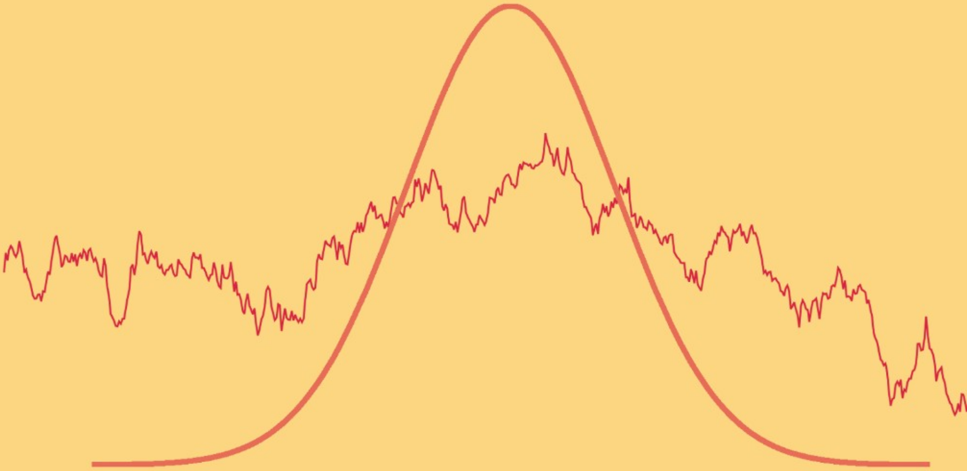
Universitext

Rabi Bhattacharya
Edward C. Waymire

A Basic Course in Probability Theory



Springer

Universitext

Universitext

Editors (North America): S. Axler and K.A. Ribet

Aguilar/Gitler/Prieto: Algebraic Topology from a Homotopical Viewpoint
Aksoy/Khamsi: Nonstandard Methods in Fixed Point Theory
Andersson: Topics in Complex Analysis
Aupetit: A Primer on Spectral Theory
Bachman/Narici/Beckenstein: Fourier and Wavelet Analysis
Badescu: Algebraic Surfaces
Balakrishnan/Ranganathan: A Textbook of Graph Theory
Balser: Formal Power Series and Linear Systems of Meromorphic Ordinary Differential Equations
Bapat: Linear Algebra and Linear Models (2nd ed.)
Berberian: Fundamentals of Real Analysis
Bhattacharya/Waymire: A Basic Course in Probability Theory
Blyth: Lattices and Ordered Algebraic Structures
Boltyanskii/Efremovich: Intuitive Combinatorial Topology (Shenitzer, trans.)
Booss/Bleecker: Topology and Analysis
Borkar: Probability Theory: An Advanced Course
Böttcher/Silbermann: Introduction to Large Truncated Toeplitz Matrices
Carleson/Gamelin: Complex Dynamics
Cecil: Lie Sphere Geometry: With Applications to Submanifolds
Chae: Lebesgue Integration (2nd ed.)
Charlap: Bieberbach Groups and Flat Manifolds
Chern: Complex Manifolds Without Potential Theory
Cohn: A Classical Invitation to Algebraic Numbers and Class Fields
Curtis: Abstract Linear Algebra
Curtis: Matrix Groups
Debarre: Higher-Dimensional Algebraic Geometry
Deitmar: A First Course in Harmonic Analysis (2nd ed.)
DiBenedetto: Degenerate Parabolic Equations
Dimca: Singularities and Topology of Hypersurfaces
Edwards: A Formal Background to Mathematics I a/b
Edwards: A Formal Background to Mathematics II a/b
Engel/Nagel: A Short Course on Operator Semigroups
Farenick: Algebras of Linear Transformations
Foulds: Graph Theory Applications
Friedman: Algebraic Surfaces and Holomorphic Vector Bundles
Fuhrmann: A Polynomial Approach to Linear Algebra
Gardiner: A First Course in Group Theory
Gårding/Tambour: Algebra for Computer Science
Goldblatt: Orthogonality and Spacetime Geometry
Gustafson/Rao: Numerical Range: The Field of Values of Linear Operators and Matrices
Hahn: Quadratic Algebras, Clifford Algebras, and Arithmetic Witt Groups
Heinonen: Lectures on Analysis on Metric Spaces
Holmgren: A First Course in Discrete Dynamical Systems
Howe/Tan: Non-Abelian Harmonic Analysis: Applications of $SL(2, \mathrm{R})$
Howes: Modern Analysis and Topology
Hsieh/Sibuya: Basic Theory of Ordinary Differential Equations
Humi/Miller: Second Course in Ordinary Differential Equations
Hurwitz/Kritikos: Lectures on Number Theory

(continued after index)

Rabi Bhattacharya
Edward C. Waymire

A Basic Course in Probability Theory

Springer

Rabi Bhattacharya
Department of Mathematics
University of Arizona
Tucson, AZ 85750
USA
rabi@math.arizona.edu

Edward C. Waymire
Department of Mathematics
Oregon State University
Corvallis, OR 97331
USA
waymire@math.orst.edu

Mathematics Subject Classification (2000): 60-xx, 60Jxx

Library of Congress Control Number: 2007928731

ISBN 978-0-387-71938-2 e-ISBN 978-0-387-71939-9

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To Gowri and Linda

PREFACE

In 1937, A.N. Kolmogorov introduced a measure-theoretic mathematical framework for probability theory in response to David Hilbert's Sixth Problem. This text provides the basic elements of probability within this framework. It may be used for a one-semester course in probability, or as a reference to prerequisite material in a course on stochastic processes. Our pedagogical view is that the subsequent applications to stochastic processes provide a continued opportunity to motivate and reinforce these important mathematical foundations. The book is best suited for students with some prior, or at least concurrent, exposure to measure theory and analysis. But it also provides a fairly detailed overview, with proofs given in appendices, of the measure theory and analysis used.

The selection of material presented in this text grew out of our effort to provide a self-contained reference to foundational material that would facilitate a companion treatise on stochastic processes that we have been developing.[1] While there are many excellent textbooks available that provide the probability background for various continued studies of stochastic processes, the present treatment was designed with this as an explicit goal. This led to some unique features from the perspective of the ordering and selection of material.

We begin with Chapter I on various measure-theoretic concepts and results required for the proper mathematical formulation of a probability space, random maps, distributions, and expected values. Standard results from measure theory are motivated and explained with detailed proofs left to an appendix.

Chapter II is devoted to two of the most fundamental concepts in probability theory: independence and conditional expectation (and/or conditional probability). This continues to build upon, reinforce, and motivate basic ideas from real analysis and measure theory that are regularly employed in probability theory, such as Carathéodory constructions, the Radon–Nikodym theorem, and the Fubini–Tonelli theorem. A careful proof of the Markov property is given for discrete-parameter random walks on $\mathbb{R}^k$ to illustrate conditional probability calculations in some generality.

Chapter III provides some basic elements of martingale theory that have evolved to occupy a significant foundational role in probability theory. In particular, optional stopping and maximal inequalities are cornerstone elements. This chapter provides sufficient martingale background, for example, to take up a course in stochastic differential equations developed in a chapter of our text on stochastic processes. A more comprehensive treatment of martingale theory is deferred to stochastic processes with further applications there as well.

The various laws of large numbers and elements of large deviation theory are developed in Chapter IV. This includes the classical 0-1 laws of Kolmogorov and

[1] Bhattacharya, R. and E. Waymire (2007): *Theory and Applications of Stochastic Processes*, Springer-Verlag, Graduate Texts in Mathematics.

Hewitt–Savage. Some emphasis is given to size-biasing in large deviation calculations which are of contemporary interest.

Chapter V analyzes in detail the topology of weak convergence of probabilities defined on metric spaces, culminating in the notion of tightness and a proof of Prohorov's theorem.

The characteristic function is introduced in Chapter VI via a first-principles development of Fourier series and the Fourier transform. In addition to the operational calculus and inversion theorem, Herglotz's theorem, Bochner's theorem, and the Cramér–Lévy continuity theorem are given. Probabilistic applications include the Chung–Fuchs criterion for recurrence of random walks on $\mathbb{R}^k$, and the classical central limit theorem for i.i.d. random vectors with finite second moments. The law of rare events (i.e., Poisson approximation to binomial) is also included as a simple illustration of the continuity theorem, although simple direct calculations are also possible.

In Chapter VII, central limit theorems of Lindeberg and Lyapounov are derived. Although there is some mention of stable and infinitely divisible laws, the full treatment of infinite divisibility and Lévy–Khinchine representation is more properly deferred to a study of stochastic processes with independent increments.

The Laplace transform is developed in Chapter VIII with Karamata's Tauberian theorem as the main goal. This includes a heavy dose of exponential size-biasing techniques to go from probabilistic considerations to general Radon measures. The standard operational calculus for the Laplace transform is developed along the way.

Random series of independent summands are treated in Chapter IX. This includes the mean square summability criterion and Kolmogorov's three series criteria based on Kolmogorov's maximal inequality. An alternative proof to that presented in Chapter IV for Kolmogorov's strong law of large numbers is given, together with the Marcinkiewicz and Zygmund extension, based on these criteria and Kronecker's lemma. The equivalence of a.s. convergence, convergence in probability, and convergence in distribution for series of independent summands is also included.

In Chapter X, Kolmogorov's consistency conditions lead to the construction of probability measures on the Cartesian product of infinitely many spaces. Applications include a construction of Gaussian random fields and discrete parameter Markov processes. The deficiency of Kolmogorov's construction of a model for Brownian motion is described, and the Lévy–Ciesielski "wavelet" construction is provided.

Basic properties of Brownian motion are taken up in Chapter XI. Included are various rescalings and time-inversion properties, together with the fine-scale structure embodied in the law of the iterated logarithm for Brownian motion.

In Chapter XII many of the basic notions introduced in the text are tied together via further considerations of Brownian motion. In particular, this chapter revisits conditional probabilities in terms of the Markov and strong Markov properties for Brownian motion, stopping times, and the optional stopping and/or sampling theorems for Brownian motion and related martingales, and leads to weak convergence of rescaled random walks with finite second moments to Brownian motion, i.e., Donsker's invariance principle or the functional central limit theorem, via the Skorokhod embedding theorem.

The text is concluded with a historical overview, Chapter XIII, on Brownian motion and its fundamental role in applications to physics, financial mathematics, and partial differential equations, which inspired its creation.

Most of the material in this book has been used by us in graduate probability courses taught at the University of Arizona, Indiana University, and Oregon State University. The authors are grateful to Virginia Jones for superb word-processing skills that went into the preparation of this text. Also, two Oregon State University graduate students, Jorge Ramirez and David Wing, did an outstanding job in uncovering and reporting various bugs in earlier drafts of this text. Thanks go to Professor Anirban Dasgupta, the editorial staff at Springer and anonymous referees for their insightful remarks. Finally, the authors gratefully acknowledge partial support from NSF grants DMS 04-06143 and CMG 03-27705, which facilitated the writing of this book.

Rabi Bhattacharya
Edward C. Waymire
March 2007

Contents

CHAPTER I

Random Maps, Distribution, and Mathematical Expectation

In the spirit of a refresher, we begin with an overview of the measure-theoretic framework for probability. Readers for whom this is entirely new material may wish to consult the appendices for statements and proofs of basic theorems from analysis. A **measure space** is a triple $(S, \mathcal{S}, \mu)$, where S is a nonempty set; $\mathcal{S}$ is a collection of subsets of S, referred to as a **σ-field**, which includes $\emptyset$ and is closed under complements and countable unions; and $\mu : \mathcal{S} \to [0, \infty]$ satisfies (i) $\mu(\emptyset) = 0$, (ii) (**countable additivity**) $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ if $A_1, A_2, \dots$ is a sequence of disjoint sets in $\mathcal{S}$. Subsets of S belonging to $\mathcal{S}$ are called **measurable sets**.. The pair $(S, \mathcal{S})$ is referred to as a **measurable space**, and the set function μ is called a **measure**. Familiar examples from real analysis are **Lebesgue measure** μ on $S = \mathbb{R}^k$, equipped with a σ-field $\mathcal{S}$ containing the class of all k-dimensional rectangles, say $R = (a_1, b_1] \times \cdots \times (a_k, b_k]$, with "volume" measure $\mu(R) = \prod_{j=1}^{k}(b_j - a_j)$; or **Dirac point mass measure** $\mu = \delta_x$ at $x \in S$ defined by $\delta_x(B) = 1$ if $x \in B$, $\delta_x(B) = 0$ if $x \in B^c$, for $B \in \mathcal{S}$. Such examples should suffice for the present, but see Appendix A for constructions of these and related measures based on the **Carathéodory extension theorem**. If $\mu(S) < \infty$ then μ is referred to as a **finite measure**. If one may write $S = \cup_{n=1}^{\infty} S_n$, where each $S_n \in \mathcal{S}(n \geq 1)$ and $\mu(S_n) < \infty, \forall n$, then μ is said to be a $\sigma-$**finite measure**.

A **probability space** is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a nonempty set, $\mathcal{F}$ is a σ-field of subsets of Ω, and P is a finite measure on the measurable space $(\Omega, \mathcal{F})$ with $P(\Omega) = 1$. The measure P is referred to as a **probability**. Intuitively, Ω represents the set of all possible "outcomes" of a random experiment, real or conceptual, for some given coding of the results of the experiment. The set Ω is referred to as the **sample space** and the elements $\omega \in \Omega$ as **sample points** or possible outcomes.

The $\sigma-$field $\mathcal{F}$ comprises "events" $A \subseteq \Omega$ whose probability $P(A)$ of occurrence is well defined.

The finite total probability and countable additivity of a probability have many important consequences such as **finite additivity**, **finite** and **countable subadditivity**, **inclusion–exclusion**, **monotonicity**, and the formulas for both relative complements and universal complements. Proofs of these properties are left to the reader and included among the exercises.

***Example 1** (Finite Sampling of a Balanced Coin).* Consider m repeated tosses of a balanced coin. Coding the individual outcomes as 1 or 0 (or, say, H,T), the possible outcomes may be represented as sequences of binary digits of length m. Let $\Omega = \{0,1\}^m$ denote the set of all such sequences and $\mathcal{F} = 2^{\Omega}$, the power set of Ω. The condition that the coin be balanced may be defined by the requirement that $P(\{\omega\})$ is the same for each sequence $\omega \in \Omega$. Since Ω has cardinality $|\Omega| = 2^m$, it follows from the finite additivity and total probability requirements that

$$P(\{\omega\}) = \frac{1}{2^m} = \frac{1}{|\Omega|}, \quad \omega \in \Omega.$$

Using finite additivity this completely and explicitly specifies the model $(\Omega, \mathcal{F}, P)$ with

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \frac{|A|}{|\Omega|}, \quad A \subseteq \Omega.$$

The so-called **continuity properties** also follow from the definition as follows: A sequence of events $A_n, n \geq 1$, is said to be **increasing** (respectively, **decreasing**) with respect to set inclusion if $A_n \subseteq A_{n+1}, \forall n \geq 1$ (respectively $A_n \supseteq A_{n+1} \forall n \geq 1$). In the former case one defines $\lim_n A_n := \cup_n A_n$, while for decreasing measurable events $\lim_n A_n := \cap_n A_n$. In either case the continuity of a probability, from below or above, respectively, is the following consequence of countable additivity[1] (Exercise 3):

$$P(\lim_n A_n) = \lim_n P(A_n). \tag{1.1}$$

A bit more generally, if $\{A_n\}_{n=1}^{\infty}$ is a sequence of measurable events one defines

$$\limsup_n A_n := \cap_{n=1} \cup_{m \geq n} A_m \tag{1.2}$$

[1] With the exception of properties for "complements" and "continuity from above," these and the aforementioned consequences can be checked to hold for any measure.

and

$$\liminf_n A_n := \cup_{n=1}^{\infty} \cap_{m \geq n} A_m. \tag{1.3}$$

The event $\limsup_n A_n$ denotes the collection of outcomes $\omega \in \Omega$ that correspond to the occurrences of A_n for infinitely many n; i.e., the events A_n occur **infinitely often**. This event is also commonly denoted by $[A_n \; i.o.] := \limsup_n A_n$. On the other hand, $\liminf_n A_n$ is the set of outcomes ω that belong to A_n for all but finitely many n. Note that $[A_n \; i.o.]^c$ is the event that A_n^c occurs for all but finitely many n and equals $\liminf_n A_n^c$.

Lemma 1 (Borel–Cantelli I). Let $(\Omega, \mathcal{F}, P)$ be a probability space and $A_n \in \mathcal{F}, n = 1, 2, \ldots$. If $\sum_{n=1}^{\infty} P(A_n) < \infty$ then $P(A_n \; i.o.) = 0$.

Proof. Apply (1.1) to the decreasing sequence of events $\cup_{m=1}^{\infty} A_m \supseteq \cup_{m=2}^{\infty} A_m \supseteq \cdots$ and then subadditivity of the probability to get

$$P(\limsup_n A_n) = \lim_{n \to \infty} P(\cup_{m=n}^{\infty} A_m) \leq \lim_{n \to \infty} \sum_{m=n}^{\infty} P(A_m) = 0. \tag{1.4}$$

■

A partial converse (Borel–Cantelli II) will be given in the next chapter.

Example 2 (Infinite Sampling of a Balanced Coin). The possible outcomes of nonterminated repeated coin tosses can be coded as infinite binary sequences of 1's and 0's. Thus the sample space is the infinite product space $\Omega = \{0, 1\}^{\infty}$. Observe that a sequence $\omega \in \Omega$ may be viewed as the digits in a binary expansion of a number x in the unit interval. The binary expansion $x = \sum_{n=1}^{\infty} \omega_n(x) 2^{-n}$, where $\omega_n(x) \in \{0, 1\}$, is not unique for binary rationals, e.g., $\frac{1}{2} = .1000000\ldots = .011111\ldots$, however it may be made unique by requiring that infinitely many 0's occur in the expansion. Thus Ω and $[0, 1)$ may be put in one-to-one correspondence. Observe that for a given specification $\varepsilon_n \in \{0, 1\}, n = 1, \ldots, m$, of the first m tosses, the event $A = \{\omega = (\omega_1, \omega_2, \ldots) \in \Omega : \omega_n = \varepsilon_n, n \leq m\}$ corresponds to the subinterval $[\sum_{n=1}^{m} \varepsilon_n 2^{-n}, \sum_{n=1}^{m} \varepsilon_n 2^{-n} + 2^{-m})$ of $[0, 1)$ of length (Lebesgue measure) 2^{-m}. Again modeling the repeated tosses of a balanced coin by the requirement that for each fixed m, $P(A)$ not depend on the specified values $\varepsilon_n \in \{0, 1\}, 1 \leq n \leq m$, it follows from finite additivity and total probability one that $P(A) = 2^{-m} = |A|$, where $|A|$ denotes the one-dimensional Lebesgue measure of A. Based on these considerations, one may use Lebesgue measure on $[0, 1)$ to define a probability model for infinitely many tosses of a balanced coin. As we will see below, this is an essentially unique choice. For now, let us exploit the model with an illustration of the Borel–Cantelli Lemma I. Fix a nondecreasing sequence r_n of positive integers and let $A_n = \{x \in [0, 1) : \omega_k(x) = 1, k = n, n+1, \ldots, n + r_n - 1\}$ denote the event that a run of 1's occurs of length at least r_n starting at the nth toss. Note that this is an interval of length 2^{-r_n}. Thus if r_n increases so quickly that $\sum_{n=1}^{\infty} 2^{-r_n} < \infty$ then the Borel–Cantelli lemma I yields that $P(A_n\ i.o.) = 0$. For a concrete illustration, let $r_n = [\theta \log_2 n]$, for fixed $\theta > 0$, with $[\cdot]$ denoting the integer part. Then $P(A_n\ i.o.) = 0$ for $\theta > 1$.

In the previous example, probability considerations led us to conclude that under the identification of sequence space with the unit interval, the probabilities of events in a certain collection $\mathcal{C}$ coincide with their Lebesgue measures. Let us pursue this situation somewhat more generally. For a given collection $\mathcal{C}$ of subsets of Ω, the smallest σ-field that contains all of the events in $\mathcal{C}$ is called the **σ-field generated by** $\mathcal{C}$ and is denoted by $\sigma(\mathcal{C})$; if $\mathcal{G}$ is any $\sigma-$field containing $\mathcal{C}$ then $\sigma(\mathcal{C}) \subseteq \mathcal{G}$. Note that, in general, if $\mathcal{F}_\lambda, \lambda \in \Lambda$, is an arbitrary collection of σ-fields of subsets of Ω, then $\cap_{\lambda\in\Lambda}\mathcal{F}_\lambda := \{F \subseteq \Omega : F \in \mathcal{F}_\lambda \,\forall\lambda \in \Lambda\}$ is a σ-field. On the other hand $\cup_{\lambda\in\Lambda}\mathcal{F}_\lambda := \{F \subseteq \Omega : F \in \mathcal{F}_\lambda \text{ for some } \lambda \in \Lambda\}$ is not generally a σ-field. Define the **join σ-field**, denoted by $\bigvee_{\lambda\in\Lambda} \mathcal{F}_\lambda$, to be the σ-field generated by $\cup_{\lambda\in\Lambda}\mathcal{F}_\lambda$.

It is not uncommon that $\mathcal{F} = \sigma(\mathcal{C})$ for a collection $\mathcal{C}$ *closed under finite intersections*; such a collection $\mathcal{C}$ is called a **π-system**, e.g., $\Omega = (-\infty, \infty)$, $\mathcal{C} = \{(a, b] : -\infty \leq a \leq b < \infty\}$, or infinite sequence space $\Omega = \mathbb{R}^\infty$, and $\mathcal{C} = \{(a_1, b_1]\times\cdots\times(a_k, b_k]\times\mathbb{R}^\infty : -\infty \leq a_i \leq b_i < \infty, i = 1, \ldots, k, k \geq 1\}$.

A **λ-system** is a collection $\mathcal{L}$ of subsets of Ω such that (i) $\Omega \in \mathcal{L}$, (ii) If $A \in \mathcal{L}$ then $A^c \in \mathcal{L}$, (iii) If $A_n \in \mathcal{L}, A_n \cap A_m = \emptyset, n \neq m, n, m = 1, 2, \ldots$, then $\cup_n A_n \in \mathcal{L}$. A $\sigma-$field is clearly also a λ-system. The following π-λ theorem provides a very useful tool for checking measurability.

Theorem 1.1 *(Dynkin's π-λ Theorem).* If $\mathcal{L}$ is a λ-system containing a π-system $\mathcal{C}$, then $\sigma(\mathcal{C}) \subseteq \mathcal{L}$.

Proof. Let $\mathcal{L}(\mathcal{C}) = \cap\mathcal{F}$, where the intersection is over all λ-systems $\mathcal{F}$ containing $\mathcal{C}$. We will prove the theorem by showing (i) $\mathcal{L}(\mathcal{C})$ is a π-system, and (ii) $\mathcal{L}(\mathcal{C})$ is a λ-system. For then $\mathcal{L}(\mathcal{C})$ is a σ-field (see Exercise 2), and by its definition $\sigma(\mathcal{C}) \subseteq \mathcal{L}(\mathcal{C}) \subseteq \mathcal{L}$. Now (ii) is simple to check. For clearly $\Omega \in \mathcal{F}$ for all $\mathcal{F}$, and hence $\Omega \in \mathcal{L}(\mathcal{C})$. If $A \in \mathcal{L}(\mathcal{C})$, then $A \in \mathcal{F}$ for all $\mathcal{F}$, and since every $\mathcal{F}$ is a λ-system, $A^c \in \mathcal{F}$ for every $\mathcal{F}$. Thus $A^c \in \mathcal{L}(\mathcal{C})$. If $A_n \in \mathcal{L}(\mathcal{C}), n \geq 1$, is a disjoint sequence, then for each $\mathcal{F}$, $A_n \in \mathcal{F}$, for all n and $A \equiv \cup_n A_n \in \mathcal{F}$ for all $\mathcal{F}$. Since this is true for every λ-system $\mathcal{F}$, one has $A \in \mathcal{L}(\mathcal{C})$. It remains to prove (i). For each set A, define the class $\mathcal{L}_A := \{B : A\cap B \in \mathcal{L}(\mathcal{C})\}$. It suffices to check that $\mathcal{L}_A \supseteq \mathcal{L}(\mathcal{C})$ for all $A \in \mathcal{L}(\mathcal{C})$. First note that if $A \in \mathcal{L}(\mathcal{C})$, then $\mathcal{L}_A$ is a λ-system, by arguments along the line of (ii) above (Exercise 2). In particular, if $A \in \mathcal{C}$, then $A\cap B \in \mathcal{C}$ for all $B \in \mathcal{C}$, since $\mathcal{C}$ is closed under finite intersections. Thus $\mathcal{L}_A \supseteq \mathcal{C}$. This implies, in turn, that $\mathcal{L}(\mathcal{C}) \subseteq \mathcal{L}_A$. This says that $A\cap B \in \mathcal{L}(\mathcal{C})$ for all $A \in \mathcal{C}$ and for all $B \in \mathcal{L}(\mathcal{C})$. Thus, if we fix $B \in \mathcal{L}(\mathcal{C})$, then $\mathcal{L}_B \equiv \{A : B\cap A \in \mathcal{L}(\mathcal{C})\} \supseteq \mathcal{C}$. Therefore $\mathcal{L}_B \supseteq \mathcal{L}(\mathcal{C})$. In other words, for every $B \in \mathcal{L}(\mathcal{C})$ and $A \in \mathcal{L}(\mathcal{C})$, one has $A\cap B \in \mathcal{L}(\mathcal{C})$. ■

In view of the additivity properties of a probability, the following is an immediate and important corollary to the π-λ theorem.

Corollary 1.2 (Uniqueness). If P_1, P_2 are two probability measures such that $P_1(C) = P_2(C)$ for all events C belonging to a $\pi-$system $\mathcal{C}$, then $P_1 = P_2$ on all of $\mathcal{F} = \sigma(\mathcal{C})$.

Proof. Check that $\{A \in \mathcal{F} : P_1(A) = P_2(A)\} \supseteq \mathcal{C}$ is a λ-system. ■

For a related application suppose that (S, ρ) is a metric space. The **Borel σ-field** of S, denoted by $\mathcal{B}(S)$, is defined as the σ-field generated by the collection $\mathcal{C} = \mathcal{T}$ of open subsets of S, the collection $\mathcal{T}$ being referred to as the topology on S specified by the metric ρ. More generally, one may specify a **topology** for a set S by a collection $\mathcal{T}$ of subsets of S that includes both $\emptyset$ and S, and is closed under arbitrary unions and finite intersections. Then $(S, \mathcal{T})$ is called a **topological space** and members of $\mathcal{T}$ define the open subsets of S. The topology is said to be **metrizable** when it may be specified by a metric ρ as above. In any case, one defines the Borel σ-field by $\mathcal{B}(S) := \sigma(\mathcal{T})$.

Definition 1.1. A class $\mathcal{C} \subseteq \mathcal{B}(S)$ is said to be **measure-determining** if for any two finite measures μ, ν such that $\mu(C) = \nu(C)\ \forall C \in \mathcal{C}$, it follows that $\mu = \nu$ on $\mathcal{B}(S)$.

One may directly apply the π-λ theorem, noting that S is both open and closed, to see that the class $\mathcal{T}$ of all open sets is measure-determining, as is the class $\mathcal{K}$ of all closed sets.

If $(S_i, \mathcal{S}_i)$, $i = 1, 2$, is a pair of measurable spaces then a function $f : S_1 \to S_2$ is said to be a **measurable map** if $f^{-1}(B) := \{x \in S_1 : f(x) \in B\} \in \mathcal{S}_1$ for all $B \in \mathcal{S}_2$. In usual mathematical discourse the σ-fields required for this definition may not be explicitly mentioned and will need to be inferred from the context. For example if $(S, \mathcal{S})$ is a measurable space, by a **Borel-measurable function** $f : S \to \mathbb{R}$ is meant measurability when $\mathbb{R}$ is given its Borel σ-field. A **random variable,** or a **random map,** X is a measurable map on a probability space $(\Omega, \mathcal{F}, P)$ into a measurable space $(S, \mathcal{S})$. Measurability of X means that each event[2] $[X \in B] := X^{-1}(B)$ belongs to $\mathcal{F}\ \forall\ B \in \mathcal{S}$. The term "random variable" is most often used to denote a real-valued random variable, i.e., where $S = \mathbb{R}$, $\mathcal{S} = \mathcal{B}(\mathbb{R})$. When $S = \mathbb{R}^k$, $\mathcal{S} = \mathcal{B}(\mathbb{R}^k)$, $k > 1$, one uses the term **random vector**.

A common alternative to the use of a metric to define a topology, is to indirectly characterize the topology by specifying what it means for a sequence to converge in the topology. That is, if $\mathcal{T}$ is a topology on S, then a sequence $\{x_n\}_{n=1}^{\infty}$ in S **converges to $x \in S$ with respect to the topology $\mathcal{T}$** if for arbitrary $U \in \mathcal{T}$ such that $x \in U$, there is an N such that $x_n \in U$ for all $n \geq N$. A topological space $(S, \mathcal{T})$, or a topology $\mathcal{T}$, is said to be **metrizable** if $\mathcal{T}$ coincides with the class of open sets defined by a metric ρ on S. Using this notion, other commonly occurring measurable

[2] Throughout, this square-bracket notation will be used to denote events defined by inverse images.

image spaces may be described as follows: (i) $S = \mathbb{R}^\infty$—the space of all sequences of reals with the (metrizable) **topology of pointwise convergence**, and $\mathcal{S} = \mathcal{B}(\mathbb{R}^\infty)$, (ii) $S = C[0,1]$—the space of all real-valued continuous functions on the interval $[0,1]$ with the (metrizable) **topology of uniform convergence**, and $\mathcal{S} = \mathcal{B}(C[0,1])$, and (iii) $S = C([0,\infty) \to \mathbb{R}^k)$—the space of all continuous functions on $[0,\infty)$ into $\mathbb{R}^k$, with the (metrizable) **topology of uniform convergence on compact subsets of** $[0,\infty)$, $\mathcal{S} = \mathcal{B}(S)$ (see Exercise 7).

The relevant quantities for a random map X on a probability space $(\Omega, \mathcal{F}, P)$ are the probabilities with which X takes sets of values. In this regard, P determines the most important aspect of X, namely, its **distribution** $Q \equiv P \circ X^{-1}$ defined on the image space $(S, \mathcal{S})$ by $Q(B) := P(X^{-1}(B)) \equiv P(X \in B)$, $B \in \mathcal{S}$. The distribution is sometimes referred to as the **induced measure** of X under P. Note that given any probability measure Q on a measurable space $(S, \mathcal{S})$ one can construct a probability space $(\Omega, \mathcal{F}, P)$ and a random map X on $(\Omega, \mathcal{F})$ with distribution Q. The simplest such construction is given by letting $\Omega = S$, $\mathcal{F} = \mathcal{S}$, $P = Q$, and X the **identity map**: $X(\omega) = \omega$, $\omega \in S$. This is often called a **canonical construction**, and $(S, \mathcal{S}, Q)$ with the identity map X is called a **canonical model**.

If $X = \sum_{j=1}^m a_j \mathbf{1}_{A_j}, A_j \in \mathcal{F}, A_i \cap A_j = \emptyset (i \neq j)$, is a **discrete random variable** or, equivalently, a **simple random variable**, then $\mathbb{E}X \equiv \int_\Omega X dP := \sum_{j=1}^m a_j P(A_j)$. If $X : \Omega \to [0,\infty)$ is a random variable, then $\mathbb{E}X$ is defined by the "simple function approximation" $\mathbb{E}X \equiv \int_\Omega X dP := \sup\{\mathbb{E}Y : 0 \leq Y \leq X, Y \text{ simple}\}$. In particular, one may apply the standard simple function approximations $X = \lim_{n\to\infty} X_n$ given by the nondecreasing sequence

$$X_n := \sum_{j=0}^{n2^n - 1} \frac{j}{2^n} \mathbf{1}_{[j2^{-n} \leq X < (j+1)2^{-n}]} + n\mathbf{1}_{[X \geq n]}, \quad n = 1, 2, \ldots, \tag{1.5}$$

to write

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n = \lim_{n\to\infty} \left\{ \sum_{j=0}^{n2^n - 1} \frac{j}{2^n} P(j2^{-n} \leq X < (j+1)2^{-n}) + nP(X \geq n) \right\}. \tag{1.6}$$

Note that if $\mathbb{E}X < \infty$, then $nP(X > n) \to 0$ as $n \to \infty$ (Exercise 16). Now, more generally, if X is a real-valued random variable, then the **expected value** (or, **mean**) of X is defined as

$$\mathbb{E}(X) \equiv \int_\Omega X dP := \mathbb{E}X^+ - \mathbb{E}X^-, \tag{1.7}$$

provided at least one of $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ is finite, where $X^+ = X\mathbf{1}_{[X \geq 0]}$ and $X^- = -X\mathbf{1}_{[X \leq 0]}$. If both $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$, or equivalently, $\mathbb{E}|X| = \mathbb{E}X^+ + \mathbb{E}X^- < \infty$, then X is said to be **integrable** with respect to the probability P. Note that if X is bounded a.s., then applying (1.5) to X^+ and X^-, one obtains a sequence

$X_n (n \geq 1)$ of simple functions that *converge uniformly* to X, outside a P-null set. (Exercise 1(i)).

If X is a random variable with values in $(S, \mathcal{S})$ and if h is a real-valued Borel-measurable function on S, then using simple function approximations to h, one may obtain the following basic **change of variables formula** (Exercise 11)

$$\mathbb{E}(h(X)) \equiv \int_\Omega h(X(\omega))P(d\omega) = \int_S h(x)Q(dx), \tag{1.8}$$

where Q is the distribution of X, provided one of the two indicated integrals may be shown to exist. If $X = (X_1, X_2, \dots, X_k)$ is a random vector, one defines $\mathbb{E}(X) = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_k))$.

This definition of expectation as an *integral in the sense of Lebesgue* is precisely the same as that used in real analysis to define $\int_S f(x)\mu(dx)$ for a real-valued Borel measurable function f on an arbitrary measure space $(S, \mathcal{S}, \mu)$; see Appendix A. One may exploit standard tools of real analysis (see Appendices A and C), such as *Lebesgue's dominated convergence theorem, Lebesgue's monotone convergence theorem, Fatou's lemma, Fubini–Tonelli theorem, Radon–Nykodym theorem*, for estimates and computations involving expected values.

Definition 1.2. A sequence $\{X_n\}_{n=1}^\infty$ of random variables on a probability space $(\Omega, \mathcal{F}, P)$ is said to **converge in probability** to a random variable X if for each $\varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$. The convergence is said to be **almost sure** if the event $[X_n \not\to X] \equiv \{\omega \in \Omega : X_n(\omega) \not\to X(\omega)\}$ has P-measure zero.

Note that almost-sure convergence always implies convergence in probability, since for arbitrary $\varepsilon > 0$ one has $0 = P(\cap_{n=1}^\infty \cup_{m=n}^\infty [|X_m - X| > \varepsilon]) = \lim_{n\to\infty} P(\cup_{m=n}^\infty [|X_m - X| > \varepsilon]) \geq \limsup_{n\to\infty} P(|X_n - X| > \varepsilon)$. An equivalent formulation of convergence in probability can be cast in terms of almost-sure convergence as follows.

Proposition 1.3. A sequence of random variables $\{X_n\}_{n=1}^\infty$ on $(\Omega, \mathcal{F}, P)$ converges in probability to a random variable X on $(\Omega, \mathcal{F}, P)$ if and only if every subsequence has an a.s. convergent subsequence to X.

Proof. Suppose that $X_n \to X$ in probability as $n \to \infty$. Let $\{X_{n_k}\}_{k=1}^\infty$ be a subsequence, and for each $m \geq 1$ recursively choose $n_{k(0)} = 1, n_{k(m)} = \min\{n_k > n_{k(m-1)} : P(|X_{n_k} - X| > 1/m) \leq 2^{-m}\}$. Then it follows from the Borel–Cantelli lemma (Part I) that $X_{n_{k(m)}} \to X$ a.s. as $m \to \infty$. For the converse suppose that X_n does not converge to X in probability. Then there exists $\varepsilon > 0$ and a sequence $n_1, n_2, \dots$ such that $\lim_k P(|X_{n_k} - X| > \varepsilon) = \alpha > 0$. Since a.s. convergence implies convergence in probability (see Appendix A, Proposition 2.4), there cannot be an a.s. convergent subsequence of $\{X_{n_k}\}_{k=1}^\infty$. ■

The notion of measure-determining classes of sets extends to classes of functions as follows. Let μ, ν be arbitrary finite measures on the Borel σ-field of a metric space S. A class Γ of real-valued bounded Borel measurable functions on S is **measure-determining** if $\int_S g\, d\mu = \int_S g\, d\nu \ \forall g \in \Gamma$ implies $\mu = \nu$.

Proposition 1.4. The class $C_b(\mathcal{S})$ of real-valued bounded continuous functions on S is measure-determining.

Proof. To prove this, it is enough to show that for each (closed) $F \in \mathcal{K}$ there exists a sequence of nonnegative functions $\{f_n\} \subseteq C_b(S)$ 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 dominated convergence theorem applies. ■

Note that the functions f_n in the proof of Proposition 1.4 are uniformly continuous, since $|f_n(x) - f_n(y)| \le (n\rho(x, y)) \wedge (2 \sup_x |f(x)|)$. It follows that the **set** $UC_b(S)$ **of bounded uniformly continuous functions on** S is measure determining.

Consider the L^p-space $L^p(\Omega, \mathcal{F}, P)$ of (real-valued) random variables X such that $\mathbb{E}|X|^p < \infty$. When random variables that differ only on a P-null set are identified, then for $p \ge 1$, it follows from Theorem 1.5(e) below that $L^p(\Omega, \mathcal{F}, P)$ is a normed linear space with norm $\|X\|_p := (\int_\Omega |X|^p dP)^{\frac{1}{p}} \equiv (\mathbb{E}|X|^p)^{\frac{1}{p}}$. It may be shown that with this norm (and distance $\|X - Y\|_p$), it is a complete metric space, and therefore a **Banach space** (Exercise 18). In particular, $L^2(\Omega, \mathcal{F}, P)$ is a **Hilbert space** with inner product (see Appendix C)

$$\langle X, Y \rangle = \mathbb{E}XY \equiv \int_\Omega XY dP, \quad ||X||_2 = \langle X, X \rangle^{\frac{1}{2}} \ . \tag{1.9}$$

The $L^2(S, \mathcal{S}, \mu)$ spaces are the only Hilbert spaces that are required in this text, where $(S, \mathcal{S}, \mu)$ is a σ-finite measure space; see Appendix C for an exposition of the essential structure of such spaces. Note that by taking S to be a countable set with counting measure μ, this includes the l^2 *sequence space*. Unlike the case of a measure space $(\Omega, \mathcal{F}, \mu)$ with an infinite measure μ, for finite measures it is always true that

$$L^r(\Omega, \mathcal{F}, P) \subseteq L^s(\Omega, \mathcal{F}, P) \quad \text{ if } r > s \ge 1, \tag{1.10}$$

as can be checked using $|x|^s < |x|^r$ for $|x| > 1$. The basic inequalities in the following Theorem 1.5 are consequences of *convexity* at some level. So let us be precise about this notion.

Definition 1.3. A function φ defined on an open interval J is said to be a **convex function** if $\varphi(ta + (1 - t)b) \le t\varphi(a) + (1 - t)\varphi(b)$, for all $a, b \in J$, $0 \le t \le 1$.

If the function φ is sufficiently smooth, one may use calculus to check convexity, see Exercise 14. The following lemma is required to establish a geometrically obvious "line of support property" of convex functions.

Lemma 2 *(Line of Support).* Suppose φ is convex on an interval J. (a) If J is open, then (i) the left-hand and right-hand derivatives φ^- and φ^+ exist and are finite and nondecreasing on J, and $\varphi^- \leq \varphi^+$. Also (ii) for each $x_0 \in J$ there is a constant $m = m(x_0)$ such that $\varphi(x) \geq \varphi(x_0) + m(x - x_0), \forall x \in J$. (b) If J has a left (or right) endpoint and the right-hand (left-hand) derivative is finite, then the line of support property holds at this endpoint x_0.

Proof. (a) In the definition of convexity, one may take $a < b$, $0 < t < 1$. Thus convexity is equivalent to the following inequality with the identification $a = x$, $b = z$, $t = (z - y)/(z - x)$: For any $x, y, z \in J$ with $x < y < z$,

$$\frac{\varphi(y) - \varphi(x)}{y - x} \leq \frac{\varphi(z) - \varphi(y)}{z - y}. \tag{1.11}$$

More generally, use the definition of convexity to analyze monotonicity and bounds on the Newton quotients (slopes of secant lines) from the right and left to see that (1.11) implies $\frac{\varphi(y)-\varphi(x)}{y-x} \leq \frac{\varphi(z)-\varphi(x)}{z-x} \leq \frac{\varphi(z)-\varphi(y)}{z-y}$ (use the fact that $c/d \leq e/f$ for $d, f > 0$ implies $c/d \leq (c + e)/(d + f) \leq e/f$). The first of these inequalities shows that $\frac{\varphi(y)-\varphi(x)}{y-x}$ decreases as y decreases, so that the right-hand derivative $\varphi^+(x)$ exists and $\frac{\varphi(y)-\varphi(x)}{y-x} \geq \varphi^+(x)$. Letting $z \downarrow y$ in (1.11), one gets $\frac{\varphi(y)-\varphi(x)}{y-x} \leq \varphi^+(y)$ for all $y > x$. Hence φ^+ is finite and nondecreasing on J. Now fix $x_0 \in J$. By taking $x = x_0$ and $y = x_0$ in turn in these two inequalities for φ^+, it follows that $\varphi(y) - \varphi(x_0) \geq \varphi^+(x_0)(y - x_0)$ for all $y \geq x_0$, and $\varphi(x_0) - \varphi(x) \leq \varphi^+(x_0)(x_0 - x)$ for all $x \leq x_0$. Thus the "line of support" property holds with $m = \varphi^+(x_0)$. (b) If J has a left (right) endpoint x_0, and $\varphi^+(x_0)$ $(\varphi^-(x_0))$ is finite, then the above argument remains valid with $m = \varphi^+(x_0)$ $(\varphi^-(x_0))$.

A similar proof applies to the left-hand derivative $\varphi^-(x)$ (Exercise 14). On letting $x \uparrow y$ and $z \downarrow y$ in (1.11), one obtains $\varphi^-(y) \leq \varphi^+(y)$ for all y. In particular, the line of support property now follows for $\varphi^-(x_0) \leq m \leq \varphi^+(x_0)$. ■

Theorem 1.5 *(Basic Inequalities).* Let X, Y be random variables on $(\Omega, \mathcal{F}, P)$.

