof. Replace g by $\lambda^2 g$ in Proposition 21.16. Then,

$$\lambda \log \int_S e^{\lambda g} dQ - \lambda D(P||P^*) = \lambda^2 \int_S g dP - \lambda D(P||Q).$$

Thus, for all $P << Q$

$$\begin{aligned}\log \int_S e^{\lambda g} dQ &= \lambda \int_S g dP - D(P||Q) + D(P||P^*)\\ &\geq \lambda \int_S g dP - D(P||Q),\end{aligned} \tag{21.71}$$

with equality if and only if $P = P^*$ (Exercise 17). The asserted identity now follows. ■

As an application of the large deviation theory one may obtain the following consequence.

Theorem 21.18 *(Varadhan's Integral Formula).* Suppose that $\{X_n : n \geq 1\}$ satisfies a LDP on S with rate function I, and let $\varphi : S \to \mathbb{R}$ be a bounded continuous function. If the level sets $\{x \in S : I(x) \leq B\}$ are compact for all $B > 0$, then

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{n\varphi(X_n)} = \sup_{x\in S}\{\varphi(x) - I(x)\}. \tag{21.72}$$

Proof. We view the asserted equality as an upper bound and a lower bound on the left hand side. Assume $|\varphi(x)| \leq M$, for all $x \in S$ for some $M > 0$. For the lower bound fix $x_0 \in S$ and $\delta > 0$. Let $G = \{x \in S : \varphi(x) > \varphi(x_0) - \delta\}$. Then G is open since φ is continuous. Now,

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{n\varphi(X_n)} &\geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{n\varphi(X_n)} \mathbf{1}_{[X_n \in G]}\\ &\geq \varphi(x_0) - \delta + \liminf_{n\to\infty} \frac{1}{n} \log P(X_n \in G)\\ &\geq \varphi(x) - I(x_0) - \delta.\end{aligned} \tag{21.73}$$

The lower bound follows by letting $\delta \downarrow 0$ since $x_0 \in S$ is arbitrary.

For the upper bound, for each $N \geq 1$, partition the interval $[-N, M]$ into a finite number m of closed subintervals $J_{N,j}$ of length $\frac{1}{N}$, and let $F_{N,j} = \varphi^{-1}(J_{N,j})$. Then each $F_{N,j}$ is closed, the oscillation of φ on $F_{N,j}$ is at most $\frac{1}{N}$, and $\varphi \leq -N$ outside $\cup_{j=1}^{N} F_{N,j}$. Now,

$$\begin{aligned}
&\limsup_n \frac{1}{n}\log\int_S e^{n\varphi(x)}P(X_n\in dx)\\
&=\limsup_n \frac{1}{n}\log\left[\int_{\cup_{j=1}^m F_{N,j}} e^{n\varphi(x)}P(X_n\in dx)+\int_{\cap_{j=1}^m F_{N,j}^c} e^{n\varphi(x)}P(X_n\in dx)\right]\\
&\le \max_{1\le j\le m}\limsup_n \frac{1}{n}\log\left[\int_{F_{N,j}} e^{n\varphi(x)}P(X_n\in dx)\right]\vee(-N)\\
&\le \max_{1\le j\le m}\limsup_n \left[\frac{1}{n}\log e^{n\sup_{x\in F_{N,j}}\varphi(x)}+\frac{1}{n}\log P(X_n\in F_{N,j})\right]\vee(-N)\\
&\le \max_{1\le j\le m}\left[\sup_{x\in F_{N,j}}\varphi(x)-\inf_{x\in F_{N,j}}I(x)\right]\vee(-N)\\
&\le \max_{1\le j\le m}\sup_{x\in F_{N,j}}\left[\varphi(x)-I(x)+\frac{1}{N}\right]\vee(-N). \qquad (21.74)
\end{aligned}$$

Let $N\to\infty$ to obtain the desired upper bound. ■

An application of major elements of this theory to a problem in cryptography is provided in the Special Topics Chapter 22. Statistical physics provides another area of application that motivated aspects of the theoretical development of large deviation theory, e.g., see Ellis (1985), Dembo and Zeitouni (1998), Den Hollander (2008).

Exercises

1. Prove Corollary 21.2. [Hint: $q(x,S) = \mathbb{E}_x e^{v(X_1)}$, $q^{(2)}(x,S) = \int_S\int_S e^{v(y)}p(x,dy)e^{v(z)}p(y,dz) = \mathbb{E}_x e^{v(X_1)+v(X_2)}$ (conditioning on X_1 first). Now use induction. For the limit use (21.3), noting $f^+\le 1$, and f^+ bounded away from zero.]
2. Complete the induction following (21.7) .
3. Show that it is sufficient that $q_h^{(n_0)}$ satisfying hypothesis of the Perron–Frobenius theorem for some $n_0\ge 1$.
4. Prove the assertion under time reversibility in Remark 21.7 .
5. Prove the Corollary 21.4. [Hint:Compactness of T_q follows from the Arzela-Ascoli theorem (BCPT p. 244.)]
6. Show that λ^+ is a simple eigenvalue, i.e., the space of associated eigenfunctions is one-dimensional. [*Hint*: Assume that f is another eigenfunction and, for fixed $x\in S$, choose $g(y)=f(y)\frac{f^+(x)}{f^+(y)}$ in the definition of convergence in total variation 21.3. With this choice show that $f(x)=cf^+(x)$ for a constant $c>0$.

7. Suppose that T is an $n \times n$ real matrix, with transpose T^t. Show that

   (a) If $T = T^t$, then, letting $\mathcal{E}$ denote the set of eigenvalues of T, one has $||T||_{op} = \max_{\lambda \in \mathcal{E}} |\lambda|$. [*Hint*: Diagonalize T.
   (b) $||T||_{op} = \sqrt{||T^tT||_{op}}$. [*Hint*: Use the Euclidean norm and Cauchy–Schwarz to derive $||Tv||^2 = \langle Tv, Tv\rangle \le ||T^tTv|| \cdot ||v|| \le ||T^tT||_{op} \cdot ||v||^2$ to see that $||T||^2_{op} \le ||T^tT||_{op}$. Then apply Gelfand's formula to $||T^tT||_{op}$.
   (c) Show that $||T||_{op}$ is the square root of the largest of magnitudes of eigenvalues of $||T^tT||$. [*Hint*: Apply the above to T^tT.

8. This exercise is in reference to Example 1.

   (a) Compute $\lambda^+(h)$.
   (b) Use Exercise 7 to calculate the operator norm $||T_{q_h}||_{op}$.
   (c) In the case $p = q = 1/2$ compute $I(a), 0 < a < 1$, and show that the lower bound on the large deviation rate furnished by the operator norm is precisely $\frac{1}{2}I(a)$.

9. Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables with common distribution function G, and let $G_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{(-\infty,t]}(X_j)$ denote the empirical distribution function. (a) Show that for each fixed $t \in \mathbb{R}$, $G_n(t) \to G(t)$ as $n \to \infty$, *with probability one*. (b) Show that

$$\sup_{-\infty < t < \infty} |G_n(t) - G(t)| \to 0 \quad \text{in probability as } n \to \infty.$$

   (c) The Glivenko–Cantelli Theorem asserts that (b) in fact holds almost surely. Prove this by completing the following steps.

   (a) For each t, the event $[G_n(t^-) \to G(t^-)]$ has probability one.
   (b) Let $\tau(y) = \inf\{t : G(t) \ge y\}, 0 < y < 1$. Then $G(\tau(y)^-) \le y \le G(\tau(y))$.
   (c) Let $D_{m,n} = \max_{1 \le k \le m}\{|G_n(\tau(k/m)) - G(\tau(k/m))|, |G_n(\tau(k/m)^-) - G(\tau(k/m)^-)|\}$. Then, by considering the cases $\tau\left(\frac{k-1}{m}\right) \le t < \tau\left(\frac{k}{m}\right)$, $\quad t < \tau\left(\frac{1}{m}\right)$ or if $t \ge \tau(1)$, show that $\sup_t |G_n(t) - G(t)| \le D_{m,n} + \frac{1}{m}$.[*Hint*: Check that both $G_n(t) - G(t) \le D_{m,n} + \frac{1}{m}$ and $G_n(t) - G(t) \ge -D_{m,n} - \frac{1}{m}$ by using monotonicity, followed by adding and subtracting appropriate terms.
   (d) $C = \bigcup_{m=1}^{\infty}\bigcup_{k=1}^{m}[G_n(\tau(k/m)) \not\to (G(\tau(k/m))] \cup [G_n(\tau(k/m)^-) \not\to G(\tau(k/m)^-)]$ has probability zero, and for $\omega \in C^c$ and each in $m \ge 1$ $D_{m,n}(\omega) \to 0 \quad$ as $n \to \infty$.
   (e) $\sup_t |G_n(t,\omega) - G(t)| \to 0 \quad$ as $n \to \infty \quad$ for $\omega \in C^c$.

   (c) For an interpretation of Sanov's theorem in the context of Glivenko–Cantelli, show that for a closed set F not containing G one has $P(G_n \in F) \to 0$ as $n \to \infty$, and Sanov's theorem gives the rate.

10. Let $f : S \to (-\infty, \infty]$ be a lower semicontinuous function on a metric space (S, ρ), f bounded below. Prove that the functions $f_n(x) = \inf_y(f(y) + n\rho(x, y))$ are (i) non-decreasing, (ii) continuous, and (iii) $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in S_x$. [*Hint*: Assume without essential loss of generality that $f \geq 0$, (i) is obvious; for (ii) note that $|f_n(x) - f_n(z)| \leq \sup_y |n\rho(x, y) - n\rho(z, y)| \leq n\rho(x, z)$. To prove (iii), fix x and $\varepsilon > 0$; $\{y : f(y) > f(x) - \varepsilon\}$ is open, so it contains a ball $B(x, \delta)$ of center x and radius $\delta > 0$. If $y \in B(x, \delta)^c$, then for all sufficiently large n, $n\rho(x, y) \geq n\delta > f(x)$. Therefore, for such n, $f_n(x) > f(x) - \varepsilon$, i.e., $f(x) - \varepsilon < f_n(x) \leq f(x)$.
11. Show that $I(\gamma)$ defined in Lemma 3 is lower semicontinuous [*Hint*: For a continuous and bounded v, $\gamma \to \int v d\gamma$ is continuous on $\mathcal{P}$ in the weak-star topology. For $v \in L(S)$, let v_n be continuous $\uparrow v$, as in Exercise 10. For $M > 0$, $\gamma \to \int (v_n \wedge M) d\gamma$ is continuous. Letting $M \uparrow \infty$, $\gamma \to \int v_n d\gamma$ is lower semicontinuous and the increasing sequence of lower semicontinuous functions is lower semicontinuous.
12. Show that the derived inequality $P_n(C_d^c) \leq 2^{-(n+1)}e^{-nd}$ for compact C_d ($d = 1, 2, \ldots$) implies exponential tightness.
13. Show that in Theorem 21.1, $\lambda^+(v)$ is the largest eigenvalue of T_q as an operator on $C_b(S)$, and f^+ is the eigenfunction, unique up to a scalar multiple.
14. Consider the Ehrenfest birth–death chain X on $S = \{0, 2, \ldots, 2d\}$ with transition probabilities $p_{i,i+1} = \beta_i = \frac{2d-i}{2d}$, and $p_{i,i-1} = \delta_i = \frac{i}{2d}$, for $i = 0, 1, \ldots, 2d$. (a) Show that $\pi_j = \binom{2d}{j}2^{-2d}$ is a time-reversible invariant probability. (b) Compute the operator norm bound on the spectral radius and corresponding large deviation rate. (c) Compute the precise large deviation rate if X is replaced by an i.i.d. sequence distributed as π.
15. (*Life Insurance Risk*) Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli $0 - 1$ valued random variables with $P(X_1 = 1) = p = 1 - q$, and $S_n = X_1 + \cdots + X_n$. Consider a portfolio of life insurance contracts containing n individuals in the same risk category. Let $p > 0$ denote the probability of an individual death in a given year of coverage. One may assume that the life spans of the insured individuals are independent. Then the total claims is given by gS_n, where $g > 0$ is the amount paid to an individual upon death, and S_n is binomial with parameters n, p. A standard problem for actuaries is to determine an annual individual life insurance premium π to be paid for coverage such that $P(gS_n > n\pi) \leq r$ for a given risk tolerance $r \in (0, 1)$. Use the upper bound on $P(\sum_{j=1}^n v(X_j) > a)$, where $v(x) = gx$, $a = n\pi$, to determine an equation sufficient to determine π for given values of the parameters g, n, r, p.
16. Let $X_1, X_2, \ldots$ be i.i.d. with lognormal distribution having parameters μ, σ^2, i.e., $X_1 = e^{\sigma Z + \mu}$, for standard normal Z. Show that there is a large deviation principle for the distributions of $\overline{X}_n = (\prod_{j=1}^n X_j)^{\frac{1}{n}}$, $n \geq 1$, and compute the rate function.
17. Show that $D(P||Q) \geq 0$ with equality if and only if $P = Q$.[*Hint*: Use Jensen's inequality.

18. Let S be finite, say $S = \{1, 2, \ldots, k\}$, and $((p_{ij}))$ a transition probability matrix on S, $p_{ij} > 0$, for all $i, j \in S$. Prove that $h \to \lambda^+(h)$ is differentiable. [*Hint*: Let $q_h(i, j) = e^{hv(j)}p_{ij}$. By the (classical) Perron–Frobenius theorem $q_h = ((q_h(i, j)))$ has, for each h, a simple eigenvalue $\lambda^+(h)$. Write det(q)= det($q_h - \lambda I$) as $\prod_{j=1}^{n}(\lambda - a_j(h))$, with each $a_j(h)$ analytic. For given h^*, $\lambda - \lambda^+(h)$ appears only once, say $\lambda^+(h) = a_j(h)$ in a neighborhood of h^*.
19. Prove that $\lambda^+(h) \to \infty$ exponentially fast as $h \to \infty$ under the hypothesis of Remark 21.8. [*Hint*: As in the proof of Theorem 21.1, $\lambda^+(h) \geq \lambda_m(h) = \inf_x q_h(x, S)$. But $\inf_x \int_S e^{hv(y)}p(x, dy) \to \infty$ exponentially fast as $h \to \infty$.
20. Derive the large deviation rate for
$$\lim_{n\to\infty} P\left(\sum_{j=1}^{n}(v(X_j) - \mathbb{E}_\pi v) < -na\right)$$
for $a > 0$, under the hypothesis of Theorem 21.1 for $q(x, dy) = e^{v(y)}p(x, dy)$.
21. Verify the hypothesis of the Cramér–Chernoff theorem for the canonical distribution as noted in the proof of Corollary 21.5.
22. Check the inequality (21.59). [Hint: For bounded v and g, $\lambda^+(v + \epsilon g) = \int_S e^{v+\epsilon g}d\mu = \int_S e^v\{(1 + \epsilon g)d\mu + o(\epsilon^2)\}$.]
23. (*Contraction Principle*) Suppose $\{P_n\}$ satisfies a LDP with rate function I, on a metric space (S, ρ). Let $\varphi : S \to S'$ be a continuous map into a metric space (S', ρ'). Prove that the family $\{P_n' = P_n \circ \varphi^{-1}\}$ satisfies the LDP with rate function $I'(x') = \inf\{I(x) : \varphi(x) = x'\}$. [Hint: Let F be a closed subset of S', then $\varphi^{-1}(F)$ is a closed subset of S. $\limsup \frac{1}{n}\log P_n'(F) = \limsup \frac{1}{n}\log P_n(\varphi^{-1}(F)) = -\inf_{x\in\varphi^{-1}(F)} I(x) = -\inf_{x:\varphi(x)\in F} I(x) = -\inf_{z\in F}\inf_{x:\varphi(x)=z} I(x)$. The lower semicontinuity of I' follows from that of I.]
24. Let $\{P_n\}$ be a sequence of probability measures on S that satisfies a LDP with rate function I.

    (a) Suppose that S, S', are both locally compact and σ-compact Polish spaces and $\varphi : S \to S'$ is Borel measurable. Suppose that there are compact sets $K_m \subset S, m \geq 1$, such that φ is continuous on S. Show that if $\lim_{m\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log P_n(K_m^c) = -\infty$, then the sequence $\{P_n \circ \varphi^{-1}\}$ satisfies the LDP with rate $I'(a) = \inf\{I(x) : \varphi(x) = a\}$.
    (b) Show for the case $S = \mathbb{R}^d$ and $\int_S e^{u\cdot x}\pi(dx) < \infty$ for all $u \in \mathbb{R}^d$, there is an increasing function $\tau : [0, \infty) \to [0, \infty)$ such that $\tau(0) = 0, t^{-1}\tau(t) \to \infty$ as $t \to \infty$, and such that $\int_S e^{\tau}(|x|)\pi(dx) < \infty$.
    (c) Use these to complete the proof of Corollary 21.13 using Sanov's theorem. [*Hint*: Consider the cases in which $K_m = \{\nu \in \mathcal{P}(\mathbb{R}^d) : \int_S \tau(|x|)\nu(dx) \leq m\}$, for each $m \geq 1$.

25. In Example 2, let $g(x) = \beta x$, $|\beta| < 1$, i.e., consider the discrete Ornstein–Uhlenbeck model $X_{n+1} = \beta X_n + \sigma Z_{n+1}$, where $\{Z_n\}$ is i.i.d. $N(0, 1)$, $\sigma > 0$.

Compute the large deviation rate $I(a) = \lim_{n\to\infty} \frac{1}{n} \log P(\sum_{j=1}^{n} X_j > na)$ of the Cramér–Chernoff theorem for this example.

26. In the context of the method of types, show that $\#\mathcal{P}_n$ is no more than $(n + 1)^r$. [Hint: Think of positioning n balls marked $1, \ldots, n$ in a row, with $n - 1$ locations between successive balls, and one to the left of ball 1 and one to the right of ball n. Distribute r sticks, one at a time, at random in these locations. Label sticks according to position (left to right). The number of balls to the left of stick 1 is the size of group 1 observations. The number of balls to the right of stick r is the size of group r. If all sticks fall to the left of ball 1, then take the size of group r as n, and if all sticks fall the right of ball n, then take the size of group 1 as n, and the number of balls between (i-1)th and ith sticks as the size of group i observations, $i = 2, \ldots, r - 1$.]
27. In the context of Cramér-Sanov large deviation theory for i.i.d. random variables with common distribution π, show that for every bounded, measurable f on S, one has $\log \int_S e^f d\pi = \sup_{\nu\in\mathcal{P}(S)}(\int_S f d\nu - D(\nu||\pi))$. [Hint: By (21.56), $\log \int_S e^f d\pi \geq \sup_{\nu\in\mathcal{P}(S)}(\int_S f d\nu - D(\nu||\pi))$. In the arguments preceding Sanov's theorem (see (21.58), (21.59), etc.), it is shown that the supremum in (21.57) is attained at $v = \log h$, where $h = \frac{d\nu}{d\mu}$. That is, given f, and ν determined by $\frac{d\nu}{d\pi} = e^f$, one has $\log \int_S e^f d\pi = \int_S f d\nu - D(\nu||\pi) \leq \sup_{\nu\in\mathcal{P}(S)}(\int_S f d\nu - D(\nu||\pi))$. On the other hand, (21.56) shows the opposite inequality.]

Chapter 22
Special Topic: Applications of Large Deviation Theory

This chapter includes two applications of the large deviation theory presented in Chapter 21. One concerns an application to a problem in cryptography in which, among other motivations, hackers attempt to break a password by guessing. The other is an application to the efficiency of large sample statistical tests of hypothesis.

Example 1 *(Encrypted Security Systems*[1]*).* The problem to be considered here is of interest to cryptographers analyzing, for example, attempts by a hacker to enter a password protected system by robotically guessing it. The problem can be abstractly stated as follows: For a given finite set $S = \{1, 2, \ldots k\}$, say, Alice randomly generates a cipher $X^{(n)} = \mathbf{x} \in S^n$ of length n, where $X^{(n)} = (X_1, \ldots, X_n) \in S^n$ has a joint probability mass function $p_{X^{(n)}}(x_1, \ldots, x_n), x_j \in S, 1 \leq j \leq n$. According to some guessing strategy, Bob systematically steps through the messages $\mathbf{y} \in S^n$ in some specified order, and Alice responds $X^{(n)} = \mathbf{y}$ with "yes" or "no," according to whether $\mathbf{y} = \mathbf{x}$ or not. The goal is to quantify the effort required by guessing. Throughout it will be assumed without further mention that the *message source* $X_1, X_2, \ldots$ is a stationary process.

Mathematically, guessing is given by a bijection $G : S^n \to \{1, 2, \ldots, |S|^n\}$ prescribing the orders in which guesses $\mathbf{y} \in S^n$ are made in the guessing strategy. $G(\mathbf{x})$ is then the number of guesses to reach the given cipher $\mathbf{x}$.

[1] This example is based on Hanawal and Sundaresan (2011).

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_22

To minimize the expected number of guesses, an optimal choice is a guessing function G^* that would therefore make the order of selections according to decreasing probabilities $f(\mathbf{y}), \mathbf{y} \in S^n$. Note that if $f(\mathbf{y}) = f(\mathbf{z})$, then the order in which $\mathbf{y}$ and $\mathbf{z}$ are guessed will not affect the number of guesses to unlock the password. In particular, an optimal G is not unique with regard to minimizing the expected number of guesses.

As a measure of the attackers effort, cryptologists consider an optimal G^* to define an optimal guessing exponent by

$$g(\rho) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}G^*(X^{(n)})^\rho, \tag{22.1}$$

when the limit exists. The primary focus of this chapter is on the computation of $g(\rho)$ in some generality via large deviation theory. This is achieved by systematically establishing a succession of equivalent computations: Proposition 22.2 recasts the problem in terms of an equivalent computation for word lengths, Proposition 22.3 recasts this in terms of a Rényi entropy computation, and finally Theorem 22.4, Corollary 22.5 in terms of a large deviation computation for the so-called information spectrum.

Remark 22.1. Calculations have been made for $g(\rho)$ in the case of i.i.d. encodings $X_1, \ldots, X_n$ by Arikan (1996), and irreducible Markov chain encodings by Malone and Sullivan (2004). These will appear as applications of the large deviation results of Hanawal and Sundaresan (2011) at the end of this example.

It will be helpful to introduce the *guessing length function* $L_G : S^n \to \mathbf{N}$ associated with G defined by

$$L_G(\mathbf{x}) = \lceil -\ln \frac{1}{CG(\mathbf{x})} \rceil, \quad \mathbf{x} \in S^n, \tag{22.2}$$

where $\lceil\cdot\rceil$ is the ceiling function, i.e., $\lceil x \rceil$ is the smallest integer not smaller than x, and $C = \sum_{\mathbf{x}\in S^n} \frac{1}{G(\mathbf{x})}$ is a normalization constant. In particular,

$$Q_G(x) = \frac{1}{CG(x)}, \quad x \in S^n, \tag{22.3}$$

defines a probability mass function on S^n. Note that since $C \geq 1$,

$$G(x) = \frac{1}{CQ_G(x)} \leq \frac{1}{Q_G(x)}. \tag{22.4}$$

Clearly, $\ln G(x) \leq L_G(x), x \in S^n$ by definition, and

$$\ln G(x) = -\ln Q_G(x) - \ln C \geq \lceil -\ln Q_G(x) \rceil - 1 - \ln C, \tag{22.5}$$

so that, in summary,

$$L_G(x) - 1 - \ln C \le \ln G(x) \le L_G(x). \tag{22.6}$$

To denote the dependence of G and L on the message length n, we write G_n, L_n, C_n, respectively, when necessary. Note that L_n satisfies the so-called *Kraft inequality* (Exercise 4)

$$\sum_{x \in S^n} \exp\{-L_n(x)\} \le 1. \tag{22.7}$$

In general any function $L : S^n \to \mathbb{N}$ satisfying the Kraft inequality will be referred to as a *length function*. We let $\mathcal{L}_n$ denote the set of all such functions on S^n. L^* will denote a length function that minimizes $\mathbb{E}\exp\{\rho L(X^{(n)})\}$.

Suppose that $X_1, X_2, \ldots$ is a stationary process and let $Q \in \mathcal{P}_n$ denote the distribution of $(X_{1+m}, \ldots, X_{n+m})$ (m=1,2,...). The *Shannon entropy*[2] expressed in nats, i.e., using natural logarithms, is defined by

$$H(X_1, \ldots, X_n) \equiv H(Q) = -\sum_{x \in S^n} Q(\{x\}) \ln Q(\{x\}).$$

Shannon's entropy of the stationary process is defined by

$$H = \lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n},$$

for which existence is a direct consequence of subadditivity using Fekete's lemma from Chapter 5. Specifically, letting Q_n denote the distribution of $(X_1, X_2, \ldots, X_n)$, one has

$$\begin{aligned} H(Q_{n+m}) &\equiv H(X_1, \ldots, X_{n+m}) \\ &\le H(X_1, \ldots, X_n) + H(X_{n+1}, \ldots, X_{n+m}) \\ &= H(X_1, \ldots, X_n) + H(X_1, \ldots, X_m) = H(Q_n) + H(Q_m), \end{aligned} \tag{22.8}$$

where the essential second line is left as Exercise 2.

Remark 22.2. Note the existence of a length function L for which the (approximate) expected lengths are minimal, i.e., the problem

$$\min_{L \in \mathcal{L}_n : \sum_x e^{-L(x)} \le 1} \sum_{x \in S^n} p_{X^{(n)}}(x) L(x)$$

[2] See Bhattacharya and Waymire (1990, 2009), pp.184–189 for a related treatment of Shannon entropy.

can be shown to have a solution by the method of Lagrange multipliers (Exercise 5) providing one permits non-integer solutions.

Theorem 22.1 *(Shannon).* For a length function L one has

$$H(X_1, \dots, X_n) \le \mathbb{E}L(X_1, \dots, X_n),$$

with equality if and only if $p_{X^{(n)}}(x) = e^{-L(x)}$. Moreover, letting $L^*(X_1, \dots, X_n)$ denote the lengths having smallest expected value possible for the word $(X_1, \dots, X_n)$, one has

$$H(X_1, \dots, X_n) \le \mathbb{E}L^*(X_1, \dots, X_n) \le H(X_1, \dots, X_n) + 1.$$

In particular,

$$\frac{H(X_1, \dots, X_n)}{n} \le \frac{\mathbb{E}L^*(X_1, \dots, X_n)}{n} \le \frac{H(X_1, \dots, X_n)}{n} + \frac{1}{n},$$

$$\lim_{n\to\infty} \mathbb{E}\frac{L^*(X_1, \dots, X_n)}{n} = H.$$

Proof. To prove the lower bound let $q(x) = \frac{e^{-L(x)}}{\sum_{y\in S^n} e^{-L(y)}}$, and $K = \sum_{y\in S^n} e^{-L(y)} \le 1$, by the Kraft inequality. Then,

$$\begin{aligned}
&\mathbb{E}L(X_1, \dots, X_n) - H(X_1, \dots, X_n) \\
&= \sum_{x\in S^n} p_{X^{(n)}}(x)L(x) - \sum_{x\in S^n} p_{X^{(n)}}(x) \ln \frac{1}{p_{X^{(n)}}(x)} \\
&= -\sum_{x\in S^n} p_{X^{(n)}}(x) \ln e^{-L(x)} + \sum_{x\in S^n} p_{X^{(n)}}(x) \ln p_{X^{(n)}}(x) \\
&= \sum_{x\in S^n} p_{X^{(n)}}(x) \ln \frac{p_{X^{(n)}}(x)}{q(x)} - \ln K \\
&= D(p_{X^{(n)}}||q) + \ln \frac{1}{K} \ge 0. \qquad (22.9)
\end{aligned}$$

Note that approximately if $L(x) = \ln \frac{1}{p_{X^{(n)}}(x)}$, then $H = L$. However, such a choice for L is not an integer. Taking $L(x) = \lceil \ln \frac{1}{p_{X^{(n)}}(x)} \rceil$, the Kraft inequality is preserved by this choice Now, for this choice of lengths, a simple calculation yields,

$$H(X_1, \dots, X_n) \le \mathbb{E}L(X_1, \dots, X_n) \le H(X_1, \dots, X_n) + 1.$$

Since $\mathbb{E}L^*(X_1, \ldots, X_n) \leq \mathbb{E}L(X_1, \ldots, X_n)$ both the lower and upper bounds are satisfied by $\mathbb{E}L^*(X_1, \ldots, X_n)$. ■

Lemma 1. Let G be a guessing function and L_G its associated length function. Then,

$$\left| \frac{1}{\rho} \ln \mathbb{E}G^*(X^{(n)})^\rho - \frac{1}{\rho} \ln \mathbb{E} \exp \left\{ \rho L^*(X^{(n)}) \right\} \right| \leq 1 + \ln C, \tag{22.10}$$

where $C = \sum_{\mathbf{x} \in S^n} \frac{1}{G(\mathbf{x})}$.

Proof. For a length function $L \in \mathcal{L}_n$, let G_L be the guessing function that guesses in the increasing order of L-lengths. Messages of the same L-length are ordered according to an arbitrary fixed rule, say lexicographical order on S^n. Define a probability mass function on S^n by

$$Q_L(x) = \frac{\exp\{-L(x)\}}{\sum_{y \in S^n} \exp\{-L(y)\}}, \quad x \in S^n. \tag{22.11}$$

Note that G_L guesses in the decreasing order of Q_L probabilities. In particular, $G_L(x) \leq \sum_{y \in S^n} \mathbf{1}[Q_L(y) \geq Q_L(x)] \leq \sum_{y \in S^n} \frac{Q_L(y)}{Q_L(x)} = \frac{1}{Q_L(x)}$, so that

$$\ln G_L(x) \leq -\ln Q_L(x) \quad x \in S^n. \tag{22.12}$$

Also, by definition of Q_L and using Kraft's inequality (22.7),

$$\frac{1}{Q_L(x)} = \exp\{L(x)\} \sum_{y \in S^n} \exp\{-L(y)\} \leq \exp\{L(x)\},$$

so that

$$-\ln Q_L(x) \leq L(x), \quad x \in S^n. \tag{22.13}$$

From these inequalities one deduces that for any $B \geq 1$,

$$\{x : L_G(x) \geq B + 1 + \ln C\} \subset \{x : G(x) \geq e^B\} \subset \{x : L_G(x) \geq B\}, \tag{22.14}$$

and

$$\{x : G_L(x) \geq e^B\} \subset \{x : L(x) \geq B\}. \tag{22.15}$$

Now, by (22.12) followed by (22.6),

$$\begin{aligned} \mathbb{E} \exp\{\rho L(X^{(n)})\} &\geq \mathbb{E}G_L(X^{(n)})^\rho \geq \mathbb{E}G^*(X^{(n)})^\rho \\ &\geq \mathbb{E} \exp\{\rho L_{G^*}(X^{(n)})\} \exp\{-\rho(1 + \ln C)\} \end{aligned}$$

$$\geq \mathbb{E}\exp\{\rho L^*(X^{(n)})\}\exp\{-\rho(1+\ln C)\}. \tag{22.16}$$

Thus,

$$\frac{\mathbb{E}G_L(X^{(n)})^\rho}{\mathbb{E}G^*(X^{(n)})^\rho} \leq \frac{\mathbb{E}\exp\{\rho L(X^{(n)})\}}{\mathbb{E}\exp\{\rho L^*(X^{(n)})\}}\exp\{\rho(1+\ln C)\}, \tag{22.17}$$

and, in terms of the length function L_G associated with G, one similarly has

$$\frac{\mathbb{E}G(X^{(n)})^\rho}{\mathbb{E}G^*(X^{(n)})^\rho} \geq \frac{\mathbb{E}\exp\{\rho L_G(X^{(n)})\}}{\mathbb{E}\exp\{\rho L^*(X^{(n)})\}}\exp\{-\rho(1+\ln C)\}. \tag{22.18}$$

The lemma now follows from these bounds upon taking logarithms with $L = L^*$ in (22.16). That is

$$1 \geq \frac{\mathbb{E}G^*(X^{(n)})^\rho}{\mathbb{E}\exp\{\rho L^*(X^{(n)})\}} \geq \exp\{-\rho(1+\ln C)\}, \tag{22.19}$$

so that $0 \geq \ln \mathbb{E}G^*(X^{(n)})^\rho - \ln \mathbb{E}\{\rho L^*(X^{(n)})\} \geq -\rho(1+\ln C)$. ■

The existence and determination of $g(\rho)$ will ultimately follow from an application of Varadhan's integral formula applied to a related function of $X_1, \ldots, X_n$ obtained from the next three propositions and their lemmas.

Proposition 22.2. The guessing exponent $g(\rho)$ exists if and only if

$$\ell(\rho) = \lim_{n\to\infty} \inf_{L\in\mathcal{L}_n} \frac{1}{n}\ln \mathbb{E}\exp\{\rho L(X^{(n)})\} \tag{22.20}$$

exists. Moreover $g(\rho) = \ell(\rho)$ when either exists.

Proof. Note that $C_n \leq 1+n\ln|S|$. Dividing both sides of the inequality in Lemma 1 by n, one has

$$\left|\frac{1}{n\rho}\ln \mathbb{E}G^{*\rho}(X^{(n)}) - \frac{1}{n\rho}\ln \mathbb{E}(\exp\{\rho L^*(X^{(n)})\})\right| \leq \frac{1}{n}(1+\ln C_n) = O\left(\frac{\ln n}{n}\right). \tag{22.21}$$

Thus the sequences differ by $o(1)$ as $n \to \infty$. ■

The next proposition requires the *Rényi entropy rate* of order $\alpha \neq 1$ defined by

$$H_\alpha(p_{X^{(n)}}) = \frac{1}{1-\alpha}\ln \sum_{\mathbf{x}\in S^n} p^\alpha_{X^{(n)}}(\mathbf{x})) \equiv \frac{1}{1-\alpha}\ln \mathbb{E}p^{\alpha-1}_{X^{(n)}}(X^{(n)}). \tag{22.22}$$

Proposition 22.3. $\lim_{n\to\infty}\inf_{L\in\mathcal{L}_n}\frac{1}{n}\ln\mathbb{E}\exp\{\rho L(X^{(n)})\}$, or equivalently $\lim_n\ln\mathbb{E}G^*(X^{(n)})^\rho$, exists if and only if $\lim_{n\to\infty}\frac{1}{n}H_\alpha(p_{X^{(n)}})$ exists for $\alpha = \frac{1}{1+\rho}$. Moreover, if the latter limit exists, then it is given by $\frac{g(\rho)}{\rho}$.