(a) *(Jensen's Inequality)* If φ is a convex function on the interval J and $P(X \in J) = 1$, then $\varphi(\mathbb{E}X) \leq \mathbb{E}(\varphi(X))$ provided that the indicated expectations exist. Moreover, if φ is strictly convex, then equality holds if and only if X is a.s. constant.

(b) *(Lyapounov Inequality)* If $0 < r < s$ then $(\mathbb{E}|X|^r)^{\frac{1}{r}} \leq (\mathbb{E}|X|^s)^{\frac{1}{s}}$.

(c) *(Hölder Inequality)* Let $p \geq 1$. If $X \in L^p, Y \in L^q$, $\frac{1}{p} + \frac{1}{q} = 1$, then $XY \in L^1$ and $\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|Y|^q)^{\frac{1}{q}}$.

(d) *(Cauchy–Schwarz Inequality)* If $X, Y \in L^2$ then $XY \in L^1$ and one has $|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}Y^2}$.

(e) *(Minkowski Triangle Inequality)* Let $p \geq 1$. If $X, Y \in L^p$ then $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$.

(f) *(Markov and Chebyshev-type Inequalities)* Let $p \geq 1$. If $X \in L^p$ then $P(|X| \geq \lambda) \leq \frac{\mathbb{E}(|X|^p \mathbf{1}_{[|X|\geq\lambda]})}{\lambda^p} \leq \frac{\mathbb{E}|X|^p}{\lambda^p}, \quad \lambda > 0$,

Proof. The proof of Jensen's inequality hinges on the line of support property of convex functions in Lemma 2 by taking $x = X(\omega), \omega \in \Omega, x_0 = \mathbb{E}X$. The Lyapounov inequality follows from Jensen's inequality by writing $|X|^s = (|X|^r)^{\frac{s}{r}}$. for $0 < r < s$. For the Hölder inequality, let $p, q > 1$ be **conjugate exponents** in the sense that $\frac{1}{p} + \frac{1}{q} = 1$. Using convexity of the function $\exp(x)$ one sees that $|ab| = \exp(\ln(|a|^p)/p + \ln(|b|^q)/q)) \leq \frac{1}{p}|a|^p + \frac{1}{q}|b|^q$. Applying this to $a = \frac{|X|}{\|X\|_p}, b = \frac{|Y|}{\|Y\|_q}$ and integrating, it follows that $\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}}(\mathbb{E}|Y|^q)^{\frac{1}{q}}$. The Cauchy–Schwarz inequality is the Hölder inequality with $p = q = 2$. For the proof of Minkowski's inequality, first use the inequality (1.21) to see that $|X+Y|^p$ is integrable from the integrability of $|X|^p$ and $|Y|^p$. Applying Hölder's inequality to each term of the expansion $\mathbb{E}(|X|+|Y|)^p = \mathbb{E}|X|(|X|+|Y|)^{p-1} + \mathbb{E}|Y|(|X|+|Y|)^{p-1}$, and solving the resulting inequality for $\mathbb{E}(|X|+|Y|)^p$ (using conjugacy of exponents), it follows that $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$. Finally, for the Markov and Chebyshev-type inequalities simply observe that since $\mathbf{1}_{\{|X|\geq\lambda\}} \leq \frac{|X|^p \mathbf{1}_{\{|X|\geq\lambda\}})}{\lambda^p} \leq \frac{|X|^p}{\lambda^p}$ on Ω, taking expectations yields $P(|X| \geq \lambda) \leq \frac{\mathbb{E}(|X|^p \mathbf{1}_{\{|X|\geq\lambda\}})}{\lambda^p} \leq \frac{\mathbb{E}|X|^p}{\lambda^p}, \quad \lambda > 0$. ■

The Markov inequality refers to the case $p = 1$ in (f). Observe from the proofs that (c–e) hold with the random variables X, Y replaced by measurable functions, in fact complex-valued, on an arbitrary (not necessarily finite) measure space $(S, \mathcal{S}, \mu)$; see Exercise 19.

The main text includes use of another limit result for Lebesgue integrals, **Scheffé's theorem**, which is more particularly suited to probability applications in which one may want to include the consequence of convergence almost everywhere in terms of convergences in other metrics. It is included here for ease of reference. To state it, suppose that $(S, \mathcal{S}, \mu)$ is an arbitrary measure space and $g : S \to [0, \infty)$ a Borel-measurable function, though not necessarily integrable. One may use g as a **density** to define another measure ν on $(S, \mathcal{S})$, i.e., with g as its Radon–Nykodym derivative $d\nu/d\mu = g$, also commonly denoted by $d\nu = g\,d\mu$, and meaning that $\nu(A) = \int_A g\,d\mu$, $A \in \mathcal{S}$; see Appendix C for a full treatment of the Radon–Nikodym theorem[3].

Recall that a sequence of measurable functions $\{g_n\}_{n=1}^{\infty}$ on S is said to **converge μ-a.e.** to a measurable function g on S if and only if $\mu(\{x \in S : \lim_n g_n(x) \neq g(x)\}) = 0$.

[3] A probabilistic proof can be given for the Radon–Nikodym theorem based on martingales. Such a proof is given in the text on stochastic processes.

Theorem 1.6 (Scheffé). Let $(S, \mathcal{S}, \mu)$ be a measure space and suppose that $\nu, \{\nu_n\}_{n=1}^{\infty}$ are measures on $(S, \mathcal{S})$ with respective nonnegative densities $g, \{g_n\}_{n=1}^{\infty}$ with respect to μ, such that

$$\int_S g_n \, d\mu = \int_S g \, d\mu < \infty, \quad \forall n = 1, 2, \ldots.$$

If $g_n \to g$ as $n \to \infty$, μ-a.e., then

$$\sup_{A \in \mathcal{S}} | \int_A g \, d\mu - \int_A g_n \, d\mu| \leq \int_S |g - g_n| \, d\mu \to 0, \text{ as } n \to \infty.$$

Proof. The indicated bound on the supremum follows from the triangle inequality for integrals. Since $\int_S (g - g_n) \, d\mu = 0$ for each n, $\int_S (g - g_n)^+ \, d\mu = \int_\Omega (g - g_n)^- \, d\mu$. In particular, since $|g - g_n| = (g - g_n)^+ + (g - g_n)^-$,

$$\int_S |g - g_n| \, d\mu = 2 \int_S (g - g_n)^+ \, d\mu.$$

But $0 \leq (g - g_n)^+ \leq g$. Since g is μ-integrable, one obtains $\int_S (g - g_n)^+ \, d\mu \to 0$ as $n \to \infty$ from Lebesgue's dominated convergence theorem. ■

For a measurable space $(S, \mathcal{S})$, a useful metric (see Exercise 1) defined on the space $\mathcal{P}(S)$ of probabilities on $\mathcal{S} = \mathcal{B}(S)$ is furnished by the **total variation distance** defined by

$$d_v(\mu, \nu) := \sup\{|\mu(A) - \nu(A)| : A \in \mathcal{B}(S)\}, \quad \mu, \nu \in \mathcal{P}(S). \tag{1.12}$$

Proposition 1.7. Suppose that $(S, \mathcal{S})$ is a measurable space. Then

$$d_v(\mu, \nu) = \frac{1}{2} \sup \left\{ \left| \int_S f \, d\mu - \int_S f \, d\nu \right| : f \in B(S), |f| \leq 1 \right\},$$

where $B(S)$ denotes the space of bounded Borel-measurable functions on S. Moreover, $(\mathcal{P}(S), d_v)$ is a complete metric space.

Proof. Let us first establish the formula for the total variation distance. By standard simple function approximation it suffices to consider bounded simple functions in the supremum. Fix arbitrary $\mu, \nu \in \mathcal{P}(S)$. Let $f = \sum_{i=1}^k a_i \mathbf{1}_{A_i} \in B(S)$ with $|a_i| \leq 1, i = 1, \ldots, k$ and disjoint sets $A_i \in \mathcal{S}$, $1 \leq i \leq k$. Let $I^+ := \{i \leq k : \mu(A_i) \geq \nu(A_i)\}$. Let I^- denote the complementary set of indices. Then by definition of the integral of a simple function and splitting the sum over $I^\pm$ one has upon twice using the triangle inequality that

$$\left| \int_S f d\mu - \int_S f d\nu \right| \leq \sum_{i \in I^+} |a_i| (\mu(A_i) - \nu(A_i)) + \sum_{i \in I^-} |a_i| (\nu(A_i) - \mu(A_i))$$

$$\begin{aligned}
&\le \sum_{i\in I^+}(\mu(A_i)-\nu(A_i)) + \sum_{i\in I^-}(\nu(A_i)-\mu(A_i)) \\
&= \mu(\cup_{i\in I^+}A_i) - \nu(\cup_{i\in I^+}A_i) + \nu(\cup_{i\in I^-}A_i) - \mu(\cup_{i\in I^-}A_i) \\
&\le 2\sup\{|\mu(A)-\nu(A)| : A \in \mathcal{S}\}. \qquad (1.13)
\end{aligned}$$

On the other hand, taking $f = \mathbf{1}_A - \mathbf{1}_{A^c}, A \in \mathcal{S}$, one has

$$\begin{aligned}
\left|\int_S f d\mu - \int_S f d\nu\right| &= |\mu(A) - \mu(A^c) - \nu(A) + \nu(A^c)| \\
&= |\mu(A) - \nu(A) - 1 + \mu(A) + 1 - \nu(A)| \\
&= 2|\mu(A) - \nu(A)|. \qquad (1.14)
\end{aligned}$$

Thus, taking the supremum over sets $A \in \mathcal{S}$ establishes the asserted formula for the total variation distance. Next, to prove that the space $\mathcal{P}(S)$ of probabilities is complete for this metric, let $\{\mu_n\}_{n=1}^\infty$ be a Cauchy sequence in $\mathcal{P}(S)$. Since the closed interval $[0,1]$ of real numbers is complete, one may define $\mu(A) := \lim_n \mu_n(A), A \in \mathcal{S}$. Because this convergence is uniform over $\mathcal{S}$, it is simple to check that $\mu \in \mathcal{P}(S)$ and $\mu_n \to \mu$ in the metric d_v; see Exercise 1. ■

So we note that Scheffé's theorem provides conditions under which a.s. convergence implies $L^1(S, \mathcal{S}, \mu)$-convergence of the densities g_n to g, and convergence in the total variation metric of the probabilities ν_n to ν.

We will conclude this chapter with some further basic convergence theorems for probability spaces. For this purpose we require a definition.

Definition 1.4. A sequence $\{X_n\}_{n=1}^\infty$ of random variables on a probability space $(\Omega, \mathcal{F}, P)$ is said to be **uniformly integrable** if $\lim_{\lambda\to\infty}\sup_n \mathbb{E}\{|X_n|\mathbf{1}_{[|X_n|\ge\lambda]}\} = 0$.

Theorem 1.8 ***(L^1–Convergence Criterion).*** Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P), X_n \in L^1$ $(n \ge 1)$. Then $\{X_n\}_{n=1}^\infty$ converges in L^1 to a random variable X if and only if (i) $X_n \to X$ in probability as $n \to \infty$, and (ii) $\{X_n\}_{n=1}^\infty$ is uniformly integrable.

Proof. *(Necessity)* If $X_n \to X$ in L^1 then convergence in probability (i) follows from the Markov inequality. Also

$$\begin{aligned}
\int_{[|X_n|\ge\lambda]} |X_n| dP &\le \int_{[|X_n|\ge\lambda]} |X_n - X| dP + \int_{[|X_n|\ge\lambda]} |X| dP \\
&\le \int_\Omega |X_n - X| dP + \int_{[|X|\ge\lambda/2]} |X| dP
\end{aligned}$$

$$+\int_{[|X|<\lambda/2,|X_n-X|\geq\lambda/2]}|X|dP. \tag{1.15}$$

The first term of the last sum goes to zero as $n\to\infty$ by hypothesis. For each $\lambda>0$ the third term goes to zero by the dominated convergence theorem as $n\to\infty$. The second term goes to zero as $\lambda\to\infty$ by the dominated convergence theorem too. Thus there are numbers $n(\varepsilon)$ and $\lambda(\varepsilon)$ such that for all $\lambda\geq\lambda(\varepsilon)$,

$$\sup_{n\geq n(\varepsilon)}\int_{[|X_n|\geq\lambda]}|X_n|dP\leq\varepsilon. \tag{1.16}$$

Since a *finite* sequence of integrable random variables $\{X_n : 1\leq n\leq n(\varepsilon)\}$ is always uniformly integrable, it follows that the full sequence $\{X_n\}$ is uniformly integrable.

(*Sufficiency)* Under the hypotheses (i), (ii), given $\varepsilon>0$ one has for all n that

$$\int_\Omega|X_n|dP\leq\int_{[|X_n|\geq\lambda]}|X_n|dP+\lambda\leq\varepsilon+\lambda(\varepsilon) \tag{1.17}$$

for sufficiently large $\lambda(\varepsilon)$. In particular, $\{\int_\Omega|X_n|dP\}_{n=1}^\infty$ is a bounded sequence. Thus $\int_\Omega|X|dP<\infty$ by Fatou's lemma. Now

$$\begin{aligned}\int_{[|X_n-X|\geq\lambda]}|X_n-X|dP &= \int_{[|X_n-X|\geq\lambda,|X_n|\geq\lambda/2]}|X_n-X|dP\\ &\quad+\int_{[|X_n|<\lambda/2,|X_n-X|\geq\lambda]}|X_n-X|dP\\ &\leq\int_{[|X_n|\geq\lambda/2]}|X_n|dP+\int_{[|X_n-X|\geq\lambda/2]}|X|dP\\ &\quad+\int_{[|X_n|<\lambda/2,|X_n-X|\geq\lambda]}(\frac{\lambda}{2}+|X|)dP.\end{aligned} \tag{1.18}$$

Now, using (ii), given $\varepsilon>0$, choose $\lambda=\lambda(\varepsilon)>0$ so large that the first term of the last sum is smaller than ε. With this value of $\lambda=\lambda(\varepsilon)$ the second and third terms go to zero as $n\to\infty$ by Lebesgue's dominated convergence theorem, using (i). Thus,

$$\limsup_{n\to\infty}\int_{[|X_n-X|\geq\lambda(\varepsilon)]}|X_n-X|dP\leq\varepsilon. \tag{1.19}$$

But again applying the dominated convergence theorem one also has

$$\limsup_{n\to\infty}\int_{[|X_n-X|<\lambda(\varepsilon)]}|X_n-X|dP=0. \tag{1.20}$$

Thus the conditions are also sufficient for L^1 convergence to X. ■

The next result follows as a corollary.

***Theorem 1.9** (L^p−Convergence Criterion).* Let $p \geq 1$. Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P), X_n \in L^p (n \geq 1)$. Then $\{X_n\}_{n=1}^\infty$ converges in L^p to a random variable X if and only if (i) $X_n \to X$ in probability as $n \to \infty$, and (ii) $\{|X_n|^p\}_{n=1}^\infty$ is uniformly integrable.

Proof. Apply the preceding result to the sequence $\{|X_n - X|^p\}_{n=1}^\infty$. The proof of necessity is analogous to (1.15) and (1.16) using the following elementary inequalities:

$$|a+b|^p \leq (|a|+|b|)^p \leq (2\max\{|a|,|b|\})^p \leq 2^p(|a|^p + |b|^p). \tag{1.21}$$

For sufficiency, note as in (1.17) that (i), (ii) imply $X \in L^p$, and then argue as in (1.18) that the uniform integrability of $\{|X_n|^p : n \geq 1\}$ implies that of $\{|X_n - X|^p : n \geq 1\}$. ■

Chebyshev-type inequalities often provide useful ways to check uniform integrability of $\{|X_n|^p\}_{n=1}^\infty$ in the case that $\{\mathbb{E}|X_n|^m\}$ can be shown to be a bounded sequence for some $m > p$ (see Exercise 15).

EXERCISES

Exercise Set I

1. Let $(S, \mathcal{S})$ be a measurable space. (i) Show that if f is a real-valued bounded measurable function, $|f(x)| \leq c$ for all x, then the standard simple function approximations (1.5) to f^+ and f^- provide a sequence of simple functions f_n converging to f *uniformly* on S, and satisfying $|f_n(x)| \leq c$ for all x and for all n. (ii) Show that (1.12) defines a metric on $\mathcal{P}(S)$ i.e., is a well-defined nonnegative symmetric function on $\mathcal{P}(S)\times\mathcal{P}(S)$ satisfying the triangle inequality with $d_v(\mu,\nu) = 0$ if and only if $\mu = \nu$. Also show for a Cauchy sequence $\{\mu_n\}_{n=1}^\infty$ in $\mathcal{P}(S)$, that the set function defined by $\mu(A) := \lim_n \mu_n(A) \in [0,1], A \in \mathcal{S}$ is a probability measure. [*Hint*: The convergence of the real numbers $\mu_n(A) \to \mu(A)$ is uniform for $A \in \mathcal{S}$.]
2. Show that if $\mathcal{L}$ is a π-system and a λ-system, then it is a σ-field. In the proof of Dynkin's π-λ theorem, show that if $A \in \mathcal{L}(\mathcal{C})$, then $\mathcal{L}_A$ is a λ-system. [*Hint:* $A \cap B^c = (A^c \cup (A \cap B))^c$.]
3. Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space and let $A_1, A_2, \ldots$ be measurable events. Prove each of the following.
   (i) (Finite Additivity). If $A_1, \ldots, A_m$ are disjoint then $P(\cup_{j=1}^m A_j) = \sum_{j=1}^m P(A_j)$.
   (ii) (Monotonicity). If $A_1 \subseteq A_2$ then $P(A_1) \leq P(A_2)$.
   (iii) (Inclusion–Exclusion). $P(\cup_{j=1}^m A_j) = \sum_{k=1}^m (-1)^{k+1} \sum_{1\leq j_1<\cdots<j_k\leq m} P(A_{j_1} \cap \cdots \cap A_{j_k})$.
   (iv) (Subadditivity). $P(\cup_j A_j) \leq \sum_j P(A_j)$.
   (v) Show that the property $\mu(A_n) \uparrow \mu(A)$ if $A_n \uparrow A$, holds for all measures μ. [*Hint*: $A = \cup_n B_n$, $B_1 = A_1, B_2 = A_1^c \cap A_2, \ldots, B_n = A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n$, so that $A_n = \cup_{j=1}^n B_j$.]

(vi) Show that the property: $\mu(A_n) \downarrow \mu(A)$ if $A_n \downarrow A$ holds for *finite* measures. Show by counterexample that it does not, in general, hold for measures μ that are not finite.

4. (*Bonferroni Inequalities*) Show that for odd $m \in \{1, 2, \ldots, n\}$, (a) $P(\cup_{j=1}^n A_j) \leq \sum_{k=1}^m \sum_{1 \leq j_1 \leq j_2 \leq \cdots \leq j_k \leq n} (-1)^{k+1} P(A_{j_1} \cap \cdots \cap A_{j_k})$, and for even $m \in \{2, \ldots, n\}$, (b) $P(\cup_{j=1}^n A_j) \geq \sum_{k=1}^m \sum_{1 \leq j_1 \leq j_2 \leq \cdots \leq j_k \leq n} (-1)^{k+1} P(A_{j_1} \cap \cdots \cap A_{j_k})$.

5. Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space and suppose $A, B \in \mathcal{F}$ are *independent* events, i.e., $P(A \cap B) = P(A)P(B)$, and $P(A) \geq \frac{1}{2} \leq P(B)$. Show that $P(A \cup B) \geq \frac{3}{4}$.

6. Show that the Borel σ-field of $\mathbb{R}$ is generated by any one of the following classes of sets: (i) $\mathcal{C} = \{(a, b) : -\infty \leq a \leq b \leq \infty\}$; (ii) $\mathcal{C} = \{(a, b] : -\infty \leq a \leq b < \infty\}$; (iii) $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\}$.

7. In each case below, show that ρ is a metric for the indicated topology.
   (i) For $S = \mathbb{R}^\infty$, $\rho(x, y) = \sum_{k=1}^\infty 2^{-k} |x_k - y_k| / (1 + |x_k - y_k|)$, for $x = (x_1, x_2, \ldots)$, $y = (y_1, y_2, \ldots) \in \mathbb{R}^\infty$ metrizes the topology of pointwise convergence: $x^{(n)} \to x$ if and only if $x_k^{(n)} \to x_k$ for each k, as $n \to \infty$.
   (ii) For $S = C[0, 1]$, $\rho(f, g) = \max\{|f(x) - g(x)| : x \in [0, 1]\}$ metrizes the topology of uniform convergence of continuous functions on $[0, 1]$.
   (iii) For $S = C([0, \infty) \to \mathbb{R}^k)$, $\rho(f, g) = \sum_{n=1}^\infty 2^{-n} \|f - g\|_n / (1 + \|f - g\|_n)$, where $\|f - g\|_n := \max\{\|f(x) - g(x)\| : x \in [0, n]\}$, $\|\cdot\|$ denoting the Euclidean norm on $\mathbb{R}^k$, metrizes the topology of uniform convergence on compacts.

8. Let X be a random map on $(\Omega, \mathcal{F}, P)$ with values in a measurable space $(S, \mathcal{S})$. Show that $\mathcal{G} := \{[X \in A] : A \in \mathcal{S}\}$ is the smallest sub-σ-field of $\mathcal{F}$ such that $X : \Omega \to S$ is a random map on $(\Omega, \mathcal{G})$, i.e., such that $[X \in A] \in \mathcal{G}$ for all $A \in \mathcal{S}$.

9. Let $\Omega = \{(1, 1), (2, 2), (1, 2), (2, 1)\}$ equipped with the power set $\mathcal{F}$. Define a simple random variable by $X(\omega) = \omega_1 + \omega_2, \quad \omega = (\omega_1, \omega_2) \in \Omega$. Give an explicit description of $\sigma(X)$ as a subcollection of sets in $\mathcal{F}$ and give an example of a set in $\mathcal{F}$ that is not in $\sigma(X)$.

10. (i) Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{P} = \{A_1, A_2, \ldots, A_m\}$, $\emptyset \neq A_j \in \mathcal{F}$, $1 \leq j \leq m$, be a disjoint partition of Ω. Let $(S, \mathcal{S})$ be an arbitrary measurable space such that $\mathcal{S}$ contains all of the singleton sets $\{x\}$ for $x \in S$. Show that a random map $X : \Omega \to S$ is $\sigma(\mathcal{P})$-measurable if and only if X is a $\sigma(\mathcal{P})$-measurable simple function. Give a counterexample in the case that $\mathcal{S}$ does not contain singletons. (ii) Let $A_1, \ldots, A_k$ be nonempty subsets of Ω. Describe the smallest σ-field containing $\{A_1, \ldots, A_k\}$ and show that its cardinality is at most 2^k.

11. Give a proof of the change of variables formula. [*Hint*: (Method of simple function approximation) Begin with h an indicator function, then h a simple function, then $h \geq 0$, and finally write $h = h^+ - h^-$.]

12. Let X_1, X_2 be real-valued random variables on $(\Omega, \mathcal{F}, P)$. Suppose that $F_i(x) = P(X_i \leq x)$, $x \in \mathbb{R} (i = 1, 2)$ are two distribution functions on $(\mathbb{R}, \mathcal{B})$ and $F_1 = F_2$. Show that X_1 and X_2 have the same distribution.

13. Suppose that X_1 and X_2 are two bounded real-valued random variables on $(\Omega, \mathcal{F}, P)$ such that $\mathbb{E}X_1^m = \mathbb{E}X_2^m$, $m = 1, 2, \ldots$. Show that X_1 and X_2 must have the same distribution. [*Hint*: According to the Weierstrass approximation theorem, a continuous function on a closed and bounded interval may be approximated by polynomials uniformly over the interval (see Appendix B).]

14. (i) Show that for a convex function φ on an open interval J, φ^- is finite and nondecreasing, and the "line of support" property holds with $m = \varphi^-(x_0)$, as well as with any $m \in [\varphi^-(x_0), \varphi^+(x_0)]$. (ii) Show that while a convex φ is continuous on an open interval, it need not be so on an interval with left-hand and/or right-hand endpoints. (iii) Show that if φ has a continuous, nondecreasing derivative φ' on J, then φ is convex. In particular, if φ is twice differentiable and $\varphi'' \geq 0$ on J, then φ is convex. [*Hint*: Use the mean value theorem from calculus.]

15. Let $p \geq 1, X_n \in L^m(\Omega, \mathcal{F}, P)$ for some $m > p$. Suppose there is an M such that $\mathbb{E}|X_n|^m \leq M, \forall n \geq 1$. Show that $\{|X_n|^p\}_{n=1}^\infty$ is uniformly integrable. [*Hint*: Use a Chebyshev-type inequality.]

16. Let X be a nonnegative random variable. (i) Show that $nP(X > n) \to 0$ as $n \to \infty$ if $\mathbb{E}X < \infty$. [*Hint*: $nP(X > n) \leq \mathbb{E}X\mathbf{1}_{[X>n]}$.] (ii) Prove that $\sum_{n=1}^\infty P(X > n) \leq \mathbb{E}X \leq \sum_{n=0}^\infty P(X > n)$. [*Hint*: $\sum_{n=1}^\infty (n-1)P(n-1 < X \leq n) \leq \mathbb{E}X \leq \sum_{n=1}^\infty nP(n-1 < X \leq n)$.]

17. Let $\{f_n : n \geq 1\}$ be a Cauchy sequence in measure: $\mu(|f_n - f_m| > \varepsilon) \to 0$ as $n, m \to \infty$, $\forall \varepsilon > 0$. Prove that there exists a measurable function f such that $f_n \to f$ in measure. [*Hint*: Find a sequence $n_1 < n_2 < \cdots$ such that $\mu(|f_{n_k} - f_{n_{k+1}}| > 2^{-k}) < 2^{-k}, k = 1, 2, \ldots$. Let $B = [|f_{n_k} - f_{n_{k+1}}| > 2^{-k} \textit{ i.o.}]$, and show that $\mu(B) = 0$. On B^c, $\{f_{n_k}\}_{k=1}^\infty$ is a Cauchy sequence, converging to some function f. Also for every $\varepsilon > 0$, $\mu(|f_n - f| > \varepsilon) \leq \mu(|f_n - f_{n_k}| > \varepsilon/2) + \mu(|f_{n_k} - f| > \varepsilon/2)$. The first term on the right of this inequality is $o(1)$ as $k \to \infty, n \to \infty$. Also, outside $B_k := \cup_{m=k}^\infty [|f_{n_m} - f_{n_{m+1}}| > 2^{-m}]$, one has $|f_{n_k} - f| \leq \sum_{m=k}^\infty 2^{-m} = 2^{-(k-1)}$. By choosing k_0 such that $2^{-(k_0-1)} < \varepsilon/2$, one gets $\mu(|f_{n_k} - f| > \varepsilon/2) \leq \mu(B_{k_0}) \leq \varepsilon/2$ for all $k \geq k_0$.]

18. Show that for every $p \geq 1$, $L^p(S, \mathcal{S}, \mu)$ is a complete metric space.

19. (*Integration of Complex-Valued Functions*) A Borel measurable function $f = g + ih$ on a measure space $(S, \mathcal{S}, \mu)$ into $\mathbb{C}$, (i.e., g, h are real-valued Borel-measurable), is said to be *integrable* if its real and imaginary parts g and h are both integrable. Since $2^{-\frac{1}{2}}(|g| + |h|) \leq |f| \equiv \sqrt{g^2 + h^2} \leq |g| + |h|$, f is integrable if and only if $|f|$ is integrable. The following extend a number of standard results for measurable real-valued functions to measurable complex-valued functions.
    (a) Extend Lebesgue's dominated convergence theorem (Appendix A) to complex-valued functions.
    (b) Extend the inequalities of Lyapounov, Hölder, Minkowski, and Markov–Chebyshev (Theorem 1.5(b),(c),(e),(f)) to complex-valued functions.
    (c) For $p \geq 1$, let the L^p-space of complex-valued functions be defined by equivalence classes of complex-valued functions f induced by equality a.e. such that $|f|^p$ is integrable. Show that this L^p-space is a Banach space over the field of complex numbers with norm $\|f\|_p = (\int_S |f|^p d\mu)^{\frac{1}{p}}$.
    (d) Show that the L^2-space of complex-valued square-integrable functions is a Hilbert space with inner product $\langle f_1, f_2 \rangle = \int_S f_1 \overline{f}_2 \, d\mu$, where $\overline{f}_2$ is the complex conjugate of f_2.
    (e) Show that for the special case of real-valued functions, the L^p-norm defined above reduces to that introduced in the text.

20. Suppose that $X_1, X_2, \ldots$ is a sequence of identically distributed random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. Show that if $\mathbb{E}e^{|X_1|} < \infty$, then a.s. $\limsup_{n\to\infty} \frac{|X_n|}{\ln n} \leq 1$.

CHAPTER II

Independence, Conditional Expectation

A **stochastic process** $\{X_t : t \in \Lambda\}$ on a probability space $(\Omega, \mathcal{F}, P)$ with values in a (measurable) space $(S, \mathcal{S})$ is a family of random maps $X_t : \Omega \to S$, $t \in \Lambda$. The index set Λ is most often of one of the following types: (i) $\Lambda = \{0, 1, 2, \ldots\}$. Then $\{X_t : t = 0, 1, 2, \ldots\}$ is referred to as a **discrete-parameter stochastic process**, usually with $S = \mathbb{R}$ or $\mathbb{R}^k$. (ii) $\Lambda = [0, \infty)$. Then $\{X_t : t \geq 0\}$ is called a **continuous-parameter stochastic process**, usually with $S = \mathbb{R}$ or $\mathbb{R}^k$.

Given an arbitrary collection of sets S_t, $t \in \Lambda$, the **product space**, denoted by $S = \prod_{t \in \Lambda} S_t \equiv \times_{t \in \Lambda} S_t$, is defined as the space of functions $\mathbf{x} = (x_t, t \in \Lambda)$ mapping Λ to $\cup_{t \in \Lambda} S_t$ such that $x_t \in S_t$ for each $t \in \Lambda$. This general definition applies to cases in which Λ is finite, countably infinite, or a continuum. In the case that each $S_t, t \in \Lambda$, is also a measurable space with respective σ-fields $\mathcal{S}_t$, the **product σ-field**, denoted by $\otimes_{t \in \Lambda} \mathcal{S}_t$, is defined as the σ-field generated by the collection $\mathcal{C}$ of **finite-dimensional rectangles** of the form $C = \{\mathbf{x} \in \prod_{t \in \Lambda} S_t : (x_{t_1}, \ldots, x_{t_k}) \in B_1 \times \cdots \times B_k\}$, for $k \geq 1, B_i \in \mathcal{S}_{t_i}, 1 \leq i \leq k$. Alternatively, the product σ-field is the smallest σ-field of subsets of $\prod_{t \in \Lambda} S_t$ which makes each of the **coordinate projections**, $X_s(\mathbf{x}) = x_s, \mathbf{x} \in \prod_{t \in \Lambda} S_t, s \in \Lambda$, a measurable map. In this case the pair $(S = \prod_{t \in \Lambda} S_t, \otimes_{t \in \Lambda} \mathcal{S}_t)$ will also be referred to as the (measure-theoretic) **product space.**

Recall that a **field** is a collection of subsets of Ω closed under complements and finite unions, and contains the empty set. A nonnegative set function μ defined on a field $\mathcal{F}_0$ of subsets of Ω is called a **measure** *defined on this field* if $\mu(\emptyset) = 0$ and $\mu(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu(A_n)$ whenever $A_1, A_2, \ldots$ is a disjoint sequence of members of $\mathcal{F}_0$ such that $\cup_{n=1}^\infty A_n \in \mathcal{F}_0$. A measure μ is said to be **σ-finite on $\mathcal{F}_0$** provided that $\Omega = \cup_{n=1}^\infty A_n$ for some $A_n \in \mathcal{F}_0, n \geq 1$, such that $\mu(A_n) < \infty$ for all n. Given a measure μ on a field $\mathcal{F}_0$ of subsets of Ω, one may define an extension of μ, denoted by μ^*, as a nonnegative function on all $A \subseteq \Omega$, by Carathéodory's extension formula

$$\mu^*(A) = \inf\{\sum_n \mu(A_n) : A_n \in \mathcal{F}_0, n \geq 1, \cup_n A_n \supseteq A\}. \tag{2.1}$$

While the set function μ^* defined for all subsets of Ω by (2.1) is generally not countably additive, the Carathéodory method of restricting μ^* to a smaller σ-field than the power set of Ω provides an important extension of μ to a measure defined on the smallest σ-field $\mathcal{F} = \sigma(\mathcal{F}_0)$ containing $\mathcal{F}_0$; see Appendix A.

Given a finite number of σ-finite measure spaces $(S_i, \mathcal{S}_i, \mu_i), i = 1, \ldots, n$, one may uniquely determine a measure $\mu_1 \times \cdots \times \mu_n$, called the **product measure** on the product space $(S_1 \times \cdots \times S_n, \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n)$, by prescribing that

$$\mu_1 \times \cdots \times \mu_n(B_1 \times \cdots \times B_n) := \prod_{i=1}^n \mu_i(B_i), B_i \in \mathcal{S}_i, \quad (1 \leq i \leq n). \tag{2.2}$$

In particular, for $A \in \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n$ one has according to the Carathéodory extension formula (see Appendix A)

$$\mu_1 \times \cdots \times \mu_n(A) = \inf \sum_{j=1}^{\infty} \mu_1(B_{1j}) \times \cdots \times \mu_n(B_{nj}), \tag{2.3}$$

where the infimum is taken over all *covers* $\cup_{j=1}^{\infty} B_{1j} \times \cdots \times B_{nj} \supseteq A$, by $B_{ij} \in \mathcal{S}_i, 1 \leq i \leq n, j \geq 1$. *Associativity* of product measure, i.e., $\mu_1 \times \mu_2 \times \mu_3 = (\mu_1 \times \mu_2) \times \mu_3 = \mu_1 \times (\mu_2 \times \mu_3)$, is another important consequence of the uniqueness of the product measure, that requires the property of σ-finiteness, and will be assumed throughout without further mention.

A central notion in probability is that of independence. Given a finite set of random variables (maps) $X_1, X_2, \ldots, X_n$, with X_i a measurable map on $(\Omega, \mathcal{F}, P)$ into $(S_i, \mathcal{S}_i)$ $(1 \leq i \leq k)$, the X_i $(1 \leq i \leq n)$, are said to be **independent** if the distribution Q of $X := (X_1, X_2, \ldots, X_n)$ on the product space $(S = S_1 \times S_2 \times \cdots \times S_n$, $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2 \otimes \cdots \otimes \mathcal{S}_n)$ is the product measure $Q = Q_1 \times Q_2 \times \cdots \times Q_n$, where Q_i is the distribution of X_i $(1 \leq i \leq n)$. In other words, $X_1, X_2, \ldots, X_n$ are independent iff $\forall B_i \in \mathcal{S}_i, 1 \leq i \leq n$,

$$Q(B_1 \times B_2 \times \cdots \times B_n) \equiv P(X_i \in B_i, 1 \leq i \leq n) = \prod_{i=1}^n P(X_i \in B_i) \equiv \prod_{i=1}^n Q_i(B_i). \tag{2.4}$$

This formulation of independence is particularly amenable to the use of the Fubini–Tonelli theorem for integration over product spaces in terms of iterated integrals. The following important application of Fubini–Tonelli is left as Exercise 4. Also see Exercise 5 for applications to sums of independent exponentially distributed random variables and Gaussian random variables.

Theorem 2.1. Suppose $\mathbf{X}_1, \mathbf{X}_2$ are independent k-dimensional random vectors having distributions Q_1, Q_2, respectively. The distribution of $\mathbf{X}_1 + \mathbf{X}_2$ is given by the **convolution** of Q_1 and Q_2:

$$Q_1 * Q_2(B) = \int_{\mathbb{R}^k} Q_1(B - y) Q_2(dy), \quad B \in \mathcal{B}^k,$$

where $B - y := \{x - y : x \in B\}$.

Observe that any subcollection of independent random variables will be independent. In particular, pairs of random variables will be independent. The converse is not true (Exercise 2). One may also observe that the definition of independence implies that the factors of the indicated joint distribution are the **marginal distributions** $Q_i = P \circ X_i^{-1}$, $i = 1, \ldots, n$, of the respective random variables comprising the vector $(X_1, \ldots, X_n)$ (Exercise 2).

In practice, one typically applies the Tonelli part to $|f|$ in order to determine whether the Fubini part is applicable to f. Let us record a useful formula for the moments of a random variable derived from the Fubini-Tonelli theorem before proceeding. Namely, if X is a random variable on $(\Omega, \mathcal{F}, P)$, then for any $p > 0$, $x \geq 0$, writing $x^p = p\int_0^x y^{p-1}dy$ in the formula $\mathbb{E}|X|^p = \int_\Omega |X(\omega)|^p P(d\omega)$ and applying the Tonelli part, one obtains

$$\mathbb{E}|X|^p = \int_\Omega \left(p \int_0^{|X(\omega)|} y^{p-1} dy \right) P(d\omega) = p \int_0^\infty y^{p-1} P(|X| > y) dy. \tag{2.5}$$

The following result is an important consequence of independence.

Theorem 2.2. If $X_1, \ldots, X_n$ are independent random variables on $(\Omega, \mathcal{F}, P)$ such that $\mathbb{E}|X_j| < \infty$, $1 \leq j \leq n$, then $\mathbb{E}|X_1 \cdots X_n| < \infty$ and

$$\mathbb{E}(X_1 \cdots X_n) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_n).$$

Proof. Let $Q_j = P \circ X_j^{-1}$, $j \geq 1$. Since by independence, $(X_1, \ldots, X_n)$ has product measure as joint distribution, one may apply a change of variables and the Tonelli part to obtain

$$\begin{aligned} \mathbb{E}|X_1 \cdots X_n| &= \int_\Omega |X_1 \cdots X_n| dP \\ &= \int_{\mathbb{R}^n} |x_1 \cdots x_n| Q_1 \times \cdots \times Q_n(dx_1 \times \cdots \times dx_n) \\ &= \prod_{j=1}^n \int_{\mathbb{R}} |x_j| Q_j(dx_j) = \prod_{j=1}^n \mathbb{E}|X_j| < \infty. \end{aligned}$$

With the integrability established one may apply the Fubini part to do the same thing for $\mathbb{E}(X_1 \cdots X_n)$ and the product measure distribution $P \circ X_1^{-1} \times \cdots \times P \circ X_n^{-1}$ of $(X_1, \ldots, X_n)$. ■

Two random variables X_1, X_2 in $L^2 = L^2(\Omega, \mathcal{F}, P)$ are said to be **uncorrelated** if their **covariance** $\operatorname{Cov}(X_1, X_2)$ is zero, where

$$\operatorname{Cov}(X_1, X_2) := \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2). \quad (2.6)$$

The **variance** $\operatorname{Var}(Y)$ of a random variable $Y \in L^2$ is defined by the average squared deviation of Y from its mean $\mathbb{E}Y$. That is, $\operatorname{Var}(Y) = \mathbb{E}(Y - \mathbb{E}Y)^2 = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$. The covariance term naturally appears in consideration of the variance of sums of random variables $X_j \in L^2(\Omega, \mathcal{F}, P), 1 \le j \le n$, i.e.,

$$\operatorname{Var}\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n \operatorname{Var}(X_j) + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j).$$

Note that if X_1 and X_2 are independent, then it follows from Theorem 2.2 that they are uncorrelated; but the converse is easily shown to be false.

Let $\{X_t : t \in \Lambda\}$ be a possibly infinite family of random maps on $(\Omega, \mathcal{F}, P)$, with X_t a measurable map into $(S_t, \mathcal{S}_t)$, $t \in \Lambda$. We will say that $\{X_t : t \in \Lambda\}$ is a **family of independent maps** if every finite subfamily is a family of independent maps. More precisely, for all $n \ge 1$ and for every n-tuple $(t_1, t_2, \ldots, t_n)$ of distinct points in Λ, the maps $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ are independent (in the sense of (2.4)). Given any family of probability measures Q_t (on $(S_t, \mathcal{S}_t)$), $t \in \Lambda$, one can construct a probability space $(\Omega, \mathcal{F}, P)$ on which are defined random maps X_t ($t \in \Lambda$) such that (i) X_t has distribution Q_t ($t \in \Lambda$) and (ii) $\{X_t : t \in \Lambda\}$ is a family of independent maps. Indeed, on the product space ($S \equiv \times_{t \in \Lambda} S_t, \mathcal{S} \equiv \otimes_{t \in \Lambda} \mathcal{S}_t$) there exists a product probability measure $Q = \prod_{t \in \Lambda} Q_t$; and one may take $\Omega = S$, $\mathcal{F} = \mathcal{S}$, $P = Q$, $X_t(\omega) = x_t$ for $\omega = (x_t, t \in \Lambda) \in S$. The existence of such product probability measures is postponed to Chapter X.

Let us briefly return to notions of uncorrelated and independent random variables. Although zero correlation is a weaker notion than statistical independence, using approximation by simple functions one may obtain the following characterization of independence (Exercise 12).

Proposition 2.3. A family of random maps $\{X_t : t \in \Lambda\}$ (with X_t a measurable map into $(S_t, \mathcal{S}_t)$) is an independent family if and only if for every pair of disjoint finite subsets Λ_1, Λ_2 of Λ, any random variable $V_1 \in L^2(\sigma\{X_t : t \in \Lambda_1\})$ is uncorrelated with any random variable $V_2 \in L^2(\sigma\{X_t : t \in \Lambda_2\})$

The important special case of a sequence $X_1, X_2, \ldots$ of **independent** and **identically distributed** random maps is referred to as an **i.i.d.** sequence. An example of the construction of an i.i.d. (coin tossing) sequence $\{X_n\}_{n=1}^\infty$ of Bernoulli valued random variables with values in $\{0,1\}$ and defined on a probability space $(\Omega,\mathcal{F},P)$ with prescribed distribution $P(X_1=1)=p=1-P(X_1=0)$, for given $p\in[0,1]$, is given in Exercise 1. The general existence of infinite product measures will also follow as a special case of the **Kolmogorov extension theorem** proved in Chapter X in the case that $(S,\mathcal{S})$ has some extra topological structure; see Exercise 1 for a simple special case illustrating how one may exploit topological considerations. Existence of an infinite-product probability measure will also be seen to follow in full measure-theoretic generality from the **Tulcea extension theorem** discussed in Chapter X.