Proof. The equivalence is the content of Proposition 22.2. We focus on the former limit. For each n the Donsker–Varadhan variational formula of Corollary 21.17 yields, upon replacing g by $L(X^{(n)})$, λ by ρ, Q by $p_{X^{(n)}}$, and P by Q, that

$$\ln\mathbb{E}\exp\{\rho L(X^{(n)})\} = \sup_{Q\in\mathcal{P}_n}\{\rho\mathbb{E}_Q L(X^{(n)}) - D(Q||p_{X^{(n)}})\}. \tag{22.23}$$

Taking the infimum on both sides over all length functions $L\in\mathcal{L}_n$ and applying Fan's minimax exchange of supremum and infimum, one has

$$\begin{aligned}\inf_{L\in\mathcal{L}_n}\ln\mathbb{E}\exp\{\rho L_n(X^{(n)})\} &= \inf_{L\in\mathcal{L}_n}\sup_{Q\in\mathcal{P}_n}\{\rho\mathbb{E}_Q L_n(X^{(n)}) - D(Q||p_{X^{(n)}})\}\\ &= \sup_{Q\in\mathcal{P}_n}\inf_{L\in\mathcal{L}_n}\{\rho\mathbb{E}_Q L_n(X^{(n)}) - D(Q||p_{X^{(n)}})\}\\ &= \sup_{Q\in\mathcal{P}_n}\{\rho H(Q) - D(Q||p_{X^{(n)}})\} + O(1)\\ &= \rho H_{\frac{1}{1+\rho}}(p_{X^{(n)}}) + O(1),\end{aligned} \tag{22.24}$$

where to justify the use of Fan's minimax formula one notes firstly convexity of the map $(Q,L)\in\mathcal{P}_n\times\mathcal{L}_n\to\mathbb{E}_Q\{\rho L(X^{(n)}) - D(Q||p_{X^{(n)}}) = \sum_{x\in S^n}\{\rho L(x) + \ln Q(x) - \ln p_{X^{(n)}}(x)\}Q(x)$, as a function of $Q\in\mathcal{P}_n$, and the linearity as a function of L. The next equation follows from Theorem 22.1, namely $\inf_{L\in\mathcal{L}_n}\mathbb{E}_{Q\in\mathcal{P}_n}\{L(X^{(n)})\} = H(Q) + O(1)$. Finally, the last equation follows by writing

$$\sup_{Q\in\mathcal{P}_n}\left\{\rho H(Q) - D(Q||p_{X^{(n)}})\right\} = (1+\rho)\sup_{Q\in\mathcal{P}_n}\left\{\mathbb{E}_Q\left[-\frac{\rho}{1+\rho}\ln p_{X^{(n)}}(X^{(n)})\right] - D(Q||p_{X^{(n)}})\right\},$$

and then applying the Donsker–Varadhan variational formula of Corollary 21.17, as in the first equation, with g replaced by $\ln p_{X^{(n)}}(X^{(n)})$, λ replaced by $\frac{1}{1+\rho}$, P replaced by Q to get the scaled Rényi entropy. That is,

$$\begin{aligned}&\sup_{Q\in\mathcal{P}_n}\{\rho H(Q) - D(Q||p_{X^{(n)}})\} + O(1)\\ &= \sup_{Q\in\mathcal{P}_n}\left\{-\rho\sum_x Q(x)\ln Q(x) - \sum_x Q(x)\ln Q(x) + \sum_x Q(x)\ln p_{X^{(n)}}(x)\right\}\\ &= \sup_{Q\in\mathcal{P}_n}\left\{\sum_x Q(x)\ln p_{X^{(n)}}(x) - (1+\rho)\sum_x Q(x)\ln Q(x)\right\}\end{aligned}$$

$$
\begin{aligned}
&= (1+\rho) \sup_{Q \in \mathcal{P}_n} \left\{ \mathbb{E}_Q \frac{1}{1+\rho} \ln p_{X^{(n)}}(X^{(n)}) - D(Q \| p_{X^{(n)}}) \right\} + O(1) \\
&= (1+\rho) \ln \mathbb{E} p_{X^{(n)}}^{\frac{1}{1+\rho}-1}(X^{(n)}) + O(1) \\
&= \rho H_{\frac{1}{1+\rho}}(p_{X^{(n)}}) + O(1).
\end{aligned}
$$

Scale by $\frac{1}{n}$ and let $n \to \infty$ to complete the proof. ■

The *information spectrum* is defined by $-\frac{1}{n} \ln p_{X^{(n)}}(X^{(n)})$. The next step is to show that the Rényi entropy rate can be computed from the distributions of the information spectra.

Theorem 22.4 *(Hanawal and Sundaresan (2011)).* Let ν_n be the distribution of the information spectrum $-\frac{1}{n} \ln p_{X^{(n)}}(X^{(n)})$. If $\nu_n, n \geq 1$, satisfy a LDP with rate function I, then the limiting Rényi entropy rate of order $\alpha = \frac{1}{1+\rho}$ exists and is given by $\beta^{-1} \sup_{t \in \mathbb{R}} \{\beta t - I(t)\}$, where $\beta = \frac{\rho}{1+\rho}$.

Proof. Let ν_n denote the distribution of the information spectrum $\frac{1}{n} \ln p_{X^{(n)}}(X^{(n)})$. Then, with $A_n = \{-\frac{1}{n} \ln p_{X^{(n)}}(x) : x \in S^n\}$, one has

$$
\begin{aligned}
\int_{\mathbb{R}} \exp(n\beta t) \nu_n(dt) &= \sum_{t \in A_n} \exp(n\beta t) \sum_{\{x : p_{X^{(n)}}(x) = \exp(-nt)\}} p_{X^{(n)}}(x) \\
&= \sum_{x \in S^n} p_{X^{(n)}}(x)^{1-\beta} \\
&= \sum_{x \in S^n} p_{X^{(n)}}(x)^{\frac{1}{1+\rho}} \\
&= \exp\{\beta H_{\frac{1}{1+\rho}}(p_{X^{(n)}})\}. \qquad (22.25)
\end{aligned}
$$

Now, scaling by $\frac{1}{n}$ and taking logarithms, one may apply the Varadhan integral formula to the left side to obtain in the limit $\beta^{-1} \sup_{t \in \mathbb{R}} \{\beta t - I(t)\}$, while one has on the right side $\beta \lim_n \frac{1}{n} H_{\frac{1}{1+\rho}}(p_{X^{(n)}})$. ■

Corollary 22.5. If the distributions of the information spectra satisfies a LDP with rate I, then the guessing exponent exists and is given by

$$
g(\rho) = (1+\rho) \sup_{t \in \mathbb{R}} \left\{ \frac{\rho}{1+\rho} t - I(t) \right\}.
$$

Proof. By Proposition 22.3 the limiting Rényi entropy is $\frac{g(\rho)}{\rho}$. Thus, one has $g(\rho) = \rho \beta^{-1} \sup_{t \in \mathbb{R}} \{\beta t - I(t)\} = (1+\rho) \sup_{t \in \mathbb{R}} \{\beta t - I(t)\}$. ■

First let us apply this theory to the case of i.i.d. message sources.

***Theorem 22.6** (I.I.D. Case).* Assume that $X_1, X_2, \dots$ is i.i.d. with common probability mass function p on the finite alphabet S. Then, the limit defining the guessing exponent $g(\rho)$ exists and is given by $g(\rho) = (1+\rho)H_{\frac{1}{1+\rho}}(p)$, where $H_\alpha(p)$ denotes the Rényi entropy rate of order α of the probability mass function p.

Proof. From Proposition 22.3 one can compute $g(\rho)$ from the Rényi entropy rate which, in turn, is given by $\frac{1+\rho}{\rho} I^*(\frac{\rho}{1+\rho})$, where $I(\cdot)$ is the large deviation rate for the energy spectrum

$$-\frac{1}{n}\ln p_{X^{(n)}}(X^{(n)}) = -\frac{1}{n}\ln \prod_{j=1}^{n} p(X_j) = -\frac{1}{n}\sum_{j=1}^{n}\ln p(X_j).$$

In particular, $I(h) = c^*(h)$ is the Legendre transform of the cumulant generating function of $-\ln p(X_1)$, namely

$$c(h) = \ln \mathbb{E}e^{h(-\ln p(X_1))} = \ln \mathbb{E}p^{-h}(X_1) = hH_{1-h}(p).$$

Since the Legendre transform operation $*$ is idempotent (see Exercise 6), it follows that

$$I^*\left(\frac{\rho}{1+\rho}\right) = (c^*)^*\left(\frac{\rho}{1+\rho}\right) = c\left(\frac{\rho}{1+\rho}\right).$$

In particular, $g(\rho) = (1+\rho)\frac{\rho}{1+\rho}H_{\frac{1}{1+\rho}} = \rho H(\frac{1}{1+\rho})$, as asserted. ■

***Theorem 22.7** (Irreducible Markov Case).* Let $X_1, X_2, \dots$ be an irreducible Markov chain on S with homogeneous transition probability matrix $p = ((p(y|x)))_{x,y\in S}$. Then the guessing exponent $g(\rho)$ exists and is given by

$$g(\rho) = (1+\rho)\lambda^+\left(\frac{\rho}{1+\rho}\right),$$

where $\lambda^+(h)$ is the largest eigenvalue of the matrix $((\pi^{1-h}(y|x)))_{x,y\in S}$.

Proof. As in the i.i.d. case, from Proposition 22.3 one can compute $g(\rho)$ from the Rényi entropy rate which, in turn, is given by $\frac{1+\rho}{\rho} I^*(\frac{\rho}{1+\rho})$, where $I(\cdot)$ is the large deviation rate for the energy spectrum

$$-\frac{1}{n}\ln p_{X^{(n)}}(X^{(n)}) = -\frac{1}{n}\left\{\ln p(X_1) + \sum_{j=1}^{n-1}\ln p(X_{j+1}|X_j)\right\}.$$

Note that $Y_j = (X_j, X_{j+1}), j = 1, 2, \dots$ is also a stationary Markov chain with one-step transition probabilities

$$\tilde{p}((w,z)|(x,y)) = \begin{cases} p(z|y), & y = w, \\ 0, & y \neq w. \end{cases}$$

To compute $I(\cdot)$ it suffices to compute the large deviation rate for $\sum_{j=1}^{n} \varphi(Y_j)$, where $g(Y_j) = -\ln p(X_{j+1}|X_j)$. Let $v(x,y) = -\ln p(y|x)$, $(x,y) \in S \times S$. Then,

$$\begin{aligned} T_h f(x,y) &= \sum_{(w,z)\in S\times S} f(w,z) e^{-h \ln p(z|w)} \tilde{p}((w,z)|(x,y)) \\ &= \sum_{z\in S} f(y,z) p^{-h}(z|y) p(z|y) = \sum_{z\in S} f(y,z) p^{1-h}(z|y). \quad (22.26) \end{aligned}$$

Observe that $T_h g(x,y) = \lambda g(x,y), (x,y) \in S \times S$ implies $g(x,y) = g(y)$, i.e., is constant in x. In particular, $\lambda^+(h) = \lambda(1-h)$, where $\lambda(a)$ is the largest eigenvalue of the matrix $((p^a(y|x))_{(x,y)\in S\times S}$. In particular, $I(h) = \lambda^*(h)$. Again using idempotency, of the Legendre transform, $I^*(t) = \lambda(t)$. It follows that the entropy rate is given by $\frac{1+\rho}{\rho} \ln \lambda(\frac{\rho}{1+\rho})$, and therefore the guessing exponent is $g(\rho) = \rho \frac{1+\rho}{\rho} \ln \lambda(\frac{\rho}{1+\rho})$ where $\lambda(\frac{\rho}{1+\rho})$ is the largest eigenvalue of the matrix $((p^{\frac{1}{1+\rho}}(y|x)))_{(x,y)\in S\times S}$. ■

Remark 22.3. Alternative representations of the guessing exponent in both of these cases can be obtained by consideration of level-2 large deviations as given in Hanawal and Sundaresan (2011). Moreover, the computation of the guessing exponent by these methods for other general classes of message sources can be found there.

The Kraft inequality for lengths plays an essential role in this application, specifically in Theorem 22.1 and its application in the proof of Proposition 22.3. In the classic monograph of Shannon (1948) messages are defined as sequences from a finite alphabet S, referred to as ciphers.[3] In the context of message compression, for a positive integer b one often defines a *b-ary coding function* as an injective map $c : S^n \to \cup_{m=1}^{\infty} \{0, 1, \ldots, b-1\}^m$ that renders a message $x \in S^n$ of length n, as a b-ary sequence $c(x)$ of length m for some m. One seeks codes c for which the average length $\mathbb{E}L(X^{(n)})$, of a message $X^{(n)}$, is minimal. A b-ary coding function is said to be *prefix-free* (or instantaneous) if for $x \neq y$ $c(x)$ is not a prefix of $c(y)$. A prefix-free code may be represented as leaves on a rooted b-ary tree obtained by coding the path from the root to the leaves (terminal vertices) with labels $\{0, 1, 2, \ldots, b-1\}$ from left to right at each level of the tree. Therefore, a prefix-free codeword can be instantaneously decoded without reference to future codewords since the end of a codeword is immediately recognizable as a leaf.

[3] The textbook by Cover and Thomas (2006) provides a good foundation for the general concepts and results encountered in information theory.

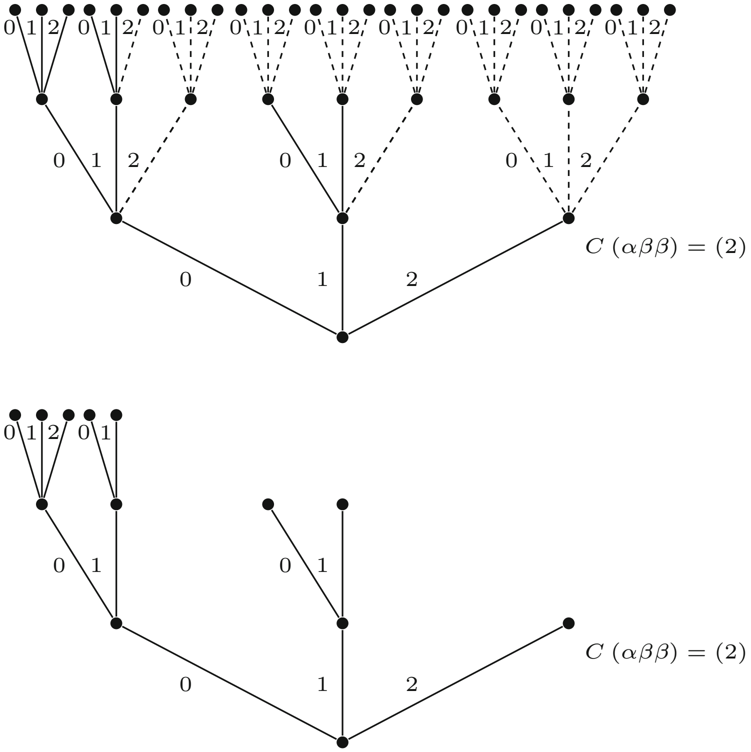


Fig. 22.1 Prefix-free code: $S = \{\alpha, \beta\}, n = 3, b = 3$; $L(\alpha, \beta, \beta) = 1$, $L(\alpha, \alpha, \alpha) = L(\alpha, \alpha, \beta) = 2$; $L(\beta, \beta, \beta) = \cdots = L(\beta, \alpha, \alpha) = 3$.

Proposition 22.8. Given any positive integers $L_1, \ldots, L_{|S|^n}$, satisfying Kraft inequality, there is a prefix-free b-ary code on S^n, $b \geq 3$, whose code words have lengths $L_1, \ldots, L_{|S|^n}$.

Proof. Observe that for positive integers $L(x), x \in S^n$, $b \geq 3$, since $2 < e < 3$, $\sum_{x \in S^n} b^{-L(x)} \leq 1$ if $\sum_{x \in S^n} \exp\{-L_(x)\} \leq 1$. Let $m = |S|^n$, $L_{\max} = \max\{L_1, \ldots, L_m\}$ and construct a full rooted b-ary tree of height $L_{\max}$ for a $b \geq 3$. Then the total number of leaves available is $b^{L_{\max}}$, at vertices of height $L_{\max}$ having height one label from $\{0, 1, \ldots, b-2\}$ (see Figure 22.1). This uses $(b-1)b^{L_{\max}-1}$ of the leaves, with $b^{L_{\max}} - (b-1)b^{L_{\max}-1} = b^{L_{\max}-1}$ remaining for coding words having lengths at most $L_{\max} - 1$. Proceed inductively. ∎

Remark 22.4. The prefix-free b-ary code constructed in the proof of Proposition 22.8 is referred to as the *Shannon code*. The units for message compression are referred to as "bits" when the logarithm is base 2, and "nats" for the natural logarithm. Natural logarithms are mathematically more convenient to the problem at hand and can be used without loss of generality.

The significance of Proposition 22.8 for the present chapter is that one may assume any given lengths, i.e., subject to the Kraft inequality, to be those of a prefix-free code.

Remark 22.5. The approximate code lengths $L(x) = \ln \frac{1}{p_{X(n)}}$ can be obtained as the solution to minimizing expected code lengths subject to Kraft inequality by the method of Lagrange multipliers (Exercise 5). However, as noted, these are not necessarily positive integers. The existence of an optimal code is a consequence of Theorem 22.1 and Proposition 22.8 by consideration of the prefix-free code associated with $L(x) = \lceil \ln \frac{1}{p_{X(n)}} \rceil$.

The next example illustrates a role for large deviation theory in large sample statistical inference.

Example 2 *(Efficiency of Statistical Tests of Hypothesis in Large Samples).* A common statistical test of hypothesis about an unknown parameter θ based on a random sample of size n from some distribution may be stated as follows:

The null hypothesis $H_0 : \theta \leq \theta_0$ is to be tested against the alternative hypothesis $H_1 : \theta > \theta_0$. The test is of the form: Reject H_0 (in favor of H_1) if $\overline{X} > a$, where $\overline{X}$ is the (sample) mean of i.i.d. variables $(X_1, \ldots, X_n)$ based on the random sample, and a is an appropriate number. The objective is to have small error probabilities $\alpha_n = P(\overline{X} > a|H_0)$, and $\beta_n = P(\overline{X} \leq a|H_1)$.

There are several competing notions for the *Asymptotic Relative Efficiency* (ARE) of such tests. For example, in the so-called *location problem*, the distribution function of X is of the form $F(x - \theta)$, $\theta \in \mathbb{R}$. In particular, F may be the normal distribution $N(\theta, 1)$. The *Normal test* M is of the form: Reject H_0 iff $\overline{X} > a$. The *t-test* T is of the form: Reject H_0 iff $\overline{X}/s > a$, where s is the sample standard deviation. The *Sign test* S is of the form: Reject H_0 iff $\frac{1}{n}\sum_{1\leq i\leq n}[X_i - \theta_0 \geq 0] > a$. The a-values of these tests are not necessarily the same.

The most commonly used test ARE is the *Pitman* ARE E_P test,[4] which fixes a "small" level $\alpha_n = \alpha$, and compares two tests A, B, say, based on the smallness of their β_n. Specifically, the *Pitman* ARE of B with respect to A is $E_P(A, B) = n/h(n)$, where $h(n)$ is the sample size needed for B to attain the same level β_n as attained by A based on a sample size n. The asymptotics here are generally based on weak convergence, especially the CLT (central limit theorem).

The two other important AREs we discuss in detail here are mainly based on large deviations. **Chernoff-**ARE:[5] Based on large deviation estimates for each test A, B, Chernoff's (modified) test picks the value of a that minimizes $\alpha_n + \lambda\beta_n$ over all a for some fixed $\lambda > 0$. (It turns out the ARE does not depend on λ). The ratio of the large deviation rates $I(A)$, $I(B)$ of decay of this minimum value δ_n, say, of $\alpha_n + \lambda\beta_n$ is compared for the tests A and B, and the *Chernoff* ARE of B with respect to A is $E_C(A, B) = I(B)/I(A)$.

[4] Serfling (1980), Chapter 10, Bhattacharya et al. (2016), Chapter 8.

[5] Serfling (1980), Chapter 10; Chernoff (1952).

Proposition 22.9. Assume $m_i(h) = \mathbb{E}(\exp\{hX_1\}|H_i) < \infty$ for all $-\infty < h < \infty (i = 0, 1)$. Let

$$c_i(a) = \sup\{ah - \ln m_i(h) : h \in \mathbb{R}\}(i = 0, 1), d(a) = \min\{c_0(a), c_1(a)\},$$
$$I = \max\{d(a) : \theta_0 \leq a \leq \theta_1\},$$

and $\rho = \exp\{-I\}$. Then

$$\lim_{n\to\infty} \frac{1}{n} \ln \delta_n = -I. \tag{22.27}$$

Proof. By the upper bound in the Cramér–Chernoff theorem (i.e., Chernoff's Inequality), $\alpha_n + \lambda\beta_n \leq \exp\{-nc_0(a)\} + \lambda \exp\{-nc_1(a)\} \leq (1+\lambda)\exp\{-nd(a)\}$. Minimizing over a, one arrives at the inequality $\delta_n \leq (1+\lambda)\rho^n$, or $\frac{1}{n}\ln\delta_n \leq -I + \frac{1}{n}\ln(1+\lambda)$, and $\limsup_n \frac{1}{n}\ln\delta_n \leq -I$. For the lower bound for δ_n, note that, by the Crameér–Chernoff theorem, $\liminf \frac{1}{n}\ln\alpha_n \geq -c_0(a)$, $\liminf_n \frac{1}{n}\ln\beta_n \geq -c_1(a)$. That is, given $\eta > 0$, for all sufficiently large n, $\min\{\alpha_n, \beta_n\} \geq \exp\{-n(d(a)+\eta)\}$, or $\alpha_n + \lambda\beta_n \geq (1+\lambda)\exp\{-n(d(a)+\eta)\}$. Hence, taking the minimum over a, $\delta_n \geq (1+\lambda)\exp\{-n(I+\eta)\}$, or $\frac{1}{n}\ln\delta_n \geq -(I+\eta) + \frac{1}{n}\ln(1+\lambda)$; so that $\liminf \frac{1}{n}\ln\delta_n \geq -(I+\eta)$ for all $\eta > 0$. Hence $\liminf_n \frac{1}{n}\ln\delta_n \geq -I$. ■

The Location Problem Consider the tests M, T, S for the location problem for $F(x-\theta)$ described in the first paragraph. Assume that F has a density f, continuous at $\theta = 0$, and a finite variance σ_f^2. Then for the test $H_0 : \theta \leq 0$, to be tested against the alternative hypothesis $H_1 : \theta \geq \theta_1 > 0$, one can show[6] that $E_P(S, M) = 4\sigma_f^2 f^2(0)$. In particular, (i) if F is $N(\theta, 1)$, then $E_P(S, M) = 2/\pi < 1$, (ii) if F is Double exponential (i.e., $f(x-\theta) = \frac{1}{2}\exp\{-|x-\theta|\}$), then $E_P(S, M) = 2$, and (iii) if f is uniform on $[-\frac{1}{2}-\theta, \frac{1}{2}-\theta]$, then $E_P(S, M) = 1/3$. In all these cases (and more broadly) $E_P(T, M) = 1$, where T is the t-test.

More interesting are Pitman comparisons among nonparametric tests for the so-called *two-sample problems*. Here two independent samples $(X_1, \ldots, X_m)$, $(Y_1, \ldots, Y_n)$ of sizes m and n are drawn from an unknown distribution whose density is of the form $f((x-\theta)/\sigma)$, $\theta \in \mathbb{R}$, $\sigma > 0$. One wishes to test $H_0 : \theta = 0$, against $H_1 : \theta > 0$. More generally, one wishes to test if the Y-distribution is stochastically larger than the X-distribution (i.e., $P(Y > z) \geq P(X > z)$ for all z, with strict inequality for at least some z). The most commonly used test for this uses the (nonparametric) statistic $T = \overline{Y} - \overline{X}$, which rejects H_0 if T exceeds a critical value. (The critical value is determined approximately by the CLT to meet the requirement $\alpha = P(Reject H_0|H_0)$). It turns out that appropriate nonparametric

[6] See Serfling (1980), Chapter 10; Bhattacharya and Waymire (2016), Chapter 8.

tests based on ranks of the combined observations X_i's and Y_j's, mostly outperform T.[7]

Chernoff Index Computation The Chernoff indices I are generally difficult to compute, since the indices try to minimize a linear combination of both error probabilities α_n and β_n. We consider the simple case where $F = N(\theta, 1)$, and $H_0 : \theta = 0$, $H_1 : \theta = \theta_1$, $\theta_1 > 0$. Again consider the test M: Reject H_0 if $\overline{X} > a$, otherwise Reject H_1. We leave it as an exercise to show that $I(M) = \theta_1^2/8$ (Exercise 8). For the sign test S: Reject H_0 if $\frac{1}{n}\sum_{1\leq i\leq n} \mathbf{1}[X_i > 0]) > a$ for some appropriate a, as considered in the discussion of the Chernoff-ARE above, one may compute the Chernoff index $I(S)$ from the distribution $B(n, p)$ with $p = \Phi(\theta_1)$, Φ being the distribution function of $N(0, 1)$. Namely,

$$I(S) = \ln\{2(b(\theta_1))^{b(\theta_1)}(1 - b(\theta_1))^{1-b(\theta_1)}\}, \tag{22.28}$$

where $b(\theta) = \ln[1 - \Phi(\theta)]/[(\ln\{(1 - \Phi(\theta))/\Phi(\theta)\}]$ (Exercise 10). The ratio $I(S)/I(M)$ provides the Chernoff-ARE $E_C(S, M)$. One may check that $E_C(S, M) \to 2/\pi = E_P(S, M)$ as $\theta_1 \downarrow 0$ (Exercise 11).

Bahadur-*ARE* As mentioned above, the Chernoff-ARE is generally difficult to compute. In addition, the threshold of the test itself is modified by the requirement of this notion of efficiency. The most popular ARE for tests based on large deviations is due to Bahadur (1960). Here is a brief description following Serfling (1980). The Bahadur-ARE is based on a large deviation rate comparison of the p-values of the tests. Consider a test of hypothesis $H_0 : \theta \in \Theta_0$, with a real-valued test statistic T_n based on observations $X_1, \ldots, X_n$, rejecting H_0 if T_n is large. The p-value of the *test* is $L_n = \sup[1 - F_{\theta_n}(T_n) : \theta \in \Theta_0] = 1 - F_{\theta_n^0}(T_n)$, say, where F_{θ_n} is the distribution function of the statistic under the parameter value θ. Thus L_n is the random quantity which is the probability (under H_0) of the statistic being larger than what is observed, i.e., of showing a discrepancy from the null hypothesis as large or larger than what is observed. Statisticians routinely use L_n to decide whether to reject H_0: smaller the p-value, stronger is the evidence against H_0. Under H_0, assuming that the distribution function $F_{\theta_n^{(0}}$ of T_n is continuous, $F_{\theta_n^{(0)}}(T_n)$ has the uniform distribution on $[0, 1]$, and so is the distribution of $L_n = 1 - F_{\theta_n^{(0)}}(T_n)$. Under fairly general conditions, $-2n^{-1} \ln L_n$ converges almost surely to a constant $c(\theta)$, which is referred to as *Bahadur's (exact) slope* for T_n, for $\theta \in \Theta_1$. The Bahadur relative efficiency of a test I with respect to test II is defined by the ratio of their corresponding slopes (a.s. large deviation rates) $e_B(I, II) = c_I(\Theta)/c_{II}(\Theta)$.

The following is a basic result which may be used to compute the slope of tests such as $H_0 : \theta \leq \theta_0$, against $H_1 : \theta > \theta_0$.[8] Write $\Theta_1 = \Theta \backslash \Theta_0$.

[7] Bhattacharya et al. (2016), Chapter 8.

[8] We follow Serfling (1980), Chapter 10, for the proof of the following result of Bahadur (1960; 1971).

***Theorem 22.10** (Bahadur (1960)).* For a test sequence T_n which rejects H_0 for large T_n, assume (i) $n^{-\frac{1}{2}}T_n$ converges a.s. (under θ) to a finite $b(\theta)$, for all $\theta \in \Theta_1$, and (ii) one has

$$\lim_{n\to\infty} -2n^{-1}\ln\sup[1 - F_{\theta_n}(n^{\frac{1}{2}}t) : \theta \in \Theta_0] = g(t), \tag{22.29}$$

where g is continuous on an open interval I containing $\{b(\theta) : \theta \in \Theta_1\}$. Then $\forall\, \theta \in \Theta_1$, with P_θ-probability one,

$$\lim_{n\to\infty} -2n^{-1}\ln L_n = g(b(\theta)) = c(\theta). \tag{22.30}$$

Proof. Fix a $\theta \in \Theta_1$, and let ω be any point in the sample space of P_θ for which the limit (i) holds. Fix $\epsilon > 0$ sufficiently small that $(b(\theta) - \epsilon, b(\theta) + \epsilon)$ is contained in I. By (i), there exists $n = n(\omega)$ such that $b(\theta) - \epsilon \le n^{-\frac{1}{2}}T_n(\omega) \le b(\theta) + \epsilon$ for all $n \ge n(\omega)$, i.e., $n^{\frac{1}{2}}(b(\theta) - \epsilon) \le T_{n(\omega)} \le n^{\frac{1}{2}}(b(\theta) + \epsilon)$ for all $n \ge n(\omega)$. Plugging these in $-2n^{-1}\ln\sup[1 - F_{\theta_n}(n^{\frac{1}{2}}t) : \theta \in \Theta_0]$, one then has

$$\begin{aligned} &-2n^{-1}\ln\sup[1 - F_{\theta_n}(b(\theta) - \epsilon)) : \theta \in \Theta_0]) \\ &\quad \le -2n^{-1}\ln L_n(\omega) \\ &\quad \le -2n^{-1}\ln\sup[1 - F_{\theta_n}(b(\theta) + \epsilon)) : \theta \in \Theta_0])\forall\, n \ge n(\omega). \end{aligned} \tag{22.31}$$

The limits as $n \to \infty$ of the two extreme sides are $g(b(\theta) - \epsilon)$ and $g(b(\theta) + \epsilon)$. Therefore, the limit points of the middle term in (22.31) all lie in this interval. By continuity of g, it follows that the middle term converges to $g(b(\theta))$. ■

The exact Bahadur slopes for the mean test M and the t-test T may be computed for testing $H_0 : \theta \le 0$, versus the alternative $H_1 : \Theta > 0$ in the model $N(\theta, 1)$, using the upper tail of the standard normal $N(0, 1)$, and that of the (Student's) t-statistic with $n - 1$ degrees of freedom. Using Bahadur's theorem, one finds $c_M(\theta) = \theta^2$, $c_T(\theta) = \ln(1 + \theta^2)$ (Exercise 12). Thus $e_B(T, M) < 1$ for all $\theta \in \Theta_1$. This is in contrast with both Pitman's ARE and Chernoff's ARE, for each of which the ARE is one.

Remark 22.6. Bahadur's ARE also distinguishes between the *frequency chi-square* and the *likelihood ratio test* in the multinomial model, showing the latter is asymptotically more efficient than the former. Again the Pitman ARE is one between the two tests.[9]

[9] Abrahamson (1965).

Exercises

1. (Shannon & Renyi Entropies) Show that the Shannon entropy H may be expressed in terms of the Renyi entropy as

$$H(X_1, \dots, X_n) = \lim_{\alpha \to 1} H_\alpha(X_1, \dots, X_n).$$

2. Complete the proof of (22.8) by showing that for random vectors $H(X, Y) \leq H(X) + H(Y)$. [*Hint*: Show how to express $H(X) + H(Y) - H(X, Y)$ as $D(p_{(X,Y)} || p_X \odot p_Y) \geq 0$, where $p_{(X,Y)}, p_X, p_Y$ are the joint and marginal distributions, respectively.
3. Let P, Q be probability measures on S^n. For convenience relabel $S^n = \{1, \dots, k\}, k = |S|^n$, $q_j = Q(\{j\}), p_j = P(\{j\})$. Prove Gibbs inequality: $\sum_j p_j \ln p_j \geq \sum_j p_j \ln q_j$. [*Hint*: Consider $\sum_j p_j \ln \frac{q_j}{p_j}$ and bound $\ln x \leq x - 1, x > 0$.
4. Give a proof of the Kraft inequality for the message length associated with G. [*Hint*: $\sum_x e^{\lceil \ln Q_G(x) \rceil} \leq \sum_x e^{\ln Q_G(x)}$.
5. Show that the problem $\min_{L \in \mathcal{L}_n : \sum_x e^{-L(x)} \leq 1} \sum_{x \in S^n} p_{X^{(n)}} L(x)$ has a solution. [*Hint*: Use Lagrange multipliers to minimize $J = \sum_{x \in S^n} p_{X^{(n)}} L(x) + \lambda \sum_{x \in S^n} e^{-L(x)}$. Derivatives with respect to each $L(x)$ are zero and λ can be determined from the constraint $\sum_x e^{-L(x)} \leq 1$.
6. Let u, v be twice-continuously differentiable functions on $\mathbb{R}$ with Legendre transforms u^*, v^*, respectively, where $f^*(x) = \sup_{h \in \mathbb{R}} \{xh - f(h)\}$, $x \in \mathbb{R}$. Show that (a) u^* is convex. (b) (Idempotency) $u^{**} = u$[*Hint*: Write $u^*(x) = xh(x) - u(h(x))$ and use the smoothness hypothesis on u to optimize.
7. Give a proof for (22.6).
8. Show that $I(M) = \theta_1^2/8$.
9. Give a proof of (22.6).
10. Verify the hypotheses (i),(ii) in Theorem 22.10 for the tests M, T, S. [Hint: For M, let $T_n = n^{\frac{1}{2}}(\overline{X} - \theta_0)$, then (i) is satisfied, since for $\theta > \theta_0$, $n^{-\frac{1}{2}} T_n \to \theta - \theta_0$. For assumption (ii) assume that X_j has a finite moment generating function, and use the Cramér-Chernoff large deviation rate. A similar, but little longer, proof applies to the statistic T, using independence of the sample mean and sample variance. For S one uses the moment generating function of Bernoulli variables.]
11. Show that $E_C(S, M) \to 2/\pi = E_P(S, M)$ as $\theta_1 \downarrow 0$.
12. Show using Bahadur's theorem, that $c_M(\theta) = \theta^2$, $c_T(\theta) = \ln(1 + \theta^2)$.

Chapter 23
Special Topic: Associated Random Fields, Positive Dependence, FKG Inequalities

The notion of *association* is a form of positive dependence among random variables independently introduced in reliability theory, percolation theory and statistical physics, where it is expressed in a form known as the "FKG-Inequalities." The main focus of this chapter is (i) a proof of Newman's central limit theorem for associated random fields with summable fast decay of correlations, and (ii) Pitt's characterization of association of multidimensional Gaussian distributions by non-negativity of covariances.

The notion of association as a form of positive dependence has proved to be of much interest in statistical physics,[1] but its potential importance goes beyond statistical physics applications. In 1980 C. M. Newman[2] announced a central limit theorem for associated random fields that will be the focus of this chapter. For stationary random fields the role of association in the asymptotic distribution of centered and scaled sums may be compared to that of martingales for stationary sequences, where only the finiteness of second moments come into play.

In their paper Esary et al. (1967), the notion is developed as a natural extension of weaker forms of positive dependence motivated by applications to reliability theory. It first appeared in Harris (1960), and was later generalized in Fortuin et al. (1971).

[1] A notion of *negative dependence* was explored by Pemantle (2000) from the perspective of statistical physics by way of stimulating examples and conjectures. However, the development of a comparable mathematical theory appears to be much less fruitful.

[2] The central limit theorem of Newman (1980) was extended to a functional central limit theorem for stationary associated sequences in Newman and Wright (1981).

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_23

We will restrict the exposition to *random fields* of real-valued random variables $\{X_x : x \in \mathbb{Z}^k\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and indexed by the k-dimensional integer lattice $\mathbb{Z}^k$. Here the natural extension of *stationarity* of sequences to that of random fields is as follows.

Definition 23.1. The random field $\mathbf{X} := \{X_x : x \in \mathbb{Z}^k\}$ is said to be *translation invariant* if for each fixed $z \in \mathbb{Z}^k$ the random field $\{X_{x+z} : x \in \mathbb{Z}^k\}$ is distributed as $\mathbf{X}$.

Definition 23.2. A finite set of random variables $X_1, \ldots, X_m$ is said to be *associated* if

$$\begin{aligned}&\mathrm{Cov}(f(X_1, \ldots, X_m), g(X_1, \ldots, X_m))\\ &\quad \equiv \mathbb{E}f(X_1, \ldots, X_m)g(X_1, \ldots, X_m) - \mathbb{E}f(X_1, \ldots, X_m)\mathbb{E}g(X_1, \ldots, X_m) \geq 0\end{aligned}$$

for any pair of bounded measurable coordinatewise non-decreasing functions f, g. An arbitrary collection $\{X_\lambda : \lambda \in \Lambda\}$ is said to be *associated* if every finite subcollection is associated.

The inequalities (23.1) are referred to as the *Fortuin–Kasteleyn–Ginbre (FKG) Inequalities*.[3] Let us begin with a useful formula for covariance in this context. The special case of this formula with $f(x) = x, g(y) = y$ was derived in Lehmann (1966) with attribution to Hoeffding (1940). Newman (1980) noticed the simple but significant extension presented here. (Recall the Definition 2.1 of the covariance of complex-valued random variables.)