A collection $\mathcal{C}$ of events $A\in\mathcal{F}$ is defined to be a set of **independent events** if the set of indicator random variables $\{\mathbf{1}_A : A\in\mathcal{C}\}$ is an independent collection. The notion of independence may also be equivalently defined in terms of sub-σ-fields of $\mathcal{F}$. Given $(\Omega,\mathcal{F},P)$, a family $\{\mathcal{F}_t : t\in\Lambda\}$ of σ-fields (contained in $\mathcal{F}$) is a **family of independent σ-fields** if for every n-tuple of distinct indices $(t_1,t_2,\dots,t_n)$ in Λ one has $P(F_{t_1}\cap F_{t_2}\cap\cdots\cap F_{t_n})=P(F_{t_1})P(F_{t_2})\cdots P(F_{t_n})$ for all $F_{t_i}\in\mathcal{F}_{t_i}$ $(1\le i\le n)$; here n is an arbitrary finite integer, $n\le$ cardinality of Λ. Note that the independence of a family $\{X_t : t\in\Lambda\}$ of random maps is equivalent to the independence of the family $\{\sigma(X_t) : t\in\Lambda\}$ of σ-fields $\sigma(X_t)\equiv\{[X_t\in B] : B\in\mathcal{S}_t\}$ generated by $X_t (t\in\Lambda)$, where $(S_t,\mathcal{S}_t)$ is the image space of X_t. The σ-field formulation of independence can be especially helpful in tracking independence, as illustrated by the following two results (Exercise 11).

It is useful to note the following consequence of the $\pi-\lambda$ theorem.

Proposition 2.4. If $\{\mathcal{C}_t\}_{t\in\Lambda}$ is a family of π-systems such that $P(C_{t_1}\cap\cdots\cap C_{t_n})=\prod_{i=1}^{n}P(C_{t_i})$, $C_{t_i}\in\mathcal{C}_{t_i}$, for any distinct $t_i\in\Lambda, n\ge 2$, then $\{\sigma(\mathcal{C}_t)\}_{t\in\Lambda}$ is a family of independent σ-fields.

The simple example in which $\Omega=\{a,b,c,d\}$ consists of four equally probable outcomes and $\mathcal{C}_1=\{\{a,b\}\}$, $\mathcal{C}_2=\{\{a,c\},\{a,d\}\}$, shows that the π-system requirement is indispensable. For a more positive perspective, note that if $A,B\in\mathcal{F}$ are independent events then it follows immediately that A,B^c and A^c,B^c are respective pairs of independent events, since $\sigma(\{A\})=\{A,A^c,\emptyset,\Omega\}$ and similarly for $\sigma(\{B\})$.

Proposition 2.5. Let $X_1,X_2,\dots$ be a sequence of independent random maps with values in measurable spaces $(S_1,\mathcal{S}_1),(S_2,\mathcal{S}_2),\dots$, respectively, and let $n_1<n_2<\cdots$ be a sequence of positive integers. Suppose that $Y_1=f_1(X_1,\dots,X_{n_1}), Y_2=f_2(X_{n_1+1},\dots,X_{n_2}),\dots$, where $f_1,f_2,\dots$ are Borel-measurable functions on the respective product measure spaces $S_1\times\cdots\times S_{n_1}, S_{n_1+1}\times\cdots\times S_{n_2},\dots$. Then $Y_1,Y_2,\dots$ is a sequence of independent random variables.

Often one also needs the notion of independence of (among) several families of σ-fields or random maps. Let Λ_i, $i\in\mathcal{I}$, be a family of index sets and, for each $i\in\mathcal{I}$, $\{\mathcal{F}_t : t\in\Lambda_i\}$ a collection of (sub) σ-fields of $\mathcal{F}$. The **families** $\{\mathcal{F}_t : t\in\Lambda_i\}_{i\in\mathcal{I}}$ are said to be **independent** (of each other) if the σ-fields $\mathcal{G}_i := \sigma(\{\mathcal{F}_t : t\in\Lambda_i\})$ generated by $\{\mathcal{F}_t : t\in\Lambda_i\}$ (i.e., $\mathcal{G}_i$ is the smallest σ-field containing $\cup_{t\in\Lambda_i}\mathcal{F}_t, i\in\mathcal{I}$, also denoted by $\mathcal{G}_i = \bigvee_{t\in\Lambda_i} \mathcal{F}_t$), are independent in the sense defined above. The corresponding definition of **independence of (among) families of random maps** $\{X_t : t \in \Lambda_i\}_{i\in\mathcal{I}}$ can now be expressed in terms of $\mathcal{F}_t := \sigma(X_t)$, $t \in \Lambda_i$, $i \in \mathcal{I}$.

We will conclude the discussion of independence with a return to considerations of a converse to the Borel–Cantelli lemma I. Clearly, by taking $A_n = A_1 \forall n$, $P(A_n \ i.o.) = P(A_1) \in [0,1]$. So there is no general theorem without some restriction on how much dependence exists among the events in the sequence. Write A_n **eventually for all** n to denote the event $[A_n^c \ i.o.]^c$, i.e., "A_n occurs for all but finitely many n."

Lemma 1 *(Borel–Cantelli II).* Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of independent events in a probability space $(\Omega, \mathcal{F}, P)$. If $\sum_{n=1}^{\infty} P(A_n) = \infty$ then $P(A_n \ i.o. \ ) = 1$.

Proof. Consider the complementary event to get from continuity properties of P, independence of complements, and the simple bound $1-x \le e^{-x}$, that $1 \ge P(A_n \ i.o.) = 1 - P(A_n^c$ eventually for all $n) = 1 - P(\cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m^c) = 1 - \lim_{n\to\infty} \prod_{m=n}^{\infty} P(A_m^c) \ge 1 - \lim_{n\to\infty} \exp\{-\sum_{m=n}^{\infty} P(A_m)\} = 1$. ■

Example 1. Suppose that $\{X_n\}_{n=1}^{\infty}$ is an i.i.d. sequence of Bernoulli 0 or 1-valued random variables with $P(X_1 = 1) = p > 0$. Then $P(X_n = 1 \ i.o.) = 1$ is a quick and easy consequence of Borel–Cantelli II.

We now come to another basic notion of fundamental importance in probability—the notion of conditional probability and conditional expectation. Since we will need to consider $\mathcal{G}$-measurable "approximations" to random variables X, where $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$, it is useful to consider the spaces $L^p(\Omega, \mathcal{G}, P)$. A little thought reveals that an element of this last (Banach) space is not in general an element of $L^p(\Omega, \mathcal{F}, P)$. For if Z is $\mathcal{G}$-measurable, then the set (equivalence class) $\tilde{Z}$ of all $\mathcal{F}$-measurable random variables each of which differs from Z on at most a P-null set may contain random variables that are not $\mathcal{G}$-measurable. However, if we denote by $L^p(\mathcal{G})$ the set of all elements of $L^p(\Omega, \mathcal{F}, P)$, each equivalent to some $\mathcal{G}$-measurable Z with $\mathbb{E}|Z|^p < \infty$, then $L^p(\mathcal{G})$ becomes a closed linear subspace of $L^p(\Omega, \mathcal{F}, P)$. In particular, under this convention, $L^2(\mathcal{G})$ is a closed linear subspace of $L^2 \equiv L^2(\mathcal{F}) \equiv L^2(\Omega, \mathcal{F}, P)$, for every σ-field $\mathcal{G} \subseteq \mathcal{F}$. The first definition below exploits the Hilbert space structure of L^2 through the projection theorem (see Appendix C) to obtain the conditional expectation of X, given $\mathcal{G}$, as the orthogonal projection of X onto $L^2(\mathcal{G})$.

Definition 2.1. *(First Definition of Conditional Expectation (on L^2)).* Let $X \in L^2$ and $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Then a **conditional expectation of** X **given** $\mathcal{G}$, denoted by $\mathbb{E}(X|\mathcal{G})$, is a $\mathcal{G}$-measurable version of the orthogonal projection of X onto $L^2(\mathcal{G})$.

Intuitively, $\mathbb{E}(X|\mathcal{G})$ is the best prediction of X (in the sense of least mean square error), given information about the experiment coded by events that constitute $\mathcal{G}$. In the case $\mathcal{G} = \sigma\{Y\}$ is a random map with values in a measurable space $(S, \mathcal{S})$, this makes $\mathbb{E}(X|\mathcal{G})$ a version of a Borel measurable function of Y. This is because of the following more general fact.

Proposition 2.6. Let $Z, Y_1, \ldots, Y_k$ be real-valued random variables on a measurable space $(\Omega, \mathcal{F})$. A random variable $Z : \Omega \to \mathbb{R}$ is $\sigma(Y_1, \ldots, Y_k)$-measurable iff there is a Borel measurable function $g : \mathbb{R}^k \to \mathbb{R}$ such that $Z = g(Y_1, \ldots, Y_k)$.

Proof. If $Z = g(Y_1, \ldots, Y_k)$, then $\sigma(Y_1, \ldots, Y_k)$-measurability is clear, since for $B \in \mathcal{B}(\mathbb{R})$, $[Z \in B] = [(Y_1, \ldots, Y_k) \in g^{-1}(B)]$ and $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^k)$ for Borel measurable g.

For the converse, suppose that Z is a simple $\sigma(Y_1, \ldots, Y_k)$-measurable random variable with *distinct* values $z_1, \ldots, z_m$. Then $[Z = z_j] \in \sigma(Y_1, \ldots, Y_k)$ implies that there is a $B_j \in \mathcal{B}(\mathbb{R}^k)$ such that $[Z = z_j] = [(Y_1, \ldots, Y_k) \in B_j]$ and $Z = \sum_{j=1}^k f_j(Y_1, \ldots, Y_k)$, where $f_j(y_1, \ldots, y_k) = z_j \mathbf{1}_{B_j}(y_1, \ldots, y_k)$, so that $Z = g(Y_1, \ldots, Y_k)$ with $g = \sum_{j=1}^k f_j$. More generally, one may use approximation by simple functions to write $Z(\omega) = \lim_{n\to\infty} Z_n(\omega)$, for each $\omega \in \Omega$, where Z_n is a $\sigma(Y_1, \ldots, Y_k)$-measurable simple function, $Z_n(\omega) = g_n(Y_1(\omega), \ldots, Y_k(\omega))$, $n \geq 1$, $\omega \in \Omega$. In particular, $g(y_1, \ldots, y_k) = \lim_{n\to\infty} g_n(y_1, \ldots, y_k)$ exists for each $(y_1, \ldots, y_k)$ in the range of $(Y_1, \ldots, Y_k)$. But since each g_n is zero off the range, the limit exists and defines g on all of $\mathbb{R}^k$. ■

As simple examples, consider the sub-σ-fields $\mathcal{G}_0 = \{\Omega, \mathcal{F}\}$, $\sigma(X)$, and $\mathcal{F}$. (The σ-field $\mathcal{G}_0$, or the one comprising only P-null sets and their complements, is called the **trivial σ-field**). One has for all $X \in L^2$,

$$\mathbb{E}(X|\mathcal{G}_0) = \mathbb{E}(X), \qquad \mathbb{E}(X|\sigma(X)) = X, \qquad \mathbb{E}(X|\mathcal{F}) = X. \tag{2.7}$$

The first of these follows from the facts that (i) the only $\mathcal{G}_0$-measurable functions are constants, and (ii) $\mathbb{E}(X - C)^2$ is minimized, uniquely, by the constant $C = \mathbb{E}X$. The other two relations in (2.7) are obvious from the definition.

In other words, if $X \in L^2$, then the orthogonal projection of X onto $1^\perp \equiv \{Y \in L^2 : Y \perp 1\} = \{Y \in L^2 : \mathbb{E}Y = 0\}$ is given by $X - \mathbb{E}(X)$, or equivalently, the projection of X onto the space of (equivalence classes of) constants is $\mathbb{E}(X)$. Thus in particular, X and Y ($\in L^2$) are uncorrelated if and only if their projections onto $1^\perp$ are orthogonal.

In addition to the intuitive interpretation of $\mathbb{E}(X|\mathcal{G})$ as a best predictor of X, there is also an interpretation based on smoothing in the sense of averages that extends beyond L^2. For example, as noted above, $\mathbb{E}(X|\{\emptyset, \Omega\}) = \mathbb{E}X = \int_\Omega X(\omega)P(d\omega)$. In particular, this may be viewed as a smoothing of the function X over all sample points $\omega \in \Omega$. Similarly, for $B \in \mathcal{F}$, $0 < P(B) < 1$, for $X \in L^2$, one may check that (Exercise 17)

$$\mathbb{E}(X|\{\emptyset, B, B^c, \Omega\}) = \left(\frac{1}{P(B)} \int_B X dP\right) \mathbf{1}_B + \left(\frac{1}{P(B^c)} \int_{B^c} X dP\right) \mathbf{1}_{B^c}. \tag{2.8}$$

It may be noted that the conditional expectation is well defined only up to a $\mathcal{G}$-measurable P-null set. That is, if X is a version of $\mathbb{E}(X|\mathcal{G})$, then so is any $\mathcal{G}$-measurable Y such that $P(Y \neq X) = 0$. Thus the conditional expectation $\mathbb{E}(X|\mathcal{G})$ is uniquely defined only as an element of $L^2(\Omega, \mathcal{G}, P)$. We will, however, continue to regard $\mathbb{E}(X|\mathcal{G})$ as a $\mathcal{G}$-measurable version of the orthogonal projection of X onto $L^2(\mathcal{G})$. The orthogonality condition is expressed by

$$\int_\Omega (X - \mathbb{E}(X|\mathcal{G}))Y dP = 0 \qquad \forall\, Y \in L^2(\Omega, \mathcal{G}, P), \tag{2.9}$$

or

$$\int_\Omega XY dP = \int_\Omega \mathbb{E}(X|\mathcal{G})Y dP \qquad \forall\, Y \in L^2(\Omega, \mathcal{G}, P). \tag{2.10}$$

In particular, with $Y = \mathbf{1}_G$ for $G \in \mathcal{G}$ in (2.10), one has

$$\int_G X dP = \int_G \mathbb{E}(X|\mathcal{G}) dP \qquad \forall\, G \in \mathcal{G}. \tag{2.11}$$

It is simple to check that for $X \in L^2(\Omega, \mathcal{F}, P)$, (2.11) is equivalent to (2.9) (or (2.10)). But (2.11) makes sense for all $X \in L^1(\Omega, \mathcal{F}, P)$, which leads to the second, more general, definition.

Definition 2.2. *(Second Definition of Conditional Expectation (on L^1)).* Let $X \in L^1(\Omega, \mathcal{F}, P)$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A $\mathcal{G}$-measurable random variable is said to be a **conditional expectation of X given** $\mathcal{G}$, denoted by $\mathbb{E}(X|\mathcal{G})$, if (2.11) holds.

That $\mathbb{E}(X|\mathcal{G})$ exists for $X \in L^1$, and is well defined a.e., may be proved by letting $X_n \in L^2$ converge to X in L^1 (i.e., $\|X_n - X\|_1 \to 0$ as $n \to \infty$), applying (2.11) to X_n, and letting $n \to \infty$. Note that L^2 is dense in L^1 (Exercise 26). Alternatively, one may apply the Radon–Nikodym theorem to the finite (signed) measure $\nu(G) := \int_G X\, dP$ on $(\Omega, \mathcal{G})$, which is absolutely continuous with respect to P (on $(\Omega, \mathcal{G})$), i.e., if $P(G) = 0$, then $\nu(G) = 0$. Hence there exists a $\mathcal{G}$-measurable function, say $\mathbb{E}(X|\mathcal{G})$, such that (2.11) holds. Viewed as an element of $L^1(\Omega, \mathcal{G}, P)$, $\mathbb{E}(X|\mathcal{G})$ is unique.

There are variations on the requirement (2.11) in the definition of conditional expectation that may be noted. In particular, a version of $\mathbb{E}(X|\mathcal{G})$ is uniquely determined by the condition that it be a $\mathcal{G}$-measurable random variable on Ω satisfying the equivalent version

$$\mathbb{E}\{Xg\} = \mathbb{E}\{\mathbb{E}(X|\mathcal{G})g\} \quad \forall g \in \Gamma, \tag{2.12}$$

of (2.11), where Γ is the set of indicator random variables $\{\mathbf{1}_B : B \in \mathcal{G}\}$, or, by simple function approximation, Γ may alternatively be taken to be (i) the collection of all bounded nonnegative $\mathcal{G}$-measurable random variables g on Ω or (ii) the collection of all bounded $\mathcal{G}$ measurable random variables g on Ω, for example, as convenient.

The following properties of $\mathbb{E}(X|\mathcal{G})$ are important and, for the most part, immediate consequences of the definitions.

Theorem 2.7. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $L^1 = L^1(\Omega, \mathcal{F}, P)$, $\mathcal{G}, \mathcal{D}$ sub-σ-fields of $\mathcal{F}$, $X, Y \in L^1$. Then the following hold almost surely (P):

(a) $\mathbb{E}(X|\{\Omega, \phi\}) = \mathbb{E}(X)$.
(b) $\mathbb{E}[\mathbb{E}(X|\mathcal{G})] = \mathbb{E}(X)$.
(c) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$.
(d) *(Linearity).* $\mathbb{E}(cX + dY|\mathcal{G}) = c\mathbb{E}(X|\mathcal{G}) + d\mathbb{E}(Y|\mathcal{G})$ for all constants c, d.
(e) *(Order).* If $X \leq Y$ a.s., then $\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(Y|\mathcal{G})$.
(f) *(Smoothing).* If $\mathcal{D} \subseteq \mathcal{G}$, then $\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{D}] = \mathbb{E}(X|\mathcal{D})$.
(g) *(Conditional Jensen's Inequality).* Let ψ be a convex function on an interval J such that ψ has finite right- (or left-)hand derivative(s) at left (or right) endpoint(s) of J if J is not open. If $P(X \in J) = 1$, and if $\psi(X) \in L^1$, then

$$\psi(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(\psi(X)|\mathcal{G}). \tag{2.13}$$

(h) *(Contraction).* For $X \in L^p(\Omega, \mathcal{F}, P)$, $p \geq 1$, $\|\mathbb{E}(X|\mathcal{G})\|_p \leq \|X\|_p \ \forall \ p \geq 1$.
(i) *(Convergences).*

(i1) If $X_n \to X$ in L^p then $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$ in L^p $(p \geq 1)$.
(i2) *(Conditional Monotone Convergence)* If $0 \leq X_n \uparrow X$ a.s., X_n and $X \in L^1$ $(n \geq 1)$, then $\mathbb{E}(X_n|\mathcal{G}) \uparrow \mathbb{E}(X|\mathcal{G})$ a.s. and $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$ in L^1.
(i3) *(Conditional Dominated Convergence)* If $X_n \to X$ a.s. and $|X_n| \leq Y \in L^1$, then $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G}) a.s.$

(j) If $XY \in L^1$ and X is $\mathcal{G}$-measurable, then $\mathbb{E}(XY|\mathcal{G}) = X\mathbb{E}(Y|\mathcal{G})$.
(k) If $\sigma(X)$ and $\mathcal{G}$ are independent then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.
(ℓ) *(Substitution Property)* Let U, V be random maps into $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, respectively. Let ψ be a measurable real-valued function on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$. If U is $\mathcal{G}$-measurable, $\sigma(V)$ and $\mathcal{G}$ are independent, and $\mathbb{E}|\psi(U, V)| < \infty$, then one has that $\mathbb{E}[\psi(U, V)|\mathcal{G}] = h(U)$, where $h(u) := \mathbb{E}\psi(u, V)$.

Proof. (a–f) follow easily from the definitions; in the case of (e) take $G = [\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(Y|\mathcal{G})] \in \mathcal{G}$ in the definition (2.11) of conditional expectation with X replaced by $Y - X$. For (g) use the line of support Lemma 2 from Chapter I. If J does not have a right endpoint, take $x_0 = \mathbb{E}(X|\mathcal{G})$, and $m = \psi^+(\mathbb{E}(X|\mathcal{G}))$, where ψ^+ is the right-hand derivative of ψ, to get $\psi(X) \geq \psi(\mathbb{E}(X|\mathcal{G})) + \psi^+(\mathbb{E}(X|\mathcal{G}))(X - \mathbb{E}(X|\mathcal{G}))$. Now take the conditional expectation, given $\mathcal{G}$, and use (e) to get (g). Similarly, if J does not have a left endpoint, take $m = \psi^-(\mathbb{E}(X|\mathcal{G}))$. If J has both right and left endpoints, say $a < b$, let $m = \psi^+(\mathbb{E}(X|\mathcal{G}))$ on $[\mathbb{E}(X|\mathcal{G}) \neq b]$ and $m = \psi^-(\mathbb{E}(X|\mathcal{G}))$ on $[\mathbb{E}(X|\mathcal{G}) \neq a]$.

The contraction property (h) follows from this by taking $\psi(x) = |x|^p$ in (2.13), and then taking expectations on both sides. The first convergence in (i) follows from (h) applied to $X_n - X$. The second convergence in (i) follows from the order property (e), and the monotone convergence theorem. The L^1 convergence in (i3) follows from (i1). For the a.s. convergence in (i3), let $Z_n := \sup\{|X_m - X| : m \geq n\}$. Then $Z_n \leq |X| + |Y|$, $|X| + |Y| - Z_n \uparrow |X| + |Y|$ a.s., so that by (i2), $\mathbb{E}(|X| + |Y| - Z_n|\mathcal{G}) \uparrow \mathbb{E}(|X| + |Y||\mathcal{G})$ a.s. Hence $\mathbb{E}(Z_n|\mathcal{G}) \downarrow 0$ a.s., and by (e), $|\mathbb{E}(X_n|\mathcal{G}) - \mathbb{E}(X|\mathcal{G})| \leq \mathbb{E}(|X_n - X||\mathcal{G}) \leq \mathbb{E}(Z_n|\mathcal{G}) \to 0$ a.s.

To prove (j), first consider the case of bounded $X \geq 0$. Let $g \in \Gamma$, the set of bounded, nonnegative $\mathcal{G}$-measurable random variables. Then $Xg \in \Gamma$, so that $\mathbb{E}(g\mathbb{E}(XY|\mathcal{G})) = \mathbb{E}(gXY) = \mathbb{E}(gX\mathbb{E}(Y|\mathcal{G}))$. Next, for $X \geq 0, X \in L^1$, apply this to $X_n = X\mathbf{1}_{[X \leq n]}$ and use $XY \in L^1$ and Lebesgue's dominated convergence theorem to get for $g \in \Gamma$, $\mathbb{E}(XYg) = \lim_{n\to\infty} \mathbb{E}(X_nYg) = \lim_{n\to\infty} \mathbb{E}(\mathbb{E}(X_nY|\mathcal{G})g) = \lim_{n\to\infty} \mathbb{E}(X_n\mathbb{E}(Y|\mathcal{G})g) = \mathbb{E}(X\mathbb{E}(Y|\mathcal{G})g)$. For the general case write $X = X^+ - X^-$ and use linearity (d).

To prove (k), again let $g \in \Gamma$ be a bounded, nonnegative $\mathcal{G}$-measurable random variable. By independence of $\sigma(X)$ and $\mathcal{G}$, one has, using Theorem 2.2, $\mathbb{E}(gX) = \mathbb{E}(g)\mathbb{E}(X) = \mathbb{E}(g\mathbb{E}(X))$. Since the constant $\mathbb{E}(X)$ is $\mathcal{G}$-measurable, indeed, constant random variables are measurable with respect to any σ-field, (k) follows by the defining property (2.12).

If one takes $\mathcal{G} = \sigma(U)$, then (ℓ) follows by the Fubini–Tonelli theorem (if one uses the change of variables formula to do integrations on the product space $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2, Q_1 \times Q_2)$, where Q_1, Q_2 are the distributions of U and V, respectively). For the general case, first consider ψ of the form $\psi(u, v) = \sum_{i=1}^n f_i(u)g_i(v)$ with f_i and g_i bounded and measurable (on $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, respectively), $1 \leq i \leq n$. In this case, for every $G \in \mathcal{G}$, one has $h(U) = \sum_{i=1}^n f_i(U)\mathbb{E}g_i(V)$, and

$$\begin{aligned}
\int_G \psi(U,V)dP &\equiv \mathbb{E}\left(\mathbf{1}_G \sum_{i=1}^n f_i(U)g_i(V)\right) \\
&= \sum_{i=1}^n \mathbb{E}(\mathbf{1}_G f_i(U) \cdot g_i(V)) = \sum_{i=1}^n \mathbb{E}(\mathbf{1}_G f_i(U)) \cdot \mathbb{E}g_i(V) \\
&= \mathbb{E}\left(\mathbf{1}_G \left\{\sum_{i=1}^n f_i(U) \cdot \mathbb{E}g_i(V)\right\}\right) = \mathbb{E}(\mathbf{1}_G h(U)) \equiv \int_G h(U)dP.
\end{aligned}$$

The case of arbitrary $\psi(U, V) \in L^1(\Omega, \mathcal{F}, P)$ follows by the convergence result (i), noting that functions of the form $\sum_{i=1}^n f_i(u)g_i(v)$ are dense in $L^1(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2, Q_1 \times Q_2)$ (Exercise 13). ■

Specializing the notion of conditional expectation to indicator functions $\mathbf{1}_A$ of sets A in $\mathcal{F}$, one defines the **conditional probability of A given $\mathcal{G}$**, denoted by $P(A|\mathcal{G})$, by

$$P(A|\mathcal{G}) := \mathbb{E}(\mathbf{1}_A|\mathcal{G}), \qquad A \in \mathcal{F}. \tag{2.14}$$

As before, $P(A|\mathcal{G})$ is a (unique) element of $L^1(\Omega, \mathcal{G}, P)$, and thus defined only up to "equivalence" by the (second) definition (2.11). That is, there are in general different versions of (2.14) differing from one another only on P-null sets in $\mathcal{G}$. In particular, the orthogonality condition may be expressed as follows:

$$P(A \cap G) = \int_G P(A|\mathcal{G})(\omega) P(d\omega), \quad \forall G \in \mathcal{G}. \tag{2.15}$$

It follows from properties (d), (e), (i) (linearity, order, and monotone convergence) in Theorem 2.7 that (outside $\mathcal{G}$-measurable P-null sets)

$$0 \le P(A|\mathcal{G}) \le 1, \quad P(\phi|\mathcal{G}) = 0, \quad P(\Omega|\mathcal{G}) = 1, \tag{2.16}$$

and that for every countable disjoint sequence $\{A_n\}_{n=1}^{\infty}$ in $\mathcal{F}$,

$$P(\cup_n A_n|\mathcal{G}) = \sum_n P(A_n|\mathcal{G}). \tag{2.17}$$

In other words, conditional probability, given $\mathcal{G}$, has properties like those of a probability measure. Indeed, under certain conditions one may choose for each $A \in \mathcal{F}$ a version of $P(A|\mathcal{G})$ such that $A \to P(A|\mathcal{G})(\omega)$ is a probability measure on $(\Omega, \mathcal{F})$ for every $\omega \in \Omega$. However, such a probability measure may not exist in the full generality in which conditional expectation is defined.[1] The technical difficulty in constructing the conditional probability measure (for each $\omega \in \Omega$) is that each one of the relations in (2.16) and (2.17) holds outside a P-null set, and individual P-null sets may pile up to a nonnull set. Such a probability measure, when it exists, is called a **regular conditional probability measure given** $\mathcal{G}$, and denoted by $P^{\mathcal{G}}(A)(\omega)$. It is more generally available as a probability measure (for each ω outside a P-null set) on appropriate sub-σ-fields of $\mathcal{F}$ (even if it is not a probability measure on all of $\mathcal{F}$). An important case occurs under the terminology of a **regular conditional distribution** of a random map Z (on $(\Omega, \mathcal{F}, P)$ into some measurable space $(S, \mathcal{S})$).

Definition 2.3. Let Y be a random map on $(\Omega, \mathcal{F}, P)$ into $(S, \mathcal{S})$. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A **regular conditional distribution of** Y **given** $\mathcal{G}$, is a function $(\omega, C) \to Q^{\mathcal{G}}(\omega, C) \equiv P^{\mathcal{G}}([Y \in C])(\omega)$ on $\Omega \times S$ such that

(i) $\forall\, C \in \mathcal{S}$, $Q^{\mathcal{G}}(\cdot, C) = P([Y \in C]|\mathcal{G})$ a.s. (and $Q^{\mathcal{G}}(\cdot, C)$ is $\mathcal{G}$-measurable),
(ii) $\forall\, \omega \in \Omega$, $C \to Q^{\mathcal{G}}(\omega, C)$ is a probability measure on $(S, \mathcal{S})$.

The following result provides a topological framework in which one can be assured of a regular version of the conditional distribution of a random map.

[1] Counterexamples have been constructed, see for example, Halmos (1950), p. 210.

Definition 2.4. A topological space S whose topology can be induced by a metric is said to be *metrizable.* If S is metrizable as a complete and separable metric space then S is referred to as a **Polish space**.

***Theorem 2.8** (Doob–Blackwell[2]).* Let Y be a random map with values in a Polish space equipped with its Borel σ-field $\mathcal{B}(S)$. Then Y has a regular conditional distribution $Q^{\mathcal{G}}$.

For our purposes in this text such an existence theorem will be unnecessary, since we will have an explicit expression of $Q^{\mathcal{G}}$ given directly when needed. Once $Q^{\mathcal{G}}$ is given, one can calculate $\mathbb{E}(f(Y)|\mathcal{G})$ (for arbitrary functions f on $(S, \mathcal{S})$ such that $f(Y) \in L^1$) as

$$\mathbb{E}(f(Y)|\mathcal{G}) = \int f(y) Q^{\mathcal{G}}(\cdot, dy). \tag{2.18}$$

This formula holds for $f(y) = \mathbf{1}_C(y)\ \forall\ C \in \mathcal{S}$ by definition. The general result follows by approximation of f by simple functions, using linearity and convergence properties of conditional expectation (and of corresponding properties of integrals with respect to a probability measure $Q^{\mathcal{G}}(\omega, \cdot)$).

The conditional Jensen inequality (g) of Theorem 2.7 follows from the existence of a regular conditional distribution of X, given $\mathcal{G}$ (see Theorem 2.8 and relation (2.18)).

The following simple examples tie up the classical concepts of conditional probability with the more modern general framework presented above.

Example 2. Let $B \in \mathcal{F}$ be such that $P(B) > 0$, $P(B^c) > 0$. Let $\mathcal{G} = \sigma(B) \equiv \{\Omega, B, B^c, \emptyset\}$. Then for every $A \in \mathcal{F}$ one has

$$P(A|\mathcal{G})(\omega) = \begin{cases} P(A|B) := \dfrac{P(A \cap B)}{P(B)}, & \text{if } \omega \in B \\ P(A|B^c) := \dfrac{P(A \cap B^c)}{P(B^c)}, & \text{if } \omega \in B^c. \end{cases} \tag{2.19}$$

More generally, let $\{B_n : n = 1, 2, \ldots\}$ be a countable disjoint sequence in $\mathcal{F}$ such that $\cup_n B_n = \Omega$, called a **partition** of Ω. Let $\mathcal{G} = \sigma(\{B_n : n \geq 1\})$ ($\mathcal{G}$ is the class of all unions of sets in this countable collection). Then for every A in $\mathcal{F}$, assuming $P(B_n) > 0$, one has

$$P(A|\mathcal{G})(\omega) = \frac{P(A \cap B_n)}{P(B_n)} \quad \text{if } \omega \in B_n. \tag{2.20}$$

[2] This result is not required for the developments in this text but is stated for sake of completeness of the discussion. For a proof, see Breiman (1968), pp. 77–80.

If $P(B_n) = 0$ then for $\omega \in B_n$, define $P(A|\mathcal{G})(\omega)$ to be some constant, say c, chosen arbitrarily (Exercise 17).

Remark 2.1. Let $Y \in L^1(\Omega, \mathcal{F}, P)$ and suppose X is a random map on $(\Omega, \mathcal{F}, P)$ with values in $(S, \mathcal{S})$. In view of Proposition 2.6, $\mathbb{E}(Y|\sigma(X))$ is a function of X, say $f(X)$, and thus constant on each event $[X = x]$, $x \in S$; i.e., $\mathbb{E}(Y|\sigma(X))(\omega) = f(X(\omega)) = f(x), \omega \in [X = x] = \{\omega \in \Omega : X(\omega) = x\}$. In particular, the notation $\mathbb{E}(Y|X = x)$ may be made precise by defining $\mathbb{E}(Y|X = x) := f(x)$, $x \in S$.

Example 3. Let $\Omega = S_1 \times S_2$, $\mathcal{F} = \mathcal{S}_1 \otimes \mathcal{S}_2$, where $(S_i, \mathcal{S}_i)$ are measurable spaces $(i = 1, 2)$. Let μ_i be a σ-finite measure on $(S_i, \mathcal{S}_i)$, $i = 1, 2$, and let f be a probability density function *(pdf)* with respect to $\mu = \mu_1 \times \mu_2$ on $S_1 \times S_2$, i.e., f is a nonnegative $\mathcal{F}$-measurable function such that $\int_\Omega f\, d\mu = 1$. Let $P(A) = \int_A f\, du$, $A \in \mathcal{F}$. Let $\mathcal{G} = \{B \times S_2 : B \in \mathcal{S}_1\}$. One may view P as the distribution of the joint coordinate maps (X, Y), where $X(\omega) = x, Y(\omega) = y$, for $\omega = (x, y) \in S_1 \times S_2$. The σ-field $\mathcal{G}$ is the σ-field generated by the first coordinate map X. For every $A = S_1 \times C$ $(C \in \mathcal{S}_2)$, one has

$$P(A|\mathcal{G})(\omega) = \frac{\int_C f(x,y)\mu_2(dy)}{\int_{S_2} f(x,y)\mu_2(dy)} \quad \text{if } \omega = (x, y') \ (\in \Omega). \tag{2.21}$$

To check this, first note that by the Fubini–Tonelli theorem, the function Z defined by the right-hand side of (2.21) is $\mathcal{G}$-measurable. Secondly, for every nonnegative bounded Borel measurable g on S_1 one has

$$\begin{aligned}
\mathbb{E}(g(X)Z) &= \int_{S_1 \times S_2} g(x) \frac{\int_C f(x,y)\mu_2(dy)}{\int_{S_2} f(x,y)\mu_2(dy)} f(x,y)\mu_1(dx)\mu_2(dy) \\
&= \int_{S_1} g(x) \left\{ \int_{S_2} \frac{\int_C f(x,y)\mu_2(dy)}{\int_{S_2} f(x,y)\mu_2(dy)} f(x,y)\mu_2(dy) \right\} \mu_1(dx) \\
&= \int_{S_1} g(x) \left\{ \frac{\int_C f(x,y)\mu_2(dy)}{\int_{S_2} f(x,y)\mu_2(dy)} \cdot \int_{S_2} f(x,y)\mu_2(dy) \right\} \mu_1(dx) \\
&= \int_{S_1} g(x) \left\{ \int_C f(x,y)\mu_2(dy) \right\} \mu_1(dx) = \mathbb{E}(\mathbf{1}_{S_1 \times C}\, g(X)) = \mathbb{E}(g(X)\mathbf{1}_A).
\end{aligned}$$

The function $f(x,y)/\int_{S_2} f(x,y)\mu_2(dy)$ is called the **conditional pdf of Y given** $X = x$, and denoted by $f(y|x)$; i.e. the conditional pdf is simply the normalization of the joint pdf to a probability density by dividing by the (marginal) pdf of X. Let $A \in \mathcal{F} = \mathcal{S}_1 \otimes \mathcal{S}_2$. By the same calculations using Fubini–Tonelli one more generally obtains (Exercise 17)

$$P(A|\mathcal{G})(\omega) = \int_{A_x} f(y|x)\mu_2(dy) \equiv \frac{\int_{A_x} f(x,y)\mu_2(dy)}{\int_{S_2} f(x,y)\mu_2(dy)} \quad \text{if } \omega \equiv (x, y'), \tag{2.22}$$

where $A_x = \{y \in S_2 : (x,y) \in A\}$.

One may change the perspective here a little and let $(\Omega, \mathcal{F}, P)$ be any probability space on which are defined two maps X and Y with values in $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, respectively. If the (joint) distribution of (X, Y) on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ has a pdf f with respect to a product measure $\mu_1 \times \mu_2$, where μ_i is a σ-finite measure on $(S_, \mathcal{S}_i)$, $i = 1, 2$, then for $\mathcal{G} = \sigma(X)$, the (regular) conditional distribution of Y given $\mathcal{G}$ (or "given X") is given by

$$P([Y \in C]|\mathcal{G})(\omega) = \frac{\int_C f(x,y)\mu_2(dy)}{\int_{S_2} f(x,y)\mu_2(dy)} \quad \text{if } X(\omega) = x,$$

i.e., if $\omega \in [X = x] \equiv X^{-1}\{x\}$, $x \in S_1$. Note that the conditional probability is constant on $[X = x]$ as required for $\sigma(X)$-measurability; cf Proposition 2.6. The earlier model $\Omega = S_1 \times S_2$, $\mathcal{F} = \mathcal{S}_1 \otimes \mathcal{S}_2$, $dP = f\, d\mu$ is a **canonical model** for this calculation.

Example 4 *(Markov Property for General Random Walks on $\mathbb{R}^k$).* Let $\{Z_n : n \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) k-dimensional random vectors defined on a probability space $(\Omega, \mathcal{F}, P)$. Let μ denote the distribution of Z_1 (hence of each Z_n). For arbitrary $x \in \mathbb{R}^k$, **a random walk starting at x with step-size distribution** μ is defined by the sequence $S_n^x := x + Z_1 + \cdots + Z_n$ $(n \geq 1)$, $S_0^x = x$.

For notational simplicity we will restrict to the case of $k = 1$ dimensional random walks, however precisely the same calculations are easily seen to hold for arbitrary $k \geq 1$ (Exercise 25). Let Q_x denote the distribution of $\{S_n^x := n \geq 0\}$ on the product space $(\mathbb{R}^\infty, \mathcal{B}^\infty)$. Here $\mathcal{B}^\infty$ is the σ-field generated by **cylinder sets** of the form $C = B_m \times \mathbb{R}^\infty := \{\mathbf{y} = (y_0, y_1, \ldots) \in \mathbb{R}^\infty; (y_0, y_1, \ldots, y_m) \in B_m\}$ with B_m a Borel subset of $\mathbb{R}^{m+1}$ $(m = 0, 1, 2, \ldots)$. Note that $Q_x(B_m \times \mathbb{R}^\infty) = P((S_0^x, S_1^x \ldots, S_m^x) \in B_m)$, so that $Q_x(B_m \times \mathbb{R}^\infty)$ may be expressed in terms of the m-fold product measure $\mu \times \mu \times \cdots \times \mu$, which is the distribution of $(Z_1, Z_2, \ldots, Z_m)$. For our illustration, let $\mathcal{G}_n = \sigma(\{S_j^x : 0 \leq j \leq n\}) = \sigma(\{Z_1, Z_2, \ldots, Z_n\})$ $(n \geq 1)$. We would like to establish the following property: **The conditional distribution of the "after-n process"** $S_n^{x+} := \{S_{n+m}^x : m = 0, 1, 2, \ldots\}$ on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ **given** $\mathcal{G}_n$ is $Q_y|_{y=S_n^x} \equiv Q_{S_n^x}$. In other words, for the random walk $\{S_n^x : n \geq 0\}$, the conditional distribution of the *future evolution* defined by S_n^{x+} given the *past states* $S_0^x, \ldots, S_{n-1}^x$ and *present state* S_n^x depends solely on the present state S_n^x, namely $Q_{S_n^x}$ i.e., it is given by the regular conditional distribution $Q^{\mathcal{G}_n}(\omega, \cdot) = Q_{S_n^x(\omega)}(\cdot)$.

Theorem 2.9 *(Markov Property).* For every $n \geq 1$, the conditional distribution of S_n^{x+} given $\sigma(S_0^x, \ldots, S_n^x)$ is a function only of S_n^x, namely, $Q_{S_n^x}$.

Proof. To prove the italicized statement above and hence the theorem, choose a cylinder set $C \in \mathcal{B}^\infty$. That is, $C = B_m \times \mathbb{R}^\infty$ for some $m \geq 0$. We want to show that

$$P([S_n^{x+} \in C]|\mathcal{G}_n) \equiv \mathbb{E}(\mathbf{1}_{[S_n^{x+} \in C]}|\mathcal{G}_n) = Q_{S_n^x}(C). \tag{2.23}$$

Now $[S_n^{x+} \in C] = [(S_n^x, S_n^x + Z_{n+1}, \ldots, S_n^x + Z_{n+1} + \cdots + Z_{n+m}) \in B_m]$, so that one may write

$$\mathbb{E}\left(\mathbf{1}_{[S_n^{x+} \in C]}|\mathcal{G}_n\right) = \mathbb{E}(\psi(U, V)|\mathcal{G}_n),$$

where $U = S_n^x$, $V = (Z_{n+1}, Z_{n+2}, \ldots, Z_{n+m})$ and, for $u \in \mathbb{R}$ and $v \in \mathbb{R}^m$, $\psi(u, v) = \mathbf{1}_{B_m}(u, u + v_1, u + v_1 + v_2, \ldots, u + v_1 + \cdots + v_m)$. Since S_n^x is $\mathcal{G}_n$-measurable and V is independent of $\mathcal{G}_n$, it follows from property (ℓ) of Theorem 2.7 that $\mathbb{E}(\psi(U, V)|\mathcal{G}_n) = h(S_n^x)$, where $h(u) = \mathbb{E}\psi(u, V)$. But

$$\begin{aligned} \mathbb{E}\psi(u, V) &= P((u, u + Z_{n+1}, u + Z_{n+1} + Z_{n+2}, \ldots, u + Z_{n+1} + \cdots + Z_{n+m}) \in B_m) \\ &= P((u, u + Z_1, u + Z_1 + Z_2, \ldots, u + Z_1 + \cdots + Z_m) \in B_m) \\ &= P((S_0^u, S_1^u, \ldots, S_m^u) \in B_m) = Q_u(C). \end{aligned}$$

Therefore, $P([S_n^{x+} \in C]|\mathcal{G}_n) = (Q_u(C))_{u = S_n^x} = Q_{S_n^x}(C)$. We have now shown that the class $\mathcal{L}$ of sets $C \in \mathcal{B}^\infty$ for which "$P([S_n^{x+} \in C]|\mathcal{G}_n) = Q_{S_n^x}(C)$ a.s." holds contains the class $\mathcal{C}$ of all cylinder sets. Since this class is a λ-system (see the convergence property (i) of Theorem 2.7) containing the π-system of cylinder sets that generate $\mathcal{B}^\infty$, it follows by the $\pi - \lambda$ theorem that $\mathcal{L} = \mathcal{B}^\infty$. ■

EXERCISES

Exercise Set II

1. Let $\Omega = \{0, 1\}^\infty$ be the space of infinite binary 0-1 sequences, and let $\mathcal{F}_0$ denote the field of finite unions of sets of the form $A_n(\varepsilon_1, \ldots, \varepsilon_n) = \{\omega = (\omega_1, \omega_2, \ldots) \in \Omega : \omega_1 = \varepsilon_1, \ldots, \omega_n = \varepsilon_n\}$ for arbitrary $\varepsilon_i \in \{0, 1\}$, $1 \leq i \leq n, n \geq 1$. Fix $p \in [0, 1]$ and define $P_p(A_n(\varepsilon_1, \ldots, \varepsilon_n)) = p^{\sum_{i=1}^n \varepsilon_i}(1 - p)^{n - \sum_{i=1}^n \varepsilon_i}$. (i) Show that the natural finitely additive extension of P_p to $\mathcal{F}_0$ defines a measure on the field $\mathcal{F}_0$. [*Hint*: By Tychonov's theorem from topology, the set Ω is compact for the product topology, see Appendix B. Check that sets $C \in \mathcal{F}_0$ are both open and closed for the product topology, so that by compactness, any countable disjoint union belonging to $\mathcal{F}_0$ must be a finite union.] (ii) Show that P_p has a unique extension to $\sigma(\mathcal{F}_0)$. This probability P_p defines the infinite product probability, also denoted by $(p\delta_1 + (1 - p)\delta_0)^\infty$. [*Hint*: Apply the Carathéodory extension theorem.] (iii) Show that the coordinate projections $X_n(\omega) = \omega_n, \omega = (\omega_1, \omega_2, \ldots) \in \Omega$, $n \geq 1$, define an i.i.d. sequence of (coin tossing) Bernoulli 0 or 1-valued random variables.