Lemma 1 *(Hoeffding-Newman Covariance Formula).* Suppose that $f(X), g(Y) \in L^2(\Omega, \mathcal{F}, P)$ and assume f, g are continuously differentiable complex-valued functions on $\mathbb{R}$ having bounded derivatives. Then,

$$\mathrm{Cov}(f(X), g(Y)) = \int_{\mathbb{R}}\int_{\mathbb{R}} f'(x)\overline{g}'(y)H_{X,Y}(x, y)dxdy,$$

where

$$\begin{aligned}H_{X,Y}(x, y) = \mathrm{Cov}(\mathbf{1}_{[X>x]}, \mathbf{1}_{[Y)>y]}) &= P(X > x, Y > y)\\ &\quad - P(X > x)P(Y > y),\ \ x, y \in \mathbb{R}.\end{aligned}$$

Proof. Let (X_1, Y_1) and (X_2, Y_2) be independent random vectors distributed as (X, Y). Note that $\mathbf{1}_{(u,\infty)}(X_1) - \mathbf{1}_{(u,\infty)}(X_2)$ is 1 if $X_2 < u < X_1$, -1 if $X_1 < u < X_2$, and 0 otherwise. Thus, by the fundamental theorem of calculus,

[3] Fortuin et al. (1971).

$$f(X_1) - f(X_2) = \int_{-\infty}^{\infty} f'(u)\{\mathbf{1}_{(u,\infty)}(X_1) - \mathbf{1}_{(u,\infty)}(X_2)\}du.$$

Similarly,

$$\overline{g}(X_1) - \overline{g}(X_2) = \int_{-\infty}^{\infty} \overline{g}'(u)\{\mathbf{1}_{(u,\infty)}(Y_1) - \mathbf{1}_{(u,\infty)}(Y_2)\}du.$$

Thus,

$$\begin{aligned}
&2\mathrm{Cov}(f(X), g(Y)) \\
&\quad = \mathbb{E}[f(X_1) - f(X_2)][\overline{g}(Y_1) - \overline{g}(Y_2)] \\
&\quad = \mathbb{E}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(\mathbf{1}_{[X_1>u]} - \mathbf{1}_{[X_2>u]}\big)\big(\mathbf{1}_{[Y_1>v]} - \mathbf{1}_{[Y_2>v]}\big) f'(u)\overline{g}'(v)dudv.
\end{aligned}$$

The formula follows by an application of Fubini's theorem to interchange expected value with integrals, after canceling the factors of 2, since expanding the product of indicators one also has by independence and the specified common joint distributions of $(X_i, Y_i), i = 1, 2$, that

$$\mathbb{E}\big(\mathbf{1}_{[X_1>u]} - \mathbf{1}_{[X_2>u]}\big)\big(\mathbf{1}_{[Y_1>v]} - \mathbf{1}_{[Y_2>v]}\big) = 2\{P(X_1 > u, Y_1 > v) - P(X_1 > u)P(Y_1 > v)\}.$$

■

Remark 23.1. Under the same conditions, the covariance formula may be expressed equivalently as

$$\mathrm{Cov}(f(X), g(Y)) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathrm{Cov}(\mathbf{1}_{[X>x]}, \mathbf{1}_{[Y>y]}) f'(x)\overline{g}'(y)dxdy.$$

Definition 23.3. A pair of real-valued random variables X, Y for which

$$P(X > u, Y > v) - P(X > u)P(Y > v) \geq 0 \quad \text{for all } u, v \in \mathbb{R},$$

is said to be *positive quadrant dependent*[4]

Proposition 23.1. Associated random variables are (pairwise) positive quadrant dependent.

Proof. Simply note that for any fixed number $a \in \mathbb{R}$, a function of the form $f(u) = \mathbf{1}_{[a,\infty)}(u)$ is non-decreasing. ■

Newman's proof of the central limit theorem exploits the covariance formulae to compare characteristic functions of sums of random variables with the corre-

[4] The notion of positive quadrant dependence was introduced by Lehmann (1966).

sponding product of characteristic functions through the following key lemma. The non-negativity of the covariance is essential to this comparison.

***Lemma 2** (Newman).* Suppose that $f(X), g(Y) \in L^2(\Omega, \mathcal{F}, P)$ where X, Y are positive quadrant dependent and f, g are continuously differentiable complex-valued functions on $\mathbb{R}$ with bounded derivatives. Then

$$|\text{Cov}\big(f(X), g(Y)\big)| \leq ||f'||_\infty ||g'||_\infty \text{Cov}(X, Y),$$

where $||\cdot||_\infty$ denotes the essential supremum norm. In particular,

$$|\mathbb{E}e^{irX+isY} - \mathbb{E}e^{irX}\mathbb{E}e^{isY}| \leq |r||s|\text{Cov}(X, Y), \quad r, s \in \mathbb{R}.$$

Proof. Using Lemma 1, the assertion follows from the triangle inequality, bounding the derivatives, and the positivity of $H(x, y)$. Specifically,

$$|\text{Cov}\big(f(X), g(Y)\big)| \leq ||f'||_\infty ||g'||_\infty \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x, y)dxdy = ||f'||_\infty ||g'||_\infty \text{Cov}(X, Y).$$

This completes the proof of the general bound. The second bound is simply an application. ■

Let us say that a collection of functions $\mathcal{C}$ is *association determining* if one may restrict the FKG inequalities to $f, g \in \mathcal{C}$ to establish association. A proof of the following proposition is left to Exercise 8.

Proposition 23.2. The collections of coordinatewise non-decreasing binary $0-$1-valued functions, and of coordinatewise non-decreasing bounded continuous functions, respectively, are each association determining.

The following properties are useful in "tracking association" and/or building examples of associated families of random variables.

Proposition 23.3.

1. Any subcollection of associated random variables is associated.
2. The union of independent collections of associated random variables is associated.
3. Measurable coordinatewise non-decreasing or coordinatewise nonincreasing functions of associated random variables are associated.
4. If for each n, $X_1^{(n)}, \ldots, X_m^{(n)}$ is associated and if $(X_1^{(n)}, \ldots, X_m^{(n)})$ converges in distribution to $(X_1, \ldots, X_m)$, then $X_1, \ldots, X_m$ is associated.
5. A singleton $\{X_1\}$ is associated.
6. Independent random variables are associated.
7. If X, Y are binary random variables, the X, Y are associated if and only if $\text{Cov}(X, Y) \geq 0$.

Proof. Part (1) follows directly from definition by considering functions whose values do not depend on variables not included in the subset. For (2), let $\mathbf{X} = (X_1, \dots, X_m)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$ be two independent sequences of associated random variables. Let $\mathbf{Z} = (X_1, \dots, X_m, Y_1, \dots Y_n)$. For non-decreasing bounded measurable functions f, g of $m + n$ variables, since the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ is a product measure by independence, one has

$$
\begin{aligned}
&\mathrm{Cov}(f(\mathbf{Z}), g(\mathbf{Z})) = \mathbb{E}f(\mathbf{Z})g(\mathbf{Z}) - \mathbb{E}f(\mathbf{Z})\mathbb{E}g(\mathbf{Z})\\
&\quad= \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(z_1, \dots, z_{m+n})g(z_1, \dots, z_{m+n})P_X(dz_1 \times \cdots \times dz_m)P_Y(dz_{m+1} \times \cdots \times dz_{m+n})\\
&\qquad- \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(z_1, \dots, z_{m+n})P_X(dz_1 \times \cdots \times dz_m)P_Y(dz_{m+1} \times \cdots \times dz_{m+n})\\
&\qquad\times \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} g(z_1, \dots, z_{m+n})P_X(dz_1 \times \cdots \times dz_m)P_Y(dz_{m+1} \times \cdots \times dz_{m+n})\\
&\quad= \int_{\mathbb{R}^n}\Big\{\int_{\mathbb{R}^m} f(z_1, \dots, z_{m+n})g(z_1, \dots, z_{m+n})dP_X\\
&\qquad- \int_{\mathbb{R}^m} f(z_1, \dots, z_{m+n})dP_X \int_{\mathbb{R}^m} g(z_1, \dots, z_{m+n})dP_X\Big\}dP_Y\\
&\qquad+ \int_{\mathbb{R}^n}\Big\{\int_{\mathbb{R}^m} f(z_1, \dots, z_{m+n})dP_X \int_{\mathbb{R}^m} g(z_1, \dots, z_{m+n})dP_X\Big\}dP_Y\\
&\qquad- \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(z_1, \dots, z_{m+n})dP_X dP_Y \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} g(z_1, \dots, z_{m+n})dP_X dP_Y\\
&\quad= \int_{\mathbb{R}^n} \mathrm{Cov}\Big(f(X_1, \dots, X_m, z_{m+1}, \dots, z_{n+m}), g(X_1, \dots, X_m, z_{m+1}, \dots, z_{n+m})\Big)dP_Y\\
&\qquad+ \mathrm{Cov}\left(\int_{\mathbb{R}^m} f(z_1, \dots, z_m, Y_1, \dots, Y_n)dP_X, \int_{\mathbb{R}^m} g(z_1, \dots, z_m, Y_1, \dots, Y_n)dP_X\right) \ge 0,
\end{aligned}
$$

where $dP_X = P_X(dz_1 \times \cdots \times dz_m), dP_Y = P_X(dz_{m+1} \times \cdots \times dz_{m+n})$. The proof of part (3) follows directly from the definition since if $X_1, \dots, X_m$ are associated and $Y_i = h_i(X_1, \dots, X_m)$ for measurable coordinatewise non-decreasing functions $h_1, \dots, h_m$, then $f(h_1, \dots, h_m)$ and $g(h_1, \dots, h_m)$ are bounded measurable coordinatewise non-decreasing whenever the same is true of f, g. For the coordinatewise nonincreasing case the composites $f(h_1, \dots, h_m)$ and $g(h_1, \dots, h_m)$ are bounded measurable coordinatewise nonincreasing for coordinatewise non-decreasing f, g. Now, $\mathrm{Cov}\big(f(h_1, \dots, h_m), g(h_1, \dots, h_m)\big) = \mathrm{Cov}\big(-f(h_1, \dots, h_m), -g(h_1, \dots, h_m)\big)$ and $-f(h_1, \dots, h_m)$ and $-g(h_1, \dots, h_m)$ are bounded measurable coordinatewise non-decreasing. For part (4), by definition of weak convergence, $\mathrm{Cov}(f(\mathbf{X}), g(\mathbf{X})) = \lim_{n\to\infty} \mathrm{Cov}(f(\mathbf{X}^{(n)}), g(\mathbf{X}^{(n)}))$ for bounded continuous functions f, g. Therefore the result follows since, by the previous proposition, bounded continuous coordinatewise non-decreasing are association determining. To prove (5) restrict to the association determining class of non-decreasing binary functions, and observe that for non-decreasing binary functions f, g of a single variable one has either $f \le g$ or $g \le f$. Without loss of generality consider the case $f \leq g$. Then $\mathrm{Cov}(f(X_1), g(X_1)) = \mathbb{E}f(X_1)g(X_1) - \mathbb{E}f(X_1)\mathbb{E}g(X_1) = \mathbb{E}f(X_1) - \mathbb{E}f(X_1)\mathbb{E}g(X_1) = \mathbb{E}f(X_1)(1-g(X_1)) \geq 0$. Property (6) follows by application of (2) and (5). Obviously $\mathrm{Cov}(X,Y) \geq 0$ is necessary for X, Y to be associated. For binary $0-1$-valued, X, Y, suppose $\mathrm{Cov}(X,Y) \geq 0$. Then the only binary coordinatewise non-decreasing $0-1$-valued functions of X, Y are $f_0 \equiv 0$, $f_1 \equiv 1$, $g_0(x,y) = x, g_1(x,y) = y$, $h_(x,y) = x \vee y, x \in \{0,1\}$. Also, $f_0 \leq g_0, g_1 \leq h \leq f_1$. In particular, this ordering trivially implies that $\mathrm{Cov}(f_j(X,Y), g_i(X,Y))] \geq 0$, $\mathrm{Cov}(g_i(X,Y), h(X,Y)) \geq 0$, $\mathrm{Cov}(f_j(X,Y), h(X,Y)) \geq 0$ for $i, j = 0, 1$. The case $\mathrm{Cov}(g_0(X,Y), g_1(X,Y)) \geq 0$ is the hypothesis. Thus, X, Y is an associated pair proving part (7). ■

Remark 23.2. An alternative proof of the association of a single random variable by *coupling* is an Exercise 2 in Chapter 24.

Example 1 *(A Tendency to Align Under Associated Dependence).* The purpose of this example[5] is to illustrate the tendency for alignment under associated dependence. Consider identically distributed Bernoulli $0-1$-valued random variables Y_0, Y_1 with distribution specified by $P(Y_0 = j) = 1/2, P(Y_1 = j|Y_0 = j) = p, j = 0, 1$ for $p \in (0,1)$. Association requires that Y_1 be most likely to align with the given value of Y_0. That is,

Proposition 23.4. Y_0, Y_1 is associated if and only if $p \geq 1/2$.

Proof. First observe that taking $f(i,j) = i$ and $g(i,j) = j, i, j = 0, 1$, one has that $\mathrm{Cov}(f(Y_0,Y_1), g(Y_0,Y_1)) = \mathrm{Cov}(Y_0,Y_1) = \frac{1}{2}p - \frac{1}{4} \geq 0$ if and only if $p \geq 1/2$. Thus $p \geq 1/2$ is necessary for association. Since Y_0, Y_1 are binary, it likewise follows from Proposition 23.3(g) that $p \geq 1/2$ is sufficient as well. ■

Remark 23.3. In the context of statistical physics association is often expressed as a property of the joint distribution μ of coordinate maps $X_x, x \in \Lambda$, on the product space $\Omega = \{-1,1\}^\Lambda$ for some finite set Λ of integer lattice points connected to the origin; i.e., $X_x(\omega) = \omega_x, \omega \in \Omega$. The probability measure μ is said to satisfy the *FKG inequalities* if for any coordinatewise non-decreasing functions f, g on Ω one has

$$\int_\Omega f(X_x)g(X_y)d\mu \geq \int_\Omega f(X_x)d\mu \int_\Omega g(X_y)d\mu, \quad x, y \in \Lambda. \tag{23.1}$$

Equivalently, the FKG inequalities are the property that the collection of spin ±1-valued random variables $X_x, x \in \Lambda$ have associated dependence. The ferromagnetic Ising model (see Chapter 13, Exercise 13) provides a well-known example in this context. The FKG inequalities for the ferromagnetic Ising model will be proved in Chapter 24, Proposition 24.12. The magnetic spin alignment reflected by association is a distinct feature of ferromagnets, responsible for their ability

[5] Also see Exercise 8 in Chapter 24 in this regard.

to be magnetized by placement in an external magnetic field. Other more general inequalities to be considered in Chapter 24 are available that imply association.[6]

The proof of the central limit theorem exploits the following basic inequality.

***Lemma 3** (Newman's Inequality).* Suppose that $X_1, \ldots, X_m$ are associated random variables having finite variance. Then for any $r_1, \ldots, r_m \in \mathbb{R}$ one has

$$|\mathbb{E}\exp\{i\sum_{j=1}^{m} r_j X_j\} - \prod_{j=1}^{m} \mathbb{E}e^{ir_jX_j}| \leq \sum_{1\leq j<k\leq m} |r_j||r_k|\operatorname{Cov}(X_j, X_k).$$

Proof. The proof is by induction on m. The case $m = 1$ is obvious and the case $m = 2$ was proven in Lemma 2. Assume the inequality holds for all $m \leq M$ and rearrange the indices (if necessary) in such a way that $\operatorname{sgn}(r_j)$ is constant, say ϵ (either $+1$ or -1), for $1 \leq j \leq m_0$, and $\operatorname{sgn}(r_j)$ is also constant, say δ, for $m_0 + 1 \leq j \leq M$. Then $\epsilon r_j \geq 0, \delta r_j \geq 0$, so that each of $X = \sum_{j=1}^{m_0} \epsilon r_j X_j$ and $Y = \sum_{j=m_0+1}^{M+1} \delta r_j X_j$ is a non-decreasing function of associated variables $X_1, \ldots, X_{M+1}$ and therefore associated. Also $\sum_{j=1}^{M+1} r_j X_j = \epsilon X + \delta Y$. Thus, applying Lemma 2 and the induction hypothesis, one has

$$\begin{aligned}
&\left|\mathbb{E}\exp\left\{i\sum_{j=1}^{M+1} r_jX_j\right\} - \prod_{j=1}^{M+1}\mathbb{E}e^{ir_jX_j}\right| \\
&\leq \left|\mathbb{E}e^{i(\epsilon X+\delta Y)} - \mathbb{E}e^{i\epsilon X}\mathbb{E}e^{i\delta Y}\right| + \left|\mathbb{E}e^{i\epsilon X}\mathbb{E}e^{i\delta Y} - \mathbb{E}e^{i\epsilon X}\prod_{j=m_0+1}^{M+1}\mathbb{E}e^{ir_jX_j}\right| \\
&\quad + \left|\mathbb{E}e^{i\epsilon X}\prod_{j=m_0+1}^{M+1}\mathbb{E}e^{ir_jX_j} - \left(\prod_{j=1}^{m_0}\mathbb{E}e^{ir_jX_j}\right)\prod_{j=m_0+1}^{M+1}\mathbb{E}e^{ir_jX_j}\right| \\
&\leq |\epsilon||\delta|\operatorname{Cov}(X, Y) + \left|\mathbb{E}e^{i\delta Y} - \prod_{j=m_0+1}^{M+1}\mathbb{E}e^{ir_jX_j}\right| + \left|\mathbb{E}e^{i\epsilon X} - \prod_{j=1}^{m_0}\mathbb{E}e^{ir_jX_j}\right| \\
&\leq \operatorname{Cov}\left(\sum_{j=1}^{m_0}\epsilon r_jX_j, \sum_{k=m_0+1}^{M+1}\delta r_kX_k\right) + \sum_{m_0+1\leq j<k\leq M+1}|r_j||r_k|\operatorname{Cov}(X_j, X_k) \\
&\quad + \sum_{1\leq j<k\leq m_0}|r_j||r_k|\operatorname{Cov}(X_j, X_k) \\
&= \sum_{1\leq j<k\leq M+1}|r_j||r_k|\operatorname{Cov}(X_j, X_k).
\end{aligned}$$

■

[6] See den Hollander and Keane (1986).

Of course, one may prefer the equivalent expression of Newman's bound as

$$\frac{1}{2}\sum_{1\le j,k\le m, j\ne k} |r_j||r_k|\operatorname{Cov}(X_j, X_k) = \sum_{1\le j<k\le M+1} |r_j||r_k|\operatorname{Cov}(X_j, X_k). \tag{23.2}$$

Lemma 4. Let $\mathbf{X} := \{X_x : x \in \mathbb{Z}^k\}$ be a translation invariant random field of associated random variables having finite second moments. Assume that

$$\gamma := \sum_{x\in\mathbb{Z}^k} \operatorname{Cov}(X_0, X_x) < \infty.$$

Let

$$B_x^{(N)} := \{y \in \mathbb{Z}^k : Nx_l \le y_l < N(x_l + 1), l = 1, \dots, k\}$$

denote a "block of lattice sites of length N located near Nx", $x = (x_1, \dots, x_k)$, and define a random field of centered and rescaled "block sum averages" by

$$A_x^{(N)} = N^{-\frac{k}{2}} \sum_{y\in B_x^{(N)}} (X_y - \mathbb{E}X_y), \qquad x \in \mathbb{Z}^k.$$

Then

$$\lim_{N\to\infty} \operatorname{Var}(A_x^{(N)}) = \gamma, \quad \text{and} \quad \lim_{N\to\infty} \operatorname{Cov}(A_x^{(N)}, A_y^{(N)}) = 0 \quad x \ne y.$$

Proof. By translation invariance it suffices to check the asserted limits for the case $x = 0$. Clearly

$$\operatorname{Var}(A_0^{(N)}) = N^{-k} \sum_{x\in B_0^{(N)}} \sum_{y\in B_0^{(N)}} \operatorname{Cov}(X_0, X_{y-x}) \le N^{-k} \sum_{x\in B_0^{(N)}} \sum_{y\in\mathbb{Z}^k} \operatorname{Cov}(X_0, X_{y-x}).$$

In particular, letting $N \to \infty$,

$$\limsup_{N\to\infty} \operatorname{Var}(A_0^{(N)}) \le \gamma.$$

For the reverse inequality let $0 < \epsilon < 1/2$ and define

$$B_0^{(N)}(\epsilon) := \{z = (z_1, \dots, z_k) : \epsilon N < z_i < (1 - \epsilon)N, i = 1, \dots, k\}.$$

Note that for $x \in B_0^{(N)}$, $y \notin B_0^{(N)}$, $|x - y| \ge \epsilon N$, so that

$$\begin{aligned}\mathrm{Var}(A_0^{(N)}) &\geq N^{-k} \sum_{x\in B_0^{(N)}(\epsilon)} \sum_{y\in B_0^{(N)}} \mathrm{Cov}(X_0, X_{y-x}) \\ &\geq N^{-k} \sum_{x\in B_0^{(N)}(\epsilon)} \sum_{|y-x|\leq \epsilon N} \mathrm{Cov}(X_0, X_{y-x}) = \frac{|B_0^{(N)}(\epsilon)|}{N^k} \sum_{|z|\leq \epsilon N} \mathrm{Cov}(X_0, X_z),\end{aligned}$$

where $|B|$ denotes cardinality of the set B. Choosing a sequence such that $\epsilon_N \downarrow 0$ and $\epsilon_N N \to \infty$, one obtains

$$\liminf_{N\to\infty} \mathrm{Var}(A_0^{(N)}) \geq \gamma.$$

This proves the asserted asymptotic variance. For the covariance decay choose a sequence $M_N \leq N$ such that $M_N/N \to 1$ and $N - M_N \to \infty$ as $N \to \infty$.

$$\begin{aligned}&\mathrm{Var}(A_0^{(N)} - A_0^{(M_N)}) \\ &= \mathrm{Var}(A_0^{(N)}) + \mathrm{Var}(A_0^{(M_N)}) - 2(NM_N)^{-\frac{k}{2}} \mathrm{Cov}(\sum_{y\in B_0^{(N)}} X_y, \sum_{y\in B_0^{(M_N)}} X_y) \\ &\leq \mathrm{Var}(A_0^{(N)}) + \mathrm{Var}(A_0^{(M_N)}) - 2(\frac{M_N}{N})^{\frac{k}{2}} \mathrm{Var}(X_0^{(M_N)}) \to 0.\end{aligned}$$

One has for $z \neq 0$,

$$\begin{aligned}\mathrm{Cov}(A_0^{(N)}, A_z^{(N)}) &= \mathrm{Cov}(A_0^{(N)} - A_0^{(M_N)}, A_z^{(N)}) + \mathrm{Cov}(A_0^{(M_N)}, A_z^{(N)}) \\ &\leq \sqrt{\mathrm{Var}(A_0^{(N)} - A_0^{(M_N)})}\sqrt{\mathrm{Var}(A_z^{(N)})} + \mathrm{Cov}(A_0^{(M_N)}, A_z^{(N)}).\end{aligned}$$

Thus the proof of covariance decay is therefore completed by the following calculation

$$\begin{aligned}\mathrm{Cov}(A_0^{(M_N)}, A_z^{(N)}) &= M_N^{-\frac{k}{2}} N^{-\frac{k}{2}} \sum_{x\in B_0^{(M_N)}} \sum_{y\in B_z^{(N)}} \mathrm{Cov}(X_0, X_{x-y}) \\ &\leq \left(\frac{M_N}{N}\right)^{\frac{k}{2}} M_N^{-k} \sum_{x\in B_0^{(M_N)}} \sum_{|y-x|\geq N-M_N} \mathrm{Cov}(X_0, X_{x-y}) \\ &= \left(\frac{M_N}{N}\right)^{\frac{k}{2}} \sum_{|y|\geq N-M_N} \mathrm{Cov}(X_0, X_y) \to 0,\end{aligned}$$

as $N \to \infty$. ■

To state Newman's central limit theorem it is helpful to have some extra notation. For $x = (x_1, \dots, x_k) \in \mathbb{Z}^k$, a "block of lattice sites of length N located near Nx" is denoted

$$B_x^{(N)} := \{y \in \mathbb{Z}^k : Nx_l \le y_l < N(x_l + 1), l = 1, \dots, k\}.$$

Given a translation invariant random field with finite second moments $\mathbf{X} := \{X_x : x \in \mathbb{Z}^k\}$, the random field of centered and rescaled "block sum averages" is denoted

$$A_x^{(N)} = N^{-\frac{k}{2}} \sum_{y \in B_x^{(N)}} (X_y - \mathbb{E}X_y).$$

Theorem 23.5 *(Newman's Central Limit Theorem).* Let $\mathbf{X} := \{X_x : x \in \mathbb{Z}^k\}$ be a translation invariant random field of associated random variables having finite second moments. Assume that

$$\gamma := \sum_{x \in \mathbb{Z}^k} \operatorname{Cov}(X_0, X_x) < \infty.$$

Then for any finite number n of lattice sites $z_1, \dots, z_n$, the (finite dimensional) distribution of $(A_{z_1}^{(N)}, A_{z_2}^{(N)}, \dots, A_{z_n}^{(N)})$ converges weakly as $N \to \infty$ to the Gaussian distribution with mean zero and covariance matrix $\operatorname{diag}(\gamma, \dots, \gamma)$.

Proof. By Newman's inequality and association inherited by the $A_z^{(N)}, z \in \mathbb{Z}^k$, and the previous lemma, it suffices to show convergence of $A_z^{(N)}$, i.e., $n = 1$, to obtain convergence for finite dimensional distributions of arbitrary size $n \ge 1$. More specifically, if one can show $\mathbb{E}e^{irA_z^{(N)}} \to e^{-\frac{\gamma}{2}r^2}$ as $N \to \infty$, then

$$\lim_{N\to\infty} \left| \mathbb{E}e^{i\sum_{j=1}^n r_j A_{z_j}^{(N)}} - \prod_{j=1}^n e^{-\frac{\gamma}{2}r_j^2} \right| \le \lim_{N\to\infty} \sum_{1 \le m < j \le n} |r_m||r_j| \operatorname{Cov}(A_{z_m}^{(N)}, A_{z_j}^{(N)}) = 0. \tag{23.3}$$

As noted earlier, by translation invariance it is sufficient to consider the case $z = 0$. For fixed $M = 1, 2, \dots$, let $M_N = M[\frac{N}{M}] \le N$, where $[\cdot]$ denotes integer-part. In the proof of the previous lemma it was shown that $\operatorname{Var}(A_0^{(N)} - A_0^{(M_N)}) \to 0$ as $N \to \infty$. Thus, one has

$$\begin{aligned} \left| \mathbb{E}e^{irA_0^{(N)}} - \mathbb{E}e^{irA_0^{(M_N)}} \right| &\le \mathbb{E}\left| e^{ir(A_0^{(N)} - A_0^{(M_N)})} - 1 \right| \\ &\le \mathbb{E}\left| A_0^{(N)} - A_0^{(M_N)} \right| \le \sqrt{\operatorname{Var}(A_0^{(N)} - A_0^{(M_N)})} \to 0. \end{aligned}$$

Next, using the simple property of the block averages that

$$A_0^{(N_1 N_2)} = N_1^{-\frac{k}{2}} \sum_{y \in B_0^{(N_1)}} A_0^{(N_2)}, \tag{23.4}$$

(for $M_N = M[\frac{N}{M}] = N_1 N_2$), one has by Newman's inequality (applied to $A_0^{(M)}$)

$$\left| \mathbb{E} e^{irA_0^{(M[\frac{N}{M}])}} - \left(\mathbb{E} e^{ir[\frac{N}{M}]^{-\frac{k}{2}} A_0^{(M)}} \right)^{([\frac{N}{M}])^k} \right| \leq \frac{1}{2} \sum_{x,y \in B_0^{([\frac{N}{M}])} \, x \neq y} r^2 \left(\left[\frac{N}{M} \right] \right)^{-k} \mathrm{Cov}\left(A_x^{(M)}, A_y^{(M)} \right).$$

This upper bound may be equivalently expressed using the block average property (23.4) as

$$\frac{r^2}{2} \left\{ \mathrm{Cov}\left(A_0^{(M[\frac{N}{M}])}, A_0^{(M[\frac{N}{M}])} \right) - \left[\frac{N}{M} \right]^{-k} \sum_{y \in B_0^{([\frac{N}{M}])}} \mathrm{Cov}\left(A_y^{(M)}, A_y^{(M)} \right) \right\}$$

$$= \frac{r^2}{2} \left\{ \mathrm{Var}\left(A_0^{(M[\frac{N}{M}])} \right) - \mathrm{Var}\left(A_0^{(M)} \right) \right\} \to \frac{r^2}{2} \left\{ \gamma - \mathrm{Var}\left(A_0^{(M)} \right) \right\}.$$

Letting $N \to \infty$ with M fixed, it follows that

$$\begin{aligned} \left(\mathbb{E} e^{ir[\frac{N}{M}]^{-\frac{k}{2}} A_0^{(M)}} \right)^{([\frac{N}{M}])^k} &= \left(1 - \frac{r^2}{2} ([\frac{N}{M}])^{-k} \mathrm{Var}(A_0^{(M)}) + o([\frac{N}{M}]^{-k}) \right) \\ &\to e^{-\frac{\mathrm{Var}(A_0^{(M)})}{2} r^2}. \end{aligned} \tag{23.5}$$

Thus, combining these estimates, one has

$$\limsup_{N \to \infty} \left| \mathbb{E} e^{irA_0^{(N)}} - e^{-\frac{\gamma}{2} r^2} \right| \leq \frac{r^2}{2} \left\{ \gamma - \mathrm{Var}(A_0^{(M)}) \right\} + \left\{ e^{-\frac{\mathrm{Var}(A_0^{(M)})}{2} r^2} - e^{-\frac{\gamma}{2} r^2} \right\}.$$

Finally, letting $M \to \infty$ completes the proof. ■

The following example[7] is a significant framework[8] in which association naturally occurs.

Example 2 *(Two-dimensional Bond Percolation Model).* The *independent bond percolation model*[9] on $\mathbb{Z}^2$ can be defined as follows: Each lattice site $x \in \mathbb{Z}^2$ has four *nearest neighbor sites* of the form $y = x \pm e$ where e is either $(1, 0)$ or $(0, 1)$. A pair of such nearest neighbor sites x, y, in turn, defines a (unoriented) *bond* $b = \{x, y\}$ of $\mathbf{Z}^2$. Let $\mathbb{L}^2$ denote the collection all such bonds of $\mathbb{Z}^2$. Let $\{Y_b : b \in \mathbb{L}^2\}$ be the i.i.d. random field of Bernoulli $0 - 1$ valued random variables with $p = P_p(Y_b = 1)$ defined by coordinate projections on the product probability space $\Omega = \{0, 1\}^{\mathbb{L}^2}$ equipped with the σ-field $\mathcal{F}$ generated by finite dimensional cylinder sets and product measure $P_p = \prod_{\mathbb{L}^2}(q\delta_{\{0\}} + p\delta_{\{1\}})$, where $q = 1 - p$. Declare the bonds b as *open* or *closed* according to whether the value of Y_b is 1 or 0, respectively. The usual interpretation of percolation is as a model for a disordered porous medium in which the open bonds permit fluid flow between nearest neighbor sites, while closed bonds block the passage of fluid. Two sites $x, z \in \mathbb{Z}^2$ are said to be *connected by an open path*, denoted $x \leftrightarrow z$ if there is a succession of sites in $\mathbb{Z}^2$, $x_0 = x, x_1, \ldots x_m = z$, $m \geq 1$, such that pairs x_i, x_{i+1} are nearest neighbor with $b_i = \{x_i, x_{i+1}\}$ open $(i = 0, \ldots, m - 1)$. A *cluster* $C(x)$ at site $x \in \mathbb{Z}^2$ is defined by the (random) set

$$C(x) := \{z \in \mathbb{Z}^2 : x \leftrightarrow z\}, \qquad x \in \mathbb{Z}^2.$$

The *cluster size* refers to the (possibly infinite) cardinality of $C(x)$ and is denoted by $|C(x)|$. The set $C(x)$ is referred to as a *percolation cluster*[10] at x if $|C(x)| = \infty$.

Definition 23.4. The existence of an infinite cluster that is the event $E := \cup_{x\in\mathbb{Z}^2}[|C(x)| = \infty]$ is referred to as the *percolation* event. Also, the *percolation probability* is defined by

$$\rho \equiv \rho(p) := P_p(E) = P_p(\cup_{x\in\mathbb{Z}^2}[|C(x)| = \infty]). \tag{23.6}$$

[7] The survey article Last et al. (2020) is a source of a wide variety of additional examples of associated stochastic random fields. Extension of Newman's central limit theorem for Poisson cluster processes and random measures was developed in Burton and Waymire (1985), and independently by Evans (1989), provides an illustrative setting for applications of positive dependence.

[8] Also see Newman (1980) for examples in the context of mathematical physics.

[9] The mathematical interest in percolation models is usually traced to Broadbent and Hammersley (1957). Broadbent's work at the British Coal Utilization Research Association involved the design of porous gas masks for coal miners. The critical nature of pore size was empirically realized in this context, motivating the subsequent development of simpler models of such phenomena of wide interest in probability and mathematical physics.

[10] Uniqueness of such percolation clusters was originally established by Aizenmann et al. (1987). A widely recognized very simple proof of uniqueness was subsequently made by Burton and Keane (1989). This has become a standard approach to uniqueness.

Proposition 23.6. Define

$$\theta \equiv \theta(p) := P_p(|C(0)| = \infty). \tag{23.7}$$

Then $\rho(p) = 0$ or $\rho(p) = 1$ if and only if $\theta(p) = 0$ or $\theta(p) > 0$, respectively.

Proof. Note that the percolation event $E = \cup_{x\in\mathbb{Z}^2}[|C(x)| = \infty]$ is a tail event for a countable collection of i.i.d. random variables $\{Y_b : b \in \mathbb{L}^2\}$. The assertion follows immediately from subadditivity and Kolmogorov's zero-one law.[11] Namely, $\rho(p) = 0$ or 1, $\theta(p) \le \rho(p)$, and $\rho(p) \le \sum_{x\in\mathbb{Z}^2}\theta(p)$. So $\rho(p) = 0$ if and only if $\theta(p) = 0$, and $\theta(p) > 0$ if and only if $\rho(p) = 1$. ■

Remark 23.4. A proof of the monotonicity of the percolation probability $p \to \theta(p)$ as a function of p by monotone coupling techniques is given for Proposition 24.3 in Chapter 24.

Definition 23.5. The *critical probability* for existence of an infinite cluster, i.e., percolation, is defined by

$$p_c = \sup\{p \in [0, 1] : \theta(p) = 0\}.$$

Remark 23.5. An important role for the FKG inequalities occurs in a simplified proof of the criticality of $p = 1/2$ for bond percolation by Bollabás and Riordan (2006). The original proof is the result of Kesten (1980), after completing the upper bound calculation from two-decades earlier by Harris (1960), that $p_c = 1/2$ for $2d$-bond percolation. The upper bound $p_c \le 1/2$ had already involved inequalities, now known as *Harris inequalities*, that may be viewed as a special case of the FKG inequalities for product measure.

For probability measures μ_1 and μ_2 on the compact space (for product topology) $S = \{0, 1\}^\Lambda$, where Λ is a finite or countably infinite set, the *Holley inequalities*[12] are a generalization of associated dependence of the form

$$\int_S f d\mu_1 \ge \int_S f d\mu_2, \tag{23.8}$$

for coordinatewise non-decreasing functions f on S; equivalently it is non-decreasing with respect to the partial order $\preceq$ on S defined by $x \preceq y$ if and only if $x_j \le y_j$, $j \in \Lambda$ for $x, y \in S$,

To see that (23.8) embodies association of a probability distribution μ on S, let f, g be nonnegative coordinatewise non-decreasing functions on S. Take $d\mu_1 = \frac{gd\mu}{\int_S gd\mu}$, $\mu_2 = \mu$. Holley's inequalities for μ_1 and μ_2 are then equivalent to the FKG

[11] See BCPT, p. 87.

[12] Holley (1974).

inequalities for μ. The so-called *log-convexity* type conditions[13] on μ, μ_1, μ_2 are available to ensure either FKG inequalities or Holley inequalities, respectively.

In the so-called *disordered phase*[14] defined by $0 < p < p_c$, the lattice a.s. consists of infinitely many disjoint *finite* random clusters of lattice sites connected by open bonds. The following simple path counting argument demonstrates the existence of a disordered phase.

Proposition 23.7. $p_c > 0$.

Proof. Consider the number N_n of open (self-avoiding) paths of length n starting at the origin. Clearly, noting that such a path can connect to any of the 4 neighbors of $(0, 0)$ and continue in $n - 1$ self-avoiding steps, $N_n \le 4(3^{n-1})$. Thus for $p < 1/3$, applying a useful but very simple inequality for nonnegative integer-valued random variables,

$$P_p(N_n \ge 1) \le \mathbb{E}_p N_n \le 4(3^{n-1})p^n \to 0 \text{ as } n \to \infty.$$

In particular,[15] since $\theta(p) \le P_p(N_n \ge 1)$ for all $n \ge 1$, one has $\theta(p) = 0$ for $p < 1/3$ and hence $p_c \ge 1/3$. ■

Lemma 5 *(Harris' Lemma[16]).* Let $X_x = \mathbf{1}_{[C(x)\neq\emptyset]}$, $x \in \mathbb{Z}^2$. Then $\{X_x : x \in \mathbb{Z}^2\}$ is a translation invariant random field of associated random variables.