2. (i) Consider three independent tosses of a balanced coin and let A_i denote the event that the outcomes of the ith and $(i + 1)$st tosses match, for $i = 1, 2$. Let A_3 be the event that

the outcomes of the third and first match. Show that A_1, A_2, A_3 are pairwise independent but not independent. (ii) Suppose that $X_1, \ldots, X_n$ are independent random maps defined on a probability space $(\Omega, \mathcal{F}, P)$. Show that the product measure $Q = P \circ (X_1, \ldots, X_n)^{-1}$ is given by $Q_1 \times \cdots \times Q_n$, where $Q_i = P \circ X_i^{-1}$. Also show that any subset of $\{X_1, \ldots, X_n\}$ comprises independent random maps.

3. Suppose that $X_1, X_2, \ldots$ is a sequence of independent random variables on $(\Omega, \mathcal{F}, P)$. Show that the two families $\{X_1, X_3, X_5, \ldots\}$ and $\{X_2, X_4, X_6, \ldots\}$ are independent.

4. Suppose $\mathbf{X}_1, \mathbf{X}_2$ are independent k-dimensional random vectors having distributions Q_1, Q_2, respectively. Prove that the distribution of $\mathbf{X}_1 + \mathbf{X}_2$ is given by the convolution $Q_1 * Q_2$ defined by $Q_1 * Q_2(B) = \int_{\mathbb{R}^k} Q_1(B - x) Q_2(dx)$, where $B - x := \{y - x : y \in B\}$ for Borel sets $B \subseteq \mathbb{R}^k$.

5. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables defined on $(\Omega, \mathcal{F}, P)$ and having (common) distribution Q.
   (i) Suppose $Q(dx) = \lambda e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x) dx$, for some $\lambda > 0$, referred to as the *exponential distribution with parameter* λ. Show that $X_1 + \cdots + X_n$ has distribution $Q^{*n}(dx) = \lambda^n \frac{x^{n-1}}{(n-1)!} e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x) dx$. This latter distribution is referred to as a *gamma distribution* with parameters n, λ.
   (ii) Suppose that $Q(dx) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}} dx$; referred to as the *Gaussian* or **normal** distribution with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Show that $X_1 + \cdots + X_n$ has a normal distribution with parameters $n\mu$ and $n\sigma^2$.
   (iii) Let X be a standard normal $N(0,1)$ random variable. Find the distribution Q of X^2, and compute Q^{*2}. [*Hint*: $\int_0^1 u^{-\frac{1}{2}}(1-u)^{-\frac{1}{2}} du = \pi$.]

6. Suppose that $X_1, X_2, \ldots$ is a sequence of random variables on $(\Omega, \mathcal{F}, P)$ each having the same distribution $Q = P \circ X_n^{-1}$. (i) Show that if $\mathbb{E}|X_1| < \infty$ then $P(|X_n| > n \textit{ i.o.}) = 0$. [*Hint*: First use (2.5) to get $\mathbb{E}|X_1| = \int_0^\infty P(|X_1| > x) dx$, and then apply Borel–Cantelli.] (ii) Assume that $X_1, X_2, \ldots$ are also independent with $\mathbb{E}|X_n| = \infty$. Show that $P(|X_n| > n \textit{ i.o.}) = 1$.

7. Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space and suppose $A_1, A_2, \ldots$ is a sequence of *independent* events in $\mathcal{F}$ with $P(A_n) < 1, \forall n$. Suppose $P(\cup_{n=1}^\infty A_n) = 1$. (i) Show that $P(A_n \textit{ i.o.}) = 1$. (ii) Give an example to show that $P(A_n) < 1\ \forall n$ is necessary in (i).

8. Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space and suppose $\{A_n\}_{n=1}^\infty$ is a sequence of *independent* events in $\mathcal{F}$ such that $\sum_{n=1}^\infty P(A_n) \geq 2$. Let E denote the event that none of the A_n's occur for $n \geq 1$. (i) Show that $E \in \mathcal{F}$.
   (ii) Show that $P(E) \leq \frac{1}{e^2}$. [*Hint*: $1 - x \leq e^{-x}$, $x \geq 0$.]

9. Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence of Bernoulli 0 or 1-valued random variables with $P(X_n = 1) = p, P(X_n = 0) = q = 1 - p$. Fix $r \geq 1$ and let $R_n := [X_n = 1, X_{n+1} = 1, \ldots, X_{n+r-1} = 1]$ be the event of a run of 1's of length at least r starting from n.
   (i) Show that $P(R_n \textit{ i.o.}) = 1$ if $0 < p \leq 1$.
   (ii) Suppose r is allowed to grow with n, say $r_n = [\theta \log_2 n]$ in defining the event R_n; here $[x]$ denotes the largest integer not exceeding x. In the case of a balanced coin ($p = 1/2$), show that if $0 < \theta \leq 1$ then $P(R_n \textit{ i.o.}) = 1$.[*Hint*: Consider a subsequence $R_{n_k} = [X_{n_k} = 1, \ldots, X_{n_k + r_{n_k} - 1} = 1]$ with n_1 sufficiently large that $\theta \log_2 n_1 > 1$,

and $n_{k+1} = n_k + r_{n_k}, k \geq 1$. Compare $\sum_{k=1}^{\infty} n_k^{-\theta} \equiv \sum_{k=1}^{\infty} \frac{n_k^{-\theta}}{n_{k+1}-n_k}(n_{k+1} - n_k)$ to an integral $\int_{n_1}^{\infty} f(x)dx$ for an appropriately selected function f.]

10. Let X_1, X_2 be random maps on $(\Omega, \mathcal{F}, P)$ taking values in the measurable spaces $(S_1, \mathcal{S}_1), (S_2, \mathcal{S}_2)$, respectively. Show that the joint distribution of (X_1, X_2) on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ is product measure if and only if $\sigma(X_1)$ and $\sigma(X_2)$ are independent σ-fields.

11. Let X_1, X_2 be random maps on $(\Omega, \mathcal{F}, P)$. (i) Show that $\sigma(X_1)$ and $\sigma(X_2)$ are independent if and only if $\mathbb{E}[g(X_1)h(X_2)] = \mathbb{E}g(X_1)\mathbb{E}h(X_2)$ for all bounded measurable functions g, h on the respective image spaces. (ii) Prove Proposition 2.5. [*Hint*: Use (i) and induction.]

12. This exercise is in reference to Proposition 2.3.
    (i) Let V_1 take values ± 1 with probability 1/4 each, and 0 with probability 1/2. Let $V_2 = V_1^2$. Show that $\mathrm{Cov}(V_1, V_2) = 0$, though they are not independent.
    (ii) Show that random maps V_1, V_2 are independent if and only if $f(V_1)$ and $g(V_2)$ are uncorrelated for all pairs of real-valued Borel-measurable functions f, g such that $f(V_1),\ g(V_2) \in L^2$.
    (iii) Show that a family of random maps $\{X_t \in \Lambda\}$ (with X_t, a measurable map into $(S_t, \mathcal{S}_t)$) is an independent family if and only if for every pair of disjoint finite subsets Λ_1, Λ_2 of Λ, any random variable $V_1 \in L^2(\sigma\{X_t : t \in \Lambda_1\})$ is uncorrelated with any random variable $V_2 \in L^2(\sigma\{X_t t \in \Lambda_2\})$.

13. Let $\mathcal{C}$ denote the collection of functions of the form $\sum_{i=1}^{n} f_i(u)g_i(v), (u, v) \in S_1 \times S_2$, where $f_i, g_i, 1 \leq i \leq n$, are bounded Borel-measurable functions on the probability spaces $(S_1, \mathcal{S}_1, Q_1)$ and $(S_2, \mathcal{S}_2, Q_2)$, respectively. Show that $\mathcal{C}$ is dense in $L^1(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2, Q_1 \times Q_2)$. [*Hint*: For $A \in \mathcal{S}_1 \otimes \mathcal{S}_2$, use the Carathéodory formula (2.3) to approximate $h = \mathbf{1}_A$ in L^1. The rest follows by the method of approximation by simple functions.]

14. Give a proof of Proposition 2.4.

15. Suppose that X, Y are independent random variables on $(\Omega, \mathcal{F}, P)$. Assume that there is a number $a < 1$ such that $P(X \leq a) = 1$. Also assume that Y is exponentially distributed with mean one. Calculate $\mathbb{E}[e^{XY}|\sigma(X)]$.

16. Suppose that (X, Y) is uniformly distributed on the unit disk $D = \{(x, y) : x^2 + y^2 \leq 1\}$, i.e., has constant pdf on D. (i) Calculate the (marginal) distribution of X. (ii) Calculate the conditional distribution of Y given $\sigma(X)$. (iii) Calculate $\mathbb{E}(Y^2|\sigma(X))$.

17. (i) Give a proof of (2.8) using the second definition of conditional expectation. [*Hint*: The only measurable random variables with respect to $\{\Omega, \emptyset, B, B^c\}$ are those of the form $c\mathbf{1}_B + d\mathbf{1}_{B^c}$, for $c, d \in \mathbb{R}$.] (ii) Prove (2.20), (2.22).

18. Suppose that X, Z are independent random variables with standard normal distribution. Let $Y = X + bZ$; i.e. X with an independent additive noise term bZ. Calculate $\mathbb{E}[X|\sigma(Y)]$.

19. Let $X_1, \ldots, X_n$ be an i.i.d. sequence of random variables on $(\Omega, \mathcal{F}, P)$ and let $S_n = X_1 + \cdots + X_n$. Assume $\mathbb{E}|X_1| < \infty$. Show that $\mathbb{E}(X_j|\sigma(S_n)) = \mathbb{E}(X_1|\sigma(S_n))$. [*Hint*: Use Fubini–Tonelli.] Calculate $\mathbb{E}(X_j|\sigma(S_n))$. [*Hint*: Add up and use properties of conditional expectation.]

20. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. exponentially distributed with mean one. Let $S_n = \sum_{j=1}^{n} Y_j$.

(i) Calculate $\mathbb{E}(Y_1^2|S_n)$. [*Hint*: Calculate the joint pdf of $(Y_1, Y_2+\cdots+Y_n)$ and then that of (Y_1, S_n) by a change of variable under the linear transformation $(y,s) \mapsto (y, y+s)$.]

(ii) Calculate $\mathbb{E}(Y_1Y_2|S_n)$. hintConsider $S_n^2 = \mathbb{E}(S_n^2|S_n)$ along with the previous exercise.

(iii) Make the above calculations in the case that $Y_1, Y_2, \ldots Y_n$ are i.i.d. with standard normal distributions.

21. (*Conditional Chebyshev-type*) For $X \in L^p$, $p \geq 1$, prove for $\lambda > 0$, $P(|X| > \lambda|\mathcal{G}) \leq \mathbb{E}(|X|^p|\mathcal{G})/\lambda^p$ a.s.

22. (*Conditional Cauchy–Schwarz*) For $X, Y \in L^2$ show that $|\mathbb{E}(XY|\mathcal{G})|^2 \leq \mathbb{E}(X^2|\mathcal{G})\mathbb{E}(Y^2|\mathcal{G})$.

23. Let Y be an exponentially distributed random variable on $(\Omega, \mathcal{F}, P)$. Fix $a > 0$.

    (i) Calculate $\mathbb{E}(Y|\sigma(Y \wedge a))$, where $Y \wedge a := \min\{Y, a\}$. [Hint: $[Y < a] = [Y \wedge a < a]$. Let g be a bounded Borel-measurable function and either make and verify an intuitive guess for $\mathbb{E}(Y|\sigma(Y \wedge a))$ (based on "lack of memory" of the exponential distribution) or calculate $\mathbb{E}(Yg(Y \wedge a))$ by integration by parts.]

    (ii) Determine the regular conditional distribution of Y given $\sigma(Y \wedge a)$.

24. Let U, V be independent random maps with values in measurable spaces $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, respectively. Let $\varphi(u, v)$ be a measurable map on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ into a measurable space $(S, \mathcal{S})$. Show that a regular conditional distribution of $\varphi(U, V)$, given $\sigma(V)$, is given by Q_V, where Q_v is the distribution of $\varphi(U, v)$. [*Hint*: Use the Fubini–Tonelli theorem or Theorem 2.7(l).]

25. Prove the Markov property for k-dimensional random walks with $k \geq 2$.

26. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Show that the set of (equivalence classes of) simple functions is dense in $L^p \equiv L^p(\Omega, \mathcal{F}, P)$, $p \geq 1$.

CHAPTER III

Martingales and Stopping Times

The notion of “martingale” has proven to be among the most powerful ideas to emerge in probability in the last century. In this section some basic foundations are presented. A more comprehensive treatment of the theory and its applications is provided in our text on stochastic processes.[1] For the prototypical illustration of the martingale property, let $Z_1, Z_2, \ldots$ be an i.i.d. sequence of integrable random variables and let $X_n = Z_1 + \cdots + Z_n,\ n \geq 1$. If $\mathbb{E}Z_1 = 0$ then one clearly has

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n, \quad n \geq 1,$$

where $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$.

Definition 3.1 *(First Definition of Martingale)***.** A sequence of integrable random variables $\{X_n : n \geq 1\}$ on a probability space $(\Omega, \mathcal{F}, P)$ is said to be a **martingale** if, writing $\mathcal{F}_n := \sigma(X_1, X_2, \ldots, X_n)$,

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n \text{ a.s. } (n \geq 1). \tag{3.1}$$

This definition extends to any (finite or infinite) family of integrable random variables $\{X_t : t \in T\}$, where T is a linearly ordered set: Let $\mathcal{F}_t = \sigma(X_s : s \leq t)$. Then

[1] Bhattacharya, R., and E. Waymire (2007): *Theory and Applications of Stochastic Processes*, Springer Graduate Texts in Mathematics.

$\{X_t : t \in T\}$ is a **martingale** if

$$\mathbb{E}(X_t|\mathcal{F}_s) = X_s \text{ a.s } \forall\ s < t\ (s, t \in T). \tag{3.2}$$

In the previous case of a sequence $\{X_n : n \geq 1\}$, as one can see by taking successive conditional expectations $\mathbb{E}(X_n|\mathcal{F}_m) = \mathbb{E}[\mathbb{E}(X_n|\mathcal{F}_{n+1})|\mathcal{F}_m] = \mathbb{E}(X_{n+1}|\mathcal{F}_m) = \cdots = \mathbb{E}(X_{m+1}|\mathcal{F}_m) = X_m$, (3.1) is equivalent to

$$\mathbb{E}(X_n|\mathcal{F}_m) = X_m \quad \text{a.s.} \quad \forall\ m < n. \tag{3.3}$$

Thus, (3.1) is a special case of (3.2). Most commonly, $T = \mathbb{N}$ or $\mathbb{Z}^+$, or $T = [0, \infty)$. Note that if $\{X_t : t \in T\}$ is a martingale, one has the *constant expectations property:* $\mathbb{E}X_t = \mathbb{E}X_s\ \forall\ s, t \in T$.

Remark 3.1. Let $\{X_n : n \geq 1\}$ be a martingale sequence. Define its associated **martingale difference sequence** by $Z_1 := X_1, Z_{n+1} := X_{n+1} - X_n$ $(n \geq 1)$. Note that for $X_n \in L^2(\Omega, \mathcal{F}, P)$, $n \geq 1$, the martingale differences are uncorrelated. In fact, for $X_n \in L^1(\Omega, \mathcal{F}, P)$, $n \geq 1$, one has

$$\begin{aligned}\mathbb{E}Z_{n+1}f(X_1, X_2, \ldots, X_n) &= \mathbb{E}[\mathbb{E}(Z_{n+1}f(X_1, \ldots, X_n)|\mathcal{F}_n)] \\ &= \mathbb{E}[f(X_1, \ldots, X_n)\mathbb{E}(Z_{n+1}|\mathcal{F}_n)] = 0 \end{aligned} \tag{3.4}$$

for all bounded $\mathcal{F}_n$ measurable functions $f(X_1, \ldots, X_n)$. If $X_n \in L^2(\Omega, \mathcal{F}, P)\ \forall\ n \geq 1$, then (3.1) implies, and is equivalent to, the fact that $Z_{n+1} \equiv X_{n+1} - X_n$ is orthogonal to $L^2(\Omega, \mathcal{F}_n, P)$. It is interesting to compare this orthogonality to that of independence of Z_{n+1} and $\{Z_m : m \leq n\}$. Recall that Z_{n+1} is independent of $\{Z_m : 1 \leq m \leq n\}$ or, equivalently, of $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ if and only if $g(Z_{n+1})$ is orthogonal to $L^2(\Omega, \mathcal{F}_n, P)$ for all bounded measurable g such that $\mathbb{E}g(Z_{n+1}) = 0$. Thus independence translates as $0 = \mathbb{E}\{[g(Z_{n+1}) - \mathbb{E}g(Z_{n+1})] \cdot f(X_1, \ldots, X_n)\} = \mathbb{E}\{g(Z_{n+1}) \cdot f(X_1, \ldots, X_n)\} - \mathbb{E}g(Z_{n+1}) \cdot \mathbb{E}f(X_1, \ldots, X_n)$, for all bounded measurable g on $\mathbb{R}$ and for all bounded measurable f on $\mathbb{R}^n$.

Example 1 ***(Independent Increment Process).*** Let $\{Z_n : n \geq 1\}$ be an independent sequence having *zero means,* and X_0 an integrable random variable independent of $\{Z_n : n \geq 1\}$. Then

$$X_0,\ X_n := X_0 + Z_1 + \cdots + Z_n \equiv X_{n-1} + Z_n\ (n \geq 1) \tag{3.5}$$

is a martingale sequence.

Definition 3.2. If with $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ one has inequality in place of (3.1), namely,

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n \text{ a.s.} \quad \forall n \geq 1, \tag{3.6}$$

then $\{X_n : n \geq 1\}$ is said to be a submartingale. More generally, if the index set T is as in (3.2), then $\{X_t : t \in T\}$ is a **submartingale** if, with $\mathcal{F}_t$ as in Definition 3.2,

$$\mathbb{E}(X_t|\mathcal{F}_s) \geq X_s \ \forall \ s < t \ (s, t \in T). \tag{3.7}$$

If instead of $\geq$, one has $\leq$ in (3.7) ((3.8)), the process $\{X_n : n \geq 1\}$ ($\{X_t : t \in T\}$) is said to be a **supermartingale**.

In Example 1, if $\mathbb{E}Z_k \geq 0 \ \forall \ k$, then the sequence $\{X_n : n \geq 1\}$ of partial sums of independent random variables is a submartingale. If $\mathbb{E}Z_k \leq 0$ for all k, then $\{X_n : n \geq 1\}$ is a supermartingale. In Example 3, it follows from $\pm X_{n+1} \leq |X_{n+1}|$ taking conditional expectations, that the sequence $\{Y_n \equiv |X_n| : \ n \geq 1\}$ is a submartingale. The following proposition provides an important generalization of this latter example.

Proposition 3.1. (a) If $\{X_n : n \geq 1\}$ is a martingale and $\varphi(X_n)$ is a convex and integrable function of X_n, then $\{\varphi(X_n) : n \geq 1\}$ is a submartingale. (b) If $\{X_n\}$ is a submartingale, and $\varphi(X_n)$ is a convex and nondecreasing integrable function of X_n, then $\{\varphi(X_n) : n \geq 1\}$ is a submartingale.

Proof. The proof is obtained by an application of the conditional Jensen's inequality given in Theorem 2.7. In particular, for (a) one has

$$\mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) = \varphi(X_n). \tag{3.8}$$

Now take the conditional expectation of both sides with respect to $\mathcal{G}_n \equiv \sigma(\varphi(X_1), \ldots, \varphi(X_n)) \subseteq \mathcal{F}_n$, to get the martingale property of $\{\varphi(X_n) : n \geq 1\}$. Similarly, for (b), for convex and nondecreasing φ one has in the case of a submartingale that

$$\mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \geq \varphi(X_n), \tag{3.9}$$

and taking conditional expectation in (3.9), given $\mathcal{G}_n$, the desired submartingale property follows. ∎

Proposition 3.1 immediately extends to martingales and submartingales indexed by an arbitrary linearly ordered set T.

Example 2. (a) If $\{X_t : t \in T\}$ is a martingale, $\mathbb{E}|X_t|^p < \infty$ $(t \in T)$ for some $p \geq 1$, then $\{|X_t|^p : t \in T\}$ is a submartingale. (b) If $\{X_t : t \in T\}$ is a submartingale, then for every real c, $\{Y_t := \max(X_t, c)\}$ is a submartingale. In particular, $\{X_t^+ := \max(X_t, 0)\}$ is a submartingale.

Remark 3.2. It may be noted that in (3.8), (3.9), the σ-field $\mathcal{F}_n$ is $\sigma(X_1, \ldots, X_n)$, and not $\sigma(\varphi(X_1), \ldots, \varphi(X_n))$, as seems to be required by the first definitions in (3.1), (3.6). It is, however, more convenient to give the definition of a **martingale (or a submartingale) with respect to a filtration** $\{\mathcal{F}_n\}$ for which (3.1) holds (or respectively, (3.6) holds) assuming at the outset that X_n is $\mathcal{F}_n$-measurable ($n \geq 1$) (or, as one often says, $\{X_n\}$ is $\{\mathcal{F}_n\}$-**adapted**). One refers to this sequence as an **$\{\mathcal{F}_n\}$-martingale** (respectively **$\{\mathcal{F}_n\}$-submartingale**). An important example of $\mathcal{F}_n$ larger than $\sigma(X_1, \ldots, X_n)$ is given by "adding independent information" via $\mathcal{F}_n = \sigma(X_1, \ldots, X_n) \vee \mathcal{G}$, where $\mathcal{G}$ is a σ-field independent of $\sigma(X_1, X_2, \ldots)$, and $\mathcal{G}_1 \vee \mathcal{G}_2$ denotes the smallest σ-field containing $\mathcal{G}_1 \cup \mathcal{G}_2$. We formalize this with the following definition; also see Exercise 10.

Definition 3.3. *(Second General Definition)* Let T be an arbitrary linearly ordered set and suppose $\{X_t : t \in T\}$ is a stochastic process with (integrable) values in $\mathbb{R}$ and defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_t : t \in T\}$ be a nondecreasing collection of sub-σ-fields of $\mathcal{F}$, referred to as a **filtration** i.e., $\mathcal{F}_s \subseteq \mathcal{F}_t$ if $s \leq t$. Assume that for each $t \in T$, X_t is **adapted** to $\mathcal{F}_t$ in the sense that X_t is $\mathcal{F}_t$ measurable. We say that $\{X_t : t \in T\}$ is a **martingale**, respectively **submartingale, supermartingale**, with respect to the filtration $\{\mathcal{F}_t\}$ if $\mathbb{E}[X_t|\mathcal{F}_s] = X_s$, $\forall s, t \in T, s \leq t$, respectively $\geq X_s, \forall s, t \in T, s \leq t$, or$\leq X_s$ $\forall s, t \in T, s \leq t$.

Example 3. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$ and let $\{\mathcal{F}_n : n \geq 1\}$ be a filtration of $\mathcal{F}$. One may check that the stochastic process defined by

$$X_n := \mathbb{E}(X|\mathcal{F}_n) \ (n \geq 1) \tag{3.10}$$

is an $\{\mathcal{F}_n\}$-martingale.

Note that for submartingales the expected values are nondecreasing, while those of supermartingales are nonincreasing. Of course, martingales continue to have constant expected values under this more general definition.

Theorem 3.2. *(Doob's Maximal Inequality).* Let $\{X_1, X_2, \ldots, X_n\}$ be an $\{\mathcal{F}_k : 1 \leq k \leq n\}$-martingale, or a nonnegative submartingale, and $\mathbb{E}|X_n|^p < \infty$ for some $p \geq 1$. Then, for all $\lambda > 0$, $M_n := \max\{|X_1|, \ldots, |X_n|\}$ satisfies

$$P(M_n \geq \lambda) \leq \frac{1}{\lambda^p} \int_{[M_n \geq \lambda]} |X_n|^p dP \leq \frac{1}{\lambda^p} \mathbb{E}|X_n|^p. \tag{3.11}$$

Proof. Let $A_1 = [|X_1| \geq \lambda]$, $A_k = [|X_1| < \lambda, \ldots, |X_{k-1}| < \lambda, \ |X_k| \geq \lambda]$ $(2 \leq k \leq n)$. Then $A_k \in \mathcal{F}_k$ and $[A_k : 1 \leq k \leq n]$ is a (disjoint) partition of $[M_n \geq \lambda]$. Therefore,

$$\begin{aligned} P(M_n \geq \lambda) &= \sum_{k=1}^n P(A_k) \leq \sum_{k=1}^n \frac{1}{\lambda^p} \mathbb{E}(\mathbf{1}_{A_k}|X_k|^p) \leq \sum_{k=1}^n \frac{1}{\lambda^p} \mathbb{E}(\mathbf{1}_{A_k}|X_n|^p) \\ &= \frac{1}{\lambda^p} \int_{[M_n \geq \lambda]} |X_n|^p dP \leq \frac{\mathbb{E}|X_n|^p}{\lambda^p}. \end{aligned}$$

■

Remark 3.3. By an obvious change in the definition of $A_k (k = 1, \ldots, n)$, one obtains (3.11) with strict inequality $M_n > \lambda$ on both sides of the asserted inequality.

Remark 3.4. The classical *Kolmogorov maximal inequality* for sums of i.i.d. mean zero, square-integrable random variables is a special case of *Doob's maximal inequality* obtained by taking $p = 2$ for the martingales of Example 1 having square-integrable increments.

Corollary 3.3. Let $\{X_1, X_2, \ldots, X_n\}$ be an $\{\mathcal{F}_k : 1 \leq k \leq n\}$-martingale such that $\mathbb{E}X_n^2 < \infty$. Then $\mathbb{E}M_n^2 \leq 4\mathbb{E}X_n^2$.

Proof. A standard application of the Fubini–Tonelli theorem (see (2.5)) provides the second moment formula

$$\mathbb{E}M_n^2 = 2\int_0^\infty \lambda P(M_n > \lambda) d\lambda. \tag{3.12}$$

Applying the first inequality in (3.11), then another application of the Fubini–Tonelli theorem, and finally the Cauchy–Schwarz inequality, it follows that

$$\begin{aligned}\mathbb{E}M_n^2 &\leq 2\int_0^\infty \mathbb{E}\left(|X_n|\mathbf{1}_{[M_n \geq \lambda]}\right) d\lambda = 2\mathbb{E}\left(|X_n|M_n\right)\\ &\leq 2\sqrt{\mathbb{E}X_n^2}\sqrt{\mathbb{E}M_n^2}.\end{aligned}$$

Divide both sides by $\sqrt{\mathbb{E}M_n^2}$ to complete the proof. ■

Corollary 3.4. Let $\{X_t : t \in [0, T]\}$ be a right-continuous nonnegative $\{\mathcal{F}_t\}$-submartingale with $\mathbb{E}|X_T|^p < \infty$ for some $p \geq 1$. Then $M_T := \sup\{X_s : 0 \leq s \leq T\}$ is $\mathcal{F}_T$-measurable and, for all $\lambda > 0$,

$$P(M_T > \lambda) \leq \frac{1}{\lambda^p}\int_{[M_T > \lambda]} X_T^p dP \leq \frac{1}{\lambda^p}\mathbb{E}X_T^p. \tag{3.13}$$

Proof. Consider the nonnegative submartingale $\{X_0, X_{T2^{-n}}, \ldots, X_{Ti2^{-n}}, \ldots, X_T\}$, for each $n = 1, 2, \ldots$, and let $M_n := \max\{X_{iT2^{-n}} : 0 \leq i \leq 2^n\}$. For $\lambda > 0, [M_n > \lambda] \uparrow [M_T > \lambda]$ as $n \uparrow \infty$. In particular, M_T is $\mathcal{F}_T$-measurable. By Theorem 3.2,

$$P(M_n > \lambda) \leq \frac{1}{\lambda^p}\int_{[M_n > \lambda]} X_T^p dP \leq \frac{1}{\lambda^p}\mathbb{E}X_T^p.$$

Letting $n \uparrow \infty$, (3.13) is obtained. ■

We finally come to the notion of **stopping times,** which provide a powerful probabilistic tool to analyze processes by viewing them at appropriate random times.

Definition 3.4. Let $\{\mathcal{F}_t : t \in T\}$ be a filtration on a probability space $(\Omega, \mathcal{F}, P)$, with T a linearly ordered index set to which one may adjoin, if necessary, a point

'∞' as the largest point of $T \cup \{\infty\}$. A random variable $\tau : \Omega \to T \cup \{\infty\}$ is an $\{\mathcal{F}_t\}$-**stopping time** if $[\tau \leq t] \in \mathcal{F}_t \ \forall \ t \in T$. If $[\tau < t] \in \mathcal{F}_t$ for all $t \in T$ then τ is called an **optional time**.

Most commonly, T in this definition is $\mathbb{N}$ or $\mathbb{Z}^+$, or $[0, \infty)$, and τ is related to an $\{\mathcal{F}_t\}$-adapted process $\{X_t : t \in T\}$.

The intuitive idea of τ as a stopping-time strategy is that to "stop by time t, or not," according to τ, is determined by the knowledge of the past up to time t, and does not require "a peek into the future."

Example 4. Let $\{X_t : t \in T\}$ be an $\{\mathcal{F}_t\}$-adapted process with values in a measurable space $(S, \mathcal{S})$, with a linearly ordered index set. (a) If $T = \mathbb{N}$ or $\mathbb{Z}^+$, then for every $B \in \mathcal{S}$,

$$\tau_B := \inf\{t \geq 0 : X_t \in B\} \tag{3.14}$$

is an $\{\mathcal{F}_t\}$-stopping time. (b) If $T = \mathbb{R}_+ \equiv [0, \infty)$, S is a metric space $\mathcal{S} = \mathcal{B}(S)$, and B is *closed*, $t \mapsto X_t$ is continuous, then τ_B is an $\{\mathcal{F}_t\}$-stopping time. (c) If $T = \mathbb{R}_+, S$ is a topological space, $t \mapsto X_t$ is right-continuous, and B is *open*, then $[\tau_B < t] \in \mathcal{F}_t$ for all $t \geq 0$, and hence τ_B is an optional time.

We leave the proofs of (a)–(c) as Exercise 2. Note that (b), (c) imply that under the hypothesis of (b), τ_B is an optional time if B is open or closed.

Definition 3.5. Let $\{\mathcal{F}_t : t \in T\}$ be a filtration on $(\Omega, \mathcal{F})$. Suppose that τ is a $\{\mathcal{F}_t\}$-stopping time. The **pre-τ σ-field** $\mathcal{F}_\tau$ comprises all $A \in \mathcal{F}$ such that $A \cap [\tau \leq t] \in \mathcal{F}_t$ for all $t \in T$.

Heuristically, $\mathcal{F}_\tau$ comprises events determined by information available only up to time τ. For example, if T is discrete with elements $t_1 < t_2 < \cdots$, and $\mathcal{F}_t = \sigma(X_s : 0 \leq s \leq t) \subseteq \mathcal{F}, \forall t$, where $\{X_t : t \in T\}$ is a process with values in some measurable space $(S, \mathcal{S})$, then $\mathcal{F}_\tau = \sigma(X_{\tau \wedge t} : t \geq 0)$; (Exercise 8). The stochastic process $\{X_{\tau \wedge t} : t \geq 0\}$ is referred to as the **stopped process**.

If τ_1, τ_2 are two $\{\mathcal{F}_t\}$-stopping times and $\tau_1 \leq \tau_2$, then it is simple to check that (Exercise 1)

$$\mathcal{F}_{\tau_1} \subseteq \mathcal{F}_{\tau_2}. \tag{3.15}$$

Suppose $\{X_t\}$ is an $\{\mathcal{F}_t\}$-adapted process with values in a measurable space $(S, \mathcal{S})$, and τ is an $\{\mathcal{F}_t\}$-stopping time. For many purposes the following notion of adapted joint measurability of $(t, \omega) \mapsto X_t(\omega)$ is important.

Definition 3.6. Let $T = [0, \infty)$ or $T = [0, t_0]$ for some $t_0 < \infty$. A stochastic process $\{X_t : t \in T\}$ with values in a measurable space $(S, \mathcal{S})$ is **progressively measurable** with respect to $\{\mathcal{F}_t\}$ if for each $t \in T$, the map $(s, \omega) \mapsto X_s(\omega)$, from $[0, t] \times \Omega$ to S is measurable with respect to the σ-fields $\mathcal{B}[0,t]\otimes\mathcal{F}_t$ (on $[0,t]\times\Omega$) and $\mathcal{S}$ (on S). Here $\mathcal{B}[0,t]$ is the Borel σ-field on $[0,t]$, and $\mathcal{B}[0,t]\otimes\mathcal{F}_t$ is the usual product σ-field.

Proposition 3.5. (a) Suppose $\{X_t : t\in T\}$ is progressively measurable, and τ is a stopping time. Then X_τ is $\mathcal{F}_\tau$-measurable, i.e., $[X_\tau\in B]\cap[\tau\le t]\in\mathcal{F}_t$ for each $B\in\mathcal{S}$ and each $t\in T$. (b) Suppose $\mathcal{S}$ is a metric space and $\mathcal{S}$ its Borel σ-field. If $\{X_t : t\in T\}$ is right-continuous, then it is progressively measurable.

Proof. (a) Fix $t\in T$. On the set $\Omega_t := [\tau\le t]$, X_τ is the composition of the maps (i) $f(\omega) := (\tau(\omega),\omega)$, from $\omega\in\Omega_t$ into $[0,t]\times\Omega_t$, and (ii) $g(s,\omega)=X_s(\omega)$ on $[0,t]\times\Omega_t$ into S. Now f is $\tilde{\mathcal{F}}_t$-measurable on Ω_t, where $\tilde{\mathcal{F}}_t := \{A\cap\Omega_t : A\in\mathcal{F}_t\}$ is the **trace σ-field** on Ω_t, and $\mathcal{B}[0,t]\otimes\tilde{\mathcal{F}}_t$ is the σ-field on $[0,t]\times\Omega_t$. Next the map $g(s,\omega)=X_s(\omega)$ on $[0,t]\times\Omega$ into S is $\mathcal{B}[0,t]\otimes\mathcal{F}_t$-measurable. Therefore, the restriction of this map to the measurable subset $[0,t]\times\Omega_t$ is measurable on the trace σ-field $\{A\cap([0,t]\times\Omega_t) : A\in\mathcal{B}[0,t]\otimes\mathcal{F}_t\}$. Therefore, the composition X_τ is $\tilde{\mathcal{F}}_t$-measurable on Ω_t, i.e., $[X_\tau\in B]\cap[\tau\le t]\in\tilde{\mathcal{F}}_t\subseteq\mathcal{F}_t$ and hence $[X_\tau\in B]\in\mathcal{F}_\tau$, for $B\in\mathcal{S}$.

(b) Fix $t\in T$. Define, for each positive integer n, the stochastic process $\{X_s^{(n)} : 0\le s\le t\}$ by

$$X_s^{(n)} := X_{j2^{-n}t}\text{ for }(j-1)2^{-n}t\le s<j2^{-n}t\quad(1\le j\le 2^n),\ X_t^{(n)}=X_t.$$

Since $\{(s,\omega)\in[0,t]\times\Omega : X_s^{(n)}(\omega)\in B\}=\cup_{j=1}^{2^n}([j-1)2^{-n}t, j2^{-n}t)\times\{\omega : X_{j2^{-n}t}(\omega)\in B\})\cup(\{t\}\times\{\omega : X_t(\omega)\in B\})\in\mathcal{B}[0,t]\otimes\mathcal{F}_t$, and $X_s^{(n)}(\omega)\to X_s(\omega)$ for all (s,ω) as $n\to\infty$, in view of the right-continuity of $s\mapsto X_s(\omega)$, it follows that $\{(s,\omega)\in[0,t]\times\Omega : X_s(\omega)\in B\}\in\mathcal{B}[0,t]\otimes\mathcal{F}_t$. ■

Remark 3.5. It is often important to relax the assumption of 'right-continuity' of $\{X_t : t\in T\}$ to "a.s. right-continuity." To ensure progressive measurability progressive measurability in this case, it is convenient to take $\mathcal{F},\mathcal{F}_t$ to be P**-complete,** i.e., if $P(A)=0$ and $B\subseteq A$ then $B\in\mathcal{F}$ and $B\in\mathcal{F}_t$ $\forall$ t. Then modify X_t to equal X_0 $\forall$ t on the P-null set $N=\{\omega : t\to X_t(\omega)\ \text{ is not right-continuous}\}$. This modified $\{X_t : t\in T\}$, together with $\{\mathcal{F}_t : t\in T\}$ satisfy the hypothesis of part (b) of Proposition 3.5.

Theorem 3.6. *(Optional Stopping).* Let $\{X_t : t\in T\}$ be a right-continuous $\{\mathcal{F}_t\}$-martingale, where $T=\mathbb{N}$ or $T=[0,\infty)$. (a) If $\tau_1\le\tau_2$ are bounded stopping times, then

$$\mathbb{E}(X_{\tau_2}|\mathcal{F}_{\tau_1})=X_{\tau_1}. \tag{3.16}$$

(b) *(Optional Sampling).* If τ is a stopping time (not necessarily finite), then $\{X_{\tau\wedge t} : t\in T\}$ is an $\{\mathcal{F}_{\tau\wedge t}\}_{t\in T}$-martingale.

(c) Suppose τ is a stopping time such that (i) $P(\tau < \infty) = 1$, and (ii) $X_{\tau\wedge t}(t \in T)$ is uniformly integrable. Then

$$\mathbb{E}X_\tau = \mathbb{E}X_0. \tag{3.17}$$

Proof. We will give a proof for the case $T = [0, \infty)$. The case $T = \mathbb{N}$ is similar and simpler (Exercise 5). Let $\tau_1 \leq \tau_2 \leq t_0$ a.s. The idea for the proof is to check that $\mathbb{E}[X_{t_0}|\mathcal{F}_{\tau_i}] = X_{\tau_i}$, for each of the stopping times $(i = 1, 2)$ simply by virtue of their being bounded. Once this is established, the result (a) follows by smoothing of conditional expectation, since $\mathcal{F}_{\tau_1} \subseteq \mathcal{F}_{\tau_2}$. That is, it will then follow that

$$\mathbb{E}[X_{\tau_2}|\mathcal{F}_{\tau_1}] = \mathbb{E}[\mathbb{E}(X_{t_0}|\mathcal{F}_{\tau_2})|\mathcal{F}_{\tau_1}] = \mathbb{E}[X_{t_0}|\mathcal{F}_{\tau_1}] = X_{\tau_1}. \tag{3.18}$$

So let τ denote either of τ_i, $i = 1, 2$, and consider $\mathbb{E}[X_{t_0}|\mathcal{F}_\tau]$. For each $n \geq 1$ consider the nth dyadic subdivision of $[0, t_0]$ and define $\tau^{(n)} = (k+1)2^{-n}t_0$ if $\tau \in [k2^{-n}t_0, (k+1)2^{-n}t_0)(k = 0, 1, \ldots, 2^n - 1)$, and $\tau^{(n)} = t_0$ if $\tau = t_0$. Then $\tau^{(n)}$ is a stopping time and $\mathcal{F}_\tau \subseteq \mathcal{F}_{\tau^{(n)}}$ (since $\tau \leq \tau^{(n)}$). For $G \in \mathcal{F}_\tau$, exploiting the martingale property $\mathbb{E}[X_{t_0}|\mathcal{F}_{(k+1)2^{-n}t_0}] = X_{t_{(k+1)2^{-n}t_0}}$, one has

$$\begin{aligned}
\mathbb{E}(\mathbf{1}_G X_{t_0}) &= \sum_{k=0}^{2^n-1} \mathbb{E}(\mathbf{1}_{G\cap[\tau^{(n)}=(k+1)2^{-n}t_0]} X_{t_0}) \\
&= \sum_{k=0}^{2^n-1} \mathbb{E}(\mathbf{1}_{G\cap[\tau^{(n)}=(k+1)2^{-n}t_0]} X_{(k+1)2^{-n}t_0}) \\
&= \sum_{k=0}^{2^n-1} \mathbb{E}(\mathbf{1}_{G\cap[\tau^{(n)}=(k+1)2^{-n}t_0]} X_{\tau^{(n)}}) = \mathbb{E}(\mathbf{1}_G X_{\tau^{(n)}}) \to \mathbb{E}(\mathbf{1}_G X_\tau).
\end{aligned} \tag{3.19}$$

The last convergence is due to the L^1-convergence criterion of Theorem 1.8 in view of the following checks: (1) X_t is right-continuous (and $\tau^{(n)} \downarrow \tau$), so that $X_{\tau^{(n)}} \to X_\tau$ a.s., and (2) $X_{\tau^{(n)}}$ is uniformly integrable, since by the submartingale property of $\{|X_t| : t \in T\}$,

$$\begin{aligned}
\mathbb{E}(\mathbf{1}_{[|X_{\tau^{(n)}}|>\lambda]}|X_{\tau^{(n)}}|) &= \sum_{k=0}^{2^n-1} \mathbb{E}(\mathbf{1}_{[\tau^{(n)}=(k+1)2^{-n}t_0]\cap[|X_{\tau^{(n)}}|>\lambda]}|X_{(k+1)2^{-n}t_0}|) \\
&\leq \sum_{k=0}^{2^n-1} \mathbb{E}(\mathbf{1}_{[\tau^{(n)}=(k+1)2^{-n}t_0]\cap[|X_{\tau^{(n)}}|>\lambda]}|X_{t_0}|) \\
&= \mathbb{E}(\mathbf{1}_{[|X_{\tau^{(n)}}|>\lambda]}|X_{t_0}|) \to \mathbb{E}(\mathbf{1}_{[|X_\tau|>\lambda]}|X_{t_0}|).
\end{aligned}$$

Since the left side of (3.19) does not depend on n, it follows that

$$\mathbb{E}(\mathbf{1}_G X_{t_0}) = \mathbb{E}(\mathbf{1}_G X_\tau) \quad \forall\, G \in \mathcal{F}_\tau,$$

i.e., $\mathbb{E}(X_{t_0}|\mathcal{F}_\tau) = X_\tau$ applies to both $\tau = \tau_1$ and $\tau = \tau_2$. The result (a) therefore follows by the smoothing property of conditional expectations noted at the start of the proof.

(b) Follows immediately from (a). For if $s < t$ are given, then $\tau \wedge s$ and $\tau \wedge t$ are both bounded by t, and $\tau \wedge s \leq \tau \wedge t$.

(c) Since $\tau < \infty$ a.s., $\tau \wedge t$ equals τ for sufficiently large t (depending on ω), outside a P-null set. Therefore, $X_{\tau\wedge t} \to X_\tau$ a.s. as $t \to \infty$. By assumption (ii), $X_{\tau\wedge t}$ $(t \geq 0)$ is uniformly integrable. Hence $X_{\tau\wedge t} \to X_\tau$ in L^1. In particular, $\mathbb{E}(X_{\tau\wedge t}) \to \mathbb{E}(X_\tau)$ as $t \to \infty$. But $\mathbb{E}X_{\tau\wedge t} = \mathbb{E}X_0$ $\forall$ t, by (b). ■

Remark 3.6. If $\{X_t : t \in T\}$ in Theorem 3.6 is taken to be a submartingale, then instead of the equality sign "=" in (3.16), (3.17), one gets "≤."

The following proposition and its corollary are often useful for verifying the hypothesis of Theorem 3.6 in examples.

Proposition 3.7. Let $\{Z_n : n \in \mathbb{N}\}$ be real-valued random variables such that for some $\varepsilon > 0$, $\delta > 0$, one has

$$P(Z_{n+1} > \varepsilon \mid \mathcal{G}_n) \geq \delta, \text{ a.s.} \quad \forall\, n = 0, 1, 2, \ldots$$

or

$$P(Z_{n+1} < -\varepsilon \mid \mathcal{G}_n) \geq \delta \text{ a.s.} \quad \forall\, n = 0, 1, 2, \ldots, \tag{3.20}$$

where $\mathcal{G}_n = \sigma\{Z_1, \ldots, Z_n\}$ $(n \geq 1)$, $\mathcal{G}_0 = \{\emptyset, \Omega\}$. Let $S_n^x = x + Z_1 + \cdots + Z_n$ $(n \geq 1)$, $S_0^x = x$, and let $a < x < b$. Let τ be the first escape time of $\{S_n^x\}$ from (a, b), i.e., $\tau = \inf\{n \geq 1 : S_n^x \in (a, b)^c\}$. Then $\tau < \infty$ a.s. and

$$\sup_{\{x: a<x<b\}} \mathbb{E}e^{\tau z} < \infty \quad \text{for} \quad -\infty < z < \frac{1}{n_0}\left(\log \frac{1}{1-\delta_0}\right), \tag{3.21}$$

where, writing $[y]$ for the integer part of y,

$$n_0 = \left[\frac{b-a}{\varepsilon}\right] + 1, \qquad \delta_0 = \delta^{n_0}. \tag{3.22}$$

Proof. Suppose the first relation in (3.20) holds. Clearly, if $Z_j > \varepsilon$ $\forall$ $j = 1, 2, \ldots, n_0$, then $S_{n_0}^x > b$, so that $\tau \leq n_0$. Therefore, $P(\tau \leq n_0) \geq P(Z_1 > \varepsilon, \ldots, Z_{n_0} > \varepsilon) \geq \delta^{n_0}$, by taking successive conditional expectations (given $\mathcal{G}_{n_0-1}, \mathcal{G}_{n_0-2}, \ldots, \mathcal{G}_0$, in that order). Hence $P(\tau > n_0) \leq 1 - \delta^{n_0} = 1 - \delta_0$. For every integer $k \geq 2$, $P(\tau > kn_0) = P(\tau > (k-1)n_0, \tau > kn_0) = \mathbb{E}[\mathbf{1}_{[\tau>(k-1)n_0]} P(\tau > kn_0|\mathcal{G}_{(k-1)n_0})] \leq (1-\delta_0)P(\tau > (k-1)n_0)$, since, on the set $[\tau > (k-1)n_0]$, $P(\tau \leq kn_0|\mathcal{G}_{(k-1)n_0}) \geq P(Z_{(k-1)n_0+1} > \varepsilon, \ldots, Z_{kn_0} > \varepsilon|\mathcal{G}_{(k-1)n_0}) \geq \delta^{n_0} = \delta_0$. By induction, $P(\tau > kn_0) \leq (1-\delta_0)^k$.

Hence $P(\tau = \infty) = 0$, and for all $z > 0$,

$$\begin{aligned}\mathbb{E}e^{z\tau} &= \sum_{r=1}^{\infty} e^{zr} P(\tau = r) \leq \sum_{k=1}^{\infty} e^{zkn_0} \sum_{r=(k-1)n_0+1}^{kn_0} P(\tau = r) \\ &\leq \sum_{k=1}^{\infty} e^{zkn_0} P(\tau > (k-1)n_0) \leq \sum_{k=1}^{\infty} e^{zkn_0} (1-\delta_0)^{k-1} \\ &= e^{zn_0}(1-(1-\delta_0)e^{zn_0})^{-1} \quad \text{if} \quad e^{zn_0}(1-\delta_0) < 1.\end{aligned}$$

An entirely analogous argument holds if the second relation in (3.20) holds. ■

The following corollary immediately follows from Proposition 3.7.

Corollary 3.8. Let $\{Z_n : n = 1, 2, \cdots\}$ be an i.i.d. sequence such that $P(Z_1 = 0) < 1$. Let $S_n^n = x + Z_1 + \cdots + Z_n$ $(n \geq 1)$, $S_0^x = x$, and $a < x < b$. Then the first escape time τ of the random walk from the interval (a, b) has a finite moment generating function in a neighborhood of 0.

Example 5. Let $Z_n (n \geq 1)$ be i.i.d. symmetric Bernoulli, $P(Z_i = +1) = P(Z_i = -1) = \frac{1}{2}$, and let $S_n^x = x + Z_1 + \cdots + Z_n (n \geq 1)$, $S_0^x = x$, be the simple symmetric random walk on the state space $\mathbb{Z}$, starting at x. Let $a \leq x \leq b$ be integers, $\tau_y := \inf\{n \geq 0 : S_n^x = y\}$, $\tau = \tau_a \wedge \tau_b = \inf\{n \geq 0 : S_n^x \in \{a, b\}\}$. Then $\{S_n^x : n \geq 0\}$ is a martingale and τ satisfies the hypothesis of Theorem 3.6 (c) (Exercise 7). Hence

$$x \equiv \mathbb{E}S_0^x = \mathbb{E}S_\tau^x = aP(\tau_a < \tau_b) + bP(\tau_b < \tau_a) = a + (b-a)P(\tau_b < \tau_a),$$

so that

$$P(\tau_b < \tau_a) = \frac{x-a}{b-a}, \quad P(\tau_a < \tau_b) = \frac{b-x}{b-a}, \quad a \leq x \leq b. \tag{3.23}$$

To illustrate the importance of the hypothesis imposed on τ in Theorem 3.6 (c), one may naively try to apply (3.17) to τ_b (see Exercise 7) and arrive at the silly conclusion $x = b$!

Example 6. One may apply Theorem 3.6 (c) to a simple asymmetric random walk with $P(Z_i = 1) = p$, $P(Z_i = -1) = q \equiv 1 - p(0 < p < 1,\ p \neq 1/2)$, so that $X_n^x := S_n^x - (2p-1)n$ $(n \geq 1)$, $X_0^x \equiv x$, is a martingale. Then with τ_a, τ_b, $\tau = \tau_a \wedge \tau_b$ as above, one gets

$$x \equiv \mathbb{E}X_0^x = \mathbb{E}X_\tau^x = \mathbb{E}S_\tau^x - (2p-1)\mathbb{E}\tau = a + (b-a)P(\tau_b < \tau_a) - (2p-1)\mathbb{E}\tau. \tag{3.24}$$

Since we do not know $\mathbb{E}\tau$ yet, we can not quite solve (3.24). We therefore use a second martingale $(q/p)^{S_n^x}$ $(n \geq 0)$. Note that $\mathbb{E}[(q/p)^{S_{n+1}^x} | \sigma\{Z_1, \ldots, Z_n\}] = (q/p)^{S_n^x} \cdot$

$\mathbb{E}[(q/p)^{Z_{n+1}}] = (q/p)^{S_n^x}[(q/p)p + (q/p)^{-1}q] = (q/p)^{S_n^x} \cdot 1 = (q/p)^{S_n^x}$, proving the martingale property of the "exponential process" $Y_n := (q/p)^{S_n^x} = \exp(cS_n^x), c = \ln(q/p), n \geq 0$. Note that $(q/p)^{S_{\tau\wedge n}^x} \leq \max\{(q/p)^y : a \leq y \leq b\}$, which is a finite number. Hence the hypothesis of uniform integrability holds. Applying (3.17) we get

$$(q/p)^x = (q/p)^a \cdot P(\tau_a < \tau_b) + (q/p)^b P(\tau_b < \tau_a),$$

or

$$P(\tau_b < \tau_a) = \frac{(q/p)^x - (q/p)^a}{(q/p)^b - (q/p)^a} \equiv \varphi(x) \quad (a \leq x \leq b). \tag{3.25}$$

Using this in (3.24) we get

$$\mathbb{E}\tau \equiv \mathbb{E}\tau_a \wedge \tau_b = \frac{x - a - (b-a)\varphi(x)}{1-2p}, \quad a \leq x \leq b. \tag{3.26}$$

EXERCISES

Exercise Set III

1. Prove (3.15). Also prove that an $\{\mathcal{F}_t\}$-stopping time is an $\{\mathcal{F}_t\}$-optional time.

2. (i) Prove that τ_B defined by (3.14) is an $\{\mathcal{F}_t\}$-stopping time if B is closed and $t \mapsto X_t$ is continuous with values in a metric space (S, ρ). [*Hint*: For $t > 0, B$ closed, $[\tau_B \leq t] = \cap_{n\in\mathbb{N}} \cup_{r\in Q\cap[0,t]} [\rho(X_r, B) \leq \frac{1}{n}]$, where Q is the set of rationals.] (ii) Prove that if $t \mapsto X_t$ is right-continuous, τ_B is an optional time for B open. [*Hint*: For B open, $t > 0$, $[\tau_B < t] = \cup_{r\in Q\cap(0,t)}[X_r \in B]$.] (iii) If $T = \mathbb{N}$ or $\mathbb{Z}^+$, prove that τ_B is a stopping time for all $B \in \mathcal{S}$.

3. (i) If τ_1 and τ_2 are $\{\mathcal{F}_t\}$-stopping times, then show that so are $\tau_1 \wedge \tau_2$ and $\tau_1 \vee \tau_2$. (ii) Show that $\tau + c$ is an $\{\mathcal{F}_t\}$-stopping time if τ is an $\{\mathcal{F}_t\}$-stopping time, $c > 0$, and $\tau + c \in T \cup \{\infty\}$. (iii) Show that (ii) is false if $c < 0$.

4. If τ is a discrete random variable with values $t_1 < t_2 < \cdots$ in a finite or countable set T in $\mathbb{R}$, then τ is an $\{\mathcal{F}_t\}_{t\in T}$-stopping time if and only if $[\tau = t] \in \mathcal{F}_t \ \forall\ t \in T$.

5. Let $\{X_n : n = 0, 1, 2, \ldots\}$ be an $\{\mathcal{F}_n\}$-martingale, and τ an $\{\mathcal{F}_n\}$-stopping time. Give simple direct proofs of the following: (i) $\mathbb{E}X_\tau = \mathbb{E}X_0$ if τ is bounded. [*Hint*: Let $\tau \leq m$ a.s. Then $\mathbb{E}X_\tau = \sum_{n=0}^m \mathbb{E}X_n \mathbf{1}_{[\tau=n]} = \sum_{n=0}^m \mathbb{E}X_m \mathbf{1}_{[\tau=n]} = \mathbb{E}X_m \mathbf{1}_{[\tau\leq m]} = \mathbb{E}X_m = \mathbb{E}X_0$.] (ii) If $\mathbb{E}\tau < \infty$ and $\mathbb{E}|X_{\tau\wedge m} - X_\tau| \to 0$ as $m \to \infty$, then $\mathbb{E}X_\tau = \mathbb{X}_0$.

6. (*Wald's Identity*) Let $\{Y_j : j \geq 1\}$ be an i.i.d. sequence with finite mean μ, and take $Y_0 = 0, a.s.$ Let τ be an $\{\mathcal{F}_n\}$-stopping time, where $\mathcal{F}_n = \sigma(Y_j : j \leq n)$. Write $S_n = \sum_{j=0}^n Y_j$. If $\mathbb{E}\tau < \infty$ and $\mathbb{E}|S_\tau - S_{\tau\wedge m}| \to 0$ as $m \to \infty$, prove that $\mathbb{E}S_\tau = \mu\mathbb{E}\tau$. [*Hint*: Apply Theorem 3.6(c) to the martingale $\{S_n - n\mu : n \geq 0\}$.]

7. In Example 5 for $\tau = \tau_a \wedge \tau_b$, show that (i) $\mathbb{E}\tau < \infty \ \forall \ a \leq x \leq b$, and $|S_{\tau\wedge n}| \leq \max\{|a|, |b|\} \ \forall \ n \geq 0$, is uniformly integrable, (ii) $P(\tau_a < \infty) = 1 \ \forall \ x, a$, but $\{S_{\tau_a \wedge n} : n \geq 0\}$ is not uniformly integrable. (iii) For Example 5 also show that $Y_n := S_n^2 - n, n \geq 0$, is a martingale and $\{Y_{\tau\wedge n} : n \geq 0\}$ is uniformly integrable. Use this to calculate $\mathbb{E}\tau$. [*Hint*: Use triangle inequality estimates on $|Y_{\tau\wedge n}| \leq |S_{\tau\wedge n}|^2 + \tau \wedge n$.]

8. Let $\{X_t : t \in T\}$ be a stochastic process on $(\Omega, \mathcal{F})$ with values in some measurable space $(S, \mathcal{S})$, T a discrete set with elements $t_1 < t_2 < \cdots$. Define $\mathcal{F}_t = \sigma(X_s : 0 \leq s \leq t) \subseteq \mathcal{F}$, $t \in T$. Assume that τ is an $\{\mathcal{F}_t\}$-stopping time and show that $\mathcal{F}_\tau = \sigma(X_{\tau\wedge t} : t \in T)$; i.e., $\mathcal{F}_\tau$ is the σ-field generated by the stopped process $\{X_{\tau\wedge t} : t \in T\}$.

9. Prove that if τ is an optional time with respect to a filtration $\{\mathcal{F}_t : 0 \leq t < \infty\}$, then τ is a stopping time with respect to $\{\mathcal{F}_{t+} : 0 \leq t < \infty\}$, where $\mathcal{F}_{t+} := \cap_{\varepsilon>0}\mathcal{F}_{t+\varepsilon}$. Deduce that under the hypothesis of Example 4(b), if B is open or closed, then τ_B is a stopping time with respect to $\{\mathcal{F}_{t+} : 0 \leq t < \infty\}$.

10. Let $\{\mathcal{F}_t : t \in T\}$ and $\{\mathcal{G}_t : t \in T\}$ be two filtrations of $(\Omega, \mathcal{F})$, each adapted to $\{X_t : t \in T\}$, and assume $\mathcal{F}_t \subseteq \mathcal{G}_t, \forall t \in T$. Show that if $\{X_t : t \in T\}$ is a $\{\mathcal{G}_t\}$-martingale (or sub or super) then it is an $\{\mathcal{F}_t\}$-martingale (or respectively sub or super).

11. Let $Z_1, Z_2, \ldots$ be i.i.d. ±1-valued Bernoulli random variables with $P(Z_n = 1) = p, P(Z_n = -1) = 1 - p, n \geq 1$, where $0 < p < 1/2$. Let $S_n = Z_1 + \cdots + Z_n, n \geq 1, S_0 = 0$.
   (i) Show that $P(\sup_{n\geq 0} S_n > y) \leq (\frac{p}{q})^y, y \geq 0$. [*Hint*: Apply a maximal inequality to $X_n = (q/p)^{S_n}$.]
   (ii) Show for $p < 1/2$ that $\mathbb{E}\sup_{n\geq 0} S_n \leq \frac{p}{q-p}$. [*Hint*: Use (2.5).]

12. Suppose that $Z_1, Z_2, \ldots$ is a sequence of independent random variables with $\mathbb{E}Z_n = 0$ such that $\sum_n \mathbb{E}Z_n^2 < \infty$. Show that $\sum_{n=1}^\infty Z_n := \lim_N \sum_{n=1}^N Z_n$ exists a.s. [*Hint*: Let $S_j = \sum_{k=1}^j Z_k$ and show that $\{S_j\}$ is a.s. a Cauchy sequence. For this note that $Y_n := \max_{k,j\geq n} |S_k - S_j|$ is a.s. a decreasing sequence and hence has a limit a.s. Apply Kolmogorov's maximal inequality to $\max_{n\leq j\leq N} |S_j - S_n|$ to show that the limit in probability is zero, and hence a.s. zero.]
   (i) For what values of θ will $\sum_{n=1}^\infty Z_n$ converge a.s. if $P(Z_n = n^{-\theta}) = P(Z_n = -n^{-\theta}) = 1/2$?
   (ii) (Random Signs) Suppose each Z_n is symmetric Bernoulli ±1-valued. Show that the series $\sum_{n=1}^\infty Z_n a_n$ converges with probability one if $\{a_n\}$ is any square-summable sequence of real numbers.
   (iii) Show that $\sum_{n=1}^\infty Z_n \sin(n\pi t)/n$ converges a.s. for each t if the Z_n's are i.i.d. standard normal.

CHAPTER IV

Classical Zero–One Laws, Laws of Large Numbers and Large Deviations

The term *law* has various meanings within probability. It is sometimes used synonymously with *distribution* of a random variable. However, it also may refer to an event or phenomenon that occurs in some predictable sense, as in a "law of averages." The latter is the context of the present section. For example, if $X_0, X_1, \ldots$ is a sequence of independent random variables and $B \in \mathcal{B}$, then, in view of the Borel–Cantelli lemmas, one may conclude that the event $[X_n \in B \ i.o.]$ will occur with probability one, or its complement is certain to occur. Before taking up the laws of large numbers, we consider two standard zero–one laws of this type. In particular, observe that the event $A = [X_n \in B \ i.o.]$ is special in that it does not depend on any finite number of values of the sequence $X_0, X_1, X_2, \ldots$. Such an event is referred to as a tail event. That is, an event $E \in \sigma(X_0, X_1, X_2, \ldots)$ is said to be a **tail event** if $E \in \sigma(X_n, X_{n+1}, \ldots)$ for every $n \geq 0$. The collection of all tail events is given by the **tail σ-field** $\mathcal{T} := \cap_{n=0}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$.

***Theorem 4.1** (Kolmogorov Zero–One Law).* A tail event for a sequence of independent random variables has probability either zero or one.

Proof. To see this first check that $\sigma(X_0, X_1, \ldots) = \sigma(\mathcal{F}_0)$, where $\mathcal{F}_0 := \cup_{k=0}^{\infty} \sigma(X_0, \ldots, X_k)$ is a field and, in particular, a π-system. For $E \in \mathcal{F}_0$, one has $E = [(X_0, \ldots, X_k) \in C]$ for some $k \geq 0, C \in \mathcal{B}^{k+1}$. Thus if A is a tail event then $A \in \sigma(X_{k+1}, \ldots)$ and hence A is independent of E; i.e., A is independent of $\mathcal{F}_0$ and hence of $\sigma(\mathcal{F}_0) = \sigma(X_0, X_1, \ldots)$ since both $\{A\}$ and $\mathcal{F}_0$ are π-systems. This makes A independent of itself and hence $P(A) = P(A \cap A) = P(A)P(A)$. The only solutions to the equation $x^2 = x$ are 0 and 1. ■

Not all tail events for the sums need be tail events for the terms of the series. Let $S_n = X_1 + \cdots + X_n, n \geq 1$. For example, an event of the form $[S_n \in B \ i.o.]$ is not covered by Kolmogorov's zero–one law since the sums $S_1, S_2, \ldots$ are not independent. However, there is a special way in which such tail events for the sums depend on the sequence $X_1, X_2, \ldots$ of i.i.d. summands captured by the following zero-one law.

Let $\mathcal{B}^\infty$ denote the (Borel) σ-field of subsets of $\mathbb{R}^\infty = \{(x_1, x_2, \ldots) : x_i \in \mathbb{R}^1\}$ generated by events depending on finitely many coordinates.

***Theorem 4.2** (Hewitt–Savage Zero–One Law).* Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables. If an event $A = [(X_1, X_2, \ldots) \in B]$, where $B \in \mathcal{B}^\infty$, is invariant under finite permutations $(X_{i_1}, X_{i_2}, \ldots)$ of terms of the sequence $(X_1, X_2, \ldots)$, that is, $A = [(X_{i_1}, X_{i_2}, \ldots) \in B]$ for any finite permutation $(i_1, i_2, \ldots)$ of $(1, 2, \ldots)$, then $P(A) = 1$ or 0.

As noted above, the symmetric dependence with respect to $\{X_n\}_{n=1}^\infty$ applies, for example, to tail events for the sums $\{S_n\}_{n=1}^\infty$.

Proof. To prove the Hewitt–Savage 0–1 law, proceed as in the Kolmogorov 0–1 law by selecting finite-dimensional approximants to A of the form $A_n = [(X_1, \ldots, X_n) \in B_n]$, $B_n \in \mathcal{B}^n$, such that $P(A \Delta A_n) \to 0$ as $n \to \infty$, where $E \Delta F := (E^c \cap F) \cup (E \cap F^c)$ is the *symmetric difference* of sets E, F; this approximation may be achieved from the Carathéodory extension formula (see Exercise 1). For each fixed n, let $(i_1, i_2, \ldots)$ be the permutation $(2n, 2n-1, \ldots, 1, 2n+1, \ldots)$ and define $\tilde{A}_n = [(X_{i_1}, \ldots, X_{i_n}) \in B_n]$. Then $\tilde{A}_n$, and A_n are independent with $P(A_n \cap \tilde{A}_n) = P(A_n)P(\tilde{A}_n) = (P(A_n))^2 \to (P(A))^2$ as $n \to \infty$. On the other hand, $P(A \Delta \tilde{A}_n) = P(A \Delta A_n) \to 0$. Note that $A \Delta \tilde{A}_n$ is obtained by a permutation from $A \Delta A_n$. Hence $P(A_n \Delta \tilde{A}_n) \leq P(A_n \Delta A) + P(\tilde{A}_n \Delta A) \to 0$ and, in particular, therefore $P(A_n \cap \tilde{A}_n) \to P(A)$ as $n \to \infty$. Thus $x = P(A)$ satisfies $x = x^2$. ■

The classical strong law of large numbers (SLLN) refers to the almost sure limit of averages of a "large number" of i.i.d. random variables having finite first moment. While the zero-one laws are not required in the following proof, they do imply that the indicated limit of the averages is either sure to exist or sure not to exist.

A good warm-up exercise is to work out a proof using the Borel–Cantelli lemma I based on Chebyshev inequality estimates, assuming finite fourth moments (Exercise 2). The proof we present in this section is due to Etemadi.[1] It is based on Part 1 of the Borel–Cantelli lemmas. Other proofs are included by other methods in other chapters of the text on stochastic processes, e.g., as consequences of the ergodic theorem and by the martingale convergence theorem.

***Theorem 4.3** (Strong Law of Large Numbers).* Let $\{X_n : n \geq 1\}$ be a sequence of pairwise independent and identically distributed random variables defined on a

[1] Etemadi, N. (1983): "On the Laws of Large Numbers for Nonnegative Random Variables," *J. Multivariate Analysis*, **13**, pp. 187–193.

probability space $(\Omega, \mathcal{F}, P)$. If $\mathbb{E}|X_1| < \infty$ then with probability 1,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mathbb{E}X_1. \tag{4.1}$$

Proof. Without loss of generality we may assume for the proof of the SLLN that the random variables X_n are nonnegative, since otherwise we can write $X_n = X_n^+ - X_n^-$, where $X_n^+ = \max(X_n, 0)$ and $X_n^- = -\min(X_n, 0)$ are both nonnegative random variables, and then the result in the nonnegative case yields that

$$\frac{S_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k^+ - \frac{1}{n}\sum_{k=1}^{n} X_k^-$$

converges to $\mathbb{E}X_1^+ - \mathbb{E}X_1^- = \mathbb{E}X_1$ with probability 1.

Truncate the variables X_n by $Y_n = X_n \mathbf{1}_{[X_n \le n]}$. Then Y_n has moments of all orders. Let $T_n = \sum_{k=1}^n Y_k$ and consider the sequence $\{T_n\}_{n=1}^\infty$ on the "fast" time scale $\tau_n = [\alpha^n]$, for a fixed $\alpha > 1$, where brackets [ ] denote the integer part. Let $\varepsilon > 0$. Then by Chebyshev's inequality and pairwise independence,

$$\begin{aligned}
P\left(\left|\frac{T_{\tau_n} - ET_{\tau_n}}{\tau_n}\right| > \varepsilon\right) &\le \frac{\operatorname{Var}(T_{\tau_n})}{\varepsilon^2\tau_n^2} = \frac{1}{\varepsilon^2\tau_n^2}\sum_{k=1}^{\tau_n} \operatorname{Var} Y_k \le \frac{1}{\varepsilon^2\tau_n^2}\sum_{k=1}^{\tau_n} \mathbb{E}Y_k^2 \\
&= \frac{1}{\varepsilon^2\tau_n^2}\sum_{k=1}^{\tau_n} \mathbb{E}\{X_k^2\mathbf{1}_{[X_k\le k]}\} = \frac{1}{\varepsilon^2\tau_n^2}\sum_{k=1}^{\tau_n} \mathbb{E}\{X_1^2\mathbf{1}_{[X_1\le k]}\} \\
&\le \frac{1}{\varepsilon^2\tau_n^2}\sum_{k=1}^{\tau_n} \mathbb{E}\{X_1^2\mathbf{1}_{[X_1\le \tau_n]}\} = \frac{1}{\varepsilon^2\tau_n^2}\tau_n\mathbb{E}\{X_1^2\mathbf{1}_{[X_1\le \tau_n]}\}. \qquad (4.2)
\end{aligned}$$

Therefore,

$$\sum_{n=1}^{\infty} P\left(\left|\frac{T_{\tau_n} - \mathbb{E}T_{\tau_n}}{\tau_n}\right| > \varepsilon\right) \le \sum_{n=1}^{\infty} \frac{1}{\varepsilon^2\tau_n}\mathbb{E}\{X_1^2\mathbf{1}_{[X_1\le\tau_n]}\} = \frac{1}{\varepsilon^2}\mathbb{E}\left\{X_1^2\sum_{n=1}^{\infty}\frac{1}{\tau_n}\mathbf{1}_{[X_1\le\tau_n]}\right\}. \tag{4.3}$$

Let $x > 0$ and let $N = \min\{n \ge 1 : \tau_n \ge x\}$. Then $\alpha^N \ge x$, and since $y \le 2[y]$ for any $y \ge 1$,

$$\sum_{n=1}^{\infty} \frac{1}{\tau_n}\mathbf{1}_{[x\le\tau_n]} = \sum_{\tau_n \ge x}\frac{1}{\tau_n} \le 2\sum_{n\ge N}\alpha^{-n} = \frac{2\alpha}{\alpha-1}\alpha^{-N} = a\alpha^{-N} \le \frac{a}{x},$$

where $a = 2\alpha/(\alpha - 1)$. Therefore,

$$\sum_{n=1}^{\infty}\frac{1}{\tau_n}\mathbf{1}_{[X_1\le\tau_n]} \le \frac{a}{X_1} \qquad \text{for } X_1 > 0.$$

So

$$\sum_{n=1}^{\infty} P\left(\left|\frac{T_{\tau_n} - \mathbb{E}T_{\tau_n}}{\tau_n}\right| > \varepsilon\right) \leq a\frac{\mathbb{E}[X_1]}{\varepsilon^2} < \infty. \tag{4.4}$$

By the Borel–Cantelli lemma I, taking a union over positive rational values of ε, with probability 1, $(T_{\tau_n} - \mathbb{E}T_{\tau_n})/\tau_n \to 0$ as $n \to \infty$. Therefore,

$$\frac{T_{\tau_n}}{\tau_n} \to \mathbb{E}X_1, \tag{4.5}$$

since $\mathbb{E}Y_n \to \mathbb{E}X_1$,

$$\lim_{n\to\infty} \frac{1}{\tau_n}\mathbb{E}T_{\tau_n} = \lim_{n\to\infty} \mathbb{E}Y_{\tau_n} = \mathbb{E}X_1.$$

Since

$$\sum_{n=1}^{\infty} P(X_n \neq Y_n) = \sum_{n=1}^{\infty} P(X_1 > n) \leq \int_0^{\infty} P(X_1 > u)\,du = \mathbb{E}X_1 < \infty,$$

we get by another application of the Borel–Cantelli lemma that, with probability 1,

$$\frac{S_n - T_n}{n} \to 0 \qquad \text{as } n \to \infty. \tag{4.6}$$

Therefore, the previous results about $\{T_n\}$ give for $\{S_n\}$ that

$$\frac{S_{\tau_n}}{\tau_n} \to \mathbb{E}X_1 \qquad \text{as } n \to \infty \tag{4.7}$$

with probability 1. If $\tau_n \leq k \leq \tau_{n+1}$, then since $X_i \geq 0$,

$$\frac{\tau_n}{\tau_{n+1}}\frac{S_{\tau_n}}{\tau_n} \leq \frac{S_k}{k} \leq \frac{\tau_{n+1}}{\tau_n}\frac{S_{\tau_{n+1}}}{\tau_{n+1}}. \tag{4.8}$$

But $\tau_{n+1}/\tau_n \to \alpha$, so that now we get with probability 1,

$$\frac{1}{\alpha}\mathbb{E}X_1 \leq \liminf_k \frac{S_k}{k} \leq \limsup_k \frac{S_k}{k} \leq \alpha\mathbb{E}X_1. \tag{4.9}$$

Take the intersection of all such events for rational $\alpha > 1$ to get $\lim_{k\to\infty} S_k/k = \mathbb{E}X_1$ with probability 1. This is the strong law of large numbers (SLLN). ■

The above proof of the SLLN is really quite remarkable, as the following observations show. First, *pairwise independence* is used only to make sure that the positive and negative parts of X_n, and their truncations, remain *(pairwise) uncorrelated* for the calculation of the variance of T_k as the sum of the variances. Observe that if the random variables all are mean zero and are uniformly bounded below, then it suffices to require that they merely be *uncorrelated* for the same proof to go through. However, this means that if the random variables are bounded, then it suffices that they be uncorrelated to get the SLLN; for one may simply add a sufficiently large constant to make them all positive. Thus, we have the following (Exercise 3).

Proposition 4.4. Let $X_1, X_2, \ldots,$ be a sequence of mean-zero uncorrelated random variables that are uniformly bounded below, or uniformly bounded above. If, in addition, $VarX_n, n \geq 1$, is a bounded sequence, then with probability 1,

$$\frac{X_1 + \cdots + X_n}{n} \to 0 \qquad \text{as } n \to \infty.$$

In particular, this holds for every bounded, mean-zero, uncorrelated sequence.

Corollary 4.5. If $X_1, X_2, \ldots$ is an i.i.d. sequence and $\mathbb{E}X_1^+ = \infty$ and $\mathbb{E}X_1^- < \infty$, then with probability 1,

$$\frac{X_1 + \cdots + X_n}{n} \to \infty \qquad \text{as } n \to \infty.$$

Similarly, if $\mathbb{E}X_1^+ < \infty$ and $\mathbb{E}X_1^- = \infty$, then the a.s. limit is $-\infty$.

As an obvious corollary, since a.s. convergence implies convergence in probability, one may conclude that the averages converge in probability as well. The latter statement is referred to as a *weak law of large numbers* (WLLN).

The proof of the Weierstrass approximation theorem given in Appendix B may be viewed as an application of the WLLN to a classic problem in calculus; namely, a continuous function f on $[0,1]$ may be uniformly approximated by polynomials $q_n(x) = \sum_{k=0}^{n} \binom{n}{k} f(\frac{k}{n}) x^k (1-x)^{n-k}$, $0 \leq x \leq 1$, referred to as **Bernstein polynomials**.

Let us now briefly turn some attention to deviations from the law of averages. Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with mean μ. Then the WLLN implies that for any $\delta > 0$, the event that the sample average $A_N := \frac{X_1+\cdots+X_N}{N}$ would fall outside the interval $\mu \pm \delta$, i.e., "would deviate from μ by a positive amount δ," is a rare event for large N. In fact, under suitable conditions one might expect the probability to be *exponentially small* for large N. The "large deviation theorem" below provides an important illustration of such conditions. We will need a few preliminaries to prepare for the statement and proof.

Definition 4.1. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ with distribution Q. The **moment generating function** of X (or Q) is defined by $m(h) = \mathbb{E}e^{hX} = \int_{\mathbb{R}} e^{hx} Q(dx)$. The **cumulant generating function** is $c(h) = \ln m(h)$, $h \in \mathbb{R}$.

Note that $m(h)$ may be infinite; see Exercise 6. The function $m(-h)$ is the **Laplace transform** of the distribution Q.

Proposition 4.6. (a) Assume that $m(h) < \infty$ for all h in a neighborhood of $h = 0$. Then $\mathbb{E}|X|^k < \infty$ for all $k \geq 1$ and $\mathbb{E}X^k = m^{(k)}(0) \equiv \frac{d^k}{dh^k} m(0)$. (b) Assume that $m(h) < \infty$ for all h in a neighborhood of $h = r \in \mathbb{R}$. Then $\mathbb{E}|X^k e^{rX}| < \infty$ and $m^{(k)}(r) \equiv \frac{d^k}{dr^k} m(r) = \mathbb{E}X^k e^{rX}$, for all $k \geq 1$.

Proof. Since $e^{|hx|} \leq e^{hx} + e^{-hx}$, it follows from the hypothesis for (a) that $\mathbb{E}e^{hX} \leq \mathbb{E}e^{|hX|} < \infty$ for all h in a neighborhood of $h = 0$. Also, since the partial sums $\sum_{k=0}^{n} \frac{|hX|^k}{k!}$ are bounded by $e^{|hX|}$, one has by the dominated convergence theorem that $\mathbb{E}e^{hX} = \mathbb{E}\sum_{k=0}^{\infty} \frac{h^k X^k}{k!} = \sum_{k=0}^{\infty} \frac{\mathbb{E}X^k}{k!} h^k$. The assertion (a) now follows from the uniqueness of the coefficients in Taylor series expansions about the origin (Exercise 7). For part (b) consider the change of measure defined by $\tilde{Q}(dx) = \frac{e^{rx}}{m(r)} Q(dx)$. Recall that the factor $\frac{1}{m(r)}$ is the normalization of $e^{rx}Q(dx)$ to a probability. If $\tilde{X}$ is a random variable with distribution $\tilde{Q}$, then its moment-generating function is given by $\tilde{m}(h) = \frac{m(h+r)}{m(r)}$. Under hypothesis (b), $\tilde{X}$ has a moment-generating function in a neighborhood of $h = 0$, so that (a) yields $\mathbb{E}X^k e^{rX}/m(r) = m^{(k)}(r)/m(r)$. Multiplying by $m(r)$ yields the assertion (b). ■

Definition 4.2. Suppose that $\hat{\mu}(b) \equiv \int e^{bx}\mu(dx) < \infty$. The change of measure defined by $\tilde{\mu}(dx) = \frac{e^{bx}}{\hat{\mu}(b)}\mu(dx)$ is called an **exponential size-bias** or **tilting** transformation.

For the rest of this section assume that X is **nondegenerate**, i.e. Q is not a Dirac measure δ_c. Assuming that $m(h)$ is finite in a neighborhood of zero, the second derivative of $m(h)$ is obviously positive, and one may use the Cauchy–Schwarz inequality to check $c^{(2)}(h) > 0$ as well; see Exercise 8.

Corollary 4.7. Suppose that $m(h) = \mathbb{E}e^{hX} < \infty$ for all $h \in \mathbb{R}$. Then both $m(h)$ and $c(h)$ are convex functions on $\mathbb{R}$.

Theorem 4.8 (Cramér-Chernoff). Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence with finite mean $\mathbb{E}X_1 = \mu$. Moreover, assume that the moment-generating function

$$m(h) := \mathbb{E}e^{hX_1}$$

is finite for all $h \in \mathbb{R}$. Let $c(h) := \ln m(h)$ denote the cumulant-generating function. Then for $A_N = \frac{X_1 + \cdots + X_N}{N}$,

$$\lim_{N\to\infty} \frac{\ln P(A_N \geq \mu + \delta)}{N} = I(\delta), \quad \delta > 0, \tag{4.10}$$

where $I(\delta) = -c^*(\mu+\delta)$ for

$$c^*(x) = \sup_{h\in\mathbb{R}}\{xh - c(h)\} \geq 0.$$

Proof. We may assume $P(X_1 \geq \mu+\delta) > 0$. For otherwise, (4.10) is trivially true, since $I(\delta) = -\infty$ and both sides of (4.10) are zero (Exercise 9). To obtain the formula (4.10), first note the simple inequality

$$\mathbb{E}e^{hNA_N} \geq \mathbb{E}\{e^{hNA_N}\mathbf{1}[A_N \geq \mu+\delta]\} \geq e^{hN(\mu+\delta)}P(A_N \geq \mu+\delta)$$

for all $h \geq 0$. Since by independence, the moment generating function of $NA_N \equiv X_1 + \cdots + X_N$ may be expressed as $e^{Nc(h)}$, one has for any $h \geq 0$,

$$P(A_N \geq \mu+\delta) \leq e^{-N((\mu+\delta)h - c(h))}.$$

Thus one obtains an (upper) bound for the rate of decay of probability in (4.10) of the form

$$\limsup_{N\to\infty} \frac{\ln P(A_N \geq \mu+\delta)}{N} \leq -c^*(\mu+\delta).$$

It suffices to prove the reverse inequality to establish (4.10). For this it is useful to exponentially size-bias the distribution of X in such a way that the deviant event is the rule, rather than the exception. For the given deviation $\mu+\delta$, suppose that the maximum defining $c^*(\mu+\delta)$ is achieved at $h = h_\delta > 0$ (see Exercise 9), with

$$c^*(\mu+\delta) = (\mu+\delta)h_\delta - c(h_\delta), \qquad \frac{d}{dh}((\mu+\delta)h - c(h))|_{h=h_\delta} = 0.$$

In particular, $\mu+\delta = \frac{d}{dh}c(h)|_{h=h_\delta}$. So define a random variable $\tilde{X}$ to have the size-biased distribution given by

$$P(\tilde{X} \in dy) = Z_\delta^{-1}e^{h_\delta y}P(X \in dy),$$

where $Z_\delta = e^{c(h_\delta)} = m(h_\delta)$ normalizes $e^{h_\delta y}P(X \in dy)$ to a probability distribution. Now observe that

$$\mathbb{E}\tilde{X} = e^{-c(h_\delta)}\int_{\mathbb{R}} ye^{h_\delta y}P(X \in dy) = \frac{d}{dh}c(h)|_{h=h_\delta} = \mu+\delta. \tag{4.11}$$

That is, for the size-biased distribution, the deviation by δ is to be expected for the average behavior. In particular, the law of large numbers yields

$$\lim_{N\to\infty}\tilde{A}_N = \lim_{N\to\infty}\frac{\tilde{X}_1 + \cdots + \tilde{X}_N}{N} = \mathbb{E}\tilde{X} = \mu+\delta.$$

From here one may obtain the reverse inequality by the law of large numbers under size biasing: Namely, let $\varepsilon > 0$, and consider deviations of size $\mu + \delta$ (to within $\pm\varepsilon$) defined by

$$D_N := \{(y_1, \ldots, y_N) \in \mathbb{R}^N : \frac{1}{N}\sum_{j=1}^{N} y_j \in (\mu + \delta - \varepsilon, \mu + \delta + \varepsilon)\}.$$

Note that for $h \geq 0$, $\exp\{-Nh(\mu + \delta + \varepsilon) + h\sum_{j=1}^{N} X_j\} \leq 1$ on the event $[(X_1, \ldots, X_N) \in D_N]$. Thus one has for $h = h_\delta \geq 0$,

$$\begin{aligned}
&P(A_N > \mu + \delta - \varepsilon)\\
&\geq P(A_N \in (\mu + \delta - \varepsilon, \mu + \delta + \varepsilon))\\
&= \mathbb{E}\mathbf{1}[(X_1, \ldots, X_N) \in D_N]\\
&\geq \mathbb{E}\mathbf{1}[(X_1, \ldots, X_N) \in D_N]\exp\{-Nh_\delta(\mu + \delta + \varepsilon) + h_\delta\sum_{j=1}^{N} X_j\}\\
&= \exp\{-Nh_\delta(\mu + \delta + \varepsilon)\}Z_\delta^N \mathbb{E}\{\mathbf{1}[(X_1, \ldots, X_N) \in D_N]\prod_{j=1}^{N} Z_\delta^{-1}e^{h_\delta X_j}\}\\
&= \exp\{-Nh_\delta(\mu + \delta + \varepsilon)\}e^{Nc(h_\delta)}P((\tilde{X}_1, \ldots, \tilde{X}_N) \in D_N)\\
&= \exp\{-(h_\delta(\mu + \delta + \varepsilon) - c(h_\delta))N\}P(\tilde{A}_N \in (\mu + \delta - \varepsilon, \mu + \delta + \varepsilon)). \qquad (4.12)
\end{aligned}$$

Now, the law of large numbers under the size-biased distribution (having mean $\mu + \delta$) makes $\tilde{A}_N \to \mu + \delta$ and hence $P(\tilde{A}_N \in (\mu + \delta - \varepsilon, \mu + \delta + \varepsilon)) \to 1$ as $N \to \infty$. In particular, it follows from (4.12) that for any $\varepsilon > 0$,

$$\liminf_{N\to\infty} \frac{\ln P(A_N > \mu + \delta - \varepsilon)}{N} \geq -c^*(\mu + \delta) - h_\delta\varepsilon.$$

Since $\liminf_{N\to\infty} \frac{\ln P(A_N > \mu + \delta - \varepsilon)}{N}$ is an increasing function of ε, the inequality follows. The case in which the supremum defining $c^*(\mu + \delta)$ is finite but not achieved is left to Exercise 7. ■

The function $I(\delta)$ is referred to as the **large deviation rate** and is computed here in terms of the so-called **Legendre transform** c^* of the cumulant-generating function c of the common distribution Q for the sequence of random variables; see Exercise 10.