Proof. Translation invariance follows directly from the definition and the fact that the distribution of the underlying random field $\{Y_b : b \in \mathbb{L}^2\}$ is invariant under translation of the lattice $\mathbb{Z}^2$. Also each X_x, $x \in \mathbb{Z}^2$, is a (coordinatewise) non-decreasing function of $\mathbf{Y} \equiv \{Y_b : b \in \mathbb{L}^2\}$. Apply Proposition 23.3. ■

For an application of the central limit theorem in this context we will establish the asymptotic normality of the cumulative size $\sum_{x\in B_0^{(N)}} |C(x)|$ of all clusters connected to points in the cube $B_0^{(N)}$, suitably centered and scaled for $0 < p < 1/3$. Additional applications[17] along these lines are given in the exercises. The conditions for the theorem will be checked in a sequence of simple lemmas, the first of which is a special case of an inequality known as the *BK Inequality* after its originators van den Berg and Kesten (1985).

[13] den Hollander and Keane (1986).

[14] Physicists often refer to the absence of long-range connectivities as "disorder."

[15] In his celebrated paper, Kesten (1980), it was proved that $p_c = 1/2$.

[16] A stronger version of this type result was first formulated and proven by Harris (1960) as a special case.

[17] The example given here serves the pedagogical purpose of simply illustrating the theorem. For more substantial applications, but requiring elements of percolation theory which are outside the scope of this exposition, consult the comprehensive text by Grimmett (1999), and numerous references therein.

To prepare for the BK inequality let us refer to a random variable X defined on Ω as *increasing* if $X(\omega_1) \le X(\omega_2)$ whenever $\omega_1, \omega_2 \in \Omega$ satisfy $\omega_1(b) \le \omega_2(b)$ for all $b \in \mathbb{L}^2$; the latter set of coordinatewise inequalities defines a partial order on Ω which we denote as $\omega_1 \preceq \omega_2$. Similarly we say that an event $A \in \mathcal{F}$ is an *increasing event* if $\mathbf{1}_A$ is an increasing random variable. Connectivity events of the form $[x \leftrightarrow y]$ are prototypical increasing events. From here out we restrict to this case.

Definition 23.6. The *disjoint occurrence* of two increasing events $A = [x \leftrightarrow y]$, $B = [z \leftrightarrow w]$ is an event denoted by $A \circ B$ and defined by

$$[x \leftrightarrow z] \circ [y \leftrightarrow w] = [x \leftrightarrow z, x \notin C(y), y \leftrightarrow w].$$

Lemma 6 *(BK Inequality-Special Case).* For $x, y, w, z \in \mathbb{Z}^2$

$$P_p([x \leftrightarrow z] \circ [y \leftrightarrow w]) \le P_p(x \leftrightarrow z) P_p(y \leftrightarrow w).$$

Proof. Observe that

$$\begin{aligned} P_p([x \leftrightarrow z] \circ [y \leftrightarrow w]) &= \mathbb{E}(\mathbf{1}_{[x \leftrightarrow z]} \mathbf{1}_{[x \notin C(y)]} \mathbf{1}_{[y \leftrightarrow w]}) \\ &= \mathbb{E}(\mathbf{1}_{[x \leftrightarrow z]} \mathbf{1}_{[x \notin C(y)]} \mathbf{1}_{[x \notin C(w)]} \mathbf{1}_{[y \leftrightarrow w]}) \\ &= P_p(x \leftrightarrow z, x \notin C(y), x \notin C(w), y \leftrightarrow w) \\ &= P_p(y \leftrightarrow w | x \leftrightarrow z, x \notin C(y), x \notin C(w)) P_p(x \leftrightarrow z, x \notin C(y), x \notin C(w)) \\ &\le P_p(y \leftrightarrow w | x \leftrightarrow z, x \notin C(y), x \notin C(w)) P_p(x \leftrightarrow z). \end{aligned} \tag{23.9}$$

So it suffices to show that

$$P_p(y \leftrightarrow w | x \leftrightarrow z, x \notin C(y), x \notin C(w)) \le P_p(y \leftrightarrow w). \tag{23.10}$$

Let A be an arbitrary but fixed finite connected subgraph of $\mathbb{L}^2$ with vertices x and z connected in A, but not connected to y nor w, i.e., having the properties of the conditioning. The graph A is referred to as a *lattice animal*. Denote the vertex and edge sets of A by A_v and A_e, respectively. Also define the *edge boundary* $\partial_e A$ as the set of (closed) edges which do not belong to A_e but have at least one endvertex in A_v. First consider the case in which y is "interior" to the lattice animal A and w is "exterior" to A in the sense that any path of bonds connecting y to w must include a bond from $\partial_e A$. Then, since on $[C(x) = A]$ the edges in $\partial_e A$ are all closed, one has for this case that

$$P_p(y \leftrightarrow w, C(x) = A | x \leftrightarrow z, x \notin C(y), x \notin C(w)) = 0.$$

On the other hand, for the case when $\partial_e A$ does not obstruct the existence of a path of open bonds connecting y to w, let us see that one may use association (FKG inequalities) to establish that

$$P_p(y \leftrightarrow w, C(x) = A | x \leftrightarrow z, x \notin C(y), x \notin C(w)) \leq P_p(y \leftrightarrow w). \qquad (23.11)$$

To prove (23.11) consider a lattice animal A such that $x \leftrightarrow z$ and $x \notin C(y), x \notin C(w)$ and for which y, w are not separated by $\partial_e A$ in the previous sense. Then $\mathbf{1}_{[y \leftrightarrow w]}$ and $\mathbf{1}_{[C(x)=A]}$ are, respectively, increasing and decreasing functions of independent random variables. Thus, by association (see Exercise 5),

$$\begin{aligned} P_p(y \leftrightarrow w, C(x) = A, x \leftrightarrow z, x \notin C(y), x \notin C(w)) &= P_p(y \leftrightarrow w, C(x) = A) \\ &\leq P_p(y \leftrightarrow w) P_p(C(x) = A) \\ &= P_p(y \leftrightarrow w) P_p(C(x) = A, x \leftrightarrow z, x \notin C(y), x \notin C(w)). \end{aligned}$$

Divide by the common (positive) probability $P_p(C(x) = A, x \leftrightarrow z, x \notin C(y), x \notin C(w))$ to obtain the bound (23.11). Then summing over such lattice animals A completes the proof of (23.10) and thus the BK inequality follows. ■

Lemma 7.

$$\mathrm{Cov}(\mathbf{1}_{[x \leftrightarrow z]}, \mathbf{1}_{[y \leftrightarrow w]}) \leq \mathbb{E}(\mathbf{1}_{[x \leftrightarrow z]} \mathbf{1}_{[x \leftrightarrow y]} \mathbf{1}_{[x \leftrightarrow w]}).$$

Proof. Let $\tau(x, z, y, w) := \mathbb{E}(\mathbf{1}_{[x \leftrightarrow z]} \mathbf{1}_{[x \leftrightarrow y]} \mathbf{1}_{[x \leftrightarrow w]})$. Note that

$$\begin{aligned} \mathbb{E}(\mathbf{1}_{[x \leftrightarrow z]} \mathbf{1}_{[y \leftrightarrow w]}) &= \tau(x, z, y, w) + \mathbb{E}(\mathbf{1}_{[x \leftrightarrow z]} \mathbf{1}_{[x \notin C(y)]} \mathbf{1}_{[y \leftrightarrow w]}) \\ &= \tau(x, z, y, w) + P_p([x \leftrightarrow z] \circ [y \leftrightarrow w]). \qquad (23.12) \end{aligned}$$

Now apply the BK inequality to the second term. Subtracting $\mathbb{E}\mathbf{1}_{[x \leftrightarrow z]} \mathbb{E}\mathbf{1}_{[y \leftrightarrow w]}$ from both sides establishes the assertion of the lemma. ■

Lemma 8. Let $U_x = |C(x)| = \sum_z \mathbf{1}_{[x \leftrightarrow z]}$. Then

$$\gamma = \sum_{x \in \mathbb{Z}^2} \mathrm{Cov}(U_0, U_x) \leq \mathbb{E}|C(0)|^3.$$

Proof. Using bi-linearity of covariance and the bound from the first lemma,

$$\mathrm{Cov}(\mathbf{1}_{[0 \leftrightarrow w]}, \mathbf{1}_{[y \leftrightarrow z]}) \leq \tau(0, w, y, z) = \mathbb{E}(\mathbf{1}_{[0 \leftrightarrow w]} \mathbf{1}_{[w \leftrightarrow y]} \mathbf{1}_{[y \leftrightarrow z]}),$$

it follows that

$$\sum_{y \in \mathbb{Z}^2} \mathrm{Cov}(U_0, U_y) \leq \sum_{y, w, z} \tau(0, w, y, z) = \sum_{y, w, z} \mathbb{E}(\mathbf{1}_{[0 \leftrightarrow w]} \mathbf{1}_{[0 \leftrightarrow y]} \mathbf{1}_{[0 \leftrightarrow z]})$$

$$= \mathbb{E}\left(\sum_{y,w,z} \mathbf{1}_{[0\leftrightarrow w]}\mathbf{1}_{[0\leftrightarrow y]}\mathbf{1}_{[0\leftrightarrow z]}\right) = \mathbb{E}|C(0)|^3.$$

■

Lemma 9. $\mathbb{E}|C(0)|^m < \infty$ for all $m \geq 0$.

Proof. Note from the proof of Proposition 23.7 that for $p < 1/3$, $\tau(0,x) = P_p(0 \leftrightarrow x) \leq \frac{4}{3}e^{-c|x|}$, where $c = -\ln(3p) > 0$. Thus, denoting by R_k the complimentary region to the (two-dimensional) square of side-lengths $2k+1$ centered at 0, one has the tail probability bound

$$P_p(|C(0)| \geq (2k+1)^2) \leq \sum_{x\in R_k} \tau(0,x) \leq \frac{4}{3}\sum_{x\in R_k} e^{-c|x|}$$
$$\leq c' \sum_{j=k+1}^{\infty} je^{-cj} \leq c''ke^{-c''k},$$

for a suitable $c'' > 0$. The second to the last inequality is a consequence of summing over x on the perimeters at respective distances j from the origin, noting that the number of sites on the perimeter is linear in j. It now follows that $\sum_{k=1}^{\infty} k^{2m-1}P(\sqrt{|C(0)|} \geq k) < \infty$, and therefore $\mathbb{E}|C(0)|^m = \mathbb{E}(\sqrt{|C(0)|})^{2m} < \infty$. ■

In view of Newman's central limit theorem this series of lemmas establishes the following fluctuation law.

Theorem 23.8. Consider two-dimensional bond percolation with $0 < p < 1/3$. Then the centered and rescaled cumulative size $\frac{1}{N}\sum_{x\in B_0^{(N)}}\{|C(x)| - \mathbb{E}|C(0)|\}$ of all clusters connected to points in the cube $B_0^{(N)}$ is asymptotically normal with mean zero and variance $0 < \gamma = \sum_{x\in\mathbb{Z}^2} \mathrm{Cov}(|C(0)|, |C(x)|) \leq \mathbb{E}|C(0)|^3 < \infty$.

We close this chapter with a celebrated result of Loren Pitt on association of positively correlated normal random variables. We provide the essence of his[18] very clever proof leaving the technical details to exercises.

Theorem 23.9 (Pitt). Let $X = (X_1, \ldots, X_k)$ be a positively correlated normal random vector. Then $\{X_1, \ldots, X_k\}$ is an associated family.

Proof. First consider the case in which the covariance matrix $\Gamma = ((\gamma_{i,j}))$ is non-singular matrix with nonnegative entries. One may show that the collection of coordinatewise non-decreasing functions f, g on $\mathbb{R}^k$ that are continuously differentiable with bounded partials is association determining (Exercise 12). As

[18] Pitt (1982).

a result, one may restrict to such functions f, g. Let $Z = (Z_1, \ldots, Z_k)$ be an independent copy of X and define

$$Y(\lambda) = \lambda X + (1 - \lambda^2)^{\frac{1}{2}} Z.$$

Then, $Y(\lambda)$ is mean-zero normal with covariance matrix

$$\mathrm{Cov}(Y_i(\lambda), Y_j(\lambda)) = \lambda^2 \gamma_{i,j} + (1 - \lambda^2)\gamma_{i,j} = \gamma_{i,j}.$$

Also, $\mathrm{Cov}(X, Y(\lambda))_{i,j} = \lambda \gamma_{i,j}$. Consider

$$F(\lambda) = \mathbb{E} f(X) g(Y(\lambda)).$$

Then, $\mathrm{Cov}(f(X), g(X)) = F(1) - F(0)$. So it suffices to show $F'(\lambda)$ exists and is positive for $0 \le \lambda < 1$. This is where the analysis is required. Namely, writing $\Gamma^{-1} = ((c_{i,j}))$, let

$$\varphi(x) = (2\pi)^{-\frac{k}{2}} (\det \Gamma)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \sum_{i,j=1}^{k} c_{i,j} x_i x_j \right\}$$

denote the Gaussian pdf[19] of X. Then the conditional pdf of $Y(\lambda)$ given $[X = x]$ is

$$p(\lambda; x, y) = (1 - \lambda^2)^{-\frac{k}{2}} \varphi((1 - \lambda^2)^{-\frac{1}{2}} (y - \lambda x)).$$

That is, $p(\lambda; x, y)$ is normal with covariance matrix $(1 - \lambda^2)\Gamma$ and mean vector λx. Thus,

$$F(\lambda) = \int_{\mathbb{R}^k} f(x) \varphi(x) g(\lambda, x) dx,$$

where

$$g(\lambda, x) = \int_{\mathbb{R}^k} g(y) p(\lambda; x, y) dy.$$

Observing that

$$g(\lambda, x) = \varphi(\lambda, \cdot) * g(\lambda x),$$

where $\varphi(\lambda, x) = (1 - \lambda^2)^{-\frac{k}{2}} \varphi((1 - \lambda^2)^{-\frac{1}{2}} x)$, one sees that $\partial g(\lambda, x)/\partial \lambda$ exists and is nonnegative, while $\partial g(\lambda, x)/\partial x_j$ exists and is bounded. To compute $\frac{\partial p}{\partial \lambda}$, let $h(t, y)$

[19] BCPT, p. 130.

denote the pdf of $\Gamma^{\frac{1}{2}}B_t$, where B is k-dimensional standard Brownian motion. Then one has

$$p(\lambda; x, y) = h(1-\lambda^2, y-\lambda x).$$

Using the chain rule and an application of the heat equation for multivariate Brownian motion[20] one arrives at

$$\frac{\partial p}{d\lambda} = -\frac{1}{\lambda}\left\{\sum_{i\neq j}\gamma_{i,j}\frac{\partial p}{\partial x_i, \partial x_j} - \sum_i x_i\frac{\partial p}{\partial x_i}\right\}.$$

Thus,

$$F'(\lambda) = -\frac{1}{\lambda}\int_{\mathbb{R}^k} f(x)\varphi(x)\left\{\sum_{i\neq j}\gamma_{i,j}\frac{\partial g(\lambda,x)}{\partial x_i, \partial x_j} - \sum_i x_i\frac{\partial g(\lambda,x)}{\partial x_i}\right\}dx.$$

Finally, with an integration by parts one arrives at

$$F'(\lambda) = \frac{1}{\lambda}\int_{\mathbb{R}^k}\varphi(x)\left\{\sum_{i\neq j}\gamma_{i,j}\frac{\partial f(x)}{\partial x_i}\frac{\partial g(\lambda,x)}{\partial x_j}\right\}dx \geq 0.$$

In the case Γ is singular one may replace Γ by the non-singular matrix $\Gamma + \epsilon\mathbf{1}_{k\times k}$, $\epsilon > 0$, and observe that for continuous f, g, $\mathrm{Cov}(f(X), g(X))$ depends continuously on Γ. Thus positivity is preserved in the limit as $\epsilon \to 0$. ■

Exercises

1. Let $X_1, \ldots, X_n$ be associated random variables, and $Y_j = f_j(X_1, \ldots, X_n)$, $j = 1, \ldots, m$ where f_j is coordinatewise non-decreasing for $j = 1, \ldots, m$. Show that (a) $P(Y_1 \leq y_1, \ldots, Y_k \leq y_k) \geq \prod_{j=1}^k P(Y_j \leq y_j)$, and (b) $P(Y_1 > y_1, \ldots, Y_k > y_k) \geq \prod_{j=1}^k P(Y_j > y_j)$. [*Hint*: Define non-decreasing functions of Z_j's by $Z_j = \mathbf{1}_{[Y_j > y_j]}$, $j = 1, \ldots, m$, and note that $Z_1 \cdots Z_i$, and $Z_{i+1} \cdots Z_m$ are non-decreasing functions of Z_j's. Apply the FKG inequalities iteratively for $i = 1, \ldots, m$, noting $\mathbb{E}Z_j = P(Z_j = 1)$. (c) Suppose that $X_1, \ldots X_n$ are independent random variables and let $S_j = \sum_{i=1}^j X_i$, $j = 1, \ldots, n$. Show that $P(S_1 \leq s_1, \ldots, S_n \leq s_n) \geq \prod_{i=1}^n P(S_i \leq s_i)$ for

[20] For the heat equation connection see Bhattacharya and Waymire (2021), Chapter 6, Remark 6.2. Pitt uses a related computational device of Plackett (1954).

all $s_1, \ldots, s_n$. [*Hint*: S_j is a non-decreasing function of $X_1, \ldots, X_n$, and independent random variables are associated.]

2. Suppose f is a continuous function on $[0, 1]$ and consider the *Bernstein polynomial* defined by $f_n(x) = \sum_{j=0}^{n} \binom{n}{j} x^j (1-x)^{n-j} = \mathbb{E} f(\frac{S_n(x)}{n}), 0 \le x \le 1$, where $S_j(x) = X_1(x) + \cdots X_j(x)$, for i.i.d. Bernoulli $0-1$-valued random variables with $P(X_j(x) = 1) = x, j = 1, \ldots, n$. (a) Show that $f_n \to f$ uniformly on $[0, 1]$ as $n \to \infty$. (b)*(Seymour-Welsh)* Show[21] that for non-decreasing f, g on $[0, 1]$, $(fg)_n(x) \ge f_n(x) g_n(x), 0 \le x \le 1$.
3. Show that binary $0-1$-valued random variables X, Y are associated if and only if they are positively quadrant dependent.
4. Complete the details for the extension of Hoeffding's lemma used in the generalization Lemma 2.
5. Suppose that $X = (X_1, \ldots, X_m)$ is a vector of associated random variables. Let f, g be, respectively, coordinatewise increasing and decreasing functions. Show that $\mathrm{Cov}(f(X), g(X)) \le 0$. Extend this to countably many associated random variables.
6. Prove the alternative formula
$$H_{X,Y}(x, y) = P(X \le x, Y \le y) - P(X \le x)P(Y \le y).$$
7. Suppose $f(X), g(Y) \in L^2(\Omega, \mathcal{F}, P)$, where X, Y are real-valued random variables bounded below by a constant $b \in \mathbb{R}$, and f, g are continuously differentiable complex functions with bounded derivatives. Prove the Hoeffding-Newman formula in this case, starting from the familiar moment formulae
$$\mathbb{E}(f(X) - f(b)) = \mathbb{E} \int_b^X f'(x) dx = \int_b^\infty P(X > x) f'(x) dx,$$
and
$$\mathbb{E}(f(X) - f(b))\overline{(g(Y) - g(b))} = \mathbb{E} \int_b^X \int_b^Y f'(x) \overline{g}'(y) dx dy.$$
8. Show that each of the collections of non-decreasing binary 0 or 1-valued functions and those of non-decreasing bounded continuous functions are association determining.
9. Consider the spatial intermittency of clusters as reflected in the density of isolated points and/or non-isolated points. A site $x \in \mathbb{Z}^2$ is *isolated* whenever $[C(x) = \emptyset]$. The numbers of isolated points and non-isolated points in a square $B_0^{(N)}$ are perfectly correlated since their total is (fixed) $|B_0^{(N)}|$. It is convenient to consider the number of non-isolated sites in the square $B_0^{(N)}$ (including surface sites for simplicity) as given by

[21] Seymour and Welsh (1975).

$$D_N = \sum_{x \in B_0^{(N)}} \mathbf{1}_{[C(x) \neq \emptyset]}.$$

Show that $\mathbb{E}D_N = (1-q^4)N^2$ and compute the asymptotic distribution of D_N, suitably centered and scaled, as $N \to \infty$.

10. Fix a positive integer k and obtain the asymptotic fluctuation law for the numbers of sites x in $B_0^{(N)}$ belonging to a cluster of size at most k, i.e., such that $|C(x)| \leq k$.
11. Let $N = \{N(A) : A \in \mathcal{B}\}$ be a Poisson point process on $\mathbb{R}^n$. Show that N is an associated family of random variables.
12. Show that the collection of coordinatewise non-decreasing functions f, g on $\mathbb{R}^k$ that are continuously differentiable with bounded partial derivatives is association determining. [*Hint*: (a) Check that any measurable increasing functions is a pointwise a.s. limit of continuous increasing functions. (b) Check that if $\rho_\epsilon, \epsilon > 0$ is a nonnegative C^∞-mollifier,[22] then $f * \rho_\epsilon$ is C^∞ increasing with bounded partials such that $f * \rho_\epsilon \to f$ pointwise, uniform boundedly.]

[22] See BCPT, p.77, for a mollifier example.

Chapter 24
Special Topic: More on Coupling Methods and Applications

Coupling methods originated as a probabilistic tool for the analysis of a given process by (possibly dependently) linking its sample path behavior to that of a "target process" whose long term properties may be better understood or known. This chapter illustrates the reach of coupling, extending well beyond Doeblin's original ideas, with a sample of applications to random fields.

The notion of *coupling* was introduced by Doeblin (1938) as a method to prove convergence to a unique invariant probability for irreducible aperiodic finite state Markov chains. Coupling was illustrated in Bhattacharya and Waymire (2021), Chapter 8, with three significant applications: (i) Error in Poisson approximation to the binomial distribution (ii) Convergence to steady state for a class of discrete parameter Markov chains on a countable state space, (iii) A renewal theorem for lattice distributions, and a fourth in Chapter 20 of the present text on (iv) Convergence of Markov chains on general state spaces. In the present text coupling also occurred in the proof of the geometric convergence rate[1] to equilibrium given in Theorem 20.6.

Definition 24.1. Let X, Y be two random maps defined on a probability space $(\Omega, \mathcal{F}, P)$ with values in the measurable space $(S, \mathcal{S})$. A coupling of X and Y,

[1] Although not covered in the present book, it is worthy of mention that the power of Doeblin's coupling was fully realized in the constructions of optimal couplings for continuous parameter Markov chains, as well as for a class of diffusions on manifolds, by Chen and Wang (1995), Chen (1996) and his student. These couplings involve sharp estimates of the L^2-spectral gap of the infinitesimal generator and are remarkable for the contribution to mathematics outside of probability theory. See Chen (1997) for an insightful overview.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_24

or a coupling of their respective distributions P_X and P_Y, is any bivariate random map $(\widetilde{X}, \widetilde{Y})$ with values in $(S \times S, \mathcal{S} \otimes \mathcal{S})$ whose marginals coincide with P_X and P_Y, respectively.

Example 1 *(Choquet–Deny Theorem for Simple Symmetric Random Walk on $\mathbb{Z}^k (k \geq 1)$).* As already noted in Chapter 12, the Choquet–Deny theorem asserts for any $k \geq 1$ that the only bounded, harmonic functions for the simple symmetric random walk are the constant functions. A simple application of this theorem was cited in Exercise 12 of Chapter 12.

Theorem 24.1 *(Choquet–Deny[2] for Simple Symmetric Random Walk on $\mathbb{Z}^k (k \geq 1)$).* The only bounded harmonic functions for the simple symmetric random walk on $\mathbb{Z}^k$, $(k \geq 1)$ are the constant functions.

Proof. Let $x, y \in \mathbb{Z}^k, |x - y| = 1$, and let h be a bounded harmonic function with respect to a simple, symmetric random walk $\{X_n : n \geq 0\}$ on $\mathbb{Z}^k$; that is, $h(x) = \mathbb{E}_x h(X_1), x \in \mathbb{Z}^k$. We wish to show that $h(x) = h(y)$. For this we construct a Markov coupling $(\widetilde{X}, \widetilde{Y}) = \{(\widetilde{X}_n, \widetilde{Y}_n) : n \geq 0\}$ starting at $(x, y) \in \mathbb{Z}^k \times \mathbb{Z}^k$ such that the marginal processes are simple symmetric random walks starting at x and y, respectively, and such that the coupling time $T = \inf\{n \geq 1 : \widetilde{X}_n = \widetilde{Y}_n\}$ is almost surely finite, i.e., the coupling is *successful*. For then one has

$$\begin{aligned} |h(x) - h(y)| &= |\mathbb{E}_x h(X_n) - \mathbb{E}_y h(X_n)| \\ &= |\mathbb{E}_{(x,y)} h(\widetilde{X}_n) - \mathbb{E}_{(x,y)} h(\widetilde{Y}_n)| \\ &\leq 2 \sup_x |h(x)| P(T > n) \to 0 \text{ as } n \to \infty. \end{aligned} \tag{24.1}$$

To make a coupling that is successful regardless of dimension one proceeds as follows: Let $\epsilon_1, \epsilon_2, \ldots$ and $\epsilon'_1, \epsilon'_2, \ldots$ be independent, i.i.d. symmetric Bernoulli ± 1-valued random variables. At each time step $n \geq 1$, select a common coordinate, say the mth, of $(\widetilde{X}^{(1)}_{n-1}, \ldots, \widetilde{X}^{(k)}_{n-1})$ and $(\widetilde{Y}^{(1)}_{n-1}, \ldots, \widetilde{Y}^{(k)}_{n-1})$ to make displacements according to the following rules: If $\widetilde{X}^{(m)}_{n-1} = \widetilde{Y}^{(m)}_{n-1}$, then take $\widetilde{X}^{(m)}_n = \widetilde{X}^{(m)}_{n-1} + \epsilon_n = \widetilde{Y}^{(m)}_{n-1} + \epsilon_n = \widetilde{Y}^{(m)}_n$. On the other hand, if $\widetilde{X}^{(m)}_{n-1} \neq \widetilde{Y}^{(m)}_{n-1}$, then take $\widetilde{X}^{(m)}_n = \widetilde{X}^{(m)}_{n-1} + \epsilon_n$ and $\widetilde{Y}^{(m)}_n = \widetilde{Y}^{(m)}_{n-1} + \epsilon'_n$. All other coordinates are left fixed at this time step. Then the Markov chain $(\widetilde{X}, \widetilde{Y})$ is easily checked to be a coupling. That this coupling is successful follows from the pointwise recurrence of 0 in the one-dimensional (lazy) simple symmetric random walk $\sum_{j=1}^n (\epsilon_j - \epsilon'_j), n \geq 1$. This shows, together with (24.1), that h does not depend on its mth coordinate. Since, the argument holds for each $m = 1, \ldots, k$, the function h is constant. ■

[2] The Choquet–Deny theorem is valid more generally for irreducible general random walks on locally compact Abelian groups; Choquet and Deny (1960).

The property that all bounded, harmonic functions are constant is sometimes referred to as the *Liouville property* in reference to its counterpart in analysis.

Remark 24.1. One may note that bounded harmonic functions for a Markov chain are constant if it is possible to construct a successful coupling with arbitrary initial states x, y. See Exercise 5 for a generalization.

Example 2 *(Association of a Singleton).* Recall that random variables X, Y are associated if $\mathrm{Cov}(f(X), g(Y)) \geq 0$ for bounded increasing functions f, g. A proof that a singleton X is associated, i.e., Proposition 23.3(e) can be made by coupling X to an independent copy Y. Then for bounded, increasing functions f, g one has, on the one hand, that

$$\mathbb{E}(f(X) - f(Y))(g(X) - g(Y)) = 2\mathbb{E}f(X)g(X) - 2\mathbb{E}f(X)\mathbb{E}g(X), \tag{24.2}$$

and, on the other hand, by monotonicity the factors $f(X) - f(Y)$ and $g(X) - g(Y)$ are of the same sign in the following decomposition. That is,

$$\begin{aligned}
&\mathbb{E}(f(X) - f(Y))(g(X) - g(Y)) \\
&\quad = \mathbb{E}\mathbf{1}_{[X \geq Y]}(f(X) - f(Y))(g(X) - g(Y)) + \mathbb{E}\mathbf{1}_{[X < Y]}(f(X) - f(Y))(g(X) - g(Y)) \\
&\quad \geq 0.
\end{aligned} \tag{24.3}$$

Thus, $0 \leq \mathbb{E}f(X)g(X) - \mathbb{E}f(X)\mathbb{E}g(X) = \mathrm{Cov}(f(X), g(X))$.

Another important context for coupling is that of *stochastic ordering*. Some related ideas pertaining to stochastic order were already considered in Chapter 23, and will be revisited here in the illustrative examples.

Definition 24.2. Suppose that X_1 and X_2 are real-valued random variables with respective distributions $\mu_i, i = 1, 2$. Then X_1 is said to be *stochastically dominated* by, or *stochastically smaller* than, X_2, denoted $X_1 \leq^s X_2$, if $\mu_1[x, \infty) \leq \mu_2[x, \infty)$ for all $x \in \mathbb{R}$. The probability measure μ_1 is said to be stochastically smaller than, or stochastically dominated by, μ_2, also denoted $\mu_1 \leq^s \mu_2$.

For real-valued random variables the following proposition is rather elementary.

Proposition 24.2 *(Coupling & Stochastic Order on* $\mathbb{R}$*).* $X_1 \leq^s X_2$ if and only if there is a coupling $(\widetilde{X}_1, \widetilde{X}_2)$ of X_1 and X_2 such that $P(\widetilde{X}_1 \leq \widetilde{X}_2) = 1$.

Proof. Assume that the coupling exists. Then

$$\mu_1[x, \infty) = P(\widetilde{X}_1 \geq x) \leq P(\widetilde{X}_2 \geq x) = \mu_2[x, \infty).$$

Suppose next that $X_1 \leq^s X_2$. Define $\widetilde{X}_i = F_i^{-1}(U), i = 1, 2$, where U is uniform on $[0, 1]$ and $F_i^{-1}(u) := \inf\{x \in \mathbb{R} : F_i(x) > u\}$. Then $\widetilde{X}_i$ has distribution μ_i, i.e., a coupling is achieved. Moreover, $F_1(x) \geq F_2(x)$ for all x is the *complimentary* *equivalent* to stochastic domination and, by the definition of the (generalized) inverse function, $P(\widetilde{X}_1 \le \widetilde{X}_2) = P(F_1^{-1}(U) \le F_2^{-1}(U)) = 1$. ■

Our goal now is to extend this to random maps (or their distributions) with values in partially ordered spaces. The cornerstone theory for this is a theorem of Strassen (1965) for the case of partially ordered Polish spaces. Strassen's proof was simplified by Lindvall (1999) using rather standard methods of functional analysis presented here and in Appendix D.

Remark 24.2. An alternative proof for compact spaces is given in Liggett (1985), that relies heavily on Nachbin (1966) interpolation theorems between semicontinuous functions on a compact partially ordered space having a closed partial order, in addition to the more standard theorems from functional analysis. In the case of finite partially ordered sets, an especially elegant (bipartite) graph theoretic presentation is given in Koperberg (2016) that relates Strassen's theorem to the Max Flow-Min Cut Theorem of Ford and Fulkerson (1956), as well as to Marriage Theorem of Philip Hall (1935).

To introduce the framework for Strassen's theorem for Polish partially ordered spaces let us first recall some basic definitions.

Definition 24.3. Let S be a set. A *partial order* $\preceq$ on S is a relation such that for every $x, y, z \in S$,

i $x \preceq x$,
ii $x \preceq y \preceq x$ iff $x = y$,
iii $x \preceq y \preceq z$ implies $x \preceq z$.

The pair $(S, \preceq)$ is referred to as a *partially ordered set (poset)*. In the case (iii) is removed, $\preceq$ is referred to as a *pre-order*. If S is a topological space and the set $\{(x, y) : x \preceq y\}$ is a closed set, then $\preceq$ is said to be closed.

Definition 24.4. Let X_1, X_2 be random variables taking values in a partially ordered space $(S, \preceq)$. Assume that $M = \{(x, y) : x \preceq y\}$ is measurable for the product space. A coupling $(\widetilde{X}_1, \widetilde{X}_2)$ of (X_1, X_2) such that $P(\widetilde{X}_1 \preceq \widetilde{X}_2) = 1$ is referred to as a *monotone coupling*. Equivalently, if X_1,X_2 have respective marginals μ_1,μ_2, then the joint distribution $\widetilde{\mu}$ of $(\widetilde{X}_1, \widetilde{X}_2)$ has marginals μ_1,μ_2 and $\widetilde{\mu}(M) = 1$. $\widetilde{\mu}$ is referred to as a *monotone coupling* of μ_1, μ_2.

To provide some orientation, the following example provides an illustration of the connection between monotone couplings on a partially ordered space and stochastic ordering for which a monotone coupling can be explicitly constructed (without appeal to Strassen's theorem for its existence).

Example 3 *(Monotonicity in Bond Percolation).* Recall the *bond percolation model* defined in Example 2 of Chapter 23.

Definition 24.5 *(Uniform Bernoulli Coupling).* Let X_1 and X_2 be Bernoulli $0-1$-valued random variables with $P(X_i = 1) = p_i, i = 1, 2$ with $p_1 < p_2$. Let U

be a uniformly distributed random variable on [0, 1]. Define $\widetilde{X}_i = \mathbf{1}_{[U \le p_i]}, i = 1, 2$. Then, the coupling $(\widetilde{X}_1, \widetilde{X}_2)$ of (X_1, X_2) is referred to as *uniform Bernoulli coupling*.

The uniform Bernoulli coupling is easily checked to be a monotone coupling, and underlies the proof of monotonicities of the following type. Denote the set of nearest neighbor bonds (edges) in the integer lattice $\mathbb{Z}^2$ by $\mathcal{E}$. The *configuration space* $S = \{0, 1\}^{\mathcal{E}}$ is a compact metric space for the product topology. Coordinatewise inequality provides a natural partial order $\preceq$ for which $x \preceq y$ iff $x_b \le y_b$ for all $b \in \mathcal{E}$. In particular, $x_b = 1$ implies $y_b = 1$.

Remark 24.3. Let Λ be a countable set. One may view the compact space $S = \{0, 1\}^{\Lambda}$ as the power set of Λ, i.e., the collection of all subsets of Λ coded by elements of Λ as being either "in(1)" or "out(0)." In this dual view, the partial order on $\{0, 1\}^{\Lambda}$, $x \preceq y$, defined coordinatewise, is equivalent to the partial order $x \subset y$ on the power set of Λ since for $j \in \Lambda$, $x_j = 1$, i.e., $j \in x$, implies $y_j = 1$, i.e., $j \in y$. In particular, note that if $x \in S$ is identified with $A = \{j \in \Lambda : x_j = 1\} \subset \Lambda$, and $y \in S$ is identified accordingly with a set $B \subset \Lambda$, then $A \cup B$ corresponds to the maximum $x \vee y \in S$, i.e., $(x \vee y)_m = x_m \vee y_m$, $m \in \Lambda$, and $A \cap B$ the minimum $x \wedge y$. The coordinatewise partial order corresponds to set inclusion. The choice of representation is generally a matter of convenience to methods.

Recall from Chapter 23 that an event $A \subset \{0, 1\}^{\mathcal{E}}$ is said to be an *increasing set* if $x \in A$ and $x \preceq y$ implies that $y \in A$. This definition extends to any poset S. Note also that the set A is increasing if and only if the indicator $\mathbf{1}_A$ is an increasing function.

Proposition 24.3 *(Monotonicity of Bond Percolation Probabilities).* Let $X^{(i)} = (X_b^{(i)})_{b \in \mathcal{E}}$, be i.i.d. Bernoulli random fields on $\{0, 1\}^{\mathcal{E}}$ with parameters p_i, respectively, where $p_i = P(X_b^{(i)} = 1)$, $i = 1, 2$. (i) Suppose that A is an increasing set in $\{0, 1\}^{\mathcal{E}}$. If $p_1 \le p_2$, then $P(X^{(1)} \in A) \le P(X^{(2)} \in A)$. In particular, (ii) $\theta(p_1) \le \theta(p_2)$ for $p_1 \le p_2$, where $\theta(p)$ denotes the percolation probability at the origin defined in Example 2 of Chapter 23.

Proof. Let $\{U_b : b \in \mathcal{E}\}$ be i.i.d. uniform random variables on [0, 1), and define respective families of i.i.d. monotone uniform Bernoulli couplings $\widetilde{X}^{(i)} := \{\widetilde{X}_b^{(i)} \equiv \mathbf{1}_{[U_b \le p_i]} : b \in \mathcal{E}\}$, $i = 1, 2$. Then, $\widetilde{X}^{(1)} \preceq \widetilde{X}^{(2)}$. Thus, by monotonicity of A, $[\widetilde{X}^{(1)} \in A] \subset [\widetilde{X}^{(2)} \in A]$, and the assertion (i) follows. For the (ii) simply note that the set A of configurations $x \in \{0, 1\}^{\mathcal{E}}$ such that the set $\{b \in \mathcal{E} : x_b = 1\}$ contains an infinite path of nearest neighbor bonds connected to the origin, is an increasing set. ■

Definition 24.6. A function $f : S \to \mathbb{R}$ is said to be increasing if $f(x) \le f(y)$ whenever $x \preceq y, x, y \in S$.

Definition 24.7. Let μ_1, μ_2 be probability distributions on a partially ordered Polish space S with partial order $\preceq$, and Borel σ-field $\mathcal{B}$. Then $\mu_1 \preceq^s \mu_2$ if and only if $\int_S f d\mu_1 \leq \int_S f d\mu_2$ for all increasing bounded measurable functions f.

We will require a couple of standard theorems from functional analysis, stated below with proofs given in Appendix D.

Theorem 24.4 *(Separation Theorem).* Suppose that A and B are disjoint, nonempty, convex sets in a topological vector space V. If A is compact, B closed, and V locally convex, then there is a $\ell_0 \in V^*$, $\gamma_1, \gamma_2 \in \mathbb{R}$, such that

$$\ell_0(u) < \gamma_1 < \gamma_2 < \ell_0(v),$$

for every $u \in A$, $v \in B$.

For the following representation theorem[3] note that if V is a topological vector space, then each $v \in V$ may be viewed as a linear functional on its dual space V^* by defining $v(\ell) = \ell(v)$, $\ell \in V^*$. The weakest topology on V^* that makes each $v \in V$ continuous as a linear functional on V^* is called the *weak* topology*.

Theorem 24.5 *(Dual Representation Theorem).* Given a vector space V, let V' be a vector space of linear functionals on V that separate points of V. Then V, equipped with the weakest topology on V that makes each $\ell \in V'$ continuous, is a locally convex space whose dual is V'.

The proof of the typical statement of Strassen's monotone coupling theorem as it usually appears in probability will follow as a corollary to the following main result of Strassen (1965).

Theorem 24.6 *(Strassen).* Assume that S is a Polish space. Let $\Pi = \mathcal{P}(S \times S)$ be the set of probability measures on $S \times S$, and $\Lambda \subset \Pi$ a convex set which is closed under weak convergence. Then for probability measures μ_1, μ_2 on S there is a probability measure $\widetilde{\mu} \in \Lambda$ with marginals μ_1, μ_2 if and only if

$$\int_S f d\mu_1 + \int_S g d\mu_2 \leq \sup_{\nu \in \Lambda} \int_{S\times S} (f(x) + g(y))\nu(dx \times dy),$$

for all continuous functions f, g on S such that $0 \leq f, g \leq 1$.

Proof. Define

$$V = \{(f, g) : f, g \in C_b(S)\} = C_b(S) \times C_b(S). \tag{24.4}$$

Then

$$||(f, g)|| = ||f||_u + ||g||_u, \quad f, g \in V, \tag{24.5}$$

[3] See Appendix D for more background from functional analysis.

defines a norm on V. Let

$$H = \mathcal{P}(S) \times \mathcal{P}(S), \tag{24.6}$$

where $\mathcal{P}(S)$ is the set of probability measures on S.

Observe that $(\mu_1, \mu_2) \in H$ defines a linear functional, denoted $\ell_{(\mu_1,\mu_2)}$, or, by an abuse of notation, simply as (μ_1, μ_2),

$$(f, g) \to \ell_{(\mu_1,\mu_2)}(f, g) := \int_S f d\mu_1 + \int_S g d\mu_2, \quad (f, g) \in V. \tag{24.7}$$

Let V' denote the space spanned by such linear functionals on V. Note that $V = (V')'$ separates points of V', where, for $(f, g) \in V$, by another abuse of notation, $(f, g)(\ell) = \ell(f, g), \ell \in V'$. In view of Theorem 24.5, V is the dual space of V'. Now define

$$H_\Lambda = \{(\mu_1, \mu_2) \in H : \exists\, \widetilde{\mu} \in \Lambda \text{ having marginals } \mu_1, \mu_2\}. \tag{24.8}$$

Due to the convexity assumption on Λ, H_Λ is convex, viewed as a subset of V' in accordance with (24.7). Applying Theorems 24.4 and 24.5 (with V' as the topological vector space), one has that

(a) If B is a closed and convex subset of V' and $\ell' \in V'\backslash B$, then there exists a weak* continuous linear functional ℓ_0 on V' such that $\ell_0(\ell') > \sup_{\beta \in B} \ell_0(\beta)$. Replacing ℓ_0 by $-\ell_0$ in Theorem 24.4 gives separation in the reverse order.

(b) All weak* continuous functionals ℓ_0 on V' are of the form $\ell_0(\ell) = \ell(f_0, g_0), \ell \in V'$, for some $(f_0, g_0) \in V$.

The aim is to show that the non-existence of a coupling probability in Λ implies that Strassen's condition does not hold. So, suppose that $(\mu_1, \mu_2) \notin \overline{H_\Lambda}$. Note again that each (μ_1, μ_2) may be viewed as a linear functional $\ell_{(\mu_1,\mu_2)} \in V'$. Taking $\ell' = \ell_{(\mu_1,\mu_2)}$, $B = \overline{H_\Lambda}$ in (a), there is a continuous linear functional ℓ_0 on V' that separates $\ell_{(\mu_1,\mu_2)}$ from $\overline{H_\Lambda}$. That is,

$$\ell_0(\ell_{(\mu_1,\mu_2)}) > \sup_{\ell \in \overline{H_\Lambda}} \ell_0(\ell).$$

Now find (f_0, g_0) representing ℓ_0 according to (b), i.e., $\ell \to \ell_0(\ell) \equiv \ell(f_0, g_0)$, $\ell \in V'$. Then, one has

$$\ell_{(\mu_1,\mu_2)}(f_0, g_0) > \sup_{\ell \in \overline{H_\Lambda}} \ell(f_0, g_0). \tag{24.9}$$

In particular, writing out the meaning of (24.9) in terms of μ_1 and μ_2, this separation is in direct contradiction to Strassen's condition that $\int_S f_0 d\mu_1 + \int_S g_0 d\mu_2 \leq$

$\sup_{\nu \in \Lambda} \int_S (f(x) + g(y))\nu(dx \times dy)$ for all continuous functions $0 \leq f, g \leq 1$ on S. Thus $(\mu_1, \mu_2) \in \overline{H_\Lambda}$.

It remains to prove that H_Λ is closed in V'. To that end, we first note that the relativized weak* topology on H is metrizable. Indeed, let $\mathcal{P}(S \times S)$ denote the set of probability measures on $S \times S$. Then the mapping $\phi : H \to \mathcal{P}(S \times S)$, defined by $\phi(\mu_1, \mu_2) = \mu_1 \times \mu_2$, is a homeomorphism from H onto a closed subset of $\mathcal{P}(S \times S)$ endowed with the weak*topology. But that topology is metrizable[4] by the Prokhorov metric, This means that in order to prove that H_Λ is weak* closed, it suffices to show that the limit of a convergent sequence of elements in H_Λ remains in H_Λ. So, let $(\mu_n^{(1)}, \mu_n^{(2)}), n \geq 1$, be a sequence in H_Λ which is weak* convergent to (μ_1, μ_2). Then $\mu_n^{(1)} \Rightarrow \mu_1$, and $\mu_n^{(2)} \Rightarrow \mu_2$. This implies that any sequence $\{\nu_n\}_{n=1}^\infty$ such that ν_n has marginals $\mu_n^{(1)}, \mu_n^{(2)}$ for $n \geq 1$ is tight. Indeed, take any $\epsilon > 0$ and let $K_\epsilon^{(1)}, K_\epsilon^{(2)}$ be compact sets such that $\inf_n \mu_n^{(1)}(K_\epsilon^{(1)}) > 1 - \epsilon/2$ and $\inf_n \mu_n^{(2)}(K_\epsilon^{(2)}) > 1 - \epsilon/2$. It follows that $\inf_n \nu_n(K_\epsilon^{(1)} \times K_\epsilon^{(2)}) > 1 - \epsilon$, and $K_\epsilon^{(1)} \times K_\epsilon^{(2)}$ is compact in $S \times S$. Now apply Prokhorov's theorem to find a cluster point $\widetilde{\mu}$ of $\{\nu_n\}_{n=1}^\infty$. Since Λ is closed, we have $\widetilde{\mu} \in \Lambda$. Moreover $\widetilde{\mu}$ has respective marginals μ_1 and μ_2 by the Mann–Wald[5] continuous mapping theorem applied to the projections $(\mu_1, \mu_2) \to \mu_1$ and $(\mu_1, \mu_2) \to \mu_2$. ■

Remark 24.4. Note that if S is compact, then H is tight, so that $H_\Lambda = \overline{H_\Lambda}$ is also compact[6] in the weak* topology.

Corollary 24.7 *(Strassen's Monotone Coupling).* Let $(S, \preceq)$ be a Polish partially ordered space with Borel σ-field. Assume that $\preceq$ is closed in the sense that $M = \{(x, y) \in S \times S : x \preceq y\}$ is a closed set. Then, for probability measures μ_1, μ_2 on $(S, \mathcal{B})$, $\mu_1 \leq^s \mu_2$ if and only if there is a monotone coupling $\widetilde{\mu}$ of μ_1, μ_2.

Proof. The sufficiency of the existence of a monotone coupling for stochastic order is obvious from the definitions. Let $\Lambda = \{\nu \in \mathcal{P}(S \times S) : \nu(M) = 1\}$. Then Λ is convex and closed in $\mathcal{P}(S \times S)$ under weak convergence. Then,

$$\int_S g d\mu_2 \leq \int_S g^* d\mu_2 \leq \int_S g^* d\mu_1,$$

where $g^*(x) = \sup\{g(y) : x \preceq y\}$, since $g \leq g^*$ and g^* is decreasing. Thus, for continuous functions f, g on S with $0 \leq f, g \leq 1$,

[4] See BCPT, Theorem 7.10, p. 144.

[5] See BCPT, Theorem 7.4, p. 140.

[6] See BCPT, Proposition 7.6, p. 142.

$$\begin{aligned}\int_S f d\mu_1 + \int_S g d\mu_2 &\leq \int_S (f + g^*) d\mu_1 \\ &\leq \sup_x (f(x) + g^*(x)) \leq \sup_{(x,y)\in M} (f(x) + g(y)) \\ &\leq \sup_{\nu\in\Lambda} \int_S (f(x) + g(y))\nu(dx \times dy). \qquad (24.10)\end{aligned}$$

Thus Strassen's main theorem provides the existence of the monotone coupling $\widetilde{\mu}$. ■

Let us now consider some applications of Strassen's theorem.

Theorem 24.8 *(Strassen's Stochastic Ordering of Markov Chains).* Let $\{X_n : n \geq 0\}$ and $\{Y_n : n \geq 0\}$ be Markov chains on a Polish poset $(S, \preceq)$ with transition probabilities $p(x, dy)$ and $q(x, dy)$, respectively. Assume that $q(y, \cdot)$ stochastically dominates $p(x, \cdot)$ for each $x, y \in S$, $x \preceq y$. Then for any $x_0 \preceq y_0$ there is a coupling $\{(\widetilde{X}_n, \widetilde{Y}_n) : n \geq 0\}$ of $\{X_n : n \geq 0\}$ started at x_0 and $\{Y_n : n \geq 0\}$ started at y_0 such that a.s. $X_n \preceq Y_n$, for each n. Furthermore, if the Markov chains have stationary distributions μ_X and μ_Y, respectively, then $\mu_X \leq^s \mu_Y$.

Proof. By Strassen's theorem there is a monotone coupling $\widetilde{p}((x_0, y_0), \cdot))$ of $p(x_0, \cdot)$ and $q(y_0, \cdot)$ such that $\widetilde{p}((x_0, y_0), \widetilde{S}) = 1$, where $\widetilde{S} = \{(x, y) \in S \times S : x \preceq y\}$. The monotone coupling is the Markov chain $\{(\widetilde{X}_n, \widetilde{Y}_n) : n = 0, 1, \ldots\}$ on $\widetilde{S}$ starting at (x_0, y_0) having transition probabilities $\widetilde{p}((x, y), \cdot)$.

To see that $\mu_X \preceq \mu_Y$, use the ergodic theorem for Markov chains to write $\mu_X(\cdot) = \lim_n p^{(n)}(x_0, \cdot)$ and $\mu_Y(\cdot) = \lim_n q^{(n)}(y_0, \cdot)$. Then for an increasing bounded continuous function f on S one has

$$\mathbb{E}_{\mu_X} f(X_0) = \lim_n \frac{1}{n}\sum_{m=0}^{n-1} f(X_m) \leq \lim_n \frac{1}{n}\sum_{m=0}^{n-1} f(Y_m) = \mathbb{E}_{\mu_Y} f(Y_0). \qquad (24.11)$$

Thus, $\mu_X \leq^s \mu_Y$. ■

Corollary 24.9. Suppose Λ is a countable set and $S = \{0, 1\}^\Lambda$ with the product topology, Borel σ-field, and coordinatewise partial order $\preceq$. If μ_1, μ_2 are probability measures on S such that $\mu_1 \leq_s \mu_2$, and if

$$\mu_1(\{x \in S : x(m) = 1\}) = \mu_2(\{x \in S : x(m) = 1\}), \qquad \text{for all } m \in \Lambda, \qquad (24.12)$$

then $\mu_1 = \mu_2$.

Proof. Let $\widehat{\mu}$ be the monotone coupling of μ_1, μ_2 furnished by Strassen's theorem. Then

$$\begin{aligned}
&\widetilde{\mu}(\{(x,y)\in S\times S : x(m)=0, y(m)=1\})\\
&\quad=\widetilde{\mu}(\{(x,y): y(m)=1\})-\widetilde{\mu}(\{(x,y): x(m)=1)\\
&\qquad+\widetilde{\mu}(\{(x,y): x(m)=1, y(m)=0\})\\
&\quad=\widetilde{\mu}(\{(x,y): y(m)=1\})-\widetilde{\mu}(\{(x,y): x(m)=1, y(m)=1\})\\
&\quad=\mu_2(\{y: y(m)=1\})-\mu_1(\{x: x(m)=1\})=0. \qquad (24.13)
\end{aligned}$$

Thus, with $E=\{(x,y)\in S\times S : x=y\}$, $\widetilde{\mu}(E)=1$, and therefore, for $B\in\mathcal{B}(S)$,

$$\mu_1(B)=\widetilde{\mu}(B\times S)=\widetilde{\mu}((B\times S)\cap E)=\widetilde{\mu}(S\times B)=\mu_2(B). \qquad \blacksquare$$

***Example 4** (Holley Inequalities).* In Chapter 23 more general inequalities, known as *Holley inequalities*,[7] were noted in the context of associated random variables. In this example we see that these inequalities can be derived from a relative *log-convexity* condition using coupling techniques. As corollaries a log-convexity condition[8] for FKG and *Harris inequalities* will follow.

Again, let S denote the power set of a finite set Λ, and define a partial order $\preceq$ on S by $A \preceq B$ iff $A\subset B$. Then S is finite, with all of its subsets measurable. Equivalently, one may view $S=\{0,1\}^{\Lambda}$ as functions $\omega:\Lambda\to\{0,1\}$, with the coordinatewise partial order $\preceq$, $x\preceq y$ iff $x_m\le y_m$, for all $m\in\Lambda, x, y\in S$.

In the following we consider probability measures in terms of their densities with respect to counting measure, i.e., their probability mass functions. In a slight abuse of notation clarified by context, notions defined for probability measures are sometimes applied to their densities. For example, if μ_1,μ_2 are probability densities, then we also write $\mu_1 \le^s \mu_2$ to indicate stochastic ordering of their corresponding probability measures. Similarly, as in the next definition, the log-convexity properties defined in terms of densities are often expressed in terms of their corresponding probability measures.

***Definition 24.8** (Relative Log-Convexity).* Consider probability measures on the power set S of Λ having densities (probability mass functions) μ,μ_1,μ_2. (i) μ_2 is log-convex with respect to μ_1 iff $\mu_2(A\cup B)\mu_1(A\cap B)\ge\mu_2(A)\mu_1(B)$, for all $A, B\subset S$. (ii) μ is said to be log-convex if $\mu_2=\mu$ is log-convex with respect to $\mu_1=\mu$.

Note that for the representation of configurations in $S=\{0,1\}^{\Lambda}$ as functions $\omega:\Lambda\to\{0,1\}$ with the coordinatewise partial, the log-convexity of a density μ_2 with respect to μ_1 takes the form

$$\mu_2(\omega\vee\eta)\mu_1(\omega\wedge\eta)\ge\mu_2(\omega)\mu_1(\eta), \quad \omega,\eta\in S,$$

where $\vee$, $\wedge$ are the max, min lattice operations.

[7] Holley (1974).

[8] Fortuin et al. (1971).

Theorem 24.10 *(Holley's Inequalities).* Let μ_1, μ_2 be probability densities on the finite poset $(S, \preceq)$, above. Suppose that μ_2 is relatively log-convex with respect to μ_1. Then $\mu_1 \leq^s \mu_2$.

Proof. The idea is to construct a monotone coupling of Markov chains having invariant probabilities with densities μ_1, μ_2, in a manner similar to that used in Markov Chain Monte Carlo (MCMC) simulations such as Propp-Wilson algorithm and the Gibbs sampler considered in Chapter 19. From here the result follows from Strassen's theorem for Markov chains (Theorem 24.8).

We will employ the representation of configurations in $S = \{0, 1\}^\Lambda$ as functions $\omega : \Lambda \to \{0, 1\}$, with the coordinatewise partial order $\preceq$. The proof that $\mu_1 \leq^s \mu_2$ is achieved by constructing a monotone coupling $(\widetilde{X}, \widetilde{Y}) = \{(\widetilde{X}_n, \widetilde{Y}_n) : n \geq 0\}$ of aperiodic, irreducible Markov chains $X = \{X_n : n \geq 0\}$ and $Y = \{Y_n : n \geq 0\}$ on S having stochastically ordered transition probabilities and invariant probabilities with densities μ_1, μ_2. The one-step transitions will involve simultaneous single coordinate changes designed so that an increase in a coordinate of $\widetilde{X}$ does not occur with a decrease in the same coordinate of $\widetilde{Y}$.

For $\omega \in S, m \in \Lambda$, let ω^m denote the configuration obtained from ω by flipping the value of ω at m from ω_m to $1 - \omega_m$.

Let $N = |\Lambda|$, and let $p_1, p_2 \in (0, 1)$ to be determined. To construct a Markov chain with invariant probability density μ_1, consider transitions with positive probability of the form $\omega \to \omega^m$ with one-step transition probabilities defined accordingly by

$$p(\omega, \omega^m) = \begin{cases} \frac{1}{N} p_1 p_2 & \text{if } \omega_m = 0 \\ \frac{1}{N} p_1 p_2 \frac{\mu_1(\omega^m)}{\mu_1(\omega)} & \text{if } \omega_m = 1, \end{cases} \tag{24.14}$$

and $p(\omega, \omega) = 1 - \sum_{\eta:|\eta-\omega|_1=1} p(\omega, \eta)$, where $|\omega|_1 = \sum_{m\in\Lambda} |\omega_m|$. The transition probability $q(\omega, \eta)$ is defined the same way with μ_1 replaced by μ_2. Naturally, p_1 must be sufficiently small to make $p(\omega, \omega), q(\omega, \omega)$ both nonnegative. The parameter p_2 will be further restricted for the coupling construction. It is straightforward to check that $p(\omega, \eta)$ and $q(\omega, \eta)$ are aperiodic, with irreducible transition probabilities having time-reversible invariant probability densities $\mu_1 \leq^s \mu_2$, respectively, e.g., using $(\omega^m)^m = \omega$, $p(\omega, \omega^m)\mu_1(\omega) = p(\omega^m, \omega)\mu_1(\omega^m)$.

To show that that for $\omega \preceq \eta$ one has $p(\omega, \cdot) \leq^s q(\eta, \cdot)$ it suffices by Theorem 24.8 to construct a monotone coupling. For this it will be useful to note that if μ_2 is log-convex with respect to μ_1 and $\omega \preceq \eta$, then when $\omega_m = 1$ and $\eta_m = 1$, one has $\omega \vee \eta^m = \eta$, and $\omega \wedge \eta^m = \omega^m$, $\mu_2(\eta)\mu_1(\omega^m) \geq \mu_2(\eta^m)\mu_1(\omega)$. In particular,

$$\frac{\mu_1(\omega^m)}{\mu_1(\omega)} \geq \frac{\mu_2(\eta^m)}{\mu_2(\eta)}. \tag{24.15}$$

The coupling construction is as follows.

Coupling Algorithm[9] Let $\omega \preceq \eta$.

1. Toss a coin with probability p_1 for heads, $1 - p_1$ for tails. On the event T that tail occurs, $(\omega, \eta) \to (\omega, \eta)$.
2. Otherwise, on the event H that head occurs, independently select $M = m \in \Lambda$ (uniformly) with probability $\frac{1}{N}$, and make a transition from (ω, η) by a double or single coordinate flip according to the following conditional probabilities given $H \cap [M = m]$ depending on (ω, η):
3. If $(\omega_m, \eta_m) = (0, 0)$, then for $0 < p_2$ sufficiently small for positivity of the indicated conditional probabilities,

$$(\omega, \eta) \to \begin{cases} (\omega^m, \eta^m) & \text{with conditional probability } p_2 \\ (\omega, \eta) & \text{with conditional probability } 1 - p_2. \end{cases}$$

4. If $(\omega_m, \eta_m) = (1, 1)$, then

$$(\omega, \eta) \to \begin{cases} (\omega^m, \eta^m) & \text{with conditional probability } p_2 \frac{\mu_2(\eta^m)}{\mu_2(\eta)} \\ (\omega^m, \eta) & \text{with conditional probability } p_2(\frac{\mu_1(\omega^m)}{\mu_1(\omega)} - \frac{\mu_2(\eta^m)}{\mu_2(\eta)}) \\ (\omega, \eta) & \text{with conditional probability } 1 - p_2 \frac{\mu_1(\omega^m)}{\mu_1(\omega)}. \end{cases}$$

5. If $(\omega_m, \eta_m) = (0, 1)$, then

$$(\omega, \eta) \to \begin{cases} (\omega^m, \eta) & \text{with conditional probability } p_2 \\ (\omega, \eta^m) & \text{with conditional probability } p_2 \frac{\mu_2(\eta^m)}{\mu_2(\eta)} \\ (\omega, \eta) & \text{with conditional probability } 1 - p_2\{1 + \frac{\mu_2(\eta^m)}{\mu_2(\eta)}\}. \end{cases}$$

It is straightforward to check, for example, that unconditionally on $H \cap [M = m]$, if $\omega_m = 0, \eta_m = 0$, then $p(\omega, \omega^m) = P_{(\omega,\eta)}(\tilde{X}_1 = \omega^m) = P_{(\omega,\eta)}(\tilde{X}_1 = \omega^m, \tilde{Y}_1 = \eta^m) = \frac{1}{N} p_1 p_2 = q(\eta, \eta^m) = P_{(\omega,\eta)}(\tilde{Y}_1 = \eta^m)$. Similarly, if $\omega_m = 1, \eta_m = 1$, then $p(\omega, \omega^m) = P_{(\omega,\eta)}(\tilde{X}_1 = \omega^m, \tilde{Y}_1 = \eta^m) + P_{(\omega,\eta)}(\tilde{X}_1 = \omega^m, \tilde{Y}_1 = \eta) = \frac{1}{N} p_1 p_2 \frac{\mu_2(\eta^m)}{\mu_2(\eta)} + \frac{1}{N} p_1 \{\frac{\mu_1(\omega^m)}{\mu_1(\omega)} - \frac{\mu_2(\omega^m)}{\mu_2(\omega)}\} = \frac{1}{N} p_1 p_2 \frac{\mu_1(\omega^m)}{\mu_1(\omega)} \leq \frac{1}{N} p_1 p_2 \frac{\mu_2(\eta^m)}{\mu_2(\eta)} = q(\eta, \eta^m) = P_{(\omega,\eta)}(\tilde{Y}_1 = \eta^m)$, and so on.

It follows from Strassen's theorem that for $\omega \preceq \eta$, $p(\omega, \cdot) \leq^s q(\eta, \cdot)$. Since μ_1 and μ_2 are densities of (time-reversible) invariant initial probabilities, for p and q, respectively, it follows that $\mu_1 \leq^s \mu_2$ by Theorem 24.8. ■

[9] The original coupling construction by Holley (1974) involved a monotone coupling of continuous parameter Markov chains. Converting it to a discrete parameter coupling was inspired by Roch (2020).

Corollary 24.11. Suppose that μ is a log-convex probability density on the finite poset S. Then the FKG inequalities are valid for the probability measure with density μ, i.e., for increasing f, g on $S = \{0, 1\}^{\Lambda}$

$$\sum_{\omega\in S} f(\omega)g(\omega)\mu(\omega) \geq \sum_{\omega,\eta\in S} f(\omega)g(\eta)\mu(\omega)\mu(\eta).$$

Proof. Without loss of generality, assume $g > 0$ by adding a constant if needed. It suffices to consider nonnegative and increasing functions f, g. Define $\mu_2(\omega) = \frac{g(\omega)\mu(\omega)}{\sum_{\gamma\in S} g(\gamma)\mu(\gamma)}$ and $\mu_1 = \mu$. Then μ_2 is easily checked to be log-convex with respect to μ_1. It follows from Theorem 24.10 that the FKG inequalities follow from the Holley inequalities. ■

Remark 24.5. The conclusion $\mu_1 \leq^s \mu_2$ of Corollary 24.11 may be expressed in the more familiar form of nonnegative covariance as:

$$\sum_{\omega\in S} f(\omega)g(\omega)\mu(\omega) - \sum_{\eta\in S} f(\eta)\mu(\eta) \sum_{\gamma\in S} g(\gamma)\mu(\gamma) \geq 0,$$

for increasing functions f, g on $(S, \preceq)$.

***Example 5** (Ising Ferromagnet).* Let Λ be a finite subset of the integer lattice $\mathbb{Z}^d$. Define $\partial\Lambda = \{m \in \mathbb{Z}^d\backslash\Lambda : |m-j| = 1 \text{ for some } j \in \Lambda\}$. The *Ising ferromagnet* on Λ with *boundary values* $\omega \in \{-1, 1\}^{\partial\Lambda}$ is the probability measure on $S = \{-1, 1\}^{\Lambda}$ defined by

$$\mu_{\Lambda}^{(\omega)}(\sigma) = Z_{\Lambda}^{-1} e^{-\beta H_{\Lambda}^{(\omega)}(\sigma)}, \quad \sigma \in S, \tag{24.16}$$

where $H_{\Lambda}^{(\omega)}(\sigma) = -\sum_{i,j\in\Lambda,|i-j|_1=1} \sigma_i\sigma_j - \sum_{i\in\Lambda,j\in\partial\Lambda,|i-j|_1=1} \sigma_i\omega_j$, $\beta > 0$, and $Z_{\Lambda} = \sum_{\sigma\in S} e^{-\beta H_{\Lambda}^{(\omega)}(\sigma)}$ normalizes the exponentials $e^{-\beta H_{\Lambda}^{(\omega)}}$ to a probability. $H_{\Lambda}^{(\omega)}$ is called the *energy Hamiltonian*, Z_{Λ} is the *partition function*, $\beta > 0$ is the *inverse temperature* parameter, , and $\sigma \in S$ is a *spin configuration.*

Proposition 24.12. The spin random variables σ_j, $j \in \Lambda$ are associated, i.e., satisfy the FKG inequalities.

Proof. The log-convexity condition for $\mu_{\Lambda}^{(\omega)}$ may be expressed as follows for $\sigma, \eta \in S$,

$$\exp\left\{-\beta\big(H_{\Lambda}^{(\omega)}(\sigma\vee\eta) + H_{\Lambda}^{(\omega)}(\sigma\wedge\eta) - H_{\Lambda}^{(\omega)}(\sigma) - H_{\Lambda}^{(\omega)}(\eta)\big)\right\} \geq 1.$$

Then it suffices to show

$$\begin{aligned}&\sum_{i,j\in\Lambda,|i-j|_1=1}\{(\sigma\vee\eta)_i(\sigma\vee\eta)_j+(\sigma\wedge\eta)_i(\sigma\wedge\eta)_j\}\\&\quad+\sum_{i\in\Lambda,j\in\partial\Lambda,|i-j|_1=1}\{(\sigma\vee\eta)_i+(\sigma\wedge\eta)_i\}\omega_j\\&\geq\sum_{i,j\in\Lambda,|i-j|_1=1}(\sigma_i\sigma_j+\eta_i\eta_j)+\sum_{i\in\Lambda,j\in\partial\Lambda,|i-j|_1=1}(\sigma_i+\eta_i)\omega_j.\end{aligned} \tag{24.17}$$

For this note that

$$\{(\sigma\vee\eta)_i+(\sigma\wedge\eta)_i\}\omega_j=(\sigma_i+\eta_i)\omega_j, \tag{24.18}$$

and for $i,j\in\Lambda$, $|i-j|_1=1$, if $\sigma_i\neq\eta_i$ and $\sigma_j\neq\eta_j$,

$$(\sigma\vee\eta)_i(\sigma\vee\eta)_j+(\sigma\wedge\eta)_i(\sigma\wedge\eta)_j=(1)(1)+(-1)(-1)=2\geq\sigma_i\sigma_j+\eta_i\eta_j, \tag{24.19}$$

while the case $\sigma_i=\eta_i$ one obtains $\sigma_i(\sigma\vee\eta)_j+\sigma_i(\sigma\wedge\eta)_j=\sigma_i\{(\sigma\vee\eta)_j+(\sigma\wedge\eta)_j\}=\sigma_i\sigma_j+\sigma_i\eta_j$. Thus $H(\sigma\vee\eta)+H(\sigma\wedge\vee)\geq H(\sigma)+H(\eta)$. ■

Remark 24.6. Note that Example 5 may be abstracted to association of a two-state, say $\alpha<\beta$, Markov chain such that $p_{\alpha,\alpha}=p_{\beta,\beta}\geq 1/2$. (Exercise 8).

In the independent case, the FKG inequalities are due to Harris (1960).

Corollary 24.13 *(Harris Inequalities).* Let f, g be nondecreasing functions on $S=\{0,1\}^\Lambda$ for a finite set Λ and let $Y=\{Y_m:m\in\Lambda\}$ be independent random variables. Then, $\mathbb{E}f(Y)g(Y)\geq\mathbb{E}f(Y)\mathbb{E}g(Y)$.

Proof. The proof is by induction on $|\Lambda|$ to check log-convexity of the distribution μ of the random field Y. For then the assertion follows from the FKG inequalities. ■

Remark 24.7. Positive dependence inequalities and coupling are also important tools for analysis of continuous time Markov processes, including interacting particle systems.[10]

Remark 24.8. Although not treated here, *path coupling*[11] is an extension[12] of Doeblin's coupling methods for Markov chains which have proved to be more

[10] Liggett (1983). Also see Burton and Waymire (1986) for a related application to renewal processes.

[11] Path coupling was introduced by Bubley and Dyer (1997).

[12] Further extensions of path coupling to aggregate path coupling were developed by Kovchegov and Otto (2018).

efficient for mixing time estimates for Markov chain sampling from models in statistical mechanics and percolation.

Exercises

1. (a) Suppose $\mathbf{p}$ is an irreducible periodic transition probability matrix on a countable state space S (if period $d > 1$). Then the Markov chain $\{\mathbf{X}_n := (X_n^{(1)}, X_n^{(2)}) : n \geq 0\}$ with $\{X_n^{(i)} : n \geq 0\}$, $i = 1, 2$, independent Markov chains each with transition probability $\mathbf{p}$, then $\{\mathbf{X}_n : n \geq 0\}$ is not irreducible and has d equivalence classes.

   (b) Example:

$$\mathbf{p} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad (d = 2).$$

2. (Maximal Coupling)[13] Suppose that μ_1, μ_2 are probability measures on defined on the power set of a finite set S. Let $\mathcal{C}$ denote the set of all couplings (X_1, X_2) of μ_1, μ_2. Show that $||\mu_1 - \mu_2||_{TV} = \inf_{(X_1,X_2)\in\mathcal{C}} P(X_1 \neq X_2)$, where $||\cdot||_{TV}$ denotes the total variation norm.[14] [*Hint*: Show that the infimum is achieved by the coupling (X_1^*, X_2^*) defined as follows: Let $S_1 = \{x \in S : \mu_1(x) > \mu_2(x)\}, S_2 = S_1^c$, $p_i^* = \sum_{x\in S_i} |\mu_1 - \mu_2|$, $p^* = p_1^* + p_2^*$. With probability p^*, choose a value $X_1^* = X_2^* = x$ from the distribution $\frac{1}{p^*}\mu_1(x) \wedge \mu_2(x)$ or, with probability $1 - p^*$ choose a value X_1^* from the distribution $\frac{1}{1-p^*}(\mu_1(x) - \mu_2(x)), x \in S_1$, and independently choose a value X_2^* from the distribution $\frac{1}{1-p^*}(\mu_2(x) - \mu_1(x)), x \in S_2$. Check that $(X_1^*, X_2^*) \in \mathcal{C}$ and $P(X_1^* \neq X_2^*) = 1 - p^* = ||\mu_1 - \mu_2||_{TV}$.]
3. Show that the bounded harmonic functions for the simple symmetric random walk on $\mathbb{Z}^k$ and those for a lazy simple symmetric random walk on $\mathbb{Z}^k$ coincide. That is, for $p(x, y) = \frac{1}{2k}, |x - y| = 1, x, y \in \mathbb{Z}^k$, or for $p_\epsilon(x, x) = \epsilon \in (0, 1), x \in \mathbb{Z}^k$, and $p_\epsilon(x, y) = \frac{1-\epsilon}{2k}, |x - y| = 1, x, y \in \mathbb{Z}^k$, one has $h(x) = \sum_y p(x, y)h(y)$, for all $x \in \mathbb{Z}^k$ if and only if $h(x) = \sum_y p_\epsilon(x, y)h(y)$, for all $x \in \mathbb{Z}^k$.
4. Show that Strassen's condition for the existence of a coupling of probability measures μ_1, μ_2 is always satisfied when $\Lambda = \mathcal{P}(S \times S)$.

[13] A more general version of this result for probabilities on Polish spaces is given in the monograph Lindvall (2002). This provides a proof of the maximality of the coupling used for the Poisson approximation.

[14] BCPT, p. 136.