EXERCISES

Exercise Set IV

1. (i) Use the definition of product probability measure via the Carathéodory construction to obtain the approximation of $A \in \mathcal{B}^\infty$ by finite-dimensional events $A_n = [(X_1, \ldots, X_n) \in B_n]$, $B_n \in \mathcal{B}^n$, such that $P(A \Delta A_n) \to 0$ as $n \to \infty$. [*Hint*: Given $\varepsilon > 0$, obtain a cover $A \subseteq \cup_{m \geq 1} R_m$, with $R_m \in \sigma(X_1, \ldots, X_m)$, such that $P(\cup_{m\geq 1} R_m \backslash A) < \varepsilon/2$. Use continuity of the probability from above to argue that $P(\cup_{m \geq 1} R_m \backslash \cup_{m=1}^n R_m) < \varepsilon/2$ for n sufficiently large.] (ii) Show that $|P(A) - P(A_n)| \leq P(A \Delta A_n)$. (iii) Show that on a finite measure space $(S, \mathcal{S}, \mu)$, (a) $\mu(B \Delta A) = \mu(A \Delta B) \geq 0$, (b) $\mu(A \Delta A) = 0$, and (c) $\mu(A \Delta B) + \mu(B \Delta C) \geq \mu(A \Delta C)$ hold for all $A, B, C \in \mathcal{S}$. That is, $(A, B) \to \mu(A \Delta B)$ is a *pseudo-metric* on $\mathcal{S}$.

2. Give a simple proof of the strong law of large numbers (SLLN) for i.i.d. random variables $Z_1, Z_2, \ldots$ having finite fourth moments. That is, for $S_n := Z_1 + \cdots + Z_n, n \geq 1$, $\lim_{n \to \infty} S_n/n \to \mathbb{E}Z_1$ a.s. as $n \to \infty$. [*Hint*: Use a fourth moment Chebyshev-type inequality and the Borel–Cantelli lemma I to check for each $\varepsilon = 1/k, k \geq 1, P(|\frac{S_n}{n} - \mathbb{E}Z_1| > \varepsilon \textit{ i.o.}) = 0$.]

3. (i) Write out proofs of Proposition 4.4 and Corollary 4.5. (ii) Suppose $\{X_n : n \geq 1\}$ is a sequence of mean zero uncorrelated random variables such that $\sum_{k=1}^n var(X_k)/n^2 \to 0$. Show that $\frac{1}{n}\sum_{k=1}^n X_k \to 0$ in probability.

4. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of positive random variables such that $\mathbb{E}|\ln X_1| < \infty$. Calculate the a.s. limiting *geometric mean* $\lim_{n\to\infty}(X_1 \cdots X_n)^{\frac{1}{n}}$. Determine the numerical value of this limit in the case of uniformly distributed random variables on $(0, 1)$.

5. (*Hausdorff's Estimate*) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with mean zero and moments of all orders. Let $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Show that given any $\varepsilon > 0$ the event $A := [|S_n| = O(n^{\frac{1}{2}+\varepsilon})$ as $n \to \infty]$ has probability one. [*Hint*: For two sequences of numbers $\{a_n\}_n$ and $\{b_n \neq 0\}_n$ one writes $a_n = O(b_n)$ as $n \to \infty$ if and only if there is a constant $C > 0$ such that $|a_n| \leq C|b_n|$ for all n. Check that $\mathbb{E}|S_n|^{2k} \leq c_k n^k$, $k = 1, 2, 3, \ldots$, and use the Borel–Cantelli lemma I to prove the assertion.]

6. Compute the moment-generating function $m(h)$ for each of the following random variables X: $P(X = n) = Q(\{n\}) = \frac{c}{n^2}, n = \pm 1, \pm 2, \ldots$, where $c^{-1} = 2\sum_{n=1}^\infty \frac{1}{n^2}$. (b) $Q(dx) = \lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}(x)dx$, where $\lambda > 0$. (c) $Q(dx) = \frac{1}{2\pi}e^{-\frac{1}{2}x^2}dx$. (d) Show that $m(h) < \infty$ for all $h \in \mathbb{R}$ if X is a bounded random variable, i.e., $P(|X| \leq B) = 1$ for some $B \geq 0$.

7. (*Interchange of the Order of Differentiation and Integration*) (i) Let $f(x, \theta)$ be a real-valued function on $S \times (c, d)$, where $(S, \mathcal{S}, \mu)$ is a measure space and (a) f is μ-integrable for all θ, (b) $\theta \mapsto f(x, \theta)$ is differentiable at $\theta = \theta_0 \in (c, d)$, and (c) $|f(x, \theta_0 + \varepsilon) - f(x, \theta_0)|/|\varepsilon| \leq g(x)$ for all $x \in S$ and for all ε such that $0 < |\varepsilon| < \varepsilon_0$ (for some $\varepsilon_0 > 0$), and $\int_S g\, d\mu < \infty$. Then show that $\frac{d}{d\theta}\int_S f(x, \theta)\mu(dx)|_{\theta=\theta_0} = \int_S (\frac{d}{d\theta} f(x, \theta))_{\theta=\theta_0}\mu(dx)$. [*Hint*: Apply Lebesgue's dominated convergence theorem to the sequence $g_n(x) := (f(x, \theta_0 + \varepsilon_n) - f(x, \theta_0))/\varepsilon_n$ $(0 \neq \varepsilon_n \to 0)$.] (ii) Verify the term-by-term differentiation of $m(h) = \sum_{k=0}^\infty \frac{h^k}{k!}\mathbb{E}X^k$ at $h = 0$ in the proof of Proposition 4.6.

8. Assume $m(h) < \infty$ for all $h \in \mathbb{R}$. If $P(X = 0) < 1$, show that $c^{(2)}(h) > 0$ for all $h \in \mathbb{R}$. [*Hint*: $c^{(2)}(h) = \mathbb{E}\tilde{X}^2 - (\mathbb{E}\tilde{X})^2$, where $\tilde{X}$ has the distribution $\frac{e^{hx}Q(dx)}{m(h)}$.]

9. For $a > \mu = \mathbb{E}X$, show that $c^*(a) = \sup_{h\in\mathbb{R}}\{ah - c(h)\}$ is attained at $h = h_a > 0$ if $P(X > a) > 0$. [*Hint*: Write $c(h) - ah = \ln m(h) + \ln e^{-ah} = \ln \mathbb{E}e^{h(X-a)} \to \infty$ as $h \to -\infty$, and it goes to ∞ as $h \to \infty$, provided $P(X > a) > 0$. If $P(X > a) = 0$, $\mathbb{E}e^{h(X-a)} \downarrow P(X = a)$ as $h \uparrow \infty$. Note that $c(0) - a \cdot 0 = 0$, and $c'(0) - a = \mathbb{E}X - a < 0$.] (ii) Show that for $a > \mu$, $P(X > a) = 0$, $c^*(a) = -\ln P(X = a)$, and it is not attained by any $h \in \mathbb{R}$. [*Hint*: In this case, $c^*(a) = \infty$ if $P(X_1 = a) = 0$. If $P(X_1 = a) > 0$, $P(A_n \geq a) = (P(X_1 = a))^n$. If $c^*(a)$ is attained by h_a, then by the proof of Theorem 4.8 one must have $c^*(a) = -\ln P(X_1 = a)$ and $c'(h_a) = a$, which implies $\mathbb{E}\tilde{X} = a$, where $\tilde{X}$ has distribution $e^{h_a x}Q(dx)/m(h_a)$. This is impossible, since the maximum value of $\tilde{X}$ is a ($\tilde{Q}$-a.s.).]

10. (*Properties of Legendre Transform*) Let u, v be smooth convex functions on $\mathbb{R}$ with Legendre transforms u^*, v^*.
    (i) (*Convexity*) Show that u^* is convex.
    (ii) (*Involution*) Show that $u^{**} = u$
    (iii) (*Young's Inequality*) Show that if $u = v^*$ then $xy \leq u(x) + v(y)$.

11. (i) Suppose that X_1 has a Gaussian distribution with mean μ and variance σ^2. Compute the large deviation rate $I(\delta)$. (ii) If X_1 is Poisson with mean μ, compute $I(\delta)$. (iii) If X_1 is Bernoulli, $P(X_1 = 0) = 1 - p$, $P(X_1 = 1) = p$ $(0 < p < 1)$, compute $I(\delta)$.

12. (i) Prove that Theorem 4.8 holds if $m(a)$ is finite in a neighborhood of zero. [*Hint*: Let (a, b) be the interval on which $m(h)$ is finite, $-\infty \leq a < 0 < b < \infty$. On $(a, b)^c$, $m(h) = \infty$, and therefore, one may restrict attention to $h \in (a, b)$ in the definition of $c^*(x)$.] (ii) Calculate $I(\delta)$ for a gamma-distributed random variable with density $f(x) = \frac{\alpha^\beta}{\Gamma(\beta)}x^{\beta-1}e^{-\alpha x}\mathbf{1}_{(0,\infty)}(x)$.

CHAPTER V

Weak Convergence of Probability Measures

Let (S, ρ) be a metric space and let $\mathcal{P}(S)$ be the set of all probability measures on $(S, \mathcal{B}(S))$. Recall the total variation metric (distance) for $\mathcal{P}(S)$ that emerged in the context of Scheffe's theorem. In this section another (weaker) form of convergence is introduced that has widespread applications pertaining to limit distributions that arise in a variety of other contexts. To fix ideas one may regard a sequence of probabilities $P_n \in \mathcal{P}(S)$, $n \geq 1$, as respective distributions of random maps X_n, $n \geq 1$, defined on some probability space and taking values in the metric space S.

A topology may be defined on $\mathcal{P}(S)$ by the following neighborhood system: For $P_0 \in \mathcal{P}(S)$, $\delta > 0$, and f_i $(1 \leq i \leq m)$ real-valued bounded continuous functions on S, define an open neighborhood of $P_0 \in \mathcal{P}(S)$ as

$$N(P_0 : f_1, f_2, \dots, f_m;\ \delta) := \{P \in \mathcal{P}(S) : \left| \int_S f_i \, dP - \int_S f_i \, dP_0 \right| < \delta \ \forall \ i = 1, \dots, m\}. \tag{5.1}$$

Here all $\delta > 0$, $m \geq 1$, and $f_i \in C_b(S)$ (the *set of all real-valued bounded continuous functions on* S), $1 \leq i \leq m$, are allowed. An **open set** of $\mathcal{P}(S)$ is defined to be a set U such that every P_0 in U has an open neighborhood of the form (5.1) contained in U. Since the neighborhoods (5.1) are taken to be open, the topology is the collection of all unions of such sets. The topology (i.e., the collection of open sets) so defined is called the **weak topology**[1] of $\mathcal{P}(S)$. We restrict the presentation to only those results that will be used in this text.

[1] Billingsley (1999) provides a detailed exposition and comprehensive account of the weak convergence theory.

Definition 5.1. A sequence of probabilities $\{P_n : n \geq 1\}$ is said to **converge weakly** to a probability P if $\int_S f\,dP_n \to \int_S f\,dP\ \forall\ f \in C_b(S)$. Denote this by $P_n \Rightarrow P$.

Recall that the collection $C_b(S)$ is a measure-determining class of functions. Thus the limit P of $\{P_n\}_{n=1}^{\infty}$ is uniquely defined by weak convergence (also see Remark 5.1 below).

Note that if P_n, P are viewed as distributions of random maps X_n, X, respectively, defined on some probability space, then the definition of weak convergence, equivalently **convergence in distribution**, takes the form

$$\lim_n \mathbb{E}f(X_n) = \mathbb{E}f(X) \qquad \forall f \in C_b. \tag{5.2}$$

There are a number of equivalent formulations of weak convergence that are useful in various contexts. We will need the following topological notions. Recall that a point belongs to the **closure** of a set A if it belongs to A or if every neighborhood of the point intersects both A and A^c. On the other hand, a point belongs to the **interior** of A if there is an open set contained in A that includes the point. Denoting the closure of a set A by A^- and the interior by A°, one defines the **boundary** by $\partial A = A^- \backslash A^\circ$. A set A in $\mathcal{B}$ whose boundary ∂A satisfies $P(\partial A) = 0$ is called a P**-continuity set**. Since ∂A is closed, it clearly belongs to the σ-field $\mathcal{S} = \mathcal{B}(S)$.

Theorem 5.1 *(Alexandrov Theorem).* Let $P_n, n \geq 1$, P be probability measures on $(S, \mathcal{B}(S))$. The following are equivalent:

(i) $P_n \Rightarrow P$.

(ii) $\lim_n \int_S f\,dP_n = \int_S f\,dP$ for all bounded, uniformly continuous real f.

(iii) $\limsup_n P_n(F) \leq P(F)$ for all closed F.

(iv) $\liminf_n P_n(G) \geq P(G)$ for all open G.

(v) $\lim_n P_n(A) = P(A)$ for all P-continuity sets A.

Proof. The plan is to first prove (i) implies (ii) implies (iii) implies (i), and hence that (i), (ii), and (iii) are equivalent. We then directly prove that (iii) and (iv) are equivalent and that (iii) and (v) are equivalent.

(i) implies (ii): This follows directly from the definition.

(ii) implies (iii): Let F be a closed set and $\delta > 0$. For a sufficiently small but fixed value of ε, $G_\varepsilon = \{x : \rho(x, F) < \varepsilon\}$ satisfies $P(G_\varepsilon) < P(F) + \delta$, by continuity of the probability measure P from above, since the sets G_ε decrease to $F = \cap_{\varepsilon\downarrow 0} G_\varepsilon$. Adopt the construction from the proof of Proposition 1.4 that $C_b(S)$ is measure-determining to produce a uniformly continuous function h on S such that $h(x) = 1$ on F, $h(x) = 0$ on the complement G_ε^c of G_ε, and $0 \leq h(x) \leq 1$ for all x. In view of (ii) one has

$\lim_n \int_S h\,dP_n = \int_S h\,dP$. In addition,

$$P_n(F) = \int_F h\,dP_n \le \int_S h\,dP_n$$

and

$$\int_S h\,dP = \int_{G_\varepsilon} h\,dP \le P(G_\varepsilon) < P(F) + \delta.$$

Thus

$$\limsup_n P_n(F) \le \lim_n \int_S h\,dP_n = \int_S h\,dP < P(F) + \delta.$$

Since δ was arbitrary this proves (iii).
(iii) implies (i): Let $f \in C_b(S)$. It suffices to prove

$$\limsup_n \int_S f\,dP_n \le \int_S f\,dP. \tag{5.3}$$

For then one also gets $\liminf_n \int_S f\,dP_n \ge \int_S f\,dP$, and hence (i), by replacing f by $-f$. But in fact, for (5.3) it suffices to consider $f \in C_b(S)$ such that $0 < f(x) < 1, x \in S$, since the more general $f \in C_b(S)$ can be reduced to this by translating and rescaling f. Fix an integer k and let F_i be the closed set $F_i = \{x : f(x) \ge i/k\},\ i = 0, 1, \dots, k$. Then taking advantage of $0 < f < 1$, one has

$$\sum_{i=1}^{k} \frac{i-1}{k} P\Big(\{x : \frac{i-1}{k} \le f(x) < \frac{i}{k}\}\Big) \le \int_S f\,dP \le \sum_{i=1}^{k} \frac{i}{k} P\left(\{x : \frac{i-1}{k} \le f(x) < \frac{i}{k}\}\right).$$

Noting that $F_0 = S$, $F_k = \emptyset$, the sum on the right telescopes as

$$\sum_{i=1}^{k} \frac{i}{k}[P(F_{i-1}) - P(F_i)] = \frac{1}{k} + \frac{1}{k}\sum_{i=1}^{k} P(F_i),$$

while the sum on the left is smaller than this by $1/k$. Hence

$$\frac{1}{k}\sum_{i=1}^{k} P(F_i) \le \int_S f\,dP < \frac{1}{k} + \frac{1}{k}\sum_{i=1}^{k} P(F_i). \tag{5.4}$$

In view of (iii), $\limsup_n P_n(F_i) \le P(F_i)$ for each i. So, using the upper bound in (5.4) with P_n in place of P and the lower bound with P, it follows that

$$\limsup_n \int_S f\,dP_n \le \frac{1}{k} + \frac{1}{k}\sum_{i=1}^{k} P(F_i) \le \frac{1}{k} + \int_S f\,dP.$$

Now let $k \to \infty$ to obtain the asserted inequality (5.3) to complete the proof of (i) from (iii).
(iii) iff (iv): This is simply due to the fact that open and closed sets are complementary.
(iii) implies (v): Let A be a P-continuity set. Since (iii) implies (iv) one has

$$\begin{aligned} P(A^-) &\geq \limsup_n P_n(A^-) \geq \limsup_n P_n(A) \\ &\geq \liminf_n P_n(A) \geq \liminf_n P_n(A^\circ) \geq P(A^\circ). \end{aligned} \tag{5.5}$$

Since $P(\partial A) = 0$, $P(A^-) = P(A^\circ)$, so that the inequalities squeeze down to $P(A)$ and $\lim_n P_n(A) = P(A)$ follows.
(v) implies (iii): Let F be a closed set. The idea is to observe that F may be expressed as the limit of a decreasing sequence of P-continuity sets as follows. Since $\partial\{x : \rho(x, F) \leq \delta\} \subseteq \{x : \rho(x, F) = \delta\}$, these boundaries are disjoint for distinct δ, (Exercise 5). Thus at most countably many of them can have positive P-measure (Exercise 5), all other, therefore, being P-continuity sets. In particular, there is a sequence of positive numbers $\delta_k \downarrow 0$ such that the sets $F_k = \{x : \rho(x, F) \leq \delta_k\}$ are P-continuity sets. From (v) one has $\limsup_n P_n(F) \leq \lim_n P_n(F_k) = P(F_k)$ for each k. Since F is closed one also has $F_k \downarrow F$, so that (iii) follows from continuity of the probability P from above. This completes the proof of the theorem. ■

The following result provides a useful tool for tracking weak convergence in a variety of settings. Note that in the case that h is continuous it follows immediately from the definition of weak convergence since compositions of bounded continuous functions with h are bounded and continuous.

Theorem 5.2. Let S_1, S_2 be a pair of metric spaces and $h : S_1 \to S_2$ a Borel-measurable map. Suppose that $\{P_n\}_{n=1}^\infty$, P are probabilities on the Borel σ-field of S_1 such that $P_n \Rightarrow P$. If h is P-a.s. continuous, then $P_n \circ h^{-1} \Rightarrow P \circ h^{-1}$.

Proof. Let F be a closed subset of S_2. Then, letting $F_h = h^{-1}(F)$, it follows from Alexandrov conditions that $\limsup_n P_n(F_h) \leq \limsup_n P_n(F_h^-) \leq P(F_h^-)$. But $P(F_h^-) = P(F_h)$ since $F_h^- \subseteq D_h \cup F_h$, where D_h denotes the set of discontinuities of h (Exercise 5) and, by hypothesis, $P(D_h) = 0$. ■

In the special finite-dimensional case $S = \mathbb{R}^k$, the following theorem provides some alternative useful conditions for weak convergence. Additional useful methods are developed in Exercises.

Theorem 5.3 *(Finite-Dimensional Weak Convergence).* Let $\{P_n\}_{n=1}^\infty, P$ be probabilities on the Borel σ-field of $\mathbb{R}^k$. The following are equivalent statements:

(i) $P_n \Rightarrow P$.
(ii) $\int_{\mathbb{R}^k} f dP_n \to \int_{\mathbb{R}^k} f dP$ for all (bounded) continuous f vanishing outside a compact set.

(iii) $\int_{\mathbb{R}^k} f dP_n \to \int_{\mathbb{R}^k} f dP$ for all infinitely differentiable functions f vanishing outside a compact set.

(iv) Let $F_n(x) := P_n((-\infty, x_1] \times \cdots \times (-\infty, x_k])$, and $F(x) := P((-\infty, x_1] \times \cdots \times (-\infty, x_k]), x \in \mathbb{R}^k, n = 1, 2, \ldots$. Then $F_n(x) \to F(x)$ as $n \to \infty$, for every point of continuity x of F.

Proof. We give the proof for the one-dimensional case $k = 1$. The case $k \geq 2$ requires no difference in proof for (i)–(iii) and is left as Exercise 1 for these parts. The equivalence of (i) and (iv) for the case $k \geq 2$ is outlined in detail in Exercise 2. First let us check that (ii) is sufficient. It is obviously necessary by definition of weak convergence. Assume (ii) and let f be an arbitrary bounded continuous function, $|f(x)| \leq c$ for all x. The idea is to construct a continuous approximation to f having compact support. For notational convenience write $\{x \in \mathbb{R}^1 : |x| \geq N\} = \{|x| \geq N\}$, etc. Given $\varepsilon > 0$ there exists N such that $P(\{|x| \geq N\}) < \varepsilon/4c$. Define θ_N by $\theta_N(x) = 1, |x| \leq N,\ \theta_N(x) = 0, |x| \geq N+1$, and linearly interpolate for $N \leq |x| \leq N+1$. Then,

$$\begin{aligned}\underline{\lim}_{n\to\infty} P_n(\{|x| \leq N+1\}) &\geq \underline{\lim}_{n\to\infty} \int \theta_N(x) dP_n(x) = \int \theta_N(x) dP(x)\\ &\geq P(\{|x| \leq N\}) > 1 - \frac{\varepsilon}{4c},\end{aligned}$$

so that

$$\overline{\lim}_{n\to\infty} P_n(\{|x| > N+1\}) \equiv 1 - \underline{\lim}_{n\to\infty} P_n(\{|x| \leq N+1\}) < \frac{\varepsilon}{4c}. \tag{5.6}$$

Now define $f_N := f\theta_{N+1}$. Noting that $f = f_N$ on $\{|x| \leq N+1\}$ and that on $\{|x| > N+1\}$ one has $|f(x)| \leq c$, upon first writing $f = f\mathbf{1}_{\{|x|\leq N+1\}} + f\mathbf{1}_{\{|x|>N+1\}}$, and then further writing $\int_{\mathbb{R}^1} f_N \mathbf{1}_{\{|x|\leq N+1\}} dP_n = \int_{\mathbb{R}^1} f_N dP_n - \int_{\{N+1<|x|\leq N+2\}} f_N dP_n$ (and similarly for the integral with respect to P), one has from the triangle inequality and the bound on f and f_N that

$$\begin{aligned}\overline{\lim}_{n\to\infty} \left| \int_{\mathbb{R}^1} f\, dP_n - \int_{\mathbb{R}^1} f\, dP \right| &\leq \overline{\lim}_{n\to\infty} \left| \int_{\mathbb{R}^1} f_N\, dP_n - \int_{\mathbb{R}^1} f_N\, dP \right|\\ &\quad + \overline{\lim}_{n\to\infty}(2cP_n(\{|x| > N+1\})\\ &\quad + 2cP(\{|x| > N+1\}))\\ &< 2c\frac{\varepsilon}{4c} + 2c\frac{\varepsilon}{4c} = \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $\int_{\mathbb{R}^1} f\, dP_n \to \int_{\mathbb{R}^1} f\, dP$. So (i) and (ii) are equivalent. Let us now show that (ii) and (iii) are equivalent. It is enough to prove (iii) is sufficient for (ii). For this we construct an approximation to f that is C^∞ and has compact support.

For each $\varepsilon > 0$ define the function

$$\rho_\varepsilon(x) = d(\varepsilon)\exp\left\{-\frac{1}{1-x^2/\varepsilon^2}\right\}\mathbf{1}_{[-\varepsilon,\varepsilon]}(x), \tag{5.7}$$

where $d(\varepsilon)$ is so chosen as to make $\int \rho_\varepsilon(x)dx = 1$. One may check that $\rho_\varepsilon(x)$ is infinitely differentiable in x. Now let f be a continuous function that vanishes outside a finite interval. Then f is uniformly continuous, and therefore, $\delta(\varepsilon) = \sup\{|f(x) - f(y)| : |x-y| \leq \varepsilon\} \to 0$ as $\varepsilon \downarrow 0$. Define

$$f^\varepsilon(x) = f * \rho_\varepsilon(x) := \int_{-\varepsilon}^{\varepsilon} f(x-y)\rho_\varepsilon(y)dy, \tag{5.8}$$

and note that $f^\varepsilon(x)$ is infinitely differentiable, vanishes outside a compact set, and is an average over values of f within the interval $(x-\varepsilon, x+\varepsilon)$, $|f^\varepsilon(x) - f(x)| \leq \delta(\varepsilon)$ for all ε. Hence,

$$\left|\int_{\mathbb{R}^1} f\,dP_n - \int_{\mathbb{R}^1} f^\varepsilon\,dP_n\right| \leq \delta(\varepsilon) \quad \text{for all } n, \qquad \left|\int_{\mathbb{R}^1} f\,dP - \int_{\mathbb{R}^1} f^\varepsilon\,dP\right| \leq \delta(\varepsilon),$$

$$\begin{aligned}\left|\int_{\mathbb{R}^1} f\,dP_n - \int_{\mathbb{R}^1} f\,dP\right| &\leq \left|\int_{\mathbb{R}^1} f\,dP_n - \int_{\mathbb{R}^1} f^\varepsilon\,dP_n\right| + \left|\int_{\mathbb{R}^1} f^\varepsilon\,dP_n - \int_{\mathbb{R}^1} f^\varepsilon\,dP\right| \\ &\quad + \left|\int_{\mathbb{R}^1} f^\varepsilon\,dP - \int_{\mathbb{R}^1} f\,dP\right| \\ &\leq 2\delta(\varepsilon) + \left|\int_{\mathbb{R}^1} f^\varepsilon\,dP_n - \int_{\mathbb{R}^1} f^\varepsilon\,dP\right| \to 2\delta(\varepsilon) \qquad \text{as } n \to \infty.\end{aligned}$$

Since $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$ it follows that $\int_{\mathbb{R}^1} f\,dP_n \to \int_{\mathbb{R}^1} f\,dP$, as claimed. Next let F_n, F be the distribution functions of P_n, P, respectively $(n = 1, 2, \ldots)$. Suppose (i) holds and observe that $(-\infty, x]$ is a P-continuity set if and only if $0 = P(\partial(-\infty, x]) = P(\{x\})$. That is, x must be a continuity point of F, so that (i) implies (iii) follows from Alexandrov's theorem. To show that the converse is also true, suppose $F_n(x) \to F(x)$ at all points of continuity of a distribution function (d.f.) F. Note that since F is nondecreasing and bounded between 0 and 1, it can have at most countably many discontinuities, i.e., only finitely many jumps of size larger than $1/n$ for any $n = 1, 2, \ldots$. Consider a continuous function f that vanishes outside the compact set K contained in an interval $[a, b]$ where a, b are selected as points of continuity of F. The idea is to construct an approximation to f by a step function with jumps at points of continuity of F. Given any $\varepsilon > 0$ one may partition $[a, b]$ into a finite number of subintervals whose endpoints are all points of continuity of F, and obtain a uniform approximation of f to within $\varepsilon > 0$ by a step function f_ε having constant values of f at the endpoint over each respective subinterval. Then, $\int_S f_\varepsilon dP_n \to \int_S f_\varepsilon dP$ as $n \to \infty$. Thus $|\int_S f dP_n - \int_S f dP| \leq \int_S |f - f_\varepsilon| dP_n + |\int_S f_\varepsilon dP_n - \int_S f_\varepsilon dP| + \int_S |f_\varepsilon - f| dP \leq 2\varepsilon + |\int_S f_\varepsilon dP_n - \int_S f_\varepsilon dP|$. Since $\varepsilon > 0$ is arbitrary, one readily obtains (i). ■

Remark 5.1. It follows from Theorem 5.1, in particular, that if (S,ρ) is a metric space then $\int_S f\,dP = \int_S f\,dQ\ \forall\ f \in C_b(S)$ implies $P = Q$. Note that by a simple rescaling this makes $\{f \in C_b(S) : \|f\|_\infty \le 1\}$ measure-determining as well. The same is true for the set $UC_b(S)$ of all bounded uniformly continuous real-valued functions on S in place of $C_b(S)$.

Using a technique from the proof of Theorem 5.1 one may also obtain the following equivalent specification of the weak topology.

Proposition 5.4. The weak topology is defined by the system of open neighborhoods of the form (5.1) with $f_1, f_2, \ldots, f_m$ bounded and uniformly continuous.

Proof. Fix $P_0 \in \mathcal{P}(S)$, $f \in C_b(S), \varepsilon > 0$. We need to show that the set $\{P \in \mathcal{P}(S) : |\int_S f\,dP - \int_S f\,dP_0| < \varepsilon\}$ contains a set of the form (5.1), but with f_i's that are uniformly continuous and bounded. Without essential loss of generality, assume $0 < f < 1$. As in the proof of Theorem 5.1, (iii) implies (i), see the relations (5.4), one may choose and fix a large integer k such that $1/k < \varepsilon/4$ and consider the sets F_i in that proof. Next, as in the proof of (ii) implies (iii) of Theorem 5.1, there exist uniformly continuous functions g_i, $0 \le g_i \le 1$, such that $g_i = 1$ on F_i and $|\int_S g_i\,dP_0 - P_0(F_i)| < \varepsilon/4, 1 \le i \le k$. Then on the set $\{P : |\int_S g_i\,dP - \int_S g_i\,dP_0| < \varepsilon/4, 1 \le i \le k\}$, one has (see (5.1))

$$\begin{aligned}\int_S f dP &\le \frac{1}{k}\sum_{i=1}^k P(F_i) + \frac{1}{k} \le \frac{1}{k}\sum_{i=1}^k \int_S g_i dP + \frac{1}{k}\\ &< \frac{1}{k}\sum_{i=1}^k \int_S g_i dP_0 + \frac{\varepsilon}{4} + \frac{1}{k} \le \frac{1}{k}\sum_{i=1}^k P_0(F_i) + \frac{2\varepsilon}{4} + \frac{1}{k}\\ &< \int_S f dP_0 + \frac{3\varepsilon}{4} < \int_S f dP_0 + \varepsilon. \end{aligned} \tag{5.9}$$

Similarly, replacing f by $1-f$ in the above argument, one may find uniformly continuous h_i, $0 \le h_i \le 1$, such that on the set $\{P : |\int_S h_i\,dP - \int_S h_i\,dP_0| < \varepsilon/4, 1 \le i \le k\}$, one has $\int_S (1-f)dP < \int_S (1-f)dP_0 + \varepsilon$. Therefore

$$\begin{aligned}&\{P \in \mathcal{P}(S) : |\int_S f\,dP - \int_S f\,dP_0| < \varepsilon\}\\ &\supseteq \left\{P : \left|\int_S g_i\,dP - \int_S g_i\,dP_0\right| < \varepsilon/4, \left|\int_S h_i\,dP - \int_S h_i\,dP_0\right| < \varepsilon/4, 1 \le i \le k\right\}.\end{aligned}$$

By taking intersections over m such sets, it follows that a neighborhood $N(P_0)$, say, of P_0 of the form (5.1) (with $f_i \in C_b(S), 1 \le i \le m$) contains a neighborhood of P_0 defined with respect to bounded uniformly continuous functions. In particular,

$N(P_0)$ is an open set defined by the latter neighborhood system. Since the latter neighborhood system is a subset of the system (5.1), the proof is complete. ■

Two points of focus for the remainder of this section are metrizability and (relative) compactness in the weak topology. Compactness in a metric space may be equivalently viewed as the existence of a limit point for any sequence in the space.

In the case that (S, ρ) is a compact metric space, $C(S) \equiv C_b(S)$ is a *complete separable metric space* under the **"sup" norm** $\|f\|_\infty := \max\{|f(x)| : x \in S\}$, i.e., under the distance $d_\infty(f, g) := \|f - g\|_\infty \equiv \max\{|f(x) - g(x)| : x \in S\}$ (see Appendix B). In this case the weak topology is metrizable, i.e., $\mathcal{P}(S)$ is a metric space with the metric

$$d_W(P, Q) := \sum_{n=1}^{\infty} 2^{-n} \left| \int_S f_n \, dP - \int_S f_n dQ \right|, \tag{5.10}$$

where $\{f_n : n \geq 1\}$ is a dense subset of $\{f \in C(S) : \|f\|_\infty \leq 1\}$. Using Cantor's diagonal procedure and the Riesz representation theorem (for bounded linear functionals on $C(S)$ in Appendix A), one may check that every sequence $\{P_n : n \geq 1\}$ has a convergent subsequence. In other words, one has the following result (Exercise 8).

Proposition 5.5. If (S, ρ) is compact, then $\mathcal{P}(S)$ is a compact metric space under the weak topology, with a metric given by (5.10).

A slightly weaker form of convergence is sometimes useful to consider within the general theme of this section, for example in analyzing the nature of certain failures of weak convergence (see Exercise 7). A function $f \in C_b(S)$ is said to **vanish at infinity** if for each $\varepsilon > 0$ there is a compact subset K_ε such that $|f(x)| < \varepsilon$ for all $x \in K_\varepsilon^c$. Let $C_b^0(S)$ denote the collection of all such functions on S.

Definition 5.2. A sequence of probability measures $\{P_n\}_{n=1}^\infty$ on $(S, \mathcal{B}(S))$ is said to **converge vaguely** to a finite measure P, not necessarily a probability, if $\lim_n \int_S f dP_n = \int_S f dP$ for all $f \in C_b^0(S)$.

Corollary 5.6 ***(Helly Selection Principle).*** Every sequence of probabilities $\mu_n, n \geq 1$, on $(\mathbb{R}, \mathcal{B})$ has a vaguely convergent subsequence.

Proof. Let $\varphi : \mathbb{R} \to (-1, 1)$ by $\varphi(x) = \frac{2}{\pi} \tan^{-1}(x), x \in \mathbb{R}$, and define a probability ν_n supported on $(-1, 1)$ by $\nu_n(A) = \mu_n(\{x : \varphi(x) \in A\})$ for Borel subsets A of $(-1, 1)$. One may regard ν_n as a probability on the compact interval $[-1, 1]$ (supported on the open interval). Thus, by Theorem 5.10, there is a probability ν on $[-1, 1]$ and a subsequence $\{\nu_{n_m} : m \geq 1\}$ such that $\nu_{n_m} \Rightarrow \nu$ as $m \to \infty$. Define $\tilde{\nu}(A) = \nu(A)$ for Borel subsets A of $(-1, 1)$. Then $\tilde{\nu}$ is a measure on $(-1, 1)$ with $\tilde{\nu}(-1, 1) \leq 1$. Let $\mu(B) = \tilde{\nu}(\{y \in (-1, 1) : \varphi^{-1}(y) \in B\})$ for Borel subsets B of $\mathbb{R}$. Since for $f \in C_b^0(\mathbb{R})$, the map $g := f \circ \varphi^{-1}$ is in $C_b([-1, 1])$, where $g(1) = g(-1) = f(\varphi^{-1}(\pm 1)) :=$

$\lim_{x\to\pm 1} f(\varphi^{-1}(x)) = 0$, one has, using the change of variable formula,

$$\begin{aligned}\int_{\mathbb{R}} f(x)\mu(dx) &= \int_{(-1,1)} f(\varphi^{-1}(y))\mu\circ\varphi^{-1}(dy) = \int_{(-1,1)} g(y)\tilde{\nu}(dy)\\ &= \int_{[-1,1]} g(y)\nu(dy) = \lim_{m\to\infty}\int_{[-1,1]} g(y)\nu_{n_m}(dy)\\ &= \lim_{m\to\infty}\int_{\mathbb{R}} f(x)\mu_{n_m}(dx),\end{aligned}$$

where the change of variable formula is again used to write the last equality. ■

For our next result we need the following lemma. Let $H = [0,1]^{\mathbb{N}}$ be the space of all sequences in $[0,1]$ with the product topology, referred to as the **Hilbert cube**

Lemma 1 *(Hilbert Cube Embedding).* Let (S,ρ) be a separable metric space. There exists a map h on S into the Hilbert cube $H \equiv [0,1]^{\mathbb{N}}$ with the product topology, such that h is a homeomorphism of S onto $h(S)$, in the relative topology of $h(S)$.

Proof. Without loss of generality, assume $\rho(x,y) \leq 1\ \forall\ x,y \in S$. Let $\{z_k : k = 1,2,\ldots\}$ be a dense subset of S. Define the map

$$h(x) = (\rho(x,z_1),\ \rho(x,z_2),\ldots,\rho(x,z_k),\ldots) \quad (x \in S). \tag{5.11}$$

If $x_n \to x$ in S, then $\rho(x_n,z_k) \to \rho(x,z_k)\ \forall\ k$, so that $h(x_n) \to h(x)$ in the (metrizable) product topology (of pointwise convergence) on $h(S)$. Also, h is one-to-one. For if $x \neq y$, one may find z_k such that $\rho(x,z_k) < \frac{1}{3}\rho(x,y)$, and hence $\rho(y,z_k) \geq \rho(y,x) - \rho(z_k,x) > \frac{2}{3}\rho(x,y)$, so that $\rho(x,z_k) \neq \rho(y,z_k)$. Finally, let $\tilde{a}_n \equiv (a_{n1},a_{n2},\ldots) \to \tilde{a} = (a_1,a_2,\ldots)$ in $h(S)$, and let $x_n = h^{-1}(\tilde{a}_n),\ x = h^{-1}(\tilde{a})$. One then has $(\rho(x_n,z_1),\ \rho(x_n,z_2),\ldots) \to (\rho(x,z_1),\ \rho(x,z_2),\ldots)$. Hence $\rho(x_n,z_k) \to \rho(x,z_k)\ \forall\ k$, implying $x_n \to x$, since $\{z_k : k \geq 1\}$ is dense in S. ■

Theorem 5.7. Let (S,ρ) be a separable metric space. Then $\mathcal{P}(S)$ is a separable metric (i.e., metrizable) space under the weak topology.

Proof. By Lemma 1, S may be replaced by its homeomorphic image $S_h \equiv h(S)$ in $[0,1]^{\mathbb{N}}$ which is compact under the product topology by Tychonov's theorem (Appendix B), and is metrizable with the metric $d(\tilde{a},\tilde{b}) := \sum_{n=1}^{\infty} 2^{-n}|a_n - b_n|(\tilde{a} = (a_1,a_2,\ldots),\ \tilde{b} = (b_1,b_2,\ldots))$. We shall consider uniform continuity of functions on S_h with respect to this metric d. Every uniformly continuous bounded f on S_h has a unique extension $\bar{f}$ to $\bar{S}_h$ ($\equiv$ closure of S_h in $[0,1]^{\mathbb{N}}$) : $\bar{f}(\tilde{a}) := \lim_{k\to\infty} f(\tilde{a}^k)$, where $\tilde{a}^k \in S_h,\ \tilde{a}^k \to \tilde{a}$. Conversely, the restriction of every $g \in C(\bar{S}_h)$ is a uniformly continuous bounded function on S_h. In other words, $UC_b(S_h)$ may be identified with $C(\bar{S}_h)$ as sets and as metric spaces under the supremum distance d_∞ between functions. Since $\bar{S}_h$ is compact, $C_b(\bar{S}_h) \equiv C(\bar{S}_h)$ is a separable metric space under the supremum distance d_∞, and therefore, so is $UC_b(S_h)$. Letting $\{f_n : n = 1, 2, \ldots\}$ be a dense subset of $UC_b(S_h)$, one now defines a metric d_W on $\mathcal{P}(S_h)$ as in (5.10). This proves metrizability of $\mathcal{P}(S_h)$.

To prove separability of $\mathcal{P}(S_h)$, for each $k = 1, 2, \ldots$, let $D_k := \{x_{ki} : i = 1, 2, \ldots, n_k\}$ be a finite $(1/k)$-net of S_h (i.e., every point of S_h is within a distance $1/k$ from some point in this net). Let $D = \{x_{ki} : i = 1, \ldots, n_k,\ k \geq 1\} = \cup_{k=1}^{\infty} D_k$. Consider the set $\mathcal{E}$ of all probabilities with finite support contained in D and having rational mass at each point of support. Then $\mathcal{E}$ is countable and is dense in $\mathcal{P}(S_h)$. To prove this last assertion, fix $P_0 \in \mathcal{P}(S_h)$. Consider the partition generated by the set of open balls $\{x \in S_h : d(x, x_{ki}) < \frac{1}{k}\}$, $1 \leq i \leq n_k$. Let P_k be the probability measure defined by letting the mass of P_0 on each nonempty set of the partition be assigned to a singleton $\{x_{ki}\}$ in D_k that is at a distance of at most $1/k$ from the set. Now construct $Q_k \in \mathcal{E}$, where Q_k has the same support as P_k but the point masses of Q_k are rational and are such that the sum of the absolute differences between these masses of P_k and the corresponding ones of Q_k is less than $1/k$. Then it is simple to check that $d_W(P_0, Q_k) \to 0$ as $k \to \infty$, that is, $\int_{S_h} g\, dQ_k \to \int_{S_h} g\, dP_0$ for every uniformly continuous and bounded g on S_h. ■

The next result is of considerable importance in probability. To state it we need a notion called "tightness."

Definition 5.3. A subset Λ of $\mathcal{P}(S)$ is said to be **tight** if for every $\varepsilon > 0$, there exists a compact subset K_ε of S such that

$$P(K_\varepsilon) \geq 1 - \varepsilon \quad \forall\ P \in \Lambda. \tag{5.12}$$

Theorem 5.8 ***(Prohorov's Theorem).*** (a) Let (S, ρ) be a separable metric space. If $\Lambda \subseteq \mathcal{P}(S)$ is tight then its weak closure $\bar{\Lambda}$ is compact (metric) in the weak topology. (b) If (S, ρ) is Polish, then the converse is true: For a set Λ to be conditionally compact (i.e., $\bar{\Lambda}$ compact) in the weak topology, it is necessary that Λ be tight.

Proof. We begin with a proof of part (a). Suppose $\Lambda \subseteq \mathcal{P}(S)$ is tight. Let $\tilde{S} = \cup_{j=1}^{\infty} K_{1/j}$, where $K_{1/j}$ is a compact set determined from (5.12) with $\varepsilon = 1/j$. Then $P(\tilde{S}) = 1\ \forall\ P \in \Lambda$. Also, $\tilde{S}$ is σ-compact, and so is its image $\tilde{S}_h$ equal to $\cup_{j=1}^{\infty} h(K_{1/j})$ under the map h (appearing in the proofs of Lemma 1 and Theorem 5.7) since the image of a compact set under a continuous map is compact. In particular, $\tilde{S}_h$ is a Borel subset of $[0, 1]^{\mathbb{N}}$ and therefore of $\bar{\tilde{S}}_h$. Let Λ_h be the image of Λ in $\tilde{S}_h$ under h, i.e., $\Lambda_h = \{P \circ h^{-1} : P \in \Lambda\} \subseteq \mathcal{P}(\tilde{S}_h)$. In view of the homeomorphism $h : \tilde{S} \to \tilde{S}_h$, it is enough to prove that Λ_h is conditionally compact as a subset of $\mathcal{P}(\tilde{S}_h)$.