5. Show that if, for any pair of initial states, there is a successful coupling of the corresponding Markov chains with transition probabilities $((p_{xy}))_{x,y\in S}$ on a countable state space S, then the only bounded, harmonic functions for the Markov chain with the given transition probabilities are constant functions.
6. (Site Percolation) Take Λ to be a finite set, and let $p \in (0, 1)$ and consider the probability measure obtained by independently assigning $Y_m \in \{0, 1\}$ values to points in $m \in \Lambda$ with respective probabilities $p, 1 - p$. Let, $\mu(A) = P(Y_m = 1, m \in A, Y_m = 0, m \in \Lambda\backslash A) = p^{|A|}(1 - p)^{|\Lambda\backslash A|}, A \subset \Lambda$. Then, show $\mu(A \cup B) \geq \mu(A)\mu(B)$. [*Hint*: Observe that for $A, B \subset \Lambda$, i.e., $A, B \in S$, $f(\cdot) = \mathbf{1}_{[\cdot\supset A]}$, $g(\cdot) = \mathbf{1}_{[\cdot\supset B]}$, are increasing functions on S, and apply the Harris inequalities.]
7. (Ising ferromagnet) In reference to the Ising ferromagnet in Example 5 Let $\omega_j^{\pm} = \pm 1, \quad \text{for all } j \in \partial\Lambda$, and show for any other boundary spin values ω one has $\mu_\Lambda^{\omega^-} \leq^s \mu_\Lambda^{\omega} \leq^s \mu_\Lambda^{\omega^+}$.
8. Prove, as asserted in Remark 24.6, the two-state Markov chain is associated if and only if $p \geq 1/2$.

Chapter 25
Special Topic: An Introduction to Kalman Filter

The Kalman filter is often heralded as among the most impactful mathematical plural concepts and algorithms of applied mathematics of the twentieth century. The basic mathematical theory is presented in this chapter, together with an often cited example to illustrate the nature of the computations required for estimation and prediction with the Kalman filter.

The Kalman[1] filter provides an approach to estimation and prediction of the state of a linear dynamical system based on indirect measurements (observations), possibly lower dimensional than the state variables. Its applications and extensions are far reaching, including navigation systems, robotics, and time series models arising in signal processing and econometrics.

Consider a state variable $x \in \mathbb{R}^p$ of interest governed by an autonomous (or time-invariant) randomly forced linear dynamical system

$$X_{t+1} = FX_t + W_{t+1} \quad t = 0, 1, 2, \dots, \tag{25.1}$$

where F is a $p \times p$ matrix and $\{W_1, W_2, \dots\}$ is a sequence of $\mathbb{R}^p$-valued i.i.d. random variables with zero mean and covariance matrix Q, independent of the initial state X_0. Our interest here lies in the situation in which the state process is not directly observable, and information about the state is to be gleaned from the $\mathbb{R}^m$-valued measurements $Z_t, t = 0, 1, \dots,$ governed by the relation

$$Z_t = HX_t + V_t \quad t = 0, 1, 2, \dots, \tag{25.2}$$

[1] Kalman (1960).

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_25

with H being a $m \times p$ matrix and $\{V_t : t = 0, 1, 2, \ldots\}$ an i.i.d. mean-zero sequence of $\mathbb{R}^m$-valued random variables with covariance matrix S, independent of the sequence $\{W_t : t \geq 0\}$ and X_0. It is assumed that F and H are known, as well as the covariance matrices Q, S of W_t and V_t, respectively.

In summary, the Kalman state space and measurement equations are given:

$$\begin{aligned} X_{t+1} &= FX_t + W_{t+1} \\ Z_t &= HX_t + V_t, \quad t = 0, 1, 2, \ldots \end{aligned} \tag{25.3}$$

The *estimation problem* is to provide an estimate $\widehat{X}_t$ of X_t from (measurements) observations $\{Z_0, Z_1, \ldots, Z_t\}$ and X_0, based on minimizing the expected squared error $\mathbb{E}|X_t - \widehat{X}_t|^2$, or equivalently, minimizing the trace $tr\, D_t$ of the *estimation error covariance matrix* $D_t = \mathbb{E}(X_t - \widehat{X}_t)(X_t - \widehat{X}_t)'$. Related *prediction problems* are to update the estimation error covariance and the state vector to the next time step $t + 1$, based on measurements up to time t.

This may be thought of as a generalized *hidden Markov model*. In the hidden Markov model, X_t is estimated from Z_t (and not $Z_0, Z_1, \ldots, Z_{t-1}, Z_t\}$; see Exercise 1). In any case, the Kalman filter is a recursive method of estimating X_t and predicting X_{t+1} as linear combinations of the measurements $Z_0, Z_1, \ldots, Z_t$ (with a given X_0) that minimizes the expected squared error of estimation. The estimate for the present state is denoted as $\widehat{X}_t$ (*fltering*), while the estimate of the next step X_{t+1} is called the *predictor*.

The procedure begins with a prior estimate of X_t, say $\widetilde{X}_t$, and corrects or updates it according to

$$\widehat{X}_t = \widetilde{X}_t + K_{g,t}(Z_t - H\widetilde{X}_t), \tag{25.4}$$

for a special determination the $p \times m$ matrix $K_{g,t}$, referred to as the *Kalman gain*, and so chosen as to minimize the expected squared error $\mathbb{E}|X_t - \widehat{X}_t|^2$ among all linear combinations of $Z_0, Z_1, \ldots Z_t$. For this purpose, (25.4) may be expressed as

$$\widehat{X}_t = \widetilde{X}_t + K_{g,t}(HX_t + V_t - H\widetilde{X}_t) = (I - K_{g,t}H)\widetilde{X}_t + K_{g,t}(HX_t + V_t). \tag{25.5}$$

Then, for the next time step, the estimate (25.4) is used to obtain the prior estimate of X_{t+1}, i.e., $\widetilde{X}_{t+1} = F\widehat{X}_t$. This completes the recursion, starting with some estimate X_0 of the initial state at time $t = 0$. This initial estimate may be a constant, e.g., a guess of the expected value of the state at time zero.

The main problem is then to find the Kalman gain $K_{g,t}$. For this, observe that using bi-linearity, the error covariance matrix $X_t - \widehat{X}_t$ can be expressed as

$$\begin{aligned} D_t &= \mathbb{E}(X_t - \widehat{X}_t)(X_t - \widehat{X}_t)' \\ &= \mathbb{E}(X_t - \widetilde{X}_t - K_{g,t}(HX_t + V_t - H\widetilde{X}_t))(X_t - \widetilde{X}_t - K_{g,t}(HX_t + V_t - H\widetilde{X}_t)' \\ &= (I - K_{g,t}H)[E(X_t - \widetilde{X}_t)(X_t - \widetilde{X}_t)'(I - K_{g,t}H)' + K_{g,t}\mathbb{E}(V_t V_t')K_{g,t}' \end{aligned}$$

$$
\begin{aligned}
&= (I - K_{g,t}H)\widetilde{D}_t(I - K_{g,t}H)' + K_{g,t}SK'_{g,t} \\
&= \widetilde{D}_t - K_{g,t}H\widetilde{D}_t - \widetilde{D}_tH'K'_{g,t} + K_{g,t}(H\widetilde{D}_tH' + S)K'_{g,t}, \qquad (25.6)
\end{aligned}
$$

where D_t is the error covariance matrix and $\widetilde{D}_t$ denotes the prior error covariance matrix $\mathbb{E}(X_t - \widetilde{X}_t)(X_t - \widetilde{X}_t)'$. The expected squared error $\mathbb{E}|X_t - \widehat{X}_t|^2$ is the trace trD_t.

$$
\mathbb{E}|X_t - \widehat{X}_t|^2 = tr\widetilde{D}_t - trK_{g,t}H\widetilde{D}_t - tr\widetilde{D}_t(K_{g,t}H)' + trK_{g,t}H\widetilde{D}_t(K_{g,t}H)' + trK_{g,t}SK'_{g,t}. \qquad (25.7)
$$

Minimization of this is achieved by differentiating this with respect to $K_{g,t}$ and setting the derivative to be zero. This equation is (Exercise 2)

$$
-2H\widetilde{D}_t + 2H\widetilde{D}_t(K_{g,t}H)' + 2SK'_{g,t} = 0, \qquad (25.8)
$$

or taking the transpose, and recalling that $\widetilde{D}_t$ and S are symmetric,

$$
-2(H\widetilde{D}_t)' + 2K_{g,t}H\widetilde{D}_tH' + 2K_{g,t}S = 0,
$$

whose solution is given by the Kalman gain formula

$$
K_{g,t} = (\widetilde{D}_tH')(H\widetilde{D}_tH' + S)^{-1}. \qquad (25.9)
$$

Using this optimal $K_{g,t}$ in (25.6), one arrives at the covariance matrix of the error as

$$
\begin{aligned}
D_t &= \widetilde{D}_t - K_{g,t}H\widetilde{D}_t - \widetilde{D}_tHK'_{g,t} + K_{g,t}(H\widetilde{D}_tH' + S)K'_{g,t} \\
&= (I - K_{g,t}H)\widetilde{D}_t - [\widetilde{D}_tH - K_{g,t}(H\widetilde{D}_tH' + S)]K'_{g,T} \\
&= (I - K_{g,t}H)\widetilde{D}_t - [\widetilde{D}_tH - \widetilde{D}_tH]K'_{g,T} \\
&= \widetilde{D}_t - K_{g,t}H\widetilde{D}_t \\
&= (I - K_{g,t}H)\widetilde{D}_t. \qquad (25.10)
\end{aligned}
$$

The recursive equation of the prior error covariance matrix $\widetilde{D}_t$ is by

$$
\begin{aligned}
\widetilde{D}_{t+1} &= \mathbb{E}(X_{t+1} - \widetilde{X}_{t+1})(X_{t+1} - \widetilde{X}_{t+1})' \\
&= \mathbb{E}(FX_t + W_{t+1} - F\widehat{X}_t)(FX_t + W_{t+1} - F\widehat{X}_t)' \\
&= \mathbb{E}(F(X_t - \widehat{X}_t) + W_{t+1})(F(X_t - \widehat{X}_t) + W_{t+1})' \\
&= \mathbb{E}F(X_t - \widehat{X}_t)(X_t - \widehat{X}_t)'F' + Q \\
&= F\widetilde{D}_tF' + Q. \qquad (25.11)
\end{aligned}
$$

From now on, we will denote by $\widetilde{D}_t$ the sequence obtained by the recursion (25.11), beginning with an *initial guess*, or estimate $\widetilde{D}_0$. Using this sequence, we will express D_t by (25.10) or, equivalently, by the last line of (25.6).

Finally, one begins with $\widetilde{X}_0 = X_0$ as a guess, perhaps a constant, e.g., what the expected value of X_0 at time zero might be. Similarly, one begins with a guess of the (prior) error covariance matrix $\widetilde{D}_0$. Putting it altogether, the results can be summarized in a theorem as follows.

Theorem 25.1 *(Kalman Recursion).* Consider the state space and measurement models defined by (25.3) with given initial state vector $\widetilde{X}_0$ and initial (prior) error covariance $\widetilde{D}_0$. The Kalman filter prediction $\widetilde{X}_{t+1}$ of the state vector at time $t+1$ can be generated recursively

$$\widetilde{X}_{t+1} = F\widehat{S}_t = F\widetilde{X}_t + F\widetilde{D}_t F'(H\widetilde{D}_t H' + S)^{-1}(Z_k - H\widetilde{X}_t), \tag{25.12}$$

where $\widetilde{D}_t$ is the error covariance of $\widehat{X}_t$ and recursively generated via

$$\widetilde{D}_{t+1} = F\widetilde{D}_t\big(I - H'(H\widetilde{D}_t H' + S)^{-1}H\widetilde{D}_t\big)F' + Q. \tag{25.13}$$

Proof. The predicted state $\widehat{X}_t$ at time t is then given by (25.4), where the Kalman gain $K_{g,t}$ is given by (25.9). The prediction error covariance matrix D_t is given by (25.10) and (25.9). ■

If the matrix F is stable (i.e., all its eigenvalues lie in the interior of the unit circle in the complex plane), then even a bad guess gets corrected pretty quickly. From then on, one uses $\widetilde{X}_t = F\widehat{X}_{t-1}$, and the recursion proceeds using (25.11), (25.10), (25.9), and (25.4). As mentioned earlier, the optimal predictor of the state X_{t+1}, based on measurements up to time t, is $F\widehat{X}_t = \widetilde{X}_{t+1}$.

Remark 25.1. If S is non-singular, then so is the matrix in (25.9). Even if $S = 0$, but Q is non-singular and H is of full rank, then $H\tilde{D}_t H'$ is also non-singular.

One may think of the optimization with respect to the Kalman gain as a convenient way to obtain the best linear predictor[2] based on $Z_1, \ldots, Z_t$.

Remark 25.2. It is possible to change the linear models (25.1), (25.2) to affine linear models by absorbing a constant in W_t (and/or in V_t). This, however, complicates the algebra, e.g., see how (25.6) changes with covariance S replaced by $\mathbb{E}(W_t + a)(W_t + a)' = \mathbb{E}W_t W_t' + a'a$.

Remark 25.3. In case of non-linearity in the deterministic part of (25.4) or (25.5), one may use linear approximations over small time steps. In such a case, one need to use time dependence of the matrices F or H (see Exercise 3).

[2] See Brockwell and Davis (1991): §12.2.

Remark 25.4. It is important to note that the derivation of the Kalman filter does not require the hypotheses of independence stated at the outset of this chapter. One only needs the random variables W_t as well as V_t to be only uncorrelated over time.[3]

Example 1 *(Fuel Tank Depths).* A common illustrative application of Kalman filters occurs in the estimation of tank depths, e.g., fuel tank level, based on depth measurements from a floating sensor. One assumes that the level increases at a constant fill rate r per unit time, so that the depth levels at successive time units Δ are given by $L_{t+1} = L_t + rt + W_{t+1}^{(1)}, t = 0, 1, \ldots, L_0 = 0$. To view this linear, rather than affine linear, dynamical system, one may consider the state variable as $X_t = (L_t, R_t)', t \geq 0$. In continuous time, $R_t = \frac{dL_t}{dt}$ is the derivative of the depth level. Then

$$X_{t+1} = \begin{pmatrix} L_{t+1} \\ R_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & \Delta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} L_t \\ R_t \end{pmatrix} + \begin{pmatrix} W_{t+1}^{(1)} \\ W_{t+1}^{(1)} \end{pmatrix} = FX_t + W_{t+1}, t = 0, 1, \ldots . \tag{25.14}$$

Suppose that the sensor measurements provide readings for L_t but not the rate R_t. So,

$$Z_{t+1} = HX_t + V_t, \quad t = 0, 1, \ldots \tag{25.15}$$

Here

$$F = \begin{pmatrix} 1 & \Delta \\ 0 & 1 \end{pmatrix} \tag{25.16}$$

is a 2 × 2 matrix with eigenvalues $\lambda = 1$, and

$$H = \begin{pmatrix} 1 & 0 \end{pmatrix} \tag{25.17}$$

is a 1 × 2 matrix. Thus, setting $\Delta = 1$,

$$\begin{pmatrix} L_{t+1} \\ R_{t+1} \end{pmatrix} = \begin{pmatrix} L_t + R_t + W_{t+1}^{(1)} \\ R_t + W_{t+1}^{(2)}, \end{pmatrix} \tag{25.18}$$

and

$$Z_t = L_t + V_t, \quad t = 0, 1, 2, \ldots \tag{25.19}$$

Assume V_t has mean zero and variance $S = s^2 > 0$, and W_t has mean zero and covariance matrix $Q = \begin{pmatrix} \sigma^2 & \gamma \\ \gamma & \sigma^2 \end{pmatrix}$. If one guesses $\widetilde{D}_0 = \begin{pmatrix} \kappa^2 & 0 \\ 0 & \kappa^2 \end{pmatrix}$, then the initial

[3] The succinct treatment presented here follows Lacey (2020).

Kalman gain is given by $K_{g,0} = H\widetilde{D}_0'(H\widetilde{D}_0H' + S)^{-1}$. In particular,

$$K_{g,0} = \kappa^2 \begin{pmatrix} (\kappa^2 + s^2)^{-1} \\ 0 \end{pmatrix}. \tag{25.20}$$

$$\begin{aligned} \widetilde{X}_1 &= F\widetilde{X}_0 + F\widetilde{D}_0H'(H\widetilde{D}_0H' + S)^{-1}(Z_0 - H\widetilde{X}_0) \\ &= \begin{pmatrix} r \\ r \end{pmatrix} + \begin{pmatrix} \kappa^2 \\ 0 \end{pmatrix} \frac{1}{\kappa^2 + s^2} z_0 \\ &= \begin{pmatrix} r + \frac{z_0\kappa^2}{\kappa^2+s^2} \\ r \end{pmatrix}. \end{aligned} \tag{25.21}$$

After further tedious matrix multiplications, one obtains

$$\begin{aligned} \widetilde{D}_1 &= F\widetilde{D}_0(I - H'(H\widetilde{D}_0H' + S)^{-1}H\widetilde{D}_0)F' + Q \\ &= \begin{pmatrix} \kappa^2 + \sigma^2 + \frac{\kappa^2 s^2}{\kappa^2+s^2} & \kappa^2 + \gamma \\ \kappa^2 + \gamma & \kappa^2 + \sigma^2 \end{pmatrix}. \end{aligned} \tag{25.22}$$

For this example, the eigenvalue 1 of F is not interior to the unit circle. Notice that the variance in the fluid level and rate are larger than σ^2.

Full implementation of the recursions will clearly be aided by computational software.[4] The intention here is to merely indicate the nature of the computations in a realistic example.

Remark 25.5. There has been a great deal of work on nonlinear filters during the past fifty years or so. We refer to Budhiraja (2003) for references and for the asymptotic properties of the filter in the model $Y_t = \int_{[0,t]} h(X_s)ds + W_t,\ t \geq 0$, where the signal process $\{X_t\}$ is a Markov process on a Polish space S, h is a function on S to $\mathbb{R}^d$, $\{W_t\}$ is a standard d-dimensional Brownian motion, and $\{Y_t\}$ is the observation process. The objective is to find the conditional distribution of X_t given the process $\{Y_s : 0 \leq s \leq t\}$.

Exercises

1. (*Hidden Markov Model*) A bivariate Markov process $(X_t, Z_t), t = 0, 1, \ldots$ such that both (the hidden component) $\{X_t : t \geq 0\}$ and the bivariate process $\{(X_t, Z_t) : t \geq 0\}$ are (homogeneous) Markov processes, and Z_t is a (possibly

[4] For example, MATLAB has special packages for Riccati equation iteration of the type required by Theorem 25.1.

random) function of X_t is referred to as a *hidden Markov process*. (a) Show that (25.1), (25.2) may be viewed as a hidden Markov model. (b) Assuming $\mathbb{E}X_t^2 < \infty$, show that the least squares estimate of X_t as a function of Z_t is $\mathbb{E}(X_t|Z_t)$.

2. Verify the critical equation (25.8) for optimality.
3. Extend the calculations for the Kalman filter with time-dependent matrices F_t, H_t in place of F, H.
4. Suppose that X is normally distributed with mean μ and variance $\sigma^2 > 0$, and the conditional distribution of Z given X is normal with mean $a + bX$ and variance $s^2 > 0$. Show that the conditional distribution of X given Z is normal with mean $\widehat{X}$ and variance v^2, where $v^2 \in (0, \infty)$ and $\widehat{X}$ are uniquely determined by
$$\frac{1}{v^2} = \frac{1}{\sigma^2} + \frac{b^2}{s^2}, \quad \frac{\widehat{X}}{v^2} = \frac{\mu}{\sigma^2} + b\frac{Z-a}{s^2}.$$
5. Consider a vehicle that moves at a constant speed v, starting from a position $X_0 = x_0$. After one unit of time, the position is given by
$$X_1 = x_0 + v + \sigma W_1,$$
where W_1 is the standard normal and $\sigma^2 \geq 0$ is the variance in the intrinsic noise term σW_1. The position is also measured, by GPS or odometer, say, to be given by
$$Z_1 = X_1 + sV_1,$$
where V_1 is the standard normal and $s^2 > 0$ is the variance in the observation error sV_1. The problem is to estimate the position. Intuitively, one might expect a convex combination $\lambda\mathbb{E}X_1 + (1-\lambda)Z_1$ to be optimal for some appropriate weighting $0 \leq \lambda \leq 1$ depending on the respective uncertainties in each. (a) Show that such intuition is correct in minimizing mean-square error with $\lambda = \frac{s^2}{\sigma^2+s^2}$. (b) Verify that this is the Kalman filter solution.
6. Suppose that X is normally distributed with mean μ and variance $\sigma^2 > 0$, and the conditional distribution of Z given X is normal with mean $a + bX$ and variance $s^2 > 0$. Show that the conditional distribution of X given Z is normal with mean $\widetilde{X}$ and variance v^2, where $v^2 \in (0, \infty)$ and $\widetilde{X}$ are uniquely determined by
$$\frac{1}{v^2} = \frac{1}{\sigma^2} + \frac{b^2}{s^2}, \quad \frac{\widetilde{X}}{v^2} = \frac{\mu}{\sigma^2} + b\frac{Z-a}{s^2}.$$

Appendix A
Spectral Theorem for Compact Self-Adjoint Operators and Mercer's Theorem

Let H be a real or complex separable Hilbert space. We will consider a special class of compact linear operators K on H in this section. The main result for this appendix is Theorem A.3, where H is the space $L^2([c, d], dx)$ of square-integrable functions (with respect to Lebesgue measure dx on $[c, d]$), and K is an integral operator of the form (A.2) in the example application. With such structure in mind we often denote elements of H by symbols $f, g, \dots$. The focus is on a spectral theorem for compact self-adjoint operators. This will be expanded to bounded self-adjoint operators in the next Appendix B.

Some basic concepts and notation are provided within the following definition.

Definition A.1. A linear operator A on H is *bounded* if $\|A\| := \sup\{\|Af\| : f \in H, \|f\| = 1\} < \infty$, in which case $\|A|\|$ defines the *norm* of A. A is *compact* if the set $A(D) = \{Af : f \in D\}$ is *precompact*, i.e., if $A(D)$ has compact closure, for every norm-bounded subset D of H.

Choosing $D = \{f \in H : \|f\| = 1\}$ makes it clear that a compact operator is a bounded operator since compact subsets of H are norm bounded.

Let H^* denote the dual space of H, i.e., the space of bounded linear functionals on H. Then, by the Riesz representation theorem,[1] for $\ell \in H^*$ there is a unique $f_\ell \in H$ such that $\ell f = \langle f, f_\ell \rangle$, for all $f \in H$. In particular, H and H^* are algebraically isomorphic and topologically isometric spaces.

Proposition A.1. Given a bounded linear operator A there is a unique bounded linear operator A^* on H, referred to as the adjoint, such that $\langle Af, g\rangle = \langle f, A^*g\rangle$, for all $f, g \in H$. Moreover $\|A\| = \|A^*\|$.

Proof. The (Banach space) adjoint, denoted $A' : H^* \to H^*$, is defined by

[1] BCPT, Theorem 1.2, p. 248–249.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

$$A'\ell(f) = \ell(Af), \quad \ell \in H^*, f \in H.$$

One may readily check that $A' \in \mathcal{H}^*$. In view of the Riesz representation theorem, there is a linear isometry $L : H \to H^*$ defined by $f \to Lf$ where $L(f)(g) := \langle f, g\rangle$, $g \in H$. With this one may define

$$A^* = L^{-1}AL$$

to obtain the (Hilbert space) adjoint map A^* with the asserted properties. ■

Definition A.2. Let $A \in \mathcal{L}(H)$. The operator A^* is referred to as the adjoint to A. A is said to be *self-adjoint* if $A = A^*$, i.e., $\langle Af, g\rangle = \langle f, Ag\rangle$ for all $f, g \in H$. A self-adjoint operator A is *positive* if $\langle Af, f\rangle \geq 0$ for all, $f \in H$. The *null space* of A is the set $\mathcal{N}_A = \{h \in H : Ah = 0\}$. The *range of* A, is the set $\mathcal{R}_A = \{Af : f \in H\}$. For $D \subseteq H$, $D^\perp \equiv \{f : \langle f, g\rangle = 0 \text{ for all, } g \in D\}$. One writes $A \perp B$ if $\langle h, g\rangle = 0$ for all, $h \in A$, for all, $g \in B$. If $\|f\| = 1$, f is called a *unit vector.* A sequence of unit orthogonal vectors is said to be *orthonormal.*

Remark A.1. More succinctly stated, for arbitrary fixed $g \in H$, the map $f \to \langle Af, g\rangle$, $f \in H$, defines a bounded linear functional on H. Thus, by the Riesz representation theorem, for each $g \in H$ there is a unique $h \in H$ such that $\langle Af, g\rangle = \langle f, h\rangle$ for all $f \in H$. In this way, $A^*g = h$ defines a bounded linear operator $A^* : H \to H$, referred to as the *adjoint* of A. The self-adjointness property, $\langle Af, g\rangle = \langle f, Ag\rangle$ for all $f, g \in H$, may also be viewed as the operator equivalence $A^* = A$.

Lemma 1. Let A be a bounded self-adjoint operator. Then

(a) $\sup\{|\langle Af, f\rangle| : \|f\| = 1\} = \|A\|$,
(b) $\mathcal{R}_A^\perp = \overline{\mathcal{R}}_A^\perp = N_A$.

Proof. (a) For every f with $\|f\| = 1$, $|\langle Af, f\rangle| \leq \|Af\| \leq \|A\|$; hence it suffices to show that $\|Af\| \leq d := \sup\{|\langle Ag, g\rangle| : \|g\| = 1\}$. For this use, for all $c > 0$ and for all f with $\|f\| = 1$, the relations

$$\begin{aligned}
\|Af\|^2 &= \frac{1}{4}\left[\left\langle A\left(cf + \frac{1}{c}Af\right), cf + \frac{1}{c}Af\right\rangle - \left\langle A\left(cf - \frac{1}{c}Af\right), cf - \frac{1}{c}Kf\right\rangle\right] \\
&\leq \frac{1}{4}\left[d\|cf + \frac{1}{c}Af\|^2 + d\|cf - \frac{1}{c}Af\|^2\right] = \frac{d}{2}\left[c^2\|f\|^2 + \frac{1}{c^2}\|Af\|^2\right] \\
&= \frac{d}{2}\left[c^2 + \frac{1}{c^2}\|Af\|^2\right]. \qquad \text{(A.1)}
\end{aligned}$$

The minimum value of the last term (as a function of $c^2 > 0$) is attained for $c^2 = \|Af\|$, so that (A.1) yields $\|Af\|^2 \leq d\|Af\|$, or, $\|Af\| \leq d$.

(b) For $h \in \mathcal{N}_A$, one has $Ah = 0$. Thus $\langle f, Ah\rangle = 0$ for all f, and therefore $\langle Af, h\rangle = 0$ for all f. In particular, $\{h\} \perp \mathcal{R}_A$. Hence $\mathcal{N}_A \subseteq R_A^\perp = \overline{\mathcal{R}}_A^\perp$. On the other hand, if $h \notin \mathcal{N}_A$, then $\|Ah\|^2 > 0$, i.e., $\langle Ah, Ah\rangle \equiv \langle h, \widetilde{A}(Ah)\rangle > 0$, so that $h \notin \mathcal{R}_{\widetilde{A}}^{\perp}$. Hence $\mathcal{N}_A \supseteq \mathcal{R}_{\widetilde{A}}^{\perp}$. ■

Theorem A.2 (Spectral Theorem for Self-Adjoint Compact Operators). Let K be a self-adjoint compact operator on H. Then the following are true:

(a) The eigenvalues of K are real and countable and the eigenspace of each nonzero eigenvalue is finite dimensional.
(b) Either the set of nonzero eigenvalues of K is finite, in which case $\mathcal{R}_K = \overline{\mathcal{R}}_K$ is finite dimensional, or the nonzero eigenvalues λ_n may be ordered by decreasing magnitude: $|\lambda_1| = \|K\| \geq |\lambda_2| \geq \ldots$, and form a denumerable sequence converging to zero.
(c) If K is also positive, then the eigenvalues are all nonnegative.
(d) Let g_i $(i \geq 1)$ denote an orthonormal sequence of eigenvectors with respective eigenvalues λ_i $(i \geq 1)$, counting multiplicities. That is, there are r mutually orthogonal unit eigenvectors (among the g_i) with a nonzero eigenvalue λ of multiplicity r. Then $\{g_i\}_{i=1}^{\infty}$ form a complete orthonormal sequence for $\overline{\mathcal{R}}_K$: so that if $g \in \mathcal{R}_K$ one has the expansion $g = \sum_i \langle g, g_i\rangle g_i$. In particular,

$$Kf = \sum_i \langle Kf, g_i\rangle g_i = \sum_i \lambda_i \langle f, g_i\rangle g_i \qquad \text{for all } f \in H.$$

Proof. We omit the trivial case $K = 0$. (a) If g is an eigenvector of norm one with a nonzero eigenvalue λ, then $\lambda = \langle \lambda g, g\rangle = \langle Kg, g\rangle = \langle g, Kg\rangle = \langle g, \lambda g\rangle = \overline{\lambda}\langle g, g\rangle = \overline{\lambda}$. Let $E = \{v_1, v_2, \cdots\}$ be an orthonormal basis of the eigenspace of $\lambda \neq 0$. If E is not finite, then the sequence $f_n = v_n/\lambda$ $(n = 1, 2, \ldots)$ is bounded (each having norm $1/|\lambda|$) so that, by compactness of K, the sequence $v_n = Kf_n$ $(n = 1, 2, \ldots)$ has a convergent (and, therefore, Cauchy) subsequence. But this is impossible since $\|v_n - v_m\|^2 = 2$ for all $n \neq m$.

(b) One can find f_n, $\|f_n\| = 1$ $(n = 1, 2, \ldots)$ such that $|\langle Kf_n, f_n\rangle| \to \|K\|$ (by Lemma 1(a)). By compactness of K, there is a subsequence $f_{n'} \to g_1 \in H$, $\|g_1\| = 1$. Then $|\langle Kg_1, g_1\rangle| = \|K\|$, so that $\langle Kg_1, g_1\rangle = \lambda_1$ with $\lambda_1 = \|K\|$ or $-\|K\|$. Now $0 \leq \|Kg_1 - \lambda_1 g_1\|^2 = \|Kg_1\|^2 + \lambda_1^2 - 2\lambda_1^2 = \|Kg_1\|^2 - \lambda_1^2 \leq \|K\|^2 - \lambda_1^2 = 0$. Hence $Kg_1 = \lambda_1 g_1$, so that λ_1 is an eigenvalue of K with a unit eigenvector g_1. Consider now the subspace $H_1 = \{g_1\}^{\perp}$ and note that if $g \in H_1$, then $\langle Kg, g_1\rangle = \langle g, Kg_1\rangle = \lambda_1\langle g, g_1\rangle = 0$; that is, K maps H_1 into H_1. Apply the above argument to H_1 to find an eigenvalue λ_2 with a unit eigenvector g_2, $|\lambda_2| \leq |\lambda_1|$. Let $H_2 = \{g_1, g_2\}^{\perp}$ and consider K on H_2 to find a unit eigenvector $g_3 \in H_2$ with eigenvalue λ_3 such that $|\lambda_3| \leq |\lambda_2|$, and so on. The process terminates after n steps if $H_{n+1} = \{g_1, g_2, \ldots, g_n\}^{\perp} = \{0\}$, or if $\langle Kg, g\rangle = 0$ for all $g \in H_{n+1}$. In either case, H_{n+1} is then the null space of K. For, by Lemma 1(a), $\|K\| = 0$ on H_{n+1}. If the process does not terminate after a finite number of steps, then there are infinitely many eigenvalues λ_i such that $|\lambda_1| \geq |\lambda_2| \geq \cdots$, with unit eigenvectors $g_1, g_2, \cdots$ as defined above. Suppose, for sake of contradiction, that λ_i does not converge to zero. Then the sequence $f_i = g_i/\lambda_i$, $i \geq 1$, is bounded, since $\|f_i\| = 1/|\lambda_i|$. By compactness of K, there exists a convergent subsequence of $g_i := K(g_i/\lambda_i)$, $i \geq 1$. But this is impossible, since $\|g_i - g_j\|^2 = 2$ for all $i \neq j$. Hence $\lambda_i \to 0$.

(c) If K is positive and λ is an eigenvalue with unit eigenvector g, then $\lambda = \langle Kg, g\rangle \geq 0$.

(d) From the procedure described in the proof of part (b), it follows that $\{g_1, g_2, \cdots\}^\perp$ is the null space $\mathcal{N}_K$ of K, whether the process terminates or not. Also, clearly, the linear span L, say, of $\{g_1, g_2, \cdots\}$ is contained in $\mathcal{R}_K$, and $L^\perp = \mathcal{N}_K$. It now follows that $\overline{L} = \overline{\mathcal{R}}_K$. Therefore, $\{g_i\}_{i=1}^\infty$ is a complete orthonormal basis of $\overline{\mathcal{R}}_K$, and for every $f \in \overline{\mathcal{R}}_K$ the expansion $f = \sum_i \langle f, g_i\rangle g_i$ holds. Also, $Kf = \sum_i \langle Kf, g_i\rangle g_i = \sum_i \lambda_i \langle f, g_i\rangle g_i$ holds for all $f \in H$. ■

An important example of a compact self-adjoint operator is an *integral operator* K defined on $H = L^2([c,d], dx) \equiv L^2$ as

$$(Kf)(x) = \int_{[c,d]} K(x,y) f(y) dy \qquad f \in L^2, \tag{A.2}$$

where the *kernel function* $K(\cdot,\cdot)$ is a real- or complex-valued continuous function on $[c,d] \times [c,d]$ satisfying

$$\overline{K(x,y)} = K(y,x). \tag{A.3}$$

It is simple to check that K is self-adjoint: $\langle Kf, g\rangle = \langle f, Kg\rangle$. To show that it is compact, note that

$$|(Kf)(x_1) - (Kf)(x_2)| \leq (d-c)^{\frac{1}{2}} \|f\| \max\{|K(x_1,y) - K(x_2,y)| : y \in [c,d]\}. \tag{A.4}$$

This shows that on any subset D of $L^2([c,d], dx)$ which is bounded in L^2-norm, Kf is equicontinuous and bounded. Therefore, by the Arzella–Ascoli Theorem (BCPT, p. 244.), $K(D)$ is precompact in the supnorm distance, and hence in the L^2-distance. Let λ_i $(i \geq 1)$ denote the nonzero eigenvalues of K (counting multiplicities) with corresponding unit eigenvectors g_i $(i \geq 1)$, as stated in Theorem A.2. Since $Kf(x)$ is continuous for all $f \in L^2$, $g_i(x) \equiv K(g_i/\lambda_i)(x)$ is continuous for all i.

To proceed we record a basic result from advanced calculus for ease of reference.

Lemma 2 (Dini's Theorem). Let $\{h_n : n \geq 1\}$ be a pointwise nondecreasing sequence of continuous real-valued functions on a compact metric space S. If $\lim_{n\to\infty} h_n(x) = h(x)$ exists for each $x \in S$ and if h is continuous, then this convergence is uniform.

Proof. For each $n \in S$ let $g_n = h - h_n$. Then $\{g_n : n \geq 1\}$ is a sequence of pointwise nonincreasing, nonnegative functions converging to zero at each $x \in S$. Let $\epsilon > 0$. For $x \in S$ there is a positive integer N_x such that $g_{N_x}(x) < \frac{\epsilon}{2}$. Since g_{N_x} is continuous at x, there is an open ball B_x centered at x such that $g_{N_x}(y) < \frac{\epsilon}{2}$ for all $y \in B_x$. By compactness of $S \subseteq \cup_{x \in S} B_x$ there is a finite subcover $S \subseteq \cup_{j=1}^{k} B_{x_j}$. Let $N = \max\{N_{x_j} : 1 \leq j \leq k\}$. Then, for arbitrary $y \in S$, one has $y \in B_{x_j}$ for some $1 \leq j \leq k$, and therefore $g_{N_{x_j}}(y) < \epsilon$. Thus, since $N \geq N_{x_j}$, $0 \leq g_N(y) < g_{N_{x_j}}(y) < \epsilon$. Since $y \in S$ is arbitrary, this proves uniform convergence of g_n to zero, and hence uniform convergence of h_n to h. ■

We may now obtain the intended main result.

Theorem A.3 (Mercer's Theorem). If, in addition to the above hypotheses of continuity of $K(\cdot,\cdot)$ and (A.3), K is positive, then $K(\cdot,\cdot)$ has the eigenfunction expansion

$$K(x,y) = \sum_i \lambda_i g_i(x)\overline{g_i(y)} \qquad (x,y) \in [c,d] \times [c,d], \tag{A.5}$$

where the convergence of the series is absolute and uniform in (x,y). Here $\|K\| = \lambda_1 \geq \lambda_2 \geq \cdots$ are the positive eigenvalues of K with corresponding complete orthonormal sequence $\{g_i\}_{i=1}^{\infty}$ of unit eigenvectors in $\mathcal{R}_K$.

Proof. Consider the kernel function

$$K_n(x,y) = K(x,y) - \sum_{i=1}^{n} \lambda_i g_i(x)\overline{g_i(y)}, \tag{A.6}$$

which is continuous in (x,y), satisfies the symmetry condition, namely $K_n(y,x) = \overline{K_n(x,y)}$. The corresponding operator K_n is nonnegative:

$$\begin{aligned}\langle K_n f, f\rangle &= \langle Kf, f\rangle - \sum_{i=1}^{n} \lambda_i \langle f, g_i\rangle\langle g_i, f\rangle \\ &= \sum_{i=1}^{\infty} \lambda_i \langle f, g_i\rangle\langle g_i, f\rangle - \sum_{i=1}^{n} \lambda_i \langle f, g_i\rangle\langle g_i, f\rangle \\ &= \sum_{i=n+1}^{\infty} \lambda_i |\langle f, g_i\rangle|^2 \geq 0,\end{aligned} \tag{A.7}$$

where we have used the expansion of Kf using Theorem A.2(d). It follows that $K_n(x,x) \geq 0$ for all x. For if $K_n(x_0,x_0) < 0$, then there exists $\varepsilon > 0$ such that $K_n(x,y) < 0$ for all x, y belonging to $(x_0 - \varepsilon, x_0 + e)$, which would imply

$$0 \leq \left\langle K_n \mathbf{1}_{(x_0-\varepsilon,x_0+\varepsilon)}, \mathbf{1}_{(x_0-\varepsilon,x_0+\varepsilon)}\right\rangle \equiv \int_{(x_0-\varepsilon,x_0+\varepsilon)^2} K_n(x,y)\,dxdy < 0,$$

a contradiction. Since, by (A.6),

$$K_n(x,x) = K(x,x) - \sum_{i=1}^{n} \lambda_i |g_i(x)|^2 \geq 0 \qquad \text{for all } x, \forall n, \tag{A.8}$$

it follows that

$$\sum_{i=1}^{\infty} \lambda_i |g_i(x)|^2 \leq K(x,x) \leq M \equiv \max\{K(x,x) : c \leq x \leq d\}. \tag{A.9}$$

As a consequence, by the Cauchy–Schwarz inequality,

$$\begin{aligned} \left| \sum_{m}^{n} \lambda_i \overline{g_i(y)} g_i(x) \right|^2 &\leq \left(\sum_{m}^{n} \lambda_i |g_i(y)|^2 \right) \left(\sum_{m}^{n} \lambda_i |g_i(x)|^2 \right) \\ &\leq M \sum_{m}^{n} \lambda_i |g_i(x)|^2. \end{aligned} \tag{A.10}$$

Thus the *series* $\sum_i x_i g_i(x) \overline{g}_i(y)$ *converges absolutely* to some quantity whose magnitude is no more than M. Let $G(x,y)$ denote the limit of this series. By (A.10), for each fixed x, the sequence of functions $y \to \sum_1^n \lambda_i g_i(x) \overline{g_i(y)}$ converges uniformly (in y) to $G(x,y)$. Hence $y \to G(x,y)$ is *continuous* for each fixed x. Now note that for all $f \in L^2$,

$$\int G(x,y) f(y) dy = \sum_i \left(\int \lambda_i \overline{g_i(y)} f(y) dy \right) g_i(x) = \sum_i \lambda_i \langle f, g_i \rangle g_i(x). \tag{A.11}$$

By Theorem A.2(d), the last series on the right equals $Kf(x)$ for all x outside a subset of $[c,d]$ of Lebesgue measure zero. We will show that the series actually converges to the (continuous) function $Kf(x)$ uniformly in x. For this write $f_n^{(x)} = \sum_1^n \langle f, g_i \rangle g_i(x)$. Then

$$|Kf_n(x) - Kf(x)|^2 = \left| \int_{[c,d]} K(x,y)[f_n(y) - f(y)] dy \right|^2 \leq M_1^2 \|f_n - f\|^2, \tag{A.12}$$

where $M_1 = \max\{|K(x,y)| : x, y \in [c,d]\}$. Hence $Kf_n(x) \to Kf(x)$ uniformly in x as $n \to \infty$. Thus the convergence of the last series in (A.11) is to $Kf(x)$ (uniformly) for all x. Therefore,

$$\int_{[c,d]} (G(x,y) - K(x,y)) f(y) dy = 0 \qquad \text{for all } f \in L^2 \quad (\forall x \in [c,d]). \tag{A.13}$$

Letting $f(y) = \overline{(G(x,y) - K(x,y))}$, it follows that $G(x,y) = K(x,y)$ for all $y \in [c,d], x \in [c,d]$. We have established the *absolute convergence of the series in (A.5) to* $K(x,y)$. To prove *uniform convergence* of the series, consider the above absolute convergence with $y = x$ to get $\sum_{i=1}^{n} \lambda_i |g_i(x)|^2 \uparrow K(x,x)$. Since $x \mapsto K(x,x)$ is continuous and $[c,d]$ compact, this convergence is also uniform in x by Dini's theorem. Hence, by the first inequality in (A.11), $|\sum_m^n \lambda_i \overline{g_i(y)} g_i(x)|$ may be made smaller than any preassigned $\varepsilon > 0$ by letting $n \geq m \geq m_\varepsilon$ for a suitable integer m_ε. Thus the convergence of the series in (A.5) is uniform for $x, y \in [c,d]$. ■

Remark A.2. Let $r(s,t)$, $c \leq s, t \leq d$, be a continuous covariance function of a mean-zero real- or complex-valued process $\{X_t : t \in [c,d]\}$. For arbitrary $t_j \in [c,d]$, $1 \leq j \leq n$, and arbitrary $a_j \in \mathbb{C}$, $1 \leq j \leq n$, one has

$$\sum_{j,k} a_j \overline{a_k} r(t_j, t_k) = \mathbb{E} \left| \sum_{1}^{n} a_j X_{t_j} \right|^2 \geq 0. \tag{A.14}$$

From this it follows by usual Riemann sum approximation of continuous functions f on $[c,d]$ that $\int_{[c,d]} r(s,t) f(t) f(s) ds dt \geq 0$. Since continuous functions are dense in L^2, $r(s,t)$ is seen to be the kernel of a nonnegative integral operator.

Remark A.3. It follows from Mercer's theorem that

$$\begin{aligned} \infty > \int_{[c,d]^2} |K(x,y)|^2 dx dy &= \int_{[c,d]^2} \left(\sum_i \lambda_i \overline{g_i(y)} g_i(x) \right) \left(\overline{\sum_j \lambda_j \overline{g_j(y)} g_j(x)} \right) dx dy \\ &= \sum \lambda_i^2. \end{aligned} \tag{A.15}$$

A compact self-adjoint operator K whose eigenvalues λ_i satisfy $\sum \lambda_i^2 < \infty$ is called a *Hilbert–Schmidt* operator.

Example 1. Consider the integral operator K on the real Hilbert space $L^2 = L^2([0,1], dx)$ with the kernel function

$$K(x,y) = x(1-y) \text{ if } 0 \leq x \leq y \leq 1, \text{ and } K(x,y) = (1-x)y \text{ if } 0 \leq y < x \leq 1. \tag{A.16}$$

By direct calculation using integration by parts, one can check that $\varphi_n(x) = \sqrt{2} \sin(n\pi x)$ is a unit eigenfunction of K with eigenvalue $\lambda_n = (n^2\pi^2)^{-1}$, for every $n = 1, 2, \cdots$. To show that $\{\varphi_n(\cdot)\}_{n=1}^{\infty}$ is a complete orthonormal sequence for $\mathcal{R}_K$ (or $\overline{\mathcal{R}}_K$) first note that $\mathcal{R}_K$ comprises twice differentiable functions vanishing at 0 and 1:

$$\frac{d^2}{dx^2}(Kf)(x) = -f(x), \quad (Kf)(0) = 0 = (Kf)(1), \quad (f \in L^2). \tag{A.17}$$

Secondly, it follows from the theory of Fourier series[2] that the functions $\sin(n\pi x)$ $(n = 1, 2, \cdots)$ are dense in the set of *odd* functions in $L^2([-1, 1], dx)$. Given an $f \in L^2([0, 1], dx))$, extend $g = Kf$ to an odd function on $[-1, 1]$ by setting $g(x) = -g(-x)$ for $-1 \leq x < 0$. Then g can be approximated on $[-1, 1]$ (and, therefore, on [0, 1]) arbitrary closely by linear combinations of the functions $\sin(n\pi x)$ $(n \geq 1)$. Therefore, by Mercer's theorem,

$$K(x, y) = 4\sum_{n=1}^{\infty}(n^2\pi^2)^{-1}\sin(n\pi x)\sin(n\pi y). \tag{A.18}$$

This kernel K is the *Green's function* (or, *fundamental solution*) of the following boundary value problem: For an arbitrarily given $f \in L^2([0, 1], dx)$ find g such that

$$g''(x) = -f(x), \qquad g(0) = g(1) = 0. \tag{A.19}$$

$x \to K(x, y)$ is the (distributional) solution of (A.19) when f is the delta function $\delta_y(\cdot)$.

[2] BCPT, Chapter VI.

Appendix B
Spectral Theorem for Bounded Self-Adjoint Operators

H will continue to denote a real or complex Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $||f|| = \sqrt{\langle f, f\rangle}$, $f \in H$. The space of bounded linear operators on H will be denoted $\mathcal{L}(H)$, with norm $\|A\| = \sup_{||f||=1} ||Af||$, $A \in \mathcal{L}(H)$; the context will be used to distinguish the operator norm from the vector space norm.

Definition B.1. Let A be a bounded linear operator on H.

(a) The resolvent set of A is $\rho(A) = \{\mu \in \mathbb{C} : \mu - A$ is a bijection with bounded inverse $\}$, where $(\mu - A)f = \mu f - Af$, $f \in H$.
(b) The complementary set $\sigma(A) = \rho^c(A)$ is called the spectrum of A.
(c) $\lambda \in \mathbb{C}$ is an eigenvalue of A if there is an $f \in H$, $f \neq 0$, such that $Af = \lambda f$. As a subset of $\sigma(A)$, the set of eigenvalues of A is referred to as the point spectrum of A.
(d) The maximum modulus of the spectrum $r(A) = \sup_{\lambda\in\sigma(A)} |\lambda|$ is referred to as the spectral radius of A.

The following propositions display additional significant consequences of self-adjoint symmetry. As noted in Appendix A, it is simple to check directly from the definition that eigenvalues (point spectra) of self-adjoint operators must be real numbers. As the first proposition shows, the same is true for all of $\sigma(A)$.

Proposition B.1. Let $A \in \mathcal{L}(H)$ be self-adjoint. Then,

(a) $\sigma(A) \subset \mathbb{R}$
(b) $\rho(A) = \{\mu = a + bi : b \neq 0\}$.

Proof. Let $a, b \in \mathbb{R}$, $\mu = a + bi$. Then let us show that if $b \neq 0$, then $\mu \notin \sigma(A)$.

$$||(A-\mu)f||^2 = ||(A-a)f||^2 + b^2||f||^2 \geq b^2||f||^2. \tag{B.1}$$

Thus if $b \neq 0$, then μ cannot be an eigenvalue, i.e., not in the point spectrum. In fact, by this inequality, if $b \neq 0$, then $A - \mu$ is injective and has a bounded inverse on its

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

range. In particular, therefore, the range of $A-\mu$ is a closed subspace. If there is a $y \in H$ such that $y \neq (A-\mu)f$ for all $f \in H$, then, using the projection theorem,[1] one can construct a bounded linear functional ℓ_y such that $\ell_y(y) = \inf_{f\in H} ||y - f|| > 0$ and ℓ_y vanishes on the range of $A-\mu$, i.e., $\ell_y(\mu f - Af) = 0$ for all $f \in H$. But this makes μ an eigenvalue of the Banach space adjoint A'. Therefore $\overline{\mu} = a - bi$ is an eigenvalue of $A = A^*$. But this violates the inequality (B.1) if $b \neq 0$. So the range of $A-\mu$ is all of H, and $A-\mu$ is bijective with a bounded inverse, i.e., $\mu \in \rho(A) = \sigma^c(A)$ if $b \neq 0$. ■

Proposition B.2. If $A \in \mathcal{L}(H)$ is self-adjoint, then

$$r(A) = \lim_{n\to\infty} \|A^n\|^{\frac{1}{n}} = \|A\| < \infty.$$

Proof. The existence of the limit is obtained by a subadditivity technique below. With this one can then identify $\|A\|$ as the limit by using self-adjointness to see by induction that $\|A^{2^n}\| = \|A\|^{2^n}$ for $n = 1, 2, \ldots$, so that $\|A\| = \|A^{2^n}\|^{\frac{1}{2^n}}, n = 1, 2, \ldots$. Now, for existence let $a_n = \|A^n\|, n \geq 1$. Since $\|A^nA^m\| \leq \|A^n\| \cdot \|A^m\|$, one has the subadditivity property $a_{n+m} \leq a_n + a_m$. In particular, $a_{mn+r} \leq na_m + ra_1$. Fix m and write $n = mq + r, 0 \leq r \leq m-1$ according to Euclidean division algorithm. Then

$$\frac{a_n}{n} \leq \frac{qa_m}{qm} + \frac{(m-1)a_1}{mq} \leq \frac{a_m}{m} + \frac{a_1}{q}.$$

Letting $q, r \to \infty$, it follows that for any $m \geq 1$ one has

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \frac{a_m}{m}.$$

Thus, $\limsup_{n\to\infty} \frac{a_n}{n} \leq \inf_m \frac{a_m}{m} \leq \liminf_{m\to\infty} \frac{a_m}{m}$ and, hence, the limit exists as asserted. ■

The essential tool required for the spectral theorem is a functional calculus for which $F(A)$ is a well-defined bounded linear operator for continuous functions f defined on $\sigma(A)$, e.g., $F(A) = \sum_{j=0}^m a_jA^j$ for the polynomial $f(\lambda) = \sum_{j=0}^m a_j\lambda^j, \lambda \in \sigma(A)$.

Theorem B.3. Let A be bounded self-adjoint linear operator on H, and let $\mathcal{L}(H)$ be the space of all bounded linear maps on H. Then there is a unique map $\varphi : C(\sigma(A) : \mathbb{C}) \to \mathcal{L}(H)$ such that for $F, G \in C(\sigma(A)) \equiv C(\sigma(A) : \mathbb{C})$,

(a) $\varphi(1) = I, \quad \varphi(\overline{F}) = \varphi(F)^*$.
(b) $\varphi(id) = A$ where $id(\lambda) = \lambda, \lambda \in \sigma(A)$.
(c) $\varphi(FG) = \varphi(F)\varphi(G), \quad \varphi(\mu F) = \mu\varphi(F)$.

[1] BCPT, p. 248.

(d) $\|\varphi(F)\| \le c||F||_\infty$, where $\|\cdot\|$ denotes the operator norm on $\mathcal{L}(H)$. In fact, $\|\varphi(F)\| = ||F||_\infty$.
(e) If $Af = \lambda f$, then $\varphi(F)f = F(\lambda)f$, $f \in H$.

Proof. The idea of the proof is to first consider polynomials $F(A) = \sum_{j=0}^{m} a_j A^j$ since by Stone–Weierstrass approximation[2] they are dense in $C(\sigma(A))$. Let us first check for polynomial F

$$\sigma(F(A)) = \{F(\lambda) : \lambda \in \sigma(A)\}. \tag{B.2}$$

To see this, let $\mu \in \sigma(F(A))$ and factor the polynomial $F(\lambda) - \mu = c(\lambda - \lambda_1)\cdots(\lambda - \lambda_m)$. Then $\lambda_j \in \sigma(A)$ for some $1 \le j \le m$, else $F(A) - \mu$ is invertible, contradicting $\mu \in \sigma(F(A))$. In particular, therefore, $\mu = F(\lambda_j)$. On the other hand, let $\lambda' \in \sigma(A)$ and factor the polynomial $F(\lambda) - F(\lambda') = (\lambda - \lambda')G(\lambda)$, where G is a polynomial. Then $F(A) - F(\lambda') = (A - \lambda')G(A)$. In particular, it follows that $F(A) - F(\lambda')$ is not invertible since $A - \lambda'$ is not invertible. Thus, $F(\lambda') \in \sigma(F(A))$. From (B.2) applied to the polynomial $\overline{F}F$ and the formula (B.2) for the spectral radius, it follows that

$$\begin{aligned}
\|F(A)\|^2 &= \|F(A)^* F(A)\| \\
&= \|\overline{F}F(A)\| \\
&= \sup_{\lambda \in \sigma(\overline{F}F(A))} |\lambda| \\
&= \sup_{\lambda \in \sigma(A)} |\overline{F}F(\lambda)| \\
&= (\sup_{\lambda \in \sigma(A)} |F(\lambda)|)^2.
\end{aligned} \tag{B.3}$$

Thus, $\|F(A)\| = \sup_{\lambda \in \sigma(A)} |F(\lambda)|$. From here one defines $\varphi(F) = F(A)$ for polynomials $F \in C(\sigma(A))$. Since $\|\varphi(F)\| = ||F||_u (\equiv \sup_{\lambda \in \sigma(A)} |F(\lambda)|$, it follows from Stone–Weierstrass approximation that φ as a unique extension to $C(\sigma(A))$. Most of the asserted properties of $\varphi(F)$ may be checked for polynomial F on $\sigma(A)$ and then extended to $F \in C(\sigma(A))$ by continuity. ■

Definition B.2. For a given self-adjoint linear operator A on H and $F \in C(\sigma(A); \mathbb{C})$, one defines $F(A) = \varphi(A)$.

An important application of the functional calculus is, for example, in noting that $\sqrt{A} \in \mathcal{L}(H)$ is definable for a bounded self-adjoint operator A. More generally, one has the following.

[2] BCPT, p. 242.

Corollary B.3. Let $A \in \mathcal{L}(H)$ be self-adjoint. If $F \in C(\sigma(A))$ is a nonnegative real-valued function, then $F(A)$ is a positive self-adjoint operator.

Proof. Note that for real-valued $F \in C(\sigma(A))$ one has $\varphi(F) = \varphi(\overline{F}) = \varphi(F)^*$, i.e., $\varphi(F)$ is self-adjoint for real F. Write $F = (\sqrt{F})^2$ for $F \geq 0$. Then $F(A) \equiv \varphi(F) = \varphi(\sqrt{F})\varphi(\sqrt{F}) = \varphi(\sqrt{F})^*\varphi(\sqrt{F})$ is a positive operator. ■

Now let us see how one may obtain a "spectral decomposition"of A using the functional calculus. Fix $g \in H$. Then, the map $F \to \langle g, F(A)g\rangle$, $F \in C(\sigma(A))$, defines a positive, bounded linear functional on $C(\sigma(A))$. Thus, by the Riesz representation theorem, for each $g \in H$, there is a unique measure ν_g on the Borel σ-field of the closed and bounded set $\sigma(A) \subset \mathbb{R}$ such that

$$\langle g, F(A)g\rangle = \int_{\sigma(A)} F(\lambda)\nu_g(d\lambda), \quad g \in H. \tag{B.4}$$

In particular, taking $F(\lambda) = \lambda^0 (= 1)$,i.e., the constant polynomial, then $F(A)$ is the identity map and

$$\int_{\sigma(A)} \nu_g(d\lambda) = \langle g, g\rangle, \quad g \in H, \tag{B.5}$$

and taking $F(\lambda) = \lambda$, one has

$$\int_{\sigma(A)} \lambda\nu_g(d\lambda) = \langle g, Ag\rangle, \quad g \in H. \tag{B.6}$$

The measure ν_g is referred to as the *spectral measure* of A. Note that one also has

$$\langle g, F(A)f\rangle = \frac{1}{4}\{\int_{\sigma(A)} F(\lambda)\nu_{g+f}(d\lambda) - \int_{\sigma(A)} F(\lambda)\nu_{g-f}(d\lambda)\} \tag{B.7}$$

from the polarization identity.

Appendix C
Borel Equivalence for Polish Spaces

Two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are regarded as *measurably equivalent* if there is bijection $h : S \to T$ such that h and h^{-1} are each measurable. In the case that S and T are topological spaces and $\mathcal{S} = \mathcal{B}(S)$, $\mathcal{T} = \mathcal{B}(T)$ are their respective Borel $\sigma-$fields then measurable equivalence is referred to as *Borel equivalence*. In particular, two homeomorphic topological spaces are Borel equivalent. A separable metric space S is homeomorphic to a subset $h(S)$ (in the relative topology) of the Hilbert cube[1] $H = [0, 1]^{\mathbb{N}}$ (given the product topology on H). If S is also complete, then we can show that $h(S)$ is a Borel subset of H. In particular, we can prove the following.

Proposition C.1. A complete and separable metric space (S, ρ) is Borel equivalent to a Borel subset of the Hilbert cube (with the product topology on H).

Proof. Without loss of generality assume $0 \leq \rho(x, y) \leq 1$ for all $x, y \in S$; else replace by $\frac{\rho(x,y)}{1+\rho(x,y)}$. As already noted it is sufficient from what has already been proven to show that completeness implies that $h(S)$ is a Borel subset of H. Recall that using separability to get a countable dense subset $\{x_1, x_2, \dots\}$ of S, the homeomorphism $h : S \to H$ is defined by $h(x) = (\rho(x, x_1), \rho(x, x_2), dots), x \in S$. To see that $h(S)$ is a Borel set, observe that since it is dense in its closure $\overline{h(S)}$, by completeness of S it is a G_δ subset of $\overline{h(S)}$; i.e., a countable intersection of open (relative to $\overline{h(S)}$) subsets, and hence a (relative) Borel subset of $\overline{h(S)}$. Finally note that the Borel subsets of any Borel set $E \subseteq H$ are simply the Borel subsets of A which are contained in E. Thus $h(S)$ is a Borel subset of H. ■

Another Borel equivalence which is somewhat standard in probability theory is that between $[0, 1]$ and the product space $\{0, 1\}^{\mathbb{N}}$ obtained by binary expansion. The proof can be made most transparent with the help of the following simple lemma.

[1] BCPT, p.143.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

Lemma 1. If S and T are metric spaces and if $S_0 \subseteq S$ and $T_0 \subseteq T$ are both countable subsets, then a Borel equivalence between the complimentary spaces $S \backslash S_0$ and $T \backslash T_0$ may be extended to a Borel equivalence between S and T.

Proof. Since S_0 and T_0 are both countable there is a bijection between them which extends any bijection h between $S \backslash S_0$ and $T \backslash T_0$. With this extension of the given Borel equivalence h, first note that any countable subset of a metric is clearly an F_σ set since points are closed; i.e., a countable union of closed sets. Similarly the union and relative complement of a Borel set with a countable set is a Borel set. In particular, therefore, a subset of S (respectively of T) is Borel if and only if its intersection with $S \backslash S_0$ (respectively $T \backslash T_0$) is a Borel set. So the extended map must be a Borel equivalence. ■

Proposition C.2. The unit interval is Borel equivalent to $\{0, 1\}^{\mathbb{N}}$.

Proof. Let $S_0 = \{(\epsilon_1, \epsilon_2, \dots) \in \{0, 1\}^{\mathbb{N}}$: either $\epsilon_j = 0$ for all but finitely many j, or $\epsilon_j = 1$ for all but finitely many $j\}$. Let $T_0 = \{x \in [0, 1] : x = m2^{-n}$ for some $m, n \in \mathbb{N}\}$. Define $h : \{0, 1\}^{\mathbb{N}} \backslash S_0 \to [0, 1] \backslash T_0$ by $h(\epsilon_1, \epsilon_2, \dots) = \sum_{j=1}^{\infty} \epsilon_j 2^{-j}$. Then since S_0 and T_0 are both countable and since h is easily checked to be a homeomorphism, the assertion follows. ■

Before coming to the main result of this appendix we require another simple observation,

Lemma 2. If S is an arbitrary topological space, then $R = S^{\mathbb{N}}$ and $T = \mathbb{R}^{\mathbb{N}}$ are homeomorphic under their respective product topologies, and hence Borel equivalent. In particular, [0, 1] is Borel equivalent to $H = [0, 1]^{\mathbb{N}}$.

Proof. Since the product $\mathbb{N} \times \mathbb{N}$ is a countable set, it may be enumerated, setting up an obvious homeomorphism between $\mathbb{R}$ and T. For the last assertion simply note that For the last assertion simply note that [0, 1] is Borel equivalent to $\mathbb{R} = \{0, 1\}^{\mathbb{N}}$ which, in turn, is Borel equivalent to $R^{\mathbb{N}}$ and hence to $[0, 1]^{\mathbb{N}}$. ■

Thus we have arrived at the main result of this appendix for Polish spaces.

Theorem C.3. Each complete and separable metric space is Borel equivalent to a Borel subset of [0, 1].

Proof. Since a complete and separable metric space is Borel equivalent to a Borel subset of H and since H is Borel equivalent to [0, 1] the assertion follows. ■

Appendix D
Hahn–Banach, Separation, and Representation Theorems in Functional Analysis

Let $(S, \preceq)$ be a partially ordered set. A *totally ordered* subset $T \subset S$ is a subset with the property that for every pair $x, y \in T$, either $x \preceq y$ or $y \preceq x$. $u \in S$ is said to be an *upper bound* for $T \subset S$ if $x \preceq u$ for all $x \in T$. If $m \in S$ has the property that $m \preceq x$ implies $m = x$, then m is referred to as a *maximal element*.

Axiom (Zorn's Lemma).[1] Let $(S, \preceq)$ be a partially ordered set. If every totally ordered subset has an upper bound, then $(S, \preceq)$ contains at least one maximal element.

Definition D.1. Let V be a vector space over the real numbers. A function $p : V \to \mathbb{R}$ such that

$$p(u+v) \le p(u) + p(v), \text{ for all } u, v \in V, \quad p(cv) = cp(v), \ c \ge 0, v \in V,$$

is referred to as a *sublinear functional* on V.

Theorem D.1 (Hahn–Banach). Let U be a subspace of a real vector space V, Suppose that $\ell : U \to \mathbb{R}$ is a linear functional such that $\ell(u) \le p(u), u \in U$, for some sublinear functional p on V. Then ℓ can extended to a linear functional $\widetilde{\ell}$ on V and such that $\widetilde{\ell}(v) \le p(v), v \in V$.

Proof. The proof has two parts. The first part shows how to linearly extend a linear functional from a proper subspace to the space spanned by adjoining a single vector, while preserving the domination by the semilinear functional p. The second part relies on Zorn's lemma to obtain a maximal extension to a subspace of V with

[1] Zorn's lemma is equivalent to the axiom of choice in Zermelo–Fraenkel (ZF) set theory. In particular, it is not constructive in producing a maximal element. See Folland, p.

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

domination by p. The first part and this maximality then show that this subspace is in fact all of V.

If $U \neq V$, then choose $v_0 \in V \backslash U$. Let $Y = \{u + tv_0 : t \in \mathbb{R}, u \in U\}$. Then Y is a subspace of V. Using sublinearity of p and linearity of ℓ, after adding and subtracting v_0 in the argument for p, one has

$$\ell(u) + \ell(v) = \ell(u + v) \leq p(u - v_0) + p(v_0 + v),$$

and therefore

$$\ell(u) - p(u - v_0) \leq p(v + v_0) - \ell(v), \quad \text{for all } u, v \in U. \tag{D.1}$$

Let a be the upper bound on the left side of (D.1) as a function of $u \in U$. Then

$$\ell(u) - a \leq p(u - v_0), \quad \ell(v) + a \leq p(v + v_0). \tag{D.2}$$

Define $\widetilde{\ell}$ on Y by

$$\widetilde{\ell}(u + tv_0) = \ell(u) + ta, \quad u \in U, t \in \mathbb{R}. \tag{D.3}$$

Then $\widetilde{\ell}$ is linear and $\widetilde{\ell} = \ell$ on U. To see that domination by p is preserved, note that from (D.3), the right side of (D.2), and Definition D.1, one has

$$\widetilde{\ell}(t^{-1}u + v_0) = \ell(t^{-1}u) + a \leq p(t^{-1}u + v_0). \tag{D.4}$$

Thus,

$$\begin{aligned} \widetilde{\ }(u + tv_0) &= t\widetilde{\ell}(t^{-1}u + v_0) \\ &\leq tp(t^{-1}u + v_0) = p(u + tv_0), \text{ for all } t > 0, u \in U. \end{aligned} \tag{D.5}$$

For $t < 0$, using the left inequality of (D.2) to get for all $u \in V, t < 0$,

$$\begin{aligned} \widetilde{\ell}(u + tv_0) &= -t\widetilde{\ell}(-t^{-1}u - v_0) = -t(\ell(t^{-1}u) - a) \\ &\leq -tp(-t^{-1}u - v_0) = p(u + tv_0). \end{aligned}$$

Let $\mathcal{L}$ denote the set of all linear functionals $\tilde{\ell}$, with respective (linear) domains $D_{\tilde{\ell}} \supseteq U$, that linearly extend ℓ and are, respectively, dominated by p on $D_{\tilde{\ell}}$. Then, in view of the first part, $\mathcal{L} \neq \emptyset$. Let us define a partial order $\preceq$ on $\mathcal{L}$ by $\tilde{\ell}_1 \preceq \tilde{\ell}_2$ if and only if $\tilde{\ell}_2$ is a linear extension of $\tilde{\ell}_1$. It is simple to check that $\preceq$ satisfies the conditions of Zorn's lemma for the existence of a maximal $\tilde{\ell} \in \mathcal{L}$. Namely, if $T = \{\hat{\ell}_t\}$, say, is a totally ordered subset of $\mathcal{L}$, then define $\widehat{\ell}$ on $\cup_t D_{\hat{\ell}_t}$ by

$$\widehat{\ell}(u) = \hat{\ell}_s(u) \quad \text{where } u \in D_{\hat{\ell}_s}. \tag{D.6}$$

The total ordering of T implies that $\cup_t D_{\hat{\ell}_t}$ is a vector space, and that $\widehat{\ell}$ is a well-defined linear functional. Thus, (D.6) defines an upper bound for T. Now it follows from Zorn's lemma that $\mathcal{L}$ has a maximal element $\widetilde{\ell} \in \mathcal{L}$, on a (linear) domain $D_{\tilde{\ell}}$, extending ℓ such that $\widetilde{\ell}(u) \leq p(u), u \in D_{\widetilde{\ell}}$. Also $D_{\widetilde{\ell}} = V$ or else, by the first part of this proof, $\widetilde{\ell}$ could be extended in contradiction to its maximality. ■

In addition to the Hahn–Banach theorem, the separation results[2] to follow are also essential tools for Lindvall's proof of Strassen's theorem.

Lemma 1. Suppose that K, F are disjoint subsets of a topological vector space V. If K is compact and F is closed, then 0 has a neighborhood U such that $(K + U) \cap (F + U) = \emptyset$.

Proof. Observe that if W is a neighborhood of 0 in V, then there is a neighborhood U of 0 such that $U = -U$ and $U + U \subset W$. This is because $0 + 0 = 0$ and addition is continuous so that there are neighborhoods U_1, U_2 of 0 such that $U_1 + U_2 \subset W$. Take $U = U_1 \cap U_2 \cap (-U_1) \cap (-U_2)$. In fact this can be iterated by replacing W by U to get $U + U + U + U \subset W$, etc. This observation can be a useful tool as follows.

Without loss of generality assume $K \neq \emptyset$, and let $v \in K$. Since $v \notin F$ and F^c is open, there exists an open neighborhood of v, say G, which is disjoint from F. One may write $G = W + v$, where W is an open neighborhood of 0. Since, by continuity of the vector space operations, the topology of V is invariant under translations, the observation above shows that 0 has a symmetric neighborhood U_v such that $(v + U_v + U_v) \cap (F + U_v) = \emptyset$. Since K is compact, there are finitely many $v_1, \dots, v_n$ in K such that $K \subset \cup_{i=1}^n (v_i + U_{v_i})$. Let $U = \cap_{i=1}^n U_{v_i}$. Then

$$K + U \subset \cup_{i=1}^n (v_i + U_{v_i} + U) \subset \cup_{i=1}^n (v_i + U_{v_i} + U_{v_i}),$$

and no terms in the last union meet $F + U$. ■

Definition D.2. Let $A \subset V$ be a convex set. Then A is said to be absorbing if for each $v \in V$ there is a $t = t_v > 0$ such that $v \in tA = \{tu : u \in A\}$. A is said to be balanced if $tA \subset A$ for all $|t| \leq 1$. The *Minkowski functional* μ_A of a convex, absorbing set A is defined by

$$\mu_A(v) = \inf\{t > 0 : t^{-1}v \in A\}, \quad v \in V.$$

Definition D.3. A topological vector space is said to be *locally convex* if every neighborhood of zero contains a convex, balanced, and absorbing open set.

Lemma 2. If $A \subset V$ is a convex, absorbing set, then μ_A is a sublinear functional.

[2] These theorems appear in Rudin (1973), pp. 55–59.

Proof. For $v \in V$, define $H_A(v) = \{t > 0 : t^{-1}v \in A\}$. Now observe that each $H_A(v)$ is a half-line whose left endpoint is $\mu_A(v)$. To see this let $t \in H_A(v)$ and $s > t$. Then, since $0 \in A$ and A is convex, it follows that $s \in H_A(v)$. Suppose that $\mu_A(u) < s$, $\mu_A(v) < t$, $r = s + t$. Then since A is convex,

$$r^{-1}(u + v) = \frac{s}{r}(s^{-1}u) + \frac{t}{r}(t^{-1}v) \in A.$$

The sublinearity of μ_A now follows. ■

Let us recall that a topological vector space is a vector space for which singletons are closed sets and addition and multiplication by scalars are continuous vector space operations. Translation invariance of the topology refers to the property that a set A is open iff all of its translates $v + A$ are open. In particular, the topology is determined by any local base.

Theorem D.2 (Separation Properties). Suppose that A and B are disjoint, nonempty, convex sets in a locally convex topological vector space V. (i) If A is open, then there is an $\ell \in V^*$ and $\gamma \in \mathbb{R}$ such that

$$\ell(u) < \gamma \leq \ell(v),$$

for all $u \in A, v \in B$. (ii) If A is compact, B closed, then there is a $\ell \in V^*$, $\gamma_1, \gamma_2 \in \mathbb{R}$, such that

$$\ell(u) < \gamma_1 < \gamma_2 < \ell(v),$$

for every $u \in A$, $v \in B$.

Proof. Consider part (i). Fix $a_0 \in A, b_0 \in B$, and let $v_0 = b_0 - a_0$. Let $G = A - B + v_0$. Then G is a convex, absorbing neighborhood of 0 in V. Let $p = \mu_G$ on V; then $p(v_0) \geq 1$ since $v_0 \notin G$, because A, B are disjoint. Define $\ell(tv_0) = t$ on the subspace $U = \{tv_0 : t \in \mathbb{R}\}$. Then ℓ is a linear map on U with $\ell \leq p$. By the Hahn–Banach theorem one may extend ℓ to V with $\ell(v) \leq p(v)$, for all $v \in V$. Thus $\ell \leq 1$ on G, and therefore $\ell \geq -1$ on $-G$. So $|\ell| \leq 1$ on the neighborhood $G \cap (-G)$ of 0. But this implies that ℓ is continuous, i.e., $\ell \in V^*$ since given $\epsilon > 0$, taking $W = \epsilon G \cap (-G)$, one has $|\ell(v)| < \epsilon$ for all $v \in W$. In view of the linearity of ℓ continuity in a neighborhood of zero implies continuity at each $v \in V$. Thus, $\ell \in V^*$. Now, if $u \in A$, $v \in B$, then

$$\ell(u) - \ell(v) + 1 = \ell(u - v + v_0) \leq p(u - v + v_0) < 1,$$

since $\ell(v_0) = 1, u - v + v_0 \in G$, and G is open. Thus $\ell(u) < \ell(v)$. It follows that $\ell(A)$, $\ell(B)$ are disjoint convex subsets of $\mathbb{R}$, with $\ell(A)$ to the left of $\ell(B)$. Since non-constant continuous linear functionals on V map open sets to open sets, $\ell(A)$ is open by hypothesis. Take γ as the right endpoint of $\ell(A)$.