Since $\tilde{S}_h$ is a Borel subset of $\bar{\tilde{S}}_h$, one may take $\mathcal{P}(\tilde{S}_h)$ as a subset of $\mathcal{P}(\bar{\tilde{S}}_h)$, extending P in $\mathcal{P}(\tilde{S}_h)$ by setting $P(\bar{\tilde{S}}_h \backslash \tilde{S}_h) = 0$. Thus $\Lambda_h \subseteq \mathcal{P}(\tilde{S}_h) \subseteq \mathcal{P}(\bar{\tilde{S}}_h)$. By Proposition 5.5, $\mathcal{P}(\bar{\tilde{S}}_h)$ is compact metric (in the weak topology). Hence every sequence $\{P_n : n = 1, 2, \ldots\}$ in Λ_h has a subsequence $\{P_{n_k} : k = 1, 2, \ldots\}$ converging weakly to some

$Q \in \mathcal{P}(\bar{\tilde{S}}_h)$. We need to show that $Q \in \mathcal{P}(\tilde{S}_h)$, that is, $Q(\tilde{S}_h) = 1$. By Theorem 5.1, $Q(h(K_{1/j})) \geq \limsup_{k\to\infty} P_{n_k}(h(K_{1/j})) \geq 1 - 1/j$. (By hypothesis, $P(h(K_{1/j})) \geq 1 - 1/j \ \forall\ P \in \Lambda_h$). Letting $j \to \infty$, one gets $Q(\tilde{S}_h) = 1$. Finally, note that if Λ is conditionally compact when considered as a subset of $\mathcal{P}(\tilde{S})$, it is also conditionally compact when regarded as a subset of $\mathcal{P}(S)$ (Exercise 12).

For part (b) suppose that (S, ρ) is separable and complete and let Λ be relatively compact in the weak topology. We will first show that given any nondecreasing sequence G_n, $n \geq 1$, of open subsets of S such that $\cup_n G_n = S$ and given any $\varepsilon > 0$, there is an $n = n(\varepsilon)$ such that $P(G_{n(\varepsilon)}) \geq 1 - \varepsilon$ for all $P \in \Lambda$. For suppose this is not true. Then there are an $\varepsilon > 0$ and $P_1, P_2, \ldots$ in Λ such that $P_n(G_n) < 1 - \varepsilon$ for all $n \geq 1$. But by the assumed compactness, there is a subsequence $P_{n(k)}$ that converges weakly to some probability $Q \in \mathcal{P}(S)$. By Alexandrov's theorem this implies, noting $G_n \subseteq G_{n(k)}$ for $n \leq n(k)$, that $Q(G_n) \leq \liminf_{k\to\infty} P_{n(k)}(G_n) \leq \liminf_{k\to\infty} P_{n(k)}(G_{n(k)}) \leq 1 - \varepsilon$, for $n \geq 1$. This leads to the contradiction $1 = Q(S) = \lim_{n\to\infty} Q(G_n) \leq 1 - \varepsilon$. Now to prove that Λ is tight, fix $\varepsilon > 0$. By separability of S for each $k \geq 1$ there is a sequence of open balls $B_{n,k}$, $n \geq 1$, having radii smaller than $1/k$ and such that $\cup_{n\geq 1} B_{n,k} = S$. Let $G_{n,k} := \cup_{m=1}^{n} B_{m,k}$. Using the first part of this proof of (b), it follows that for each k there is an $n = n(k)$ such that $P(G_{n(k),k}) \geq 1 - 2^{-k}\varepsilon$ for all $P \in \Lambda$. Define $G := \cap_{k=1}^{\infty} G_{n(k),k}$. Then its closure $\overline{G}$ is totally bounded, since for each k there is a finite cover of $\overline{G}$ by $n(k)$ closed balls $\overline{B_{n,k}}$ of diameter smaller than $1/k$. Thus completeness of S implies that $\overline{G}$ is compact[2]. But $P(\overline{G}) \geq P(G) \geq 1 - \sum_{k=1}^{\infty} 2^{-k}\varepsilon = 1 - \varepsilon$ for all $P \in \Lambda$. ■

Corollary 5.9. Let (S, ρ) be a Polish space. Then any finite collection Λ of probabilities on $(S, \mathcal{B}(S))$ is tight.

Remark 5.2. The compactness asserted in part (a) of Theorem 5.8 remains valid without the requirement of separability for the metric space (S, ρ). To see this, simply note that the set $\tilde{S} = \cup_{j=1}^{\infty} K_{1/j}$ is σ-compact metric whether S is separable or not. However, in this case $\mathcal{P}(S)$ may not be metric under the weak topology. Nonetheless, the relative weak topology on Λ (and $\bar{\Lambda}$) is metrizable.

In applications one might have $\Lambda = \{P_n\}_{n=1}^{\infty}$, where $P_n = P \circ X_n^{-1}$ is the distribution of a random map X_n. If X_n is real-valued, for example, then one might try to check tightness by a Chebyshev-type inequality, see, for example, Exercise 9.

The following definition and proposition provide a frequently used metrization in weak convergence theory.

Definition 5.4. The **Prohorov metric** d_π on $\mathcal{P}(S)$ is defined by

$$d_\pi(P, Q) := \inf\{\varepsilon > 0 : P(A) \leq Q(A^\varepsilon) + \varepsilon, Q(A) \leq P(A^\varepsilon) + \varepsilon, \forall A \in \mathcal{B}(S)\}.$$

[2] See, for example, Royden, H. L. (1968), p. 164.

Remark 5.3. Essentially using the symmetry that $A \subseteq S \backslash B^{\varepsilon}$ if and only if $B \subseteq S \backslash A^{\varepsilon}$, one may check that if $P(A) \leq Q(A^{\varepsilon}) + \varepsilon$ for all $A \in \mathcal{B}(S)$ then $d_{\pi}(P, Q) \leq \varepsilon$. That is it suffices to check that one of the inequalities holds for all $A \in \mathcal{B}(S)$ to get the other. For if the first inequality holds for all A, taking $B = S \backslash A^{\varepsilon}$, one has $P(A^{\varepsilon}) = 1 - P(B) \geq 1 - Q(B^{\varepsilon}) - \varepsilon = Q(S \backslash B^{\varepsilon}) - \varepsilon \geq Q(A) - \varepsilon$.

Proposition 5.10. Let (S, ρ) be a separable metric space. Then d_{π} metrizes the weak topology on $\mathcal{P}(S)$ in the sense that:

(i) d_{π} defines a metric on $\mathcal{P}(S)$.
(ii) If $d_{\pi}(P_n, P) \to 0$ as $n \to \infty$ then $P_n \Rightarrow P$.
(iii) If $P_n \Rightarrow P$, then $d_{\pi}(P_n, P) \to 0$ as $n \to \infty$.

Proof. Suppose that $d_{\pi}(P, Q) = 0$. Then from the definition of d_{π} one arrives for all closed sets F, letting $\varepsilon \downarrow 0$ with $A = F$ in the definition, at $P(F) \leq Q(F)$ and $Q(F) \leq P(F)$. Symmetry and nonnegativity are obvious. For the triangle inequality let $d_{\pi}(P_i, P_{i+1}) = \varepsilon_i, i = 1, 2$. Then $P_1(A) \leq P_2(A^{\varepsilon'_1}) + \varepsilon'_1 \leq P_3((A)^{\varepsilon'_1})^{\varepsilon'_2}) + \varepsilon'_1 + \varepsilon'_2$, for all $\varepsilon'_i > \varepsilon_i, i = 1, 2$. Thus $d_{\pi}(P_1, P_3) \leq \varepsilon'_1 + \varepsilon'_2$. Since this is true for all $\varepsilon'_i > \varepsilon_i$, $i = 1, 2$, the desired triangle inequality follows. Next suppose that $d_{\pi}(P_n, P) \to 0$ as $n \to \infty$. Let $\varepsilon_n \to 0$ be such that $d_{\pi}(P_n, P) < \varepsilon_n$. Then, by definition, $P_n(F) \leq P(F^{\varepsilon_n}) + \varepsilon_n$ for all closed F. Thus $\limsup_n P_n(F) \leq P(F)$ for all closed F, and weak convergence follows from Alexandrov's conditions. For the converse, fix an $\varepsilon > 0$. In view of the remark following the definition of d_{π} it suffices to show that for all n sufficiently large, say $n \geq n_0$, one has for any Borel set A that $P(A) \leq P_n(A^{\varepsilon}) + \varepsilon$. By separability, S is the union of countably many open balls B_i, $i \geq 1$, of diameter smaller than ε. Choose N such that $P(S \backslash \cup_{m=1}^{N} B_m) \leq P(\cup_{m \geq N+1} B_m) < \varepsilon$. Now by Alexandrov's conditions, $P_n \Rightarrow P$ implies that for any of the finitely many open sets of the form $G := B_{i_1} \cup \cdots \cup B_{i_m}, 1 \leq i_1 < \cdots < i_m \leq N$, there is an n_0 such that $P_n(G) > P(G) - \varepsilon$ for all $n \geq n_0$. For $A \in \mathcal{B}(S)$ let $\hat{A} = \cup_{i=1}^{N} \{B_i : B_i \cap A \neq \emptyset\}$. Then consider the special choice $G = \hat{A}^{\varepsilon} := \{x \in S : \rho(x, \hat{A}) < \varepsilon\}$. In particular, one has for $n > n_0$ that $P(A) \leq P(\hat{A}) + P(\cup_{i > N} B_i) \leq P(\hat{A}) + \varepsilon < P_n(\hat{A}) + 2\varepsilon \leq P_n(\hat{A}^{\varepsilon}) + 2\varepsilon \leq P_n(A^{2\varepsilon}) + 2\varepsilon$, since $\hat{A} \subseteq A^{\varepsilon}$, so that $\hat{A}^{\varepsilon} \subseteq A^{2\varepsilon}$. Thus $d_{\pi}(P_n, P) \leq 2\varepsilon$ for all $n \geq n_0$. ■

EXERCISES

Exercise Set V

1. Prove the equivalence of (i)–(iii) of Theorem 5.3 in the case $k \geq 2$.
2. Complete the following steps to prove the equivalence of (i) and (iv) of Theorem 5.3 in the case $k \geq 2$.
   (i) Show that F is *continuous from above* at $\mathbf{x}$ in the sense that given $\varepsilon > 0$ there is a $\delta > 0$ such that $|F(\mathbf{x}) - F(\mathbf{y})| < \varepsilon$ whenever $x_i \leq y_i < x_i + \delta, i = 1, \ldots, k$. [*Hint*: Use the continuity of probability measure from above.]

(ii) Say that F is *continuous from below* at $\mathbf{x}$ if given $\varepsilon > 0$ there is a $\delta > 0$ such that $|F(\mathbf{x}) - F(\mathbf{y})| < \varepsilon$ whenever $x_i - \delta < y_i \leq x_i, i = 1, \dots, k$. Show that $\mathbf{x}$ is a continuity point of F if and only if continuity holds from above and below. Moreover, $\mathbf{x}$ is a continuity point of F if and only if $F(\mathbf{x}) = P(\cap_{i=1}^k \{\mathbf{y} \in \mathbb{R}^k : y_i < x_i\})$.

(iii) Show that $\mathbf{x}$ is a continuity point of F if and only if $\cap_{i=1}^k \{\mathbf{y} \in \mathbb{R}^k : y_i \leq x_i\}$ is a P-continuity set. [*Hint*: The boundary of $\cap_{i=1}^k \{\mathbf{y} \in \mathbb{R}^k : y_i \leq x_i\}$ is the relative complement $\cap_{i=1}^k \{\mathbf{y} \in \mathbb{R}^k : y_i \leq x_i\} \setminus \cap_{i=1}^k \{\mathbf{y} \in \mathbb{R}^k : y_i < x_i\}$.]

(iv) Show that if $P_n \Rightarrow P$ then $F_n(\mathbf{x}) \to F(\mathbf{x})$ at all continuity points $\mathbf{x}$ of F.

(v) Let $\mathcal{A}$ be a π-system of Borel subsets of S, i.e., closed under finite intersections. Assume that each open subset of S is a finite or countable union of elements of $\mathcal{A}$. Show that if $P_n(A) \to P(A)$ for each $A \in \mathcal{A}$ then $P_n \Rightarrow P$. [*Hint*: Use the inclusion–exclusion principle to show that $P_n(\cup_{m=1}^N A_m) \to P(\cup_{m=1}^N A_m)$ if $A_m \in \mathcal{A}$ for $m = 1, \dots, m$. Verify for $\varepsilon > 0$ and open $G = \cup_m A_m, A_m \in \mathcal{A}$, that there is an N such that $P(G) - \varepsilon \leq P(\cup_{m=1}^N A_m) = \lim_n P_n(\cup_{m=1}^N A_m) \leq \liminf_n P_n(G)$.]

(vi) Let $\mathcal{A}$ be a π-system of sets such that for each $x \in S$ and every $\varepsilon > 0$ there is an $A \in \mathcal{A}$ such that $x \in A^\circ \subseteq A \subseteq B_\varepsilon(x) := \{y \in S : d(y, x) < \varepsilon\}$, where A° denotes the set of points belonging to the *interior* of A. Show that if S is a separable metric space and $P_n(A) \to P(A)$ for all $A \in \mathcal{A}$ then $P_n \Rightarrow P$. [*Hint*: Check that $\mathcal{A}$ satisfies the conditions required in the previous step.]

(vii)Show that if $F_n(\mathbf{x}) \to F(\mathbf{x})$ at each point $\mathbf{x}$ of continuity of F then $P_n \Rightarrow P$. [*Hint*: Take $\mathcal{A}$ to be the collection of sets of the form $A = \{\mathbf{x} : a_i < x_i \leq b_i, i = 1, \dots, k\}$ for which the $2k$ $(k-1)$-dimensional hyperplanes determining each of its faces has P-measure zero. The P, P_n-probabilities of $A \in \mathcal{A}$ are sums and differences of values of $F(\mathbf{x}), F_n(\mathbf{x})$, respectively, as $\mathbf{x}$ varies over the 2^k vertices of A. Moreover, vertices of $A \in \mathcal{A}$ are continuity points of F, and at most countably many parallel hyperplanes can have positive P-measure.]

3. Use Prohorov's theorem to give a simple derivation for Exercise 2. [*Hint*: Suppose that $F_n(x) \to F(x)$ at all points x of continuity of F. Show that $\{P_n : n \geq 1\}$ is tight, using $P_n((a, b]) \geq F_n(b) - \sum_{i=1}^k F_n(b_1, \dots, b_{i-1}, a_i, b_{i+1}, \dots, b_k), 1 \leq i \leq k$, for $a_i < b_i, \forall i$, where $a = (a_1, \dots, a_k), b = (b_1, \dots, b_k)$.]

4. Suppose that $\{(X_n, Y_n)\}_{n=1}^\infty$ is a sequence of pairs of real-valued random variables that converge in distribution to (X, Y). Show that $X_n + Y_n$ converges in distribution to $X + Y$. [*Hint*: The map $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by $h(x, y) = x + y$ is continuous.]

5. (i) Show that if F is closed, $\delta > 0$, then $\partial\{x : \rho(x, F) \leq \delta\} \subseteq \{x : \rho(x, F) = \delta\}$. [*Hint*: If y belongs to the set on the left, there is a sequence $y_n \to y$ such that $\rho(y_n, F) \geq \delta$.] (ii) Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space. Suppose A_δ, $\delta > 0$, is a collection of disjoint measurable sets. Show that $P(A_\delta) > 0$ for at most countably many δ. [*Hint*: For each positive integer n, the set $\{\delta > 0 : P(A_\delta) > 1/n\}$ must be a finite set.] (iii) Let $h : S_1 \to S_2$ be Borel measurable and P-a.s. continuous. With F_h as in Theorem 5.2, show that $F_h^- \subseteq F_h \cup D_h$. [*Hint*: If $y \in F_h^- \setminus F_h \subseteq \partial F_h$, then $h(y) \notin F$, but there is a sequence $y_n \to y$ such that $h(y_n) \in F$ for all $n \geq 1$.]

6. Let $\{X_n\}_{n=1}^\infty$ be a sequence of random maps with values in a metric space S with metric ρ and Borel σ-field $\mathcal{S} = \mathcal{B}(S)$.

(i) Show that X_n converges in probability to an a.s. constant c if and only if the sequence of probabilities $Q_n := P \circ X_n^{-1}$ converge weakly to δ_c. [Here *convergence in probability* means that given $\varepsilon > 0$ one has $P(\rho(X_n, c) > \varepsilon) \to 0$ as $n \to \infty$.]

(ii) Show that convergence in probability to a random map X implies $P\circ X_n^{-1} \Rightarrow P\circ X^{-1}$ as $n \to \infty$.

7. Let S be a metric space with Borel σ-field $\mathcal{B}(S)$. (a) Give an example to show that vague convergence does not imply weak convergence, referred to as *escape of probability mass to infinity*. [*Hint*: Consider, for example, $P_n = \frac{2}{3}\delta_{\{\frac{1}{n}\}} + \frac{1}{3}\delta_{\{n\}}$.] (b) Show that if $\{P_n\}_{n=1}^{\infty}$ is tight, then vague convergence and weak convergence are equivalent for $\{P_n\}_{n=1}^{\infty}$.

8. Let (S,ρ) be a compact metric space.
   (i) Show that S is separable. [*Hint*: For each integer $n \geq 1$, consider the open cover of S by open balls of radii 2^{-n} centered at $x \in S$.]
   (ii) Give a proof of Proposition 5.5. [*Hint*: Let $\{f_n\}$ be a countable dense sequence in $C(S)$. For a sequence of probabilities P_n, first consider the bounded sequence of numbers $\int_S f_1 dP_n, n \geq 1$. Extract a subsequence $P_{n_{1k}}$ such that $L(f_1) := \lim \int_S f_1 dP_{n_{1k}}$ exists. Next consider the bounded subsequence $\int_S f_2 dP_{n_{1k}}$, etc. Use Cantor's diagonalization to obtain a densely defined bounded linear functional L (on the linear span of $\{f_k : k \geq 1\}$) and extend by continuity to $C(S)$. Use the Riesz representation theorem (Appendix A) to obtain the weak limit point.]

9. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of real-valued random variables on $(\Omega, \mathcal{F}, P)$.
   (i) Suppose that each X_n is in L^p, $n \geq 1$, for some $p \geq 1$, and $\sup_n \mathbb{E}|X_n|^p < \infty$. Show that $\{Q_n = P \circ X_n^{-1}\}_{n=1}^{\infty}$ is a tight sequence. [*Hint*: Use a Chebyshev-type inequality.]
   (ii) Suppose there is a $\delta > 0$ such that for each $-\delta \leq t \leq \delta$. $E^{tX_n} < \infty$ for each n, and $\lim_{n\to\infty} \mathbb{E}e^{tX_n} = m(t)$ exists and is finite. Show that $\{Q_n = P \circ X_n^{-1}\}_{n=1}^{\infty}$ is tight. [*Hint*: Apply the Markov inequality to the event $[e^{\delta|X_n|} > e^{\delta a}]$.]

10. Define probabilities on $\mathbb{R}$ absolutely continuous with respect to Lebesgue measure with density $P_\varepsilon(dx) = \rho_\varepsilon(x)dx$, where $\rho_\varepsilon(x)$ was introduced to obtain C^∞-approximations with compact support in (5.7). Let $\delta_{\{0\}}$ denote the Dirac probability concentrated at 0, and show that $P_\varepsilon \Rightarrow \delta_{\{0\}}$ as $\varepsilon \downarrow 0$. [*Hint*: Consider probabilities of open sets in Alexandrov's theorem.]

11. Suppose that P_n, $n \geq 1$, is a sequence of probabilities concentrated on $[a,b]$. Suppose that one may show for each positive integer r that $\int_{[a,b]} x^r P_n(dx) \to m_r \in \mathbb{R}$ as $n \to \infty$. Show that there is a probability P such that $P_n \Rightarrow P$ as $n \to \infty$ and $\int_{[a,b]} x^r P(dx) = m_r$ for each $r \geq 1$.

12. Let (S,ρ) be a metric space and B a Borel subset of S given the relative (metric) topology. Let $\{P_n : n \geq 1\}$ be a sequence of probabilities in $\mathcal{P}(S)$ such that $P_n(B) = 1$ for all n. If the restrictions of P_n, $n \geq 1$, to B converge weakly to a probability $P \in \mathcal{P}(B)$, show that $P_n \Rightarrow P$, when considered in $\mathcal{P}(S)$, i.e., extending P to S by setting $P(S\backslash B) = 0$.

CHAPTER VI

Fourier Series, Fourier Transform, and Characteristic Functions

Consider a real- or complex-valued periodic function on the real line. By changing the scale if necessary, one may take the period to be 2π. Is it possible to represent f as a superposition of the periodic functions ("waves") $\cos nx$, $\sin nx$ of *frequency* n ($n = 0, 1, 2, \ldots$)? In view of the **Weierstrass approximation theorem** (Theorem 6.1 below), every continuous periodic function f of period 2π is the limit (in the sense of uniform convergence of functions) of a sequence of **trigonometric polynomials**, i.e., functions of the form

$$\sum_{n=-T}^{T} c_n e^{inx} = c_0 + \sum_{n=1}^{T} (a_n \cos nx + b_n \sin nx).$$

The theory of Fourier series says, among other things, that with the weaker notion of L^2-convergence the approximation holds for a wider class of functions, namely for all square-integrable functions f on $[-\pi, \pi]$; here square-integrability means that f is measurable and that $\int_{-\pi}^{\pi} |f(x)|^2 \, dx < \infty$. This class of functions is denoted by $L^2[-\pi, \pi]$. The successive coefficients c_n for this approximation are the so-called *Fourier coefficients*:

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, dx \qquad (n = 0, \pm 1, \pm 2, \ldots). \tag{6.1}$$

It should be noted that in general, we consider integrals of complex-valued functions in this section, and the $L^p = L^p(dx)$ spaces are those of complex-valued functions (See Exercise 19 of Chapter I).

The functions e^{inx} $(n = 0, \pm 1, \pm 2, \ldots)$ form an **orthonormal set**:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{inx}e^{-imx}\,dx = \begin{cases} 0, & \text{for } n \neq m, \\ 1 & \text{for } n = m, \end{cases} \tag{6.2}$$

so that the **Fourier series of** f written formally, without regard to convergence for the time being, as

$$\sum_{n=-\infty}^{\infty} c_n e^{inx} \tag{6.3}$$

is a representation of f as a superposition of orthogonal components. To make matters precise we first prove the following theorem; see Exercise 1 for an alternative approach.

Theorem 6.1. Let f be a continuous periodic function of period 2π. Then, given $\delta > 0$, there exists a trigonometric polynomial, specifically a **Fejér average** $\sum_{n=-N}^{N} d_n e^{inx}$, such that

$$\sup_{x\in\mathbb{R}^1}\left| f(x) - \sum_{n=-N}^{N} d_n e^{inx}\right| < \delta.$$

Proof. For each positive integer N, introduce the **Fejér kernel**

$$k_N(x) := \frac{1}{2\pi}\sum_{n=-N}^{N}\left(1 - \frac{|n|}{N+1}\right)e^{inx}. \tag{6.4}$$

This may also be expressed as

$$\begin{aligned} 2\pi(N+1)k_N(x) &= \sum_{0\le j,k\le N} e^{i(j-k)x} = \left|\sum_{j=0}^{N} e^{ijx}\right|^2 \\ &= \frac{2\{1-(\cos(N+1)x\}}{2(1-\cos x)} = \left(\frac{\sin\{\frac{1}{2}(N+1)x\}}{\sin\frac{1}{2}x}\right)^2. \end{aligned} \tag{6.5}$$

At $x = 2n\pi$ $(n = 0, \pm 1, \pm 2, \ldots)$, the right side is taken to be $(N+1)^2$. The first equality in (6.5) follows from the fact that there are $N+1-|n|$ pairs (j,k) in the sum such that $j - k = n$. It follows from (6.5) that k_N is a positive continuous periodic function with period 2π. Also, k_N is a pdf on $[-\pi, \pi]$, since nonnegativity follows from (6.5) and normalization from (6.4) on integration. For every $\varepsilon > 0$ it follows from (6.5) that $k_N(x)$ goes to zero uniformly on $[-\pi,-\varepsilon]\cup[\varepsilon,\pi]$, so that

$$\int_{[-\pi,-\varepsilon]\cup[\varepsilon,\pi]} k_N(x)dx \to 0 \qquad \text{as } N\to\infty. \tag{6.6}$$

In other words, $k_N(x)dx$ converges weakly to $\delta_0(dx)$, the point mass at 0, as $N\to\infty$.

Consider now the approximation f_N of f defined by

$$f_N(x) := \int_{-\pi}^{\pi} f(y)k_N(x-y)dy = \sum_{n=-N}^{N}\left(1-\frac{|n|}{N+1}\right)c_n e^{inx}, \tag{6.7}$$

where c_n is the n**th Fourier coefficient** of f. By changing variables and using the periodicity of f and k_N, one may express f_N as

$$f_N(x) = \int_{-\pi}^{\pi} f(x-y)k_N(y)dy.$$

Therefore, writing $M=\sup\{|f(x)| : x\in\mathbb{R}\}$, and $\delta_\varepsilon = \sup\{|f(y)-f(y')| : |y-y'|<\varepsilon\}$, one has

$$|f(x)-f_N(x)| \le \int_{-\pi}^{\pi} |f(x-y)-f(x)|k_N(y)dy \le 2M\int_{[-\pi,-\varepsilon]\cup[\varepsilon,\pi]} k_N(y)dy+\delta_\varepsilon. \tag{6.8}$$

It now follows from (6.6) that $f-f_N$ converges to zero uniformly as $N\to\infty$. Now write $d_n=(1-|n|/(N+1))c_n$. ■

The next task is to establish the convergence of the Fourier series (6.3) to f in L^2. Here the norm $\|\cdot\|$ is $\|\cdot\|_2$ as defined by (6.10) below.

Theorem 6.2.

- **a.** For every f in $L^2[-\pi,\pi]$, the Fourier series of f converges to f in L^2-norm, and the identity $\|f\| = (\sum_{-\infty}^{\infty}|c_n|^2)^{1/2}$ holds for its Fourier coefficients c_n.
- **b.** If (i) f is differentiable, (ii) $f(-\pi)=f(\pi)$, and (iii) f' is square-integrable, then the Fourier series of f also converges uniformly to f on $[-\pi,\pi]$.

Proof. (a) Note that for every square-integrable f and all positive integers N,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(f(x)-\sum_{-N}^{N}c_n e^{inx}\right)e^{-imx}dx = c_m-c_m=0 \qquad (m=0,\pm1,\ldots,\pm N). \tag{6.9}$$

Therefore, if one defines the **norm** (or "length") of a function g in $L^2[-\pi,\pi]$ by

$$\|g\| = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|g(x)|^2 dx\right)^{1/2} \equiv \|g\|_2, \tag{6.10}$$

then, writing $\bar{z}$ for the complex conjugate of z,

$$\begin{aligned}
0 &\leq \|f - \sum_{-N}^{N} c_n e^{in\cdot}\|^2 \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(f(x) - \sum_{-N}^{N} c_n e^{inx}\right)\left(\bar{f}(x) - \sum_{-N}^{N} \bar{c}_n e^{-inx}\right) dx \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}(f(x) - \sum_{-N}^{N} c_n e^{inx})\bar{f}(x)dx \\
&= \|f\|^2 - \sum_{-N}^{N} c_n\bar{c}_n = \|f\|^2 - \sum_{-N}^{N}|c_n|^2.
\end{aligned} \tag{6.11}$$

This shows that $\|f - \sum_{-N}^{N} c_n e^{in\cdot}\|^2$ *decreases* as N increases and that

$$\lim_{N\to\infty}\|f - \sum_{-N}^{N} c_n e^{in\cdot}\|2 = \|f\|^2 - \sum_{-\infty}^{\infty}|c_n|^2. \tag{6.12}$$

To prove that the right side of (6.12) vanishes, first assume that f is continuous and $f(-\pi) = f(\pi)$. Given $\varepsilon > 0$, there exists, by Theorem 6.1, a trigonometric polynomial $\sum_{-N_0}^{N_0} d_n e^{inx}$ such that

$$\max_x\left|f(x) - \sum_{-N_0}^{N_0} d_n e^{inx}\right| < \varepsilon.$$

This implies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|f(x) - \sum_{-N_0}^{N_0} d_n e^{inx}\right|^2 dx < \varepsilon^2. \tag{6.13}$$

But by (6.9), $f(x) - \sum_{-N_0}^{N_0} c_n \exp\{inx\}$ is orthogonal to e^{imx} $(m = 0, \pm1, \ldots, \pm N_0)$, so that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|f(x) - \sum_{-N_0}^{N_0} d_n e^{inx}\right|^2 dx$$

$$
\begin{aligned}
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|f(x)-\sum_{-N_0}^{N_0}c_n e^{inx}+\sum_{-N_0}^{N_0}(c_n-d_n)e^{inx}\right|^2 dx\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|f(x)-\sum_{-N_0}^{N_0}c_n e^{inx}\right|^2 dx\\
&\quad + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{-N_0}^{N_0}(c_n-d_n)e^{inx}\right|^2 dx. \qquad (6.14)
\end{aligned}
$$

Hence, by (6.13), (6.14), and (6.11),

$$
\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|f(x)-\sum_{-N_0}^{N_0}c_n e^{inx}\right|^2 dx<\varepsilon^2, \qquad \lim_{N\to\infty}\left\|f-\sum_{-N}^{N}c_n e^{in\cdot}\right\|^2\le\varepsilon^2. \qquad (6.15)
$$

Since $\varepsilon>0$ is arbitrary, it follows that

$$
\lim_{N\to\infty}\left\|f(x)-\sum_{-N}^{N}c_n e^{inx}\right\|=0, \qquad (6.16)
$$

and by (6.12),

$$
\|f\|^2=\sum_{-\infty}^{\infty}|c_n|^2. \qquad (6.17)
$$

This completes the proof of convergence for continuous periodic f. Now it may be shown that given a square-integrable f and $\varepsilon>0$, there exists a continuous periodic g such that $\|f-g\|<\varepsilon/2$ (Exercise 1). Also, letting $\sum d_n e^{inx}$, $\sum c_n e^{inx}$ be the Fourier series of g, f, respectively, there exists N_1 such that

$$
\left\|g-\sum_{-N_1}^{N_1}d_n e^{in\cdot}\right\|<\frac{\varepsilon}{2}.
$$

Hence (see (6.14))

$$
\begin{aligned}
\left\|f-\sum_{-N_1}^{N_1}c_n e^{in\cdot}\right\|&\le\left\|f-\sum_{-N_1}^{N_1}d_n e^{in\cdot}\right\|\le\|f-g\|+\left\|g-\sum_{-N_1}^{N_1}d_n e^{in\cdot}\right\|\\
&<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon. \qquad (6.18)
\end{aligned}
$$

Since $\varepsilon > 0$ is arbitrary and $\|f(x) - \sum_{-N}^{N} c_n e^{inx}\|^2$ decreases to $\|f\|^2 - \sum_{-\infty}^{\infty} |c_n|^2$ as $N \uparrow \infty$ (see (6.12)), one has

$$\lim_{N\to\infty} \left\| f - \sum_{-N}^{N} c_n e^{in\cdot} \right\| = 0; \quad \|f\|^2 = \sum_{-\infty}^{\infty} |c_n|^2. \tag{6.19}$$

To prove part (b), let f be as specified. Let $\sum c_n e^{inx}$ be the Fourier series of f, and $\sum c_n^{(1)} e^{inx}$ that of f'. Then

$$\begin{aligned} c_n^{(1)} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f'(x) e^{-inx}\, dx = \frac{1}{2\pi} f(x) e^{-inx} \Big|_{-\pi}^{\pi} + \frac{in}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx}\, dx \\ &= 0 + inc_n = inc_n. \end{aligned} \tag{6.20}$$

Since f' is square-integrable,

$$\sum_{-\infty}^{\infty} |nc_n|^2 = \sum_{-\infty}^{\infty} |c_n^{(1)}|^2 < \infty. \tag{6.21}$$

Therefore, by the Cauchy–Schwarz inequality,

$$\sum_{-\infty}^{\infty} |c_n| = |c_0| + \sum_{n\neq 0} \frac{1}{|n|} |nc_n| \leq |c_0| + \left(\sum_{n\neq 0} \frac{1}{n^2} \right)^{1/2} \left(\sum_{n\neq 0} |nc_n|^2 \right)^{1/2} < \infty. \tag{6.22}$$

But this means that $\sum c_n e^{inx}$ is uniformly absolutely convergent, since

$$\max_x \left| \sum_{|n|>N} c_n e^{inx} \right| \leq \sum_{|n|>N} |c_n| \to 0 \qquad \text{as } N \to \infty.$$

Since the continuous functions $\sum_{-N}^{N} c_n e^{inx}$ converge uniformly (as $N \to \infty$) to $\sum_{-\infty}^{\infty} c_n e^{inx}$, the latter must be a continuous function, say h. Uniform convergence to h also implies convergence in norm to h. Since $\sum_{-\infty}^{\infty} c_n e^{inx}$ also converges in norm to f, $f(x) = h(x)$ for all x. For if the two continuous functions f and h are not identically equal, then

$$\int_{-\pi}^{\pi} |f(x) - h(x)|^2 dx > 0. \qquad \blacksquare$$

Definition 6.1. For a finite measure (or a finite signed measure) μ on the circle $[-\pi,\pi)$ (identifying $-\pi$ and π), the nth **Fourier coefficient of** μ is defined by

$$c_n = \frac{1}{2\pi}\int_{[-\pi,\pi)} e^{-inx}\mu(dx) \qquad (n = 0, \pm 1, \ldots). \tag{6.23}$$

If μ has a density f, then (6.23) is the same as the nth Fourier coefficient of f given by (6.1).

Proposition 6.3. A finite measure μ on the circle is determined by its Fourier coefficients.

Proof. Approximate the measure $\mu(dx)$ by $g_N(x)\,dx$, where

$$g_N(x) := \int_{[-\pi,\pi)} k_N(x-y)\mu(dy) = \sum_{-N}^{N}\left(1 - \frac{|n|}{N+1}\right) c_n e^{inx}, \tag{6.24}$$

with c_n defined by (6.23). For every continuous periodic function h (i.e., for every continuous function on the circle),

$$\int_{[-\pi,\pi)} h(x)g_N(x)\,dx = \int_{[-\pi,\pi)}\left(\int_{[-\pi,\pi)} h(x)k_N(x-y)\,dx\right)\mu(dy). \tag{6.25}$$

As $N \to \infty$, the probability measure $k_N(x-y)\,dx = k_N(y-x)\,dx$ on the circle converges weakly to $\delta_y(dx)$. Hence, the inner integral on the right side of (6.25) converges to $h(y)$. Since the inner integral is bounded by $\sup\{|h(y)| : y \in \mathbb{R}\}$, Lebesgue's dominated convergence theorem implies that

$$\lim_{N\to\infty}\int_{[-\pi,\pi)} h(x)g_N(x)\,dx = \int_{[-\pi,\pi)} h(y)\mu\,(dy). \tag{6.26}$$

This means that μ is determined by $\{g_N : N \geq 1\}$. The latter in turn are determined by $\{c_n\}_{n\in\mathbb{Z}}$. ■

We are now ready to answer an important question: When is a given sequence $\{c_n : n = 0, \pm 1, \ldots\}$ the sequence of Fourier coefficients of a finite measure on the circle? A sequence of complex numbers $\{c_n : n = 0, \pm 1, \pm 2, \ldots\}$ is said to be **positive-definite** if for any finite sequence of complex numbers $\{z_j : 1 \leq j \leq N\}$, one has

$$\sum_{1\leq j,k\leq N} c_{j-k} z_j \bar{z}_k \geq 0. \tag{6.27}$$

Theorem 6.4 *(Herglotz Theorem).* $\{c_n : n = 0, \pm 1, \ldots\}$ is the sequence of Fourier coefficients of a probability measure on the circle if and only if it is positive-definite, and $c_0 = \frac{1}{2\pi}$.

Proof.
Necessity. If μ is a probability measure on the circle, and $\{z_j : 1 \le j \le N\}$ a given finite sequence of complex numbers, then

$$\begin{aligned}\sum_{1\le j,k\le N} c_{j-k} z_j \bar{z}_k &= \frac{1}{2\pi} \sum_{1\le j,k\le N} z_j \bar{z}_k \int_{[-\pi,\pi)} e^{i(k-j)x} \mu(dx) \\ &= \frac{1}{2\pi} \int_{[-\pi,\pi)} \left(\sum_1^N z_j e^{ikx}\right) \left(\sum_1^N \bar{z}_k e^{-ijx}\right) \mu(dx) \\ &= \frac{1}{2\pi} \int_{[-\pi,\pi)} \left|\sum_1^N z_j e^{ijx}\right|^2 \mu(dx) \ge 0. \qquad (6.28)\end{aligned}$$

Also,

$$c_0 = \frac{1}{2\pi} \int_{[-\pi,\pi)} \mu(dx) = \frac{1}{2\pi}.$$

Sufficiency. Take $z_j = e^{i(j-1)x}$, $j = 1, 2, \ldots, N+1$, in (6.27) to get

$$g_N(x) := \frac{1}{N+1} \sum_{0\le j,k\le N} c_{j-k} e^{i(j-k)x} \ge 0. \qquad (6.29)$$

Again, since there are $N + 1 - |n|$ pairs (j, k) such that $j - k = n$ $(-N \le n \le N)$ it follows that (6.29) becomes

$$0 \le g_N(x) = \sum_{-N}^{N} \left(1 - \frac{|n|}{N+1}\right) e^{inx} c_n. \qquad (6.30)$$

In particular, using (6.2),

$$\int_{[-\pi,\pi)} g_N(x) dx = 2\pi c_0 = 1. \qquad (6.31)$$

Hence g_N is a pdf on $[-\pi, \pi]$. By Proposition 5.5, there exists a subsequence $\{g_{N'}\}$ such that $g_{N'}(x)\, dx$ converges weakly to a probability measure $\mu(dx)$ on $[-\pi, \pi]$ as $N' \to \infty$. Also, again using (6.2) yields

$$\int_{[-\pi,\pi)} e^{-inx} g_N(x) dx = 2\pi \left(1 - \frac{|n|}{N+1}\right) c_n \qquad (n = 0, \pm 1, \ldots, \pm N). \qquad (6.32)$$

For each fixed n, restrict to the subsequence $N = N'$ in (6.32) and let $N' \to \infty$. Then, since for each n, $\cos(nx), \sin(nx)$ are bounded continuous functions,

$$2\pi c_n = \lim_{N'\to\infty} 2\pi\left(1 - \frac{|n|}{N'+1}\right) c_n = \int_{[-\pi,\pi)} e^{-inx}\mu(dx) \qquad (n = 0, \pm1, \ldots). \quad (6.33)$$

In other words, c_n is the nth Fourier coefficient of μ. ■

Corollary 6.5. A sequence $\{c_n : n = 0, \pm1, \ldots\}$ of complex numbers is the sequence of Fourier coefficients of a finite measure on the circle $[-\pi, \pi)$ if and only if $\{c_n : n = 0, \pm1, \ldots\}$ is positive-definite.

Proof. Since the measure $\mu = 0$ has Fourier coefficients $c_n = 0$ for all n, and the latter is trivially a positive-definite sequence, it is enough to prove the correspondence between nonzero positive-definite sequences and nonzero finite measures. It follows from Theorem 6.4, by normalization, that this correspondence is 1–1 between positive-definite sequences $\{c_n : n = 0, \pm1, \ldots\}$ with $c_0 = c > 0$ and measures on the circle having total mass $2\pi c$. ■

Definition 6.2. The **Fourier transform** of an integrable (real- or complex-valued) function f on $(-\infty, \infty)$ is the function $\hat{f}$ on $(-\infty, \infty)$ defined by

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} e^{i\xi y} f(y)\, dy, \qquad -\infty < \xi < \infty. \quad (6.34)$$

As a special case take $f = \mathbf{1}_{(c,d]}$. Then,

$$\hat{f}(\xi) = \frac{e^{i\xi d} - e^{i\xi c}}{i\xi}, \quad (6.35)$$

so that $\hat{f}(\xi) \to 0$ as $|\xi| \to \infty$. This convergence to zero as $\xi \to \pm\infty$ is clearly valid for arbitrary **step functions**, i.e., finite linear combinations of indicator functions of finite intervals. Now let f be an arbitrary integrable function. Given $\varepsilon > 0$ there exists a step function f_ε such that (see Remark following Proposition 2.6, Appendix A)

$$\|f_\varepsilon - f\|_1 := \int_{-\infty}^{\infty} |f_\varepsilon(y) - f(y)|\, dy < \varepsilon. \quad (6.36)$$

Now it follows from (6.34) that $|\hat{f}_\varepsilon(\xi) - \hat{f}(\xi)| \le \|f_\varepsilon - f\|_1$ for all ξ. Since $\hat{f}_\varepsilon(\xi) \to 0$ as $\xi \to \pm\infty$, one has $\limsup_{|\xi|\to\infty} |\hat{f}(\xi)| \le \varepsilon$. Since $\varepsilon > 0$ is arbitrary,

$$\hat{f}(\xi) \to 0 \qquad \text{as } |\xi| \to \infty. \quad (6.37)$$

Thus we have proved the following result.

Proposition 6.6 ***(Riemann–Lebesgue Lemma).*** The Fourier transform $\hat{f}(\xi)$ of an integrable function f tends to zero in the limit as $|\xi| \to \infty$.

If f is continuously differentiable and f, f' are both integrable, then integration by parts yields (Exercise 2(b))

$$\hat{f'}(\xi) = -i\xi \hat{f}(\xi). \tag{6.38}$$

The boundary terms in deriving (6.38) vanish, for if f' is integrable (as well as f) then $f(x) \to 0$ as $x \to \pm\infty$. More generally, if f is r-times continuously differentiable and $f^{(j)}$, $0 \le j \le r$, are all integrable, then one may repeat the relation (6.38) to get by induction (Exercise 2(b))

$$\widehat{f^{(r)}}(\xi) = (-i\xi)^r \hat{f}(\xi). \tag{6.39}$$

In particular, (6.39) implies that if f, f', f'' are integrable then $\hat{f}$ is integrable.