For part (ii), use Lemma 1 to obtain a convex neighborhood U of 0 in V such that $(A+U)\cap B = \emptyset$. Now apply part (i) with $A+U$ in place of A, to obtain $\ell \in V^*$ such that $\ell(A+U)$ and $\ell(B)$ are disjoint convex subsets of $\mathbb{R}$, with $\ell(A+U)$ open and to the left of $\ell(B)$. The assertion follows since $\ell(A)$ is a compact subset of $\ell(A+U)$.■

Lemma 3. Let $\ell_1, \ldots, \ell_n$ and ℓ be linear functionals on a real vector space V. Let

$$N = \cap_{i=1}^{n}\{v \in V : \ell_i(v) = 0\}.$$

The following are equivalent:

(a) Thee are scalars $\alpha_1, \ldots, \alpha_n$ such that $\ell = \sum_{i=1}^{n} \alpha_i \ell_i$.
(b) There is an $r < \infty$ such that $|\ell(v)| \le r \max_{1\le i\le n} |\ell_i(v)|, \ v \in V$.
(c) $\ell(v) = 0$ for all $v \in N$.

Proof. Clearly (a) implies (b) implies (c). So it is sufficient to show that (c) implies (a). Define $\pi : V \to \mathbb{R}^n$ by

$$\pi(v) = (\ell_1(v), \ldots, \ell_n(v)), \quad v \in V.$$

Then ℓ is constant on $\{v : \pi(v) = c\}, c \in \mathbb{R}^n$, implying that $\ell(v) = \ell_n \circ \pi$ for a linear functional ℓ_n on $\mathbb{R}^n$. Thus,

$$\ell_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} \alpha_i x_i, \ (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

for some $\alpha_1, \ldots, \alpha_n$. But this proves (a) since

$$\ell(v) = \ell_n(\pi(v)) = \sum_{i=1}^{n} \alpha_i \alpha_i(v), \ v \in V.$$

■

Theorem D.3. Suppose that V is a vector space and V' is a vector space of linear functionals on V that separate points of V. Then V, equipped with the weakest topology on V that makes each $\ell \in V'$ continuous, is a locally convex space whose dual is V'.

Proof. The linearity of $\ell \in V'$ shows that the topology is translation invariant. If $\ell_1, \ldots, \ell_n \in V'$, then for positive numbers $r_1, \ldots, r_n$

$$U = \{v : \ell_i(v) < r_i, i = 1, \ldots, n\} \tag{D.7}$$

is a convex, balanced, absorbing, and open set for the topology. Since V' separates points, the topology is Hausdorff. In fact the collection of all open sets U of this form provide a local base for the topology, making it a locally convex topology. Since $\frac{1}{2}U + \frac{1}{2}U = U$, addition is continuous. To see that multiplication by scalars is continuous, let $v \in V$ and $\alpha \in \mathbb{R}$. Then $v \in sU$ for some $s > 0$. If $|\beta - \alpha| < r$ and $u - v \in rU$, then, for suitably small r that $r(s + r) + |\alpha|r < 1$, one has

$$\beta u - \alpha v = (\beta - \alpha)u + \alpha(u - v) \in U.$$

Thus, multiplication by scalars is continuous. Finally, to see that V' is the dual space it is sufficient to show that if ℓ is a continuous linear functional for this topology, then $\ell \in V'$, since, by hypothesis, each $\ell \in V'$ is continuous. If ℓ is a continuous linear functional, then $|\ell(u)| < 1$ for all u in some set U of the form (D.7). Thus, by Lemma 3, there are scalars α_i such that $\ell = \sum_i \alpha_i \ell_i$, $\ell_i \in V'$. Since V' is a vector space it follows that $\ell \in V'$. ■

Note that if V is a topological vector space, then each $v \in V$ may be viewed as a linear functional on its dual space V^* by defining $v(\ell) = \ell(v)$, $\ell \in V^*$. So viewed, V clearly separates points. The weakest topology on V^* that makes each $v \in V$ continuous as a linear functional on V^* is called the *weak* topology*. From this perspective one has the following representation as a corollary to Theorem D.3 by simply replacing V by V^* and V' by V there.

Corollary D.1 (Dual Representation Theorem). Let V be a topological vector space with dual V^*, and give V^* the weak* topology. Then V^* is a locally convex topological vector space and every weak* continuous linear functional ℓ^* on the dual space V^* has the form $\ell^*(\ell) = \ell(v)$, $\ell \in V^*$, for some $v \in V$.

References

Abrahamson IG (1965) On the stochastic comparison of tests of statistics, PhD Dissertation, University of Chicago

Adler RJ, Taylor JE (2007) Random fields and geometry. Springer, New York

Aldous D, Diaconis P (1986) Shuffling cards and stopping times. Am Math Monthly 93:333–348

Aizenmann M, Kesten H, Newman CM (1987) Uniqueness of the infinite cluster and continuity of connectivity functions for short and long-range percolation. Commun Math Phys 111:505–532

Arikan E (1996) An inequality on guessing and its application to sequential decoding. IEEE Trans Inf Theory 42:99–105

Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404

Assimakis N, Adam M (2013) Kalman filter Riccati equation for the prediction, estimation, and smoothing error covariance matrices. Comput Math 203:7p

Athreya KB, Ney P (1978) A new approach to the limit theory of recurrent Markov chains. Trans Am Math Soc 245:493–501

Athreya KB, Dai J (2000) Random logistic maps I. J Theor Probab 13(2):595–608

Auffinger A, Damron M, Hanson J (2017) 50 years of first-passage percolation, vol 68. Am Math Soc

Bahadur R (1960) Stochastic comparison of tests. Ann Math Statist **31**:276–295

Bahadur R (1971) Some limit theorems in statistics. SIAM, Philadelphia

Becher V, Figueira S (2002) An example of a computable absolutely normal number. Theor Comput Sci 270:947–958

Berestycki N, Powell E (2021) Gaussian free field, Liouville quantum gravity and Gaussian multiplicative chaos. Open Math Notes. Amer. Math. Soc. https://www.ams.org/open-math-notes/omn-view-listing?listingId=111291

Van Den Berg J, Kesten H (1985) Inequalities with applications to percolation and reliability. J Appl Probab 22:556–569

Bhattacharya RN (1982) On the functional central limit theorem and the law of the iterated logarithm for Markov processes. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 60(2):185–201

Bhattacharya RN, Waymire E (1990) An extension of the classical method of images for the construction of reflecting diffusions. In: Bahadur RR, Gosh JK, Sen PK (eds) Proc. R.C. Bose Symp. on Probab., Math. Statist. and Design of Experiments. Wiley Eastern, New Delhi, pp 157–164

Bhattacharya R, Lee C (1995) On geometric ergodicity of nonlinear autoregressive models. Stat Probab Lett 22(4):311–315

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

Bhattacharya R, Lee C (1999) On geometric ergodicity of nonlinear autoregressive models. Stat Probab Lett 4(41):439–440

Bhattacharya R, Lee O (1988) Asymptotics of a class of Markov processes which are not in general irreducible. Ann Probab **16**(3):1333–1347

Bhattacharya R, Lee O (1997) Correction: asymptotics of a class of Markov processes which are not in general irreducible [Ann. Probab. **16**(3):1333–1347] Ann Probab 25(3):1541–1543

Bhattacharya R, Lin L, Patrangenaru V (2016) A course in mathematical statistics and large sample theory. Springer series in statistics. Springer, New York

Bhattacharya R, Majumdar M (2004) Stability in distribution of randomly per-turbed quadratic maps as Markov processes. Ann Appl Probab, pp 1802–1809

Bhattacharya R, Majumdar M (2007) Random dynamical systems: theory and applications. Cambridge University Press, Cambridge

Bhattacharya R, Majumdar M (2010a) Random iterates of monotone maps. Rev Econ Res 14:185–192

Bhattacharya R, Majumdar M (2010b) Limit theorems for monotone Markov processes. Sankhya 72A:170–190

Bhattacharya R, Waymire E (2016) A basic course in probability. Springer, New York. (ERRATA: https://sites.science.oregonstate.edu/~waymire/)

Bhattacharya R, Waymire E (2002) An approach to the existence of unique invariant probabilities for Markov processes. In: Limit theorems in probability and statistics, vol I. János Bolyai Math. Soc., pp 181–200

Bhattacharya R, Waymire E (1990, 2009) Stochastic processes with applications. Wiley, New York; Reprinted in SIAM Classics in Applied Mathematics 61

Bhattacharya R, Waymire E (2021) Random walk, Brownian motion, and martingales. Graduate text in mathematics. Springer, New York

Billingsley P (1968) Convergence of probability measures. Wiley, New York

Billingsley P (1986) Probability and measure, 2nd edn. Wiley, NY

Birkhoff GD (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17(12):656–660

Blachère S (2003) Cut times for random walks on the discrete Heisenberg group. Annales de l'IHP Probabilités et Statistiques 39(4):621–638

Bollabás B, Riordan O (2006) A short proof of the Harris-Kesten theorem. Bull Lond Math Soc 38(3):470–484

Borwein JM, Zhuang D (1986) On Fan's minimax theorem. Math Prog 34:232–234

Bradley RC (2003) Introduction to strong mixing conditions, vols 1–3. Indiana University, Bloomington

Broadbent SR, Hammersley JM (1957) Percolation processes. Math Proc Camb Philos Soc 53:629–641

Brockwell P, Davis R (1991) Time series: theory and methods. Springer series in statistics. Springer, New York

Bronski JC (2003) Asymptotics of Karhunen–Loéve eigenvalues and tight constants for probability distributions of passive scalar transport. Commun Math Phys 238:563–582

Brown BM (1971) Martingale central limit theorems. Ann Math Statist 42(1):59–66

Bubley R, Dyer M (1997) Path coupling: a technique for proving rapid mixing in Markov chains. In: Proc. 38th annual IEEE symposium on foundations of computer science, pp 223–231

Budhiraja A (2003) Asymptotic stability, ergodicity and other asymptotic properties of the nonlinear filter. Annales de l'IHP Probabilités et Statistiques 39(6):919–941

Burton RM, Keane M (1989) Density and uniqueness in percolation. Commun Math Phys

Burton RM, Waymire E (1985) Scaling limits for associated random measures. Ann Probab 13(4):1267–1278

Burton RM, Waymire E (1986) A sufficient condition for association of a renewal process. Ann Probab 14(4):1272–1276

Cannings C (1973) The equivalence of some overlapping and non-overlapping generation models for the study of genetic drift. J Appl Probab 10(2):432–436

Cannings C (1974) The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models, Advances in Applied Probability 6(2):260–290

Chen MF (1996) Estimation of spectral gap for Markov chains. Acta Math Sin New Ser 12(4):337–360

Chen MF, Wang FY (1995) Estimation of the first eigenvalue of second order elliptic operators. J Funct Anal 131(2):345–363

Chen MF (1997) Coupling, spectral gap and related topics (I-III). Chin Sci Bull 42(16):1321–1327; 42(17):1409–1416; 42(18):1497–1505

Chan KS, Tong H (1985) On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. Adv Appl Probab 17(3):666–678

Chakroborty S, Rao BV (1998) Completeness of the Bhattacharya metric on the space of probabilities. Statist Probab Lett 36:321–326

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Statist 23:493–507

Chib S, Greenberg E (1995) Understanding the . Am Statist 49(4):327–335

Choquet G, Deny J (1960) Sur lequation de convolution $\mu = \mu^{*}\sigma$. Compt Ren Heb de l'Acad des Sci 250(5):799–801

Cogburn R (1960) Asymptotic properties of stationary sequences. Univ Calif Publ Statist 3:99–146

Collet P, Eckmann L-P (1980) Iterated maps on the interval as dynamic systems. In: Jaffe A, Ruelle D (eds). Birkhäuser, Boston

Comtet A, Texier C, Touriguy Y (2013) Lyapunov exponents, one-dimensional Anderson localization and products of random matrices. J Phys A: Math Theor 46:254003

Cover JA, Thomas TM (2006) Elements of information theory. Wiley series in telecommunications and signal processing, 2nd edn

Dascaliuc R, Pham T, Thomann E, Waymire E (2022a) Doubly stochastic yule cascades (part I): the explosion problem in the time-reversible case. J Funct Anal. in press

Dascaliuc R, Pham T, Thomann E, Waymire E (2022b) Doubly stochastic yule cascades (part II): the explosion problem in the non-reversible case, Annales de l'institut Henri Poincaré (B). Probabilités et Statistiques. in press

Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd edn. Springer, New York

Denker M (1986) Uniform integrability and uniform integrability for strongly mixing processes. In: Eberlein E, Taqqu MS (eds) Dependence in probability and statistics. Birkhäuser, Boston

Den Hollander F (2008) Large deviations, vol 14 American Mathematical Society, Providence

Derrida B, Flyvbjerg B (1987) The random map model: a disordered model with deterministic dynamics. J Physique **48**(6):971–978

Devaney R (1989) An introduction to chaotic dynamical systems, 2nd edn. Addison-Wesley, Reading

Diaconis P (1988) Group representations in probability and statistics. IMS lecture notes-monograph series, vol 11, pp 1–192

Diaconis P, Freedman D (1980) de Finetti's theorem for Markov chains. Ann Probab 115–130

Diaconis P (1996) The cutoff phenomenon in finite Markov chains. Proc Nat Acad Sci 93(4):1659–1664

Dietz, Z., Sethuraman, S. (2005): Large deviations for a class of nonhomogeneous Markov chains, *Ann. Appld. Probab.*, **15**(1A), 42–486.

Dieudonné J (1960) Foundations of modern analysis. Academic Press, New York

Doeblin W (1938) Exposé de la théorie des chaınes simples constantes de Markova un nombre fini d'états. Mathématique de l'Union Interbalkanique 2:77–105, 78–80

Donsker MD, Varadhan SS (1975a) Asymptotic evaluation of certain Markov process expectations for large time, I. Commun Pure Appl Math 28(1):1–47

Donsker MD, Varadhan SRS (1975b) Asymptotic evaluation of certain Markov process expectations for large time-II. Commun Pure Appl Math 28(2):279–301

Donsker MD, Varadhan SS (1976) Asymptotic evaluation of certain Markov process expectations for large time-III. Commun Pure Appl Math 29(4):389–461

Donsker MD, Varadhan SS (1983) Asymptotic evaluation of certain Markov process expectations for large time-IV. Commun Pure Appl Math 36(2):183–212

Doyle P, Snell L (1984) Random walks and electrical networks. Carus mathematical Monographs. The Math. Assoc. America, Washington, DC

Dubins LE, Freedman DA (1966) Invariant probabilities for certain Markov processes. Ann Math Statist 37(4):837–848

Durrett R (1991) Probability theory and examples, 2nd edn. Wadsworth, Brooks & Cole, Pacific, Grove

Dvoretzky A, Erdos P (1951) Some problems on random walk in space. Proc. 2nd Berkeley Symp. on Math. Stat. and Probab. University of California Press, California, pp 353–367

Ellis RS (1985) Entropy, large deviations, and statistical mechanics. Springer, New York

Erdos P (1937) Some problems and results in elementary number theory. Proc Camb Phil Soc 33:6–12

Esary JD, Proschan F, Walkup DW (1967) Association of random variables, with applications. Ann Math Stat 38(5):1466–1474

Evans S (1989) Association and random measures. Probab Theory Rel Fields 86:1–19

Fan K (1953) Minimax theorems. Nat Acad Sci 53:42–47

Fekete M (1923) Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. Mathematische Zeitschrift 17(1):228–249

Feller W (1951) Diffusion processes in genetics. University of California Press, Berkeley, pp 227–246

Folland G (1984) Real analysis. Wiley, New York

Fortuin C, Kasteleyn P, Ginibre J (1971) Correlation inequalities on some partially ordered sets. Commun Math Phys 22:89–103

Furstenberg H, Kesten H (1960) Products of random matrices. Ann Math Statist 31:451–469

Gantert N, Ramanan K, Rembart F (2014) Large deviations for weighted sums of stretched exponential random variables. Electron Comm Probab 19

Garcia AM (1965) A Simple Proof of E. Hopf's Maximal Ergodic Theorem. J Math Mech 14:381–382

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman and Hall/CRC, London

Gihman II, Skokohod AV (1974) The theory of stochastic processes I. Springer, New York. English translation by S. Kotz of the original Russian published in 1971 by Nauka, Moscow

Glaffig C, Waymire E (1987) Infinite divisibility of a Bethe lattice Ising model. J Statist Phys 47(1):185–192

Glavaski S, Marsden JE, Murray RM (1998) Model reduction, centering, and the Karhunen–Loève expansion. In: IEEE conference on decision and control, and references therein

Gordin MI, Lifsic D (1978) The central limit theorem for stationary, ergodic Markov processes. Doklady 19:392–394

Graczyk J, Swiatek G (1997) Generic hyperbolicity in the logistic family. Ann Math 146:1–52

Grenander U (1981) Abstract inference. Wiley, New York

Grimmett G (1999) Percolation, 2nd edn. Springer, Berlin

Halmos PR (2017) Finite-dimensional vector spaces. Courier Dover Publications. Reprinted from 1958, D. Van Nostrand Co., Princeton

Hanawal MK, Sundaresan R (2011) Guessing revisited: a large deviations approach. IEEE Trans Inform Theory 57(1):70–78

Haggstrom O, Nelander K (1998) Exact sampling from anti-monotone systems. Statistica Neerlandica52:360

Hammersley JM, Welsh DJA (1965) First passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In: Bernoulli–Bayes–Laplace Anniversary Volume. Springer, Berlin, pp 61–110

Harris TE (1956) The existence of stationary measures for certain Markov processes. In: Berkeley Symp. on Math. Statist. and Prob. Proc. Third Berkeley Symp. on Math. Statist. and Prob., vol 2 Univ. of Calif. Press, pp 113–124

Harris TE (1960) A lower bound for the critical probability in a certain percolation process. Proc Camb Philos Soc 59:13–20

Hoeffding W (1940) Masstabinvariante Korrelationstheorie. Schriften des Mathematischen Instituts und des Instituts f ur Angewandte Mathematik der Universit at Berlin 5:181–233

den Hollander WTF, Keane M (1986) Inequalities of FKG type. Phys A: Stat Mech Appl 138(1–2):167–182

Holley R (1974) Remarks on the FKG inequalities. Commun Math Phys 36(3):227–231

Handel van R (2013) Stochastic analysis seminar. Princeton U., Quentin Berthet (scribe), https://blogs.princeton.edu/sas/2013/10/10/lecture-3-sanovs-theorem

Hurewicz W (1958) Lectures on ordinary differential equations. Cambridge Tech Press of the Mass. Inst. of Tech

Ibragimov IA, Linnik Yu V (1971) Independent and stationary sequences of random variables. Wolters-Noordhoff, Groningen

Ikeda N, Watanabe S (1981) Stochastic differential equations and diffusion processes. North-Holland, Kodansha

James N, Peres Y (1997) Cutpoints and exchangeable events for random walks. Thry Probab Appl 41(4):666–677

James N, Lyons R, Peres Y (2007) A transient Markov chain with finitely many cutpoints, arXiv:0706.2013

Jeffries C, Perez J (1982) Observation of a Pomeau-Manneville intermittent route to chaos in a nonlinear oscillator. Phys Rev A 26(4):2117

Kalman RE (1960) A new approach to linear filtering and prediction problem. Trans ASME Ser D: J Basic Eng 82:3445

Karhunen K (1946) Uber lineare methoden in der wahrscheinlichkeitsrechnung. Annales Academiae Scientarum Fennicae 37:3–79

Kesten H (1980) The critical probability of bond percolation on the square lattice equals $1/2$. Commun Math Phys 74:41–59

Key E (1987) Computable examples of the maximal Lyapunov exponent. Probab Theory Rel. Fields 75(1):97–107

Kingman JFC (1976) Subadditive ergodic theory. Ann Probab 883–909

Kolmogorov A (1939) Sur l'interpolation et extrapolation des suites stationnaires. CR Acad Sci Paris 208:2043–2045

Kolmogorov A (1941) Stationary sequences in Hilbert space (Russian). Bull Math Univ Moscow 2(6):40pp

Kolmogorov AN (1941) Interpolation and extrapolation of stationary sequences. Izvestiya the Academy of Sciences of the USSR Ser. Math., No. 5, 3–14

Koperberg VT (2016) On the equivalence of Strassen's theorem and some combinatorial theorems. Undergraduate Thesis. Mathematisch Instituut, Universiteit Leiden

Kovchegov Y, Otto P (2018) Path coupling and aggregate path coupling. Springer Briefs in Probab. and Math. Stat. Springer, New York

Krein MGE, Rutman MA (1948) Linear operators leaving invariant a cone in a Banach space. Uspekhi Matematicheskikh Nauk 3(1):3–95

Lacey T (2020) http://web.mit.edu/kirtley/kirtley/binlustuff/literature/control/Kalman20filter.pdf

Last G, Szekli R, Yogeshwaran D (2020) Some remarks on associated random fields, random measures and point processes. ALEA Lat Am J Probab Math Stat 17:355–374

Lawler GF (2013) Intersections of random walks. Springer, Berlin

Lehmann E (1966) Some concepts of dependence. Ann Math Stat 37:1137–115

Lévy P (1953) Random functions: general theory with special reference to Laplacian random functions (Vol. 1, No. 12). University of California Press

Liggett TM (1985) Interacting particle systems. Springer, New York

Liggett TM (1985) An improved subadditive ergodic theorem. Ann Probab 13(5):1279–1285

Liggett TM (1983) Attractive nearest particle systems. Ann Probab 11:16–33

Lindvall T (1999) On Strassen's theorem on stochastic domination. Electron Commun Probab 4:51–59

Lindvall T (2002) Lectures on the coupling method. Courier Corporation

Loève M (1948) Fonctions aleatoires du second ordre. In: Lévy P (ed) Processus Stochastiques et Mouvement Brownien. Gauther-Villars, pp 299–352

Lorenz EN (1963) Deterministic non-periodic flow. J Atmos Sci 20:130–141

Lund RB, Tweedie RL (1996) Geometric convergence rates for stochastically ordered Markov chains. Math Oper Res 21:182–196

Malone D, Sullivan WG (2004) Guesswork and entropy. IEEE Trans Inf Theory 50(4):525–526

May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467

Meyn S, Tweedie RL (1993) Markov chains and stochastic stability. Cambridge Univ. Press, Cambridge

Mirman LJ (1980) One sector economic growth and uncertainty: a survey. Stoch Program 537–567

Nagaev AV (1969) Integral limit theorems for large deviations when Cramér's condition is not fulfilled. Theory Probab Appl 14(1):51–64

Newman CM (1980) Normal fluctuations and the FKG inequalities. Commun Math Phys 74:119–128

Newman CM, Wright AL (1981) An invariance principle for certain dependent sequences. Ann Prob 9(4):671–675

Neveu J (1971) Mathematical foundations of the calculus of probability. Holden-Day, San Francisco

Noda A (1987) Generalized Radon transform and Lévy's Brownian motion. Nagoya Math J 105:71–87

Nummelin E (1978) Splitting technique for Harris recurrent Markov chains. Z Wahrs Verw Geb 43:309–318

Nummelin E, Tuominen P (1983) The rate of convergence in Orey's theorem for Harris recurrent Markov chains with applications to renewal theory. Stoch Proc Appl 15:295–311

Orey S (1971) Limit theorems for Markov chain transition probabilities. Math Studies, vol 34. Van Nostrand Reinhold, London

Peckham SD, Waymire EC, De Leenheer P (2018) Critical thresholds for eventual extinction in randomly disturbed population growth models. J Math Biol 77(2):495–525

Pemantle R (2000) Towards a theory of negative dependence. J Math Phys 41(3):1371–1390

Peres Y, Shlag W, Solomyak B (1999) Absolute continuity of Bernoulli convolutions, a simple proof. Math Res Lett 3:231–239

Pitt L (1982) Positively correlated normal variables are associated. Ann Probab 10(2):496–499

Plackett RL (1954) A reduction formula for normal multivariate integrals. Biometrika 41:351–360

Propp J, Wilson D (1998) Coupling from the past: a user's guide. Microsurveys in discrete probability (Princeton, NJ, 1997), DIMACS Ser.Discrete Math. Theoret. Comput. Sci., vol 41. American Mathematical Society, Providence, pp 181–192

Rachev ST (1991) Probability metrics and stability of stochastic models. Wiley, New York

Ramasubramanian S (2009) Lectures on insurance models. American Mathematical Society, Providence

Rezakhanlou F (2017) Lectures on the large deviation principle. https://math.berkeley.edu/~rezakhan/LD.pdf

Roberts GO, Rosenthal JS (2004) General state space Markov chain and MCMC algorithms. Probab Surv 1:20–71

Roch S (2020) textitModern discrete probability: an essential toolkit. https://www.math.wisc.edu/~roch/mdp/roch-mdp-toc.pdf

Rogers LCG (1997) Arbitrage with fractional Brownian motion. Math Financ 7(1):95–105

Rolski T, Schmidli H, Schimidt V, Teugels JL (2010) Stochastic processes for insurance & finance. Wiley, New York

Rosén B (1967) On the central limit theorem for sums of dependent random variables. Wahr und Verw Gebiete 7:48–52

Rosenblatt M (1956) A central limit theorem and a strong mixing condition. Proc Natl Acad Sci USA 42:43–47

Rosenthal JS (2002) Quantitative convergence rates of Markov chains: a simple account. Elec Comm Prob 7(13):123–128

Rudin W (1974) Real and complex analysis. McGraw-Hill, New York

Rudin W (1973) Functional analysis, 2nd ed. McGraw-Hill, New York

Samorodnitsky G, Taqqu M (1994) Stable non-gaussian random processes: stochastic models with infinite variance. Chapman and Hall, New York

Sanov IN (1957) On the probability of large deviations of random variables. Mat Sbornik 42(84):11–44

Seppalainen T (1994) Large deviations for Markov chains with random transitions. Ann Probab 22(2):713–748

Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York

Seymour PD, Welsh DJA (1975) Combinatorial applications of an inequality from statistical mechanics. Math Proc Camb Philos Soc 77:485–497

Shannon C (1948) A mathematical theory of communication. Bell Labs Tech J 27(3):379–423

Sheffield S (2007) Gaussian free fields for mathematicians. Prob Theory Rel Fields 139:521–541

Sierpinski MW (1917) Demonstration élémentaire du théoreme de M. Borel sur les nombres absolument normaux et détermination effective d'un tel nombre. Bull Soc Math France 45:127–132

Solomyak B (1995) On the random series $\sum \pm\lambda^n$ (an Erdos problem). Ann Math 142:611–625

Spitzer F (1976) Principles of random walk, 2nd edn. Springer, New York

Strassen V (1965) The existence of probability measures with given marginals. Ann Math Statist 36:423–439

Szegö G (1920) Theorie der Toeplitzschen Formen. Math Z 6:167–202

Ulam S, von Neumann J (1947) On Combination of Stochastic and Deterministic Processes. Bull Am Math Soc 53(11):1120–1127

Varadhan SRS (1984) Large deviations and applications. SIAM, Philadelphia

Varadhan SRS (2008) Large deviations, Ann Probab 36(2):397–419

Waymire E (1984) Infinitely divisible Gibbs states. Rocky Mount J Math 665–678

Wiener N (1923) Differential space. J Math Phys 2:131–174

Wiener N (1949) Extrapolation, interpolation and smoothing of stationary time series, with engineering applications. Cambridge, New York (Reprinted from a publication issued with restricted circulation in 1942)

Wold H (1938) A study in the analysis of stationary time series, Dissertation, (Stockholm) Uppsala

Yu L, Ott I, Chen Q (1990) Transition to chaos for random dynamical systems. Phys Rev Lett 65:2935–2938

Related Textbooks and Monographs

The following is list of some supplementary and/or follow-up reading, including the books cited in the text.

Adler RJ, Taylor JE (2007) Random fields and geometry. Springer, New York

Aldous D (1989) Probability approximations via the poisson clumping heuristic. Springer, New York

Asmussen S, Hering H (1983) Branching processes. Birkhäuser, Boston

Athreya KB, Lahiri SN (2006) Measure theory and probability theory. Springer texts in Statistics. Springer, New York

Athreya KB, Ney PE (1972) Branching processes. Springer, New York

Bahadur R (1971) Some limit theorems in statistics. SIAM, Philadelphia

Bass R (1995) Probabilistic Techniques in Analysis. Springer, New York

Bauer H (1972) Probability theory and elements of measure theory. English transl., Holt-Rinehart-Winston, New York

Bhattacharya R, Lin L, Patrangenaru V (2016) A course in mathematical statistics and large sample theory. Springer Series in Statistics. Springer, New York

Bhattacharya R, Waymire E (2016) A basic course in probability. Springer, New York (ERRATA: https://sites.science.oregonstate.edu/waymire/)

Bhattacharya R, Majumdar M (2007) Random dynamical systems: theory and applications. Cambridge University Press, Cambridge

Bhattacharya R, Waymire E (1990, 2009) Stochastic processes with applications. Wiley, New York; Reprinted in SIAM Classics in Applied Mathematics, vol 61

Bhattacharya R, Waymire E (2021) Random walk, Brownian motion, and martingales. Graduate text in mathematics. Springer, New York

Billingsley P (1968) Convergence of probability measures. Wiley, New York

Billingsley P (1986) Probability and measure, 2nd edn. Wiley, New York

Bingham NH, Goldie CM, Teugels JL (1987) Regular variation. Encyclopedia of mathematics and its applications. Cambridge University Press, Cambridge

Bradley RC (2003) Introduction to strong mixing conditions, vol 1–3. Indiana University, Bloomington

Breiman L (1968) Probability. Addison Wesley, Reading, MA. Reprint SIAM, Philadelphia

Brockwell P, Davis R (1991) Time series: theory and methods. Springer series in statistics. Springer New York

Chung KL (1974) A course in probability theory, 2nd edn. Academic Press, New York

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

Collet P, Eckmann L-P (1980) Iterated maps on the interval as dynamic systems. In: Jaffe A, Ruelle D. Birkhäuser, Boston

Cover JA, Thomas TM (2006) Elements of information theory. Wiley series in telecommunications and signal processing, 2nd edn.

Davis M, Etheridge A (2006) Louis Bachelier's theory of speculation: the origins of modern finance. Princeton University Press, Princeton

Den Hollander F (2008) Large deviations, vol 14. American Mathematical Society, Providence

Devaney R (1989) An introduction to chaotic dynamical systems 2nd edn. Addison-Wesley, Reading

Dieudonné J (1960) Foundations of modern analysis. Academic Press, New York

Durrett R (1984) Brownian motion and martingales in analysis. Wadsworth, Belmont

Durrett R (1995) Probability theory and examples, 2nd edn. Wadsworth, Brooks & Cole, Pacific, Grove

Durrett R (2008) Probability models for DNA sequence evolution, 2nd edn. Springer, New York

Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd edn. Springer, New York

Deuschel JD, Stroock DW (1989) Large deviations. Academic Press, Boston

Diaconis P (1988) Group representations in probability and statistics. IMS lecture notes-monograph series, vol 11 pp i–192

Doob JL (1953) Stochastic processes. Wiley, New York

Doyle P, Snell L (1984) Random walks and electrical networks. Carus mathematical monographs. The Math. Assoc. America, Washington, DC

Dym H, McKean HP (1972) Fourier series and integrals. Academic Press, New York

Ellis RS (1985)Entropy, large deviations, and statistical mechanics. Springer, New York

Ethier SN, Kurtz TG (1985) Markov processes: characterization and convergence. Wiley, New York

Faris WG ed (2014) Diffusion, quantum theory, and radically elementary mathematics, vol. 47. Princeton University Press, Princeton

Feller W (1968, 1971) An introduction to probability theory and its applications, vol 1, 3rd edn., vol 2, 2nd edn. Wiley, New York

Folland G (1984) Real analysis. Wiley, New York

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman and Hall/CRC, London

Gihman II, Skokohod AV (1974) The theory of stochastic processes I. Springer, New York. English translation by S. Kotz of the original Russian published in 1971 by Nauka, Moscow

Grenander U (1981) Abstract inference. Wiley, New York

Grimmett G (1999) Percolation, 2nd edn. Springer, Berlin

Hall P, Heyde CC (1980) Martingale limit theory and its application. Academic Press, New York

Halmos PR (2017) Finite-dimensional vector spaces. Courier Dover Publications. Reprinted from 1958, D. Van Nostrand Co. Inc. Princeton, NJ

Harris TE (1963) The theory of branching processes. Springer, Berlin

Herglotz G (1911) Uber potenzreihen mit positivem, reelen teil im einheitskreis. Ber Verhandl Sachs Akad Wiss Leipzig Math-Phys Kl 63:501–511

Hurewicz W (1958) Lectures on ordinary differential equations. CambridgeTech Press of the Mass. Inst. of Tech

Ibragimov IA, Linnik Yu V (1971) Independent and stationary sequences of random variables. Wolters-Noordhoff, Groningen

Ikeda N, Watanabe S (1981) Stochastic differential equations and diffusion processes. North-Holland, Kodansha

Jacod J, Protter P (2003) Probability essentials, 2nd edn. Springer universitext series. Springer, New York

Jagers P (1975) Branching processes with applications to biology. Wiley, New York

Kac M (1979) In: Baclawski K, Donsker MD (eds). Probability, number theory and statistical physics: selected papers. MIT Press, Cambridge

Kallenberg O (2001) Foundations of modern probability, 2nd edn. Springer, New York
Kallenberg O (2002) Foundations of modern probability, 2nd edn. Springer, NY
Karlin S, Taylor HM (1975) A first course in stochastic processes. Academic Press, New York
Karlin S, Taylor HM (1981) A second course in stochastic processes. Academic Press, New York
Laplace P-S (1878–1912) Théorie Analytique des Probabilités. Reprinted in Oeuvres Complète de Laplace, vol. 7. Gauthier-Villars, Paris
Lawler GF (2005) Conformally invariant processes in the plane. American Mathematical Society, Providence
Lawler GF (2013) Intersections of random walks. Springer, Berlin
Lévy P (1925) Calcul des Probabilités. Gauthier–Villars, Paris
Lévy P (1954) Théorie de l'addition des variables aléatories, 2nd edn. Gauthier–Villars, Paris (1st ed. 1937)
Liggett TM (1985) Interacting particle systems. Springer, New York
Lindvall T (2002) Lectures on the coupling method. Courier Corporation
Meyn S, Tweedie RL (1993) Markov chains and stochastic stability. Cambridge Univ. Press, Cambridge
Neveu J (1971) Mathematical foundations of the calculus of probability. Holden-Day, San Francisco
Neveu J (1975) Discrete parameter martingales. North-Holland, Amsterdam
Orey S (1971) Limit theorems for Markov chain transition probabilities. Math. studies, vol 34. Van Nostrand Reinhold, London
Parthasarathy KR (1967) Probability measures on metric spaces. Academic Press, New York
Pitman J (1993) Probability. Springer, New York
Pollard D (2002) A user's guide to measure theoretic probability. Cambridge University Press, Cambridge
Rachev ST (1991) Probability metrics and stability of stochastic models. Wiley, New York
Ramasubramanian S (2009) Lectures on insurance models. American Mathematical Society, Providence
Resnick S (1987) Extreme values, regular variation, and point processes. Springer, New York
Revuz D, Yor M (1999) Continuous martingales and Brownian motion, 3rd edn. Springer, Berlin
Rogers LCG, Williams D (2000) Diffusions, Markov processes and martingales, vol 1, 2nd edn. vol. 2. Cambridge
Rolski T, Schmidli H, Schimidt V, Teugels JL (2010) Stochastic processes for insurance & finance. Wiley, New York
Royden HL (1988) Real analysis, 3rd edn. MacMillan, New York
Rudin W (1974) Real and complex analysis. McGraw-Hill, New York
Rudin W (1991) Functional analysis, vol 45, 46. McGraw-Hill, New York
Samuelson P (1947) Foundations of economic analysis. Harvard University Press, MA
Spitzer F (1976) Principles of random walk, 2nd edn. Springer, New York
Samorodnitsky G, Taqqu M (1994)Stable non-Gaussian random processes: stochastic models with infinite variance. Chapman and Hall, New York
Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York
Shannon C (1948) A mathematical theory of communication. Bell Labs Tech J 27(3):379–423
Spitzer F (1976) Principles of random walk, 2nd edn. Springer, New York
Varadhan SRS (1984) Large deviations and applications. SIAM, Philadelphia
Wiener N (1949) Extrapolation, interpolation and smoothing of stationary time series, with engineering applications. Cambridge, New York (Reprinted from a publication issued with restricted circulation in 1942)
Williams D (1991) Probability with martingales. Cambridge Univ. Press, Cambridge

Author Index

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3

Subject Index

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3