Definition 6.3. The **Fourier transform $\hat{\mu}$ of a finite measure** μ on $\mathbb{R}$ is defined by

$$\hat{\mu}(\xi) = \int_{-\infty}^{\infty} e^{i\xi x}\, d\mu(x). \tag{6.40}$$

If μ is a *finite signed measure*, i.e., $\mu = \mu_1 - \mu_2$ where μ_1, μ_2 are finite measures, then also one defines $\hat{\mu}$ by (6.40) directly, or by setting $\hat{\mu} = \hat{\mu}_1 - \hat{\mu}_2$. In particular, if $\mu(dx) = f(x)\, dx$, where f is real-valued and integrable, then $\hat{\mu} \;= \hat{f}$. If μ is a probability measure, then $\hat{\mu}$ is also called the **characteristic function** of μ, or of any random variable X on $(\Omega, \mathcal{F}, P)$ whose distribution is $\mu = P \circ X^{-1}$. In this case, by the change of variable formula, one has the equivalent definition

$$\hat{\mu}(\xi) = \mathbb{E}e^{i\xi X}. \tag{6.41}$$

We next consider the **convolution** of two integrable functions f, g:

$$f * g(x) = \int_{-\infty}^{\infty} f(x-y) g(y)\, dy \qquad (-\infty < x < \infty). \tag{6.42}$$

Since by the Tonelli part of the Fubini–Tonelli theorem,

$$\begin{aligned} \int_{-\infty}^{\infty} |f * g(x)|\, dx &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f(x-y)||g(y)|\, dy\, dx \\ &= \int_{-\infty}^{\infty} |f(x)|\, dx \int_{-\infty}^{\infty} |g(y)|\, dy, \end{aligned} \tag{6.43}$$

$f * g$ is integrable. Its Fourier transform is

$$\begin{aligned}(f*g)\hat{}(\xi) &= \int_{-\infty}^{\infty} e^{i\xi x}\left(\int_{-\infty}^{\infty} f(x-y)g(y)\,dy\right)dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\xi(x-y)}e^{i\xi y} f(x-y)g(y)\,dy\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\xi z}e^{i\xi y} f(z)g(y)\,dy\,dz = \hat{f}(\xi)\hat{g}(\xi), \end{aligned} \tag{6.44}$$

a result of importance in probability and analysis. By iteration, one defines the *n-fold convolution* $f_1 * \cdots * f_n$ of n integrable functions $f_1, \dots, f_n$ and it follows from (6.44) that $(f_1 * \cdots * f_n)\hat{} = \hat{f}_1\hat{f}_2\cdots\hat{f}_n$. Note also that if f, g are real-valued integrable functions and one defines the measures μ, ν by $\mu(dx) = f(x)\,dx$, $\nu(dx) = g(x)\,dx$, and $\mu * \nu$ by $(f*g)(x)\,dx$, then

$$\begin{aligned}(\mu*\nu)(B) &= \int_B (f*g)(x)\,dx = \int_{-\infty}^{\infty}\left(\int_B f(x-y)\,dx\right)g(y)\,dy \\ &= \int_{-\infty}^{\infty} \mu(B-y)g(y)\,dy \int_{-\infty}^{\infty} \mu(B-y)d\nu(y), \end{aligned} \tag{6.45}$$

for every interval (or, more generally, for every Borel set) B. Here $B - y$ is the *translate* of B by $-y$, obtained by subtracting from each point in B the number y. Also $(\mu*\nu)\hat{} = (f*g)\hat{} = \hat{f}\hat{g} = \hat{\mu}\hat{\nu}$. In general (i.e., whether or not finite signed measures μ and/or ν have densities), the last expression in (6.45) defines the *convolution* $\mu * \nu$ of finite signed measures μ and ν. The Fourier transform of this finite signed measure is still given by $(\mu*\nu)\hat{} = \hat{\mu}\hat{\nu}$. Recall that if X_1, X_2 are independent random variables on some probability space $(\Omega, \mathcal{A}, P)$ and have distributions Q_1, Q_2, respectively, then the distribution of X_1+X_2 is Q_1*Q_2. The characteristic function (i.e., Fourier transform) may also be computed from

$$(Q_1 * Q_2)\hat{}(\xi) = \mathbb{E}e^{i\xi(X_1+X_2)} = \mathbb{E}e^{i\xi X_1}\mathbb{E}e^{i\xi X_2} = \hat{Q}_1(\xi)\hat{Q}_2(\xi). \tag{6.46}$$

This argument extends to finite signed measures, and is an alternative way of thinking about (or deriving) the result $(\mu * \nu)\hat{} = \hat{\mu}\hat{\nu}$.

Theorem 6.7 (Uniqueness). Let P, Q be probabilities on the Borel σ-field of $\mathbb{R}^1$. Then $\hat{P}(\xi) = \hat{Q}(\xi)$ for all $\xi \in \mathbb{R}$ if and only if $P = Q$.

Proof. For each $\xi \in \mathbb{R}$, one has by definition of the characteristic function that $e^{-i\xi x}\hat{P}(\xi) = \int_{\mathbb{R}} e^{i\xi(y-x)}P(dy)$. Thus, integrating with respect to Q, one obtains the duality relation

$$\int_{\mathbb{R}} e^{-i\xi x}\hat{P}(\xi)Q(d\xi) = \int_{\mathbb{R}} \hat{Q}(y-x)P(dy). \tag{6.47}$$

Let $\varphi_{1/\sigma^2}(x) = \frac{\sigma}{\sqrt{2\pi}}e^{-\frac{\sigma^2x^2}{2}}$, $x \in \mathbb{R}$, denote the Gaussian pdf with variance $1/\sigma^2$ centered at 0, and take $Q(dx) \equiv \Phi_{1/\sigma^2}(dx) := \varphi_{1/\sigma^2}(x)dx$ in (6.47). Then $\hat{Q}(\xi) = \hat{\Phi}_{1/\sigma^2}(\xi) = e^{-\frac{\xi^2}{2\sigma^2}} = \sqrt{2\pi\sigma^2}\varphi_{\sigma^2}(\xi)$ so that the right-hand side may be expressed as $\sqrt{2\pi\sigma^2}$ times the pdf of $\Phi_{\sigma^2} * P$. In particular, one has

$$\frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x}\hat{P}(\xi)e^{-\frac{\sigma^2\xi^2}{2}}\,d\xi = \int_{\mathbb{R}} \varphi_{\sigma^2}(y-x)P(dy).$$

The right-hand side may be viewed as the pdf of the distribution of the sum of independent random variables $X_{\sigma^2} + Y$ with respective distributions Φ_{σ^2} and P. Also, by the Chebyshev inequality, $X_{\sigma^2} \to 0$ in probability as $\sigma^2 \to 0$. Thus the distribution of $X_\sigma^2 + Y$ converges weakly to P. Equivalently, the pdf of $X_{\sigma^2} + Y$ is given by the expression on the left side, involving P only through $\hat{P}$. In this way $\hat{P}$ uniquely determines P. ■

Remark 6.1. The equation (6.47) may be viewed as a form of *Parseval's relation.*

At this point we have established that the map $P \in \mathcal{P}(\mathbb{R}) \to \hat{P} \in \widehat{\mathcal{P}}(\mathbb{R})$ is one-to-one, and transforms convolution as pointwise multiplication. Some additional basic properties are presented in the exercises. We next consider important special cases of an *inversion formula* for absolutely continuous finite (signed) measures $\mu(dx) = f(x)dx$ on $\mathbb{R}$. This is followed by a basic result on the *continuity* of the map $P \to \hat{P}$ for respectively the weak topology on $\mathcal{P}(\mathbb{R})$ and the topology of pointwise convergence on $\widehat{\mathcal{P}}(\mathbb{R})$, and an identification of the *range* of the Fourier transform of finite positive measures.

It is instructive to consider the Fourier transform as a limiting version of a Fourier series. In particular, if f is differentiable and vanishes outside a finite interval, and if f' is square-integrable, then one may use the Fourier series of f (scaled to be defined on $(-\pi, \pi]$) to obtain (see Exercise 6) the *Fourier inversion formula,*

$$f(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(y)e^{-izy}\,dy. \tag{6.48}$$

Moreover, any f that vanishes outside a finite interval and is square-integrable is automatically integrable, and for such an f one has the *Plancherel identity* (see Exercise 6)

$$\|\hat{f}\|_2^2 := \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2\,d\xi = 2\pi\int_{-\infty}^{\infty} |f(y)|^2\,dy = 2\pi\|f\|_2^2. \tag{6.49}$$

Let us now check that (6.48), (6.49), in fact, hold under the following more general conditions.

Theorem 6.8.

a. If f and $\hat{f}$ are both integrable, then the Fourier inversion formula (6.48) holds.
b. If f is integrable as well as square-integrable, then the Plancherel identity (6.49) holds.

Proof. (a) Let $f, \hat{f}$ be integrable. Assume for simplicity that f is continuous. Note that this assumption is innocuous since the inversion formula yields a continuous (version of) f (see Exercise 7(i) for the steps of the proof without this a priori continuity assumption for f). Let φ_{ε^2} denote the pdf of the Gaussian distribution with mean zero and variance $\varepsilon^2 > 0$. Then writing Z to denote a standard normal random variable,

$$f * \varphi_{\varepsilon^2}(x) = \int_{\mathbb{R}^1} f(x-y)\varphi_{\varepsilon^2}(y)dy = \mathbb{E}f(x - \varepsilon Z) \to f(x), \tag{6.50}$$

as $\varepsilon \to 0$. On the other hand, using the easily verifiable inversion formula for φ_{ε^2} (see Exercise 3),

$$\begin{aligned} f * \varphi_{\varepsilon^2}(x) &= \int_{\mathbb{R}} f(x-y)\varphi_{\varepsilon^2}(y)dy = \int_{\mathbb{R}} f(x-y)\left\{\frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi y}e^{-\varepsilon^2\xi^2/2}d\xi\right\}dy \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\varepsilon^2\xi^2/2}\left\{\int_{\mathbb{R}} e^{i\xi(x-y)}f(x-y)dy\right\}e^{-i\xi x}d\xi \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x}e^{-\varepsilon^2\xi^2/2}\hat{f}(\xi)d\xi \to \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x}\hat{f}(\xi)d\xi \end{aligned} \tag{6.51}$$

as $\varepsilon \to 0$. The inversion formula (6.48) follows from (6.50), (6.51). For part (b) see Exercise 7(ii). ■

The above results and notions may be extended to higher dimensions $\mathbb{R}^k$. The Fourier series of a square-integrable function f on $[-\pi,\pi) \times [-\pi,\pi) \times \cdots \times [-\pi,\pi) = [-\pi,\pi)^k$ is defined by $\sum_v c_v \exp\{iv \cdot x\}$, where the summation is over all *integral vectors* (or *multi-indices*) $v = (v^{(1)}, v^{(2)}, \ldots, v^{(k)})$, each $v^{(i)}$ being an integer. Also, $v \cdot x = \sum_{i=1}^k v^{(i)}x^{(i)}$ is the usual Euclidean inner product on $\mathbb{R}^k$ between two vectors $v = (v^{(1)}, \ldots, v^{(k)})$ and $x = (x^{(1)}, x^{(2)}, \ldots, x^{(k)})$. The *Fourier coefficients* are given by

$$c_v = \frac{1}{(2\pi)^k}\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi} f(x)e^{-iv\cdot x}\,dx. \tag{6.52}$$

The extensions of Theorems 6.1, 6.2, 6.7, 6.8 and Propositions 6.1, 6.6 are fairly obvious. Similarly, the *Fourier transform* of an integrable function (with respect to Lebesgue measure on $\mathbb{R}^k$) f is defined by

$$\hat{f}(\xi) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{i\xi\cdot y}f(y)\,dy \qquad (\xi \in \mathbb{R}^k), \tag{6.53}$$

and the Fourier inversion formula becomes

$$f(z) = \frac{1}{(2\pi)^k} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{f}(\xi) e^{-iz\cdot\xi}\, d\xi, \tag{6.54}$$

which holds when $f(x)$ and $\hat{f}(\xi)$ are integrable. The Plancherel identity (6.49) becomes

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2\, d\xi = (2\pi)^k \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |f(y)|^2\, dy, \tag{6.55}$$

which holds whenever f is integrable and square-integrable, i.e., Theorem 6.8 to $\mathbb{R}^k$. The definitions of the Fourier transform and convolution of finite signed measures on $\mathbb{R}^k$ are as in (6.40) and (6.45) with integrals over $(-\infty, \infty)$ being replaced by integrals over $\mathbb{R}^k$. The proof of the property $(\mu_1 * \mu_2)\hat{} = \hat{\mu}_1\hat{\mu}_2$ is unchanged. The following Parseval relation is easily established by an application of the Fubini–Tonelli theorem and definition of characteristic function.

***Proposition 6.9** (Parseval Relation).* Let μ and ν be probabilities on $\mathbb{R}^k$ with characteristic functions $\hat{\mu}$ and $\hat{\nu}$, respectively. Then

$$\int_{\mathbb{R}^k} \hat{\mu}(x)\nu(dx) = \int_{\mathbb{R}^k} \hat{\nu}(x)\mu(dx).$$

Next we will see that the correspondence $P \mapsto \hat{P}$, on the set of probability measures with the weak topology onto the set of characteristic functions with the topology of pointwise convergence is *continuous*, thus providing a basic tool for obtaining weak convergence of probabilities on the finite-dimensional space $\mathbb{R}^k$.

***Theorem 6.10** (Cramér–Lévy Continuity Theorem).* Let $P_n (n \geq 1)$ be probability measures on $(\mathbb{R}^k, \mathcal{B}^k)$.

- **a.** If P_n converges weakly to P, then $\hat{P}_n(\xi)$ converges to $\hat{P}(\xi)$ for every $\xi \in \mathbb{R}^k$.
- **b.** If for some continuous function φ one has $\hat{P}_n(\xi) \to \varphi(\xi)$ for every ξ, then φ is the characteristic function of a probability P, and P_n converges weakly to P.

Proof. (a) Since $\hat{P}_n(\xi)$, $\hat{P}(\xi)$ are the integrals of the bounded continuous function $\exp\{i\xi\cdot x\}$ with respect to P_n and P, it follows from the definition of weak convergence that $\hat{P}_n(\xi) \to \hat{P}(\xi)$. (b) We will show that $\{P_n : n \geq 1\}$ is tight. First let $k = 1$. For $\delta > 0$ one has

$$\begin{aligned} \frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \hat{P}_n(\xi)) d\xi &= \frac{1}{2\delta} \int_{\mathbb{R}} \left\{ \int_{-\delta}^{\delta} (1 - e^{i\xi x}) d\xi \right\} P_n(dx) \\ &= \frac{1}{2\delta} \int_{\mathbb{R}} \left(2\delta - \xi \frac{\sin(\xi x)}{\xi x} \bigg|_{-\delta}^{\delta} \right) P_n(dx) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\delta}\int_{\mathbb{R}}\left(2\delta - 2\delta\frac{\sin(\delta x)}{\delta x}\right)P_n(dx)\\
&= \int_{\mathbb{R}}\left(1 - \frac{\sin(\delta x)}{\delta x}\right)P_n(dx)\\
&\geq \frac{1}{2}P_n(\{x : |\delta x| \geq 2\}) = \frac{1}{2}P_n\left(\left\{x : |x| \geq \frac{2}{\delta}\right\}\right). \qquad (6.56)
\end{aligned}
$$

Hence, by assumption,

$$
P_n\left(\left\{x : |x| \geq \frac{2}{\delta}\right\}\right) \leq \frac{2}{2\delta}\int_{-\delta}^{\delta}(1 - \hat{P}_n(\xi))d\xi \to \frac{2}{2\delta}\int_{-\delta}^{\delta}(1 - \varphi(\xi))d\xi, \qquad (6.57)
$$

as $n \to \infty$. Since φ is continuous and $\varphi(0) = 1$, given any $\varepsilon > 0$ one may choose $\delta > 0$ such that $(1 - \varphi(\xi)) \leq \varepsilon/4$ for $|\xi| \leq \delta$. Then the limit in (6.57) is no more than $\varepsilon/2$, proving tightness. For $k > 1$, consider the distribution $P_{j,n}$ under P_n of the one-dimensional projections $x = (x_1, \ldots, x_k) \mapsto x_j$ for each $j = 1, \ldots, k$. Then $\hat{P}_{j,n}(\xi_j) = \hat{P}_n(0, \ldots, 0, \xi_j, 0, \ldots, 0) \to \varphi_j(\xi_j) := \varphi(0, \ldots, 0, \xi_j, 0, \ldots, 0)$ for all $\xi_j \in \mathbb{R}^1$. The previous argument shows that $\{P_{j,n} : n \geq 1\}$ is a tight family for each $j = 1, \ldots, k$. Hence there is a $\delta > 0$ such that $P_n(\{x \in \mathbb{R}^k : |x_j| \leq 2/\delta, j = 1, \ldots, k\}) \geq 1 - \sum_{j=1}^{k} P_{j,n}(\{x_j : |x_j| \geq 2/\delta\}) \geq 1 - k\varepsilon/2$ for all sufficiently large n, establishing the desired tightness. By Prohorov's Theorem (Theorem 5.8), there exists a subsequence of $\{P_n\}_{n=1}^{\infty}$, say $\{P_{n_m}\}_{m=1}^{\infty}$, that converges weakly to some probability P. By part (a), $\hat{P}_{n_m}(\xi) \to \hat{P}(\xi)$, so that $\hat{P}(\xi) = \varphi(\xi)$ for all $\xi \in \mathbb{R}^k$. Since the limit characteristic function $\varphi(\xi)$ is the same regardless of the subsequence $\{P_{n_m}\}_{m=1}^{\infty}$, it follows that P_n converges weakly to P as $n \to \infty$. ■

The *law of rare events*, or *Poisson approximation to the binomial distribution*, provides a simple illustration of the Cramér–Lévy continuity theorem 6.10.

Proposition 6.11 *(Law of Rare Events)***.** For each $n \geq 1$, suppose that $X_{n,1}, \ldots, X_{n,n}$ is a sequence of n i.i.d. 0 or 1-valued random variables with $p_n = P(X_{n,k} = 1)$, $q_n = P(X_{n,k} = 0)$, where $\lim_{n\to\infty} np_n = \lambda > 0$, $q_n = 1 - p_n$. Then $Y_n = \sum_{k=1}^{n} X_{n,k}$ converges in distribution to Y, where Y is distributed by the Poisson law

$$
P(Y = m) = \frac{\lambda^m}{m!}e^{-\lambda},
$$

$m = 0, 1, 2, \ldots$.

Proof. Using the basic fact that $\lim_{n\to\infty}(1 + \frac{a_n}{n})^n = e^{\lim_n a_n}$ whenever $\{a_n\}_{n=1}^{\infty}$ is a sequence of complex numbers such that $\lim_n a_n$ exists, one has by independence, and in the limit as $n \to \infty$,

$$\mathbb{E}e^{i\xi Y_n} = \left(q_n + p_n e^{i\xi}\right)^n = \left(1 + \frac{np_n(e^{i\xi} - 1)}{n}\right)^n \to \exp(\lambda(e^{i\xi} - 1)), \quad \xi \in \mathbb{R}.$$

One may simply check that this is the characteristic function of the asserted limiting Poisson distribution. ■

The development of tools for Fourier analysis of probabilities is concluded with an application of the Herglotz theorem (Theorem 6.4) to identify the **range** of the Fourier transform of finite positive measures.

Definition 6.4. A complex-valued function φ on $\mathbb{R}$ is said to be **positive-definite** if for every positive integer N and finite sequences $\{\xi_1, \xi_2, \ldots, \xi_N\} \subset \mathbb{R}$ and $\{z_1, z_2, \ldots, z_N\} \subseteq \mathbb{C}$ (the set of complex numbers), one has

$$\sum_{1 \le j,k \le N} z_j \bar{z}_k \varphi(\xi_j - \xi_k) \ge 0. \tag{6.58}$$

Theorem 6.12 ***(Bochner's Theorem).*** A function φ on $\mathbb{R}$ is the Fourier transform of a finite measure on $\mathbb{R}$ if and only if it is positive-definite and continuous.

Proof. The proof of necessity is entirely analogous to (6.28). It is sufficient to consider the case $\varphi(0) = 1$. For each positive integer N, $c_{j,N} := \varphi(-j2^{-N})$, $j = 0, \pm 1, \pm 2, \ldots$, is positive-definite in the sense of (6.27). Hence, by the Herglotz theorem, there exists a probability γ_N on $[-\pi, \pi)$ such that $c_{j,N} = (2\pi)^{-1} \int_{[-\pi,\pi)} e^{-ijx} \gamma_N(dx)$ for each j. By the change of variable $x \to 2^N x$, one has $\varphi(j2^{-N}) = (2\pi)^{-1} \int_{[-2^N\pi, 2^N\pi)} e^{ij2^{-N}x} \mu_N(dx)$ for some probability $\mu_N(dx)$ on $[-2^N\pi, 2^N\pi)$. The characteristic function $\hat{\mu}_N(\xi) := \int_{\mathbb{R}^1} e^{i\xi x} \mu_N(dx)$ agrees with φ at all dyadic rational points $j2^{-N}$, $j \in \mathbb{Z}$, dense in $\mathbb{R}$. To conclude the proof we note that one may use the continuity of $\varphi(\xi)$ to see that the family of functions $\hat{\mu}_N(\xi)$ is equicontinuous (see Exercise 27). Thus it follows by the Arzelà-Ascoli theorem (Appendix B) that there is a subsequence that converges pointwise to a continuous function g on $\mathbb{R}$. Since g and φ agree on a dense subset of $\mathbb{R}$, it follows that $g = \varphi$. ■

We will illustrate the use of characteristic functions in two probability applications. For the first, let us recall the general random walk on $\mathbb{R}^k$ from Chapter II. A basic consideration in the probabilistic analysis of the long-run behavior of a stochastic evolution involves frequencies of visits to specific states.

Let us consider the random walk $S_n := Z_1 + \cdots + Z_n, n \ge 1$, starting at $S_0 = 0$. The state 0 is said to be **neighborhood recurrent** if for every $\varepsilon > 0$, $P(S_n \in B_\varepsilon \text{ } i.o.) = 1$, where $B_\varepsilon = \{x \in \mathbb{R}^k : |x| < \varepsilon\}$. It will be convenient for the calculations to use the **rectangular norm** $|x| := \max\{|x_j| : j = 1, \ldots, k\}$, for $x = (x_1, \ldots, x_k)$. All finite-dimensional norms being equivalent, there is no loss of generality in this choice.

Observe that if 0 is not neighborhood recurrent, then for some $\varepsilon > 0$, $P(S_n \in B_\varepsilon \; i.o.) < 1$, and therefore by the Hewitt–Savage 0-1 law, $P(S_n \in B_\varepsilon \; i.o.) = 0$. Much more may be obtained with regard to recurrence dichotomies, expected return times, nonrecurrence, etc., which is postponed to a fuller treatment of stochastic processes. However, the following lemma is required for the result given here. As a warm-up, note that by the Borel–Cantelli lemma I, if $\sum_{n=1}^{\infty} P(S_n \in B_\varepsilon) < \infty$ for some $\varepsilon > 0$ then 0 cannot be neighborhood recurrent. In fact one has the following basic result.

Lemma 1 *(Chung–Fuchs).* If 0 is not neighborhood recurrent then for all $\varepsilon > 0$, $\sum_{n=1}^{\infty} P(S_n \in B_\varepsilon) < \infty$.

Proof. The proof is based on establishing the following two calculations:

$$\text{(A)} \qquad \sum_{n=0}^{\infty} P(S_n \in B_\varepsilon) = \infty \Rightarrow P(S_n \in B_{2\varepsilon} \; i.o.) = 1,$$

$$\text{(B)} \qquad \sum_{n=0}^{\infty} P(S_n \in B_\varepsilon) \geq \frac{1}{(2m)^k} \sum_{n=0}^{\infty} P(S_n \in B_{m\varepsilon}), \; m \geq 2.$$

In particular, it will follow that if $\sum_{n=0}^{\infty} P(S_n \in B_\varepsilon) = \infty$ for some $\varepsilon > 0$, then from (B), $\sum_{n=0}^{\infty} P(S_n \in B_{\varepsilon'}) = \infty$ for all $\varepsilon' < \varepsilon$. In view of (A) this would make 0 neighborhood recurrent. To prove (A), let $N_\varepsilon := \mathit{card}\{n \geq 0 : S_n \in B_\varepsilon\}$ count the number of visits to B_ε. Also let $T_\varepsilon := \sup\{n : S_n \in B_\varepsilon\}$ denote the (possibly infinite) time of the last visit to B_ε. To prove (A) we will show that if $\sum_{m=0}^{\infty} P(S_m \in B_\varepsilon) = \infty$, then $P(T_{2\varepsilon} = \infty) = 1$. Let r be an arbitrary positive integer. Notice that for arbitrary $m = 0, 1, 2, \ldots$, the event $A_m := [S_m \in B_\varepsilon, |S_n| \geq \varepsilon \; \forall \; n \geq r + m]$ is disjoint from $\cup_{j=m+r}^{\infty} [S_j \in B_\varepsilon, |S_n| \geq \varepsilon \; \forall \; n \geq r + j]$. Thus it follows that $\mathbf{1}_{A_0} + \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_{r-1}} + \cdots \leq r$. Taking expectations, it follows from Lebesgue's monotone convergence theorem that $\sum_{m=0}^{\infty} P(S_m \in B_\varepsilon, |S_n| \geq \varepsilon \; \forall \; n \geq m + r) \leq r$. Thus, using the Markov property,

$$\begin{aligned} r &\geq \sum_{m=0}^{\infty} P(S_m \in B_\varepsilon, |S_n| \geq \varepsilon \; \forall \; n \geq m + r) \\ &\geq \sum_{m=0}^{\infty} P(S_m \in B_\varepsilon, |S_n - S_m| \geq 2\varepsilon \; \forall \; n \geq m + r) \\ &= \sum_{m=0}^{\infty} P(S_m \in B_\varepsilon) P(|S_n| \geq 2\varepsilon \; \forall \; n \geq r). \end{aligned} \tag{6.59}$$

Assuming $\sum_{m=0}^{\infty} P(S_m \in B_\varepsilon) = \infty$, one must therefore have $P(T_{2\varepsilon} \le r) \le P(|S_n| \ge 2\varepsilon \ \forall \ n \ge r) = 0$. Thus $P(T_{2\varepsilon} < \infty) = 0$. For the proof of (B), let $m \ge 2$ and for $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$, define $\tau_x = \inf\{n \ge 0 : S_n \in R_\varepsilon(x)\}$, where $R_\varepsilon(x) := [0, \varepsilon)^k + x := \{y \in \mathbb{R}^k : 0 \le y_i - x_i < \varepsilon, i = 1, \ldots, k\}$ is the translate of $[0, \varepsilon)^k$ by x, i.e., "square with lower left corner at x of side lengths ε." For arbitrary fixed $x \in \{-m\varepsilon, -(m-1)\varepsilon, \ldots, (m-1)\varepsilon\}^k$,

$$\begin{aligned}
\sum_{n=0}^{\infty} P(S_n \in R_\varepsilon(x)) &= \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} P(S_n \in R_\varepsilon(x), \tau_x = m) \\
&\le \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} P(|S_n - S_m| < \varepsilon, \tau_x = m) \\
&= \sum_{m=0}^{\infty} P(\tau_x = m) \sum_{j=0}^{\infty} P(S_j \in B_\varepsilon) \\
&\le \sum_{j=0}^{\infty} P(S_j \in B_\varepsilon).
\end{aligned}$$

Thus, it now follow that

$$\begin{aligned}
\sum_{n=0}^{\infty} P(S_n \in B_{m\varepsilon}) &\le \sum_{n=0}^{\infty} \sum_{x \in \{-m\varepsilon, -(m-1)\varepsilon, \ldots, (m-1)\varepsilon\}^k} P(S_n \in R_\varepsilon(x)) \\
&= \sum_{x \in \{-m\varepsilon, -(m-1)\varepsilon, \ldots, (m-1)\varepsilon\}^k} \sum_{n=0}^{\infty} P(S_n \in R_\varepsilon(x)) \\
&\le (2m)^k \sum_{n=0}^{\infty} P(S_n \in B_\varepsilon).
\end{aligned}$$

■

We turn now to conditions on the distribution of the displacements for neighborhood recurrence. If, for example, $\mathbb{E}Z_1$ exists and is nonzero, then it follows from the strong law of large numbers that a.s. $|S_n| \to \infty$. The following is a complete characterization of neighborhood recurrence in terms of the distribution of the displacements.

Theorem 6.13 (Chung–Fuchs Recurrence Criterion). Let $Z_1, Z_2, \ldots$ be an i.i.d. sequence of random vectors in $\mathbb{R}^k$ with common distribution Q. Let $\{S_n = Z_1 + \cdots + Z_n : n \ge 1\}$, $S_0 = 0$, be a random walk on $\mathbb{R}^k$ starting at 0. Then 0 is a neighborhood-recurrent state if and only if for every $\varepsilon > 0$,

$$\sup_{0<r<1} \int_{B_\varepsilon} \mathrm{Re}\left(\frac{1}{1 - r\hat{Q}(\xi)}\right) d\xi = \infty.$$

Proof. First observe that the "triangular probability density function" $\hat{f}(\xi) = (1-|\xi|)^+$, $\xi \in \mathbb{R}$, has the characteristic function $f(x) = 2\frac{1-\cos(x)}{x^2}$, $x \in \mathbb{R}$, and therefore, $\frac{1}{2\pi}f(x)$ has characteristic function $\hat{f}(\xi)$ (Exercise 22). One may also check that $f(x) \geq 1/2$ for $|x| \leq 1$ (Exercise 22). Also with $\mathbf{f}(\mathbf{x}) := \prod_{j=1}^k f(x_j)$, $\mathbf{x} = (x_1, \ldots, x_k)$, $\hat{\mathbf{f}}(\xi) := \prod_{j=1}^k \hat{f}(\xi_j)$, $\hat{\mathbf{f}}$ has characteristic function $\mathbf{f}$. In view of Parseval's relation (Proposition 6.9), one may write

$$\int_{\mathbb{R}^k} \mathbf{f}\left(\frac{\mathbf{x}}{\lambda}\right) Q^{*n}(d\mathbf{x}) = \lambda^k \int_{\mathbb{R}^k} \hat{\mathbf{f}}(\lambda\xi)\hat{Q}^n(\xi)d\xi,$$

for any $\lambda > 0$, $n \geq 1$. Using the Fubini–Tonelli theorem one therefore has for $0 < r < 1$ that

$$\int_{\mathbb{R}^k} \mathbf{f}(\frac{\mathbf{x}}{\lambda}) \sum_{n=0}^{\infty} r^n Q^{*n}(d\mathbf{x}) = \lambda^k \int_{\mathbb{R}^k} \frac{\hat{\mathbf{f}}(\lambda\xi)}{1 - r\hat{Q}(\xi)} d\xi.$$

Also, since the integral on the left is real, the right side must also be a real integral. For what follows note that when an indicated integral is real, one may replace the integrand by its respective real part. Suppose that for some $\varepsilon > 0$,

$$\sup_{0<r<1} \int_{B_{\frac{1}{\varepsilon}}} \mathrm{Re}\left(\frac{1}{1 - r\hat{Q}(\xi)}\right) d\xi < \infty.$$

Then, it follows that

$$\begin{aligned}\sum_{n=1}^{\infty} P(S_n \in B_\varepsilon) = \sum_{n=1}^{\infty} Q^{*n}(B_\varepsilon) &\leq 2^k \int_{\mathbb{R}^k} \mathbf{f}(\frac{\mathbf{x}}{\varepsilon}) \sum_{n=0}^{\infty} Q^{*n}(d\mathbf{x}) \\ &\leq 2^k \varepsilon^k \sup_{0<r<1} \int_{\mathbb{R}^k} \frac{\hat{\mathbf{f}}(\varepsilon\xi)}{1 - r\hat{Q}(\xi)} d\xi \\ &\leq 2^k \varepsilon^k \sup_{0<r<1} \int_{B_{\frac{1}{\varepsilon}}} \mathrm{Re}\left(\frac{1}{1 - r\hat{Q}(\xi)}\right) d\xi < \infty.\end{aligned}$$

Thus, in view of of Borel–Cantelli I, 0 cannot be neighborhood recurrent.

For the converse, suppose that 0 is not neighborhood recurrent. Then, by Lemma 1, one must have for any $\varepsilon > 0$ that $\sum_{n=1}^{\infty} Q^{*n}(B_\varepsilon) < \infty$.

Let $\varepsilon > 0$. Then, again using the Parseval relation with $(2\pi)^k \hat{\mathbf{f}}$ as the Fourier transform of $\mathbf{f}$,

$$\sup_{0<r<1} \int_{B_\varepsilon} \mathrm{Re}\left(\frac{1}{1 - r\hat{Q}(\xi)}\right) d\xi \leq 2^k \sup_{0<r<1} \int_{B_\varepsilon} \mathrm{Re}\left(\frac{\mathbf{f}(\frac{\mathbf{x}}{\varepsilon})}{1 - r\hat{Q}(\mathbf{x})}\right) d\mathbf{x}$$

$$\le 2^k(2\pi)^k\varepsilon^k \sup_{0<r<1}\int_{\mathbb{R}^k} \hat{\mathbf{f}}(\varepsilon\mathbf{x})\sum_{n=0}^{\infty} r^n Q^{*n}(dx)$$

$$\le 2^k(2\pi)^k\varepsilon^k \int_{B_{\varepsilon^{-1}}} \hat{\mathbf{f}}(\varepsilon\mathbf{x})\sum_{n=0}^{\infty} Q^{*n}(dx)$$

$$\le (4\varepsilon\pi)^k \sum_{n=1}^{\infty} Q^{*n}(B_{\varepsilon^{-1}}) < \infty. \qquad \blacksquare$$

We now turn to a hallmark application of Theorem 6.10 in probability to prove the celebrated Theorem 6.14 below. First we need an estimate on the error in the Taylor polynomial approximation to the exponential function. The following lemma exploits the special structure of the exponential to obtain two bounds: a "good small x bound" and a "good large x bound", each of which is valid for all x.

***Lemma 2** (Taylor Expansion of Characteristic Function).* Suppose that X is a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$ such that $\mathbb{E}|X|^m < \infty$. Then

$$\left|\mathbb{E}e^{i\xi X} - \sum_{k=0}^{m}\frac{(i\xi)^k}{k!}\mathbb{E}X^k\right| \le \mathbb{E}\min\left\{\frac{|\xi|^{m+1}|X|^{m+1}}{(m+1)!}, 2\frac{|\xi|^m|X|^m}{m!}\right\}, \qquad \xi \in \mathbb{R}.$$

Proof. Let $f_m(x) = e^{ix} - \sum_{j=0}^{m}\frac{(ix)^j}{j!}$. Note that $f_m(x) = i\int_0^x f_{m-1}(y)dy$. Iteration yields a succession of $m-1$ iterated integrals with integrand of modulus $|f_1(y_{m-1})| = |e^{iy_{m-1}} - 1| \le 2$. The iteration of the integrals is therefore at most $2\frac{|x|^m}{m!}$. To obtain the other bound note the following integration by parts identity:

$$\int_0^x (x-y)^m e^{iy}dy = \frac{x^{m+1}}{m+1} + \frac{i}{m+1}\int_0^x (x-y)^{m+1}e^{iy}dy.$$

This defines a recursive formula that by induction leads to the expansion

$$e^{ix} = \sum_{j=0}^{m}\frac{(ix)^j}{j!} + \frac{i^{m+1}}{m!}\int_0^x (x-y)^m e^{iy}dy.$$

For $x \ge 0$, bound the modulus of the integrand by $|x-y|^m \le y^m$ to get the bound on the modulus of the integral term by $\frac{|x|^{m+1}}{(m+1)!}$. Similarly for $x < 0$. Since both bounds hold for all x, the smaller of the two also holds for all x. Now replace x by $|\xi X|$ and take expected values to complete the proof. $\blacksquare$

***Theorem 6.14** (The Classical Central Limit Theorem).* Let $\mathbf{X}_n, n \ge 1$, be i.i.d. k-dimensional random vectors with (common) mean μ and a finite covariance matrix

D. Then the distribution of $(\mathbf{X}_1 + \cdots + \mathbf{X}_n - n\mu)/\sqrt{n}$ converges weakly to Φ_D, the normal distribution on $\mathbb{R}^k$ with mean zero and covariance matrix D.

Proof. It is enough to prove the result for $\mu = \mathbf{0}$ and $D = I$, the $k \times k$ identity matrix I, since the general result then follows by an affine linear (and hence continuous) transformation. First consider the case $k = 1$, $\{X_n : n \geq 1\}$ i.i.d. $\mathbb{E}X_n = 0$, $\mathbb{E}X_n^2 = 1$. Let φ denote the (common) characteristic function of X_n. Then the characteristic function, say φ_n, of $(X_1 + \cdots + X_n)/\sqrt{n}$ is given at a fixed ξ by

$$\varphi_n(\xi) = \varphi^n(\xi/\sqrt{n}) = \left(1 - \frac{\xi^2}{2n} + o\left(\frac{1}{n}\right)\right)^n, \tag{6.60}$$

where $no(\frac{1}{n}) = o(1) \to 0$ as $n \to \infty$. The limit of (6.60) is $e^{-\frac{\xi^2}{2}}$, the characteristic function of the standard normal distribution, which proves the theorem for the case $k = 1$, using Theorem 6.10(b).

For $k > 1$, let $\mathbf{X}_n, n \geq 1$, be i.i.d. with mean zero and covariance matrix I. Then for each fixed $\xi \in \mathbb{R}^k$, $\xi \neq \mathbf{0}$, $Y_n = \xi \cdot \mathbf{X}_n$, $n \geq 1$, defines an i.i.d. sequence of real-valued random variables with mean zero and variance $\sigma_\xi^2 = \xi \cdot \xi$. Hence by the preceding, $Z_n := (Y_1 + \cdots + Y_n)/\sqrt{n}$ converges in distribution to the one-dimensional normal distribution with mean zero and variance $\xi \cdot \xi$, so that the characteristic function of Z_n converges to the function $\eta \mapsto \exp\{-(\xi \cdot \xi)\eta^2/2\}$, $\eta \in \mathbb{R}$. In particular, at $\eta = 1$, the characteristic function of Z_n is

$$\mathbb{E}e^{i\xi\cdot(\mathbf{X}_1+\cdots+\mathbf{X}_n)/\sqrt{n}} \to e^{-\xi\cdot\xi/2}. \tag{6.61}$$

Since (6.61) holds for every $\xi \in \mathbb{R}^k$, the proof is complete by the Cramér–Lévy continuity theorem. ■

EXERCISES

Exercise Set VI

1. Prove that given $f \in L^2[-\pi,\pi]$ and $\varepsilon > 0$, there exists a continuous function g on $[-\pi,\pi]$ such that $g(-\pi) = g(\pi)$ and $\|f - g\| < \varepsilon$, where $\|\|$ is the L^2-norm defined by (6.10). [*Hint*: By Proposition 2.6 in Appendix A, there exists a continuous function h on $[-\pi,\pi]$ such that $\|f - h\| < \frac{\varepsilon}{2}$. If $h(-\pi) \neq h(\pi)$, modify it on $[\pi - \delta, \pi]$ by a linear interpolation with a value $h(\pi - \delta)$ at $\pi - \delta$ and a value $h(-\pi)$ at π, where $\delta > 0$ is suitably small.]

2. (a) Prove that if $\mathbb{E}|X|^r < \infty$ for some positive integer r, then the characteristic function $\varphi(\xi)$ of X has a continuous rth order derivative $\varphi^{(r)}(\xi) = i^r \int_{\mathbb{R}} x^r e^{i\xi x} P_X(dx)$, where P_X is the distribution of X. In particular, $\varphi^{(r)}(0) = i^r \mathbb{E}X^r$. (b) Prove (6.39) assuming that f and $f^{(j)}$, $1 \leq j \leq r$, are integrable. [*Hint*: Prove (6.38) and use induction.] (c) If $r \geq 2$ in (b), prove that $\hat{f}$ is integrable.

3. This exercise concerns the *normal* (or *Gaussian*) distribution.

(i) Prove that for every $\sigma \neq 0$, $\varphi_{\sigma^2,\mu}(x) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$, is a probability density function (pdf). The probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with this pdf is called the *normal* (or *Gaussian*) *distribution with mean* μ *variance* σ^2, denoted by $\Phi_{\sigma^2,\mu}$. [*Hint*: Let $c = \int_{-\infty}^{\infty} e^{-x^2/2}dx$. Then $c^2 = \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}dxdy = \int_0^\infty \int_0^{2\pi} re^{-r^2/2}d\theta dr = 2\pi$.]

(ii) Show that $\int_{-\infty}^{\infty} x\varphi_{\sigma^2,\mu}(x)dx = \mu$, $\int_{-\infty}^{\infty}(x-\mu)^2\varphi_{\sigma^2,\mu}(x)dx = \sigma^2$. [*Hint*: $\int_{-\infty}^{\infty}(x-\mu)\varphi_{\sigma^2,\mu}(x)dx = 0$, $\int_{-\infty}^{\infty} x^2e^{-x^2/2}dx = 2\int_0^\infty x(-de^{-x^2/2}) = 2\int_0^\infty e^{-x^2/2}dx = \sqrt{2\pi}$.]

(iii) Write $\varphi = \varphi_{1,0}$, the *standard normal density*. Show that its odd-order moments vanish and the even-order moments are given by $\mu_{2n} = \int_{-\infty}^{\infty} x^{2n}\varphi(x)dx = (2n-1)\cdot(2n-3)\cdots 3\cdot 1$ for $n = 1, 2, \ldots$. [*Hint*: Use integration by parts to prove the recursive relation $\mu_{2n} = (2n-1)\mu_{2n-2}, n = 1, 2\ldots$, with $\mu_0 = 1$.]

(iv) Show that $\hat{\Phi}_{\sigma^2,\mu}(\xi) = e^{i\xi\mu-\sigma^2\xi^2/2}$, $\hat{\varphi}(\xi) = e^{-\xi^2/2}$. [*Hint*: $\hat{\varphi}(\xi) = \int_{-\infty}^{\infty}(\cos(\xi x))$ $\varphi(x)dx$. Expand $\cos(\xi x)$ in a power series and integrate term by term using (iii).]

(v) (*Fourier Inversion for* $\varphi_{\sigma^2} \equiv \varphi_{\sigma^2,0}$). Show that $\varphi_{\sigma^2}(x) = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{-i\xi x}\hat{\varphi}_{\sigma^2}$ $(\xi)d\xi$. [*Hint*: $\hat{\varphi}_{\sigma^2}(\xi) = \sqrt{\frac{2\pi}{\sigma^2}}\varphi_{\frac{1}{\sigma^2}}(\xi)$. Now use (iv).]

(vi) Let $\mathbf{Z} = (Z_1, \ldots, Z_k)$ be a random vector where $Z_1, Z_2, \ldots, Z_k$ are i.i.d. random variables with standard normal density φ. Then $\mathbf{Z}$ is said to have the *k-dimensional standard normal distribution.* Its pdf (with respect to Lebesgue measure on $\mathbb{R}^k$) is $\varphi_I(\mathbf{x}) = \varphi(x_1)\cdots\varphi(x_k) = (2\pi)^{-\frac{k}{2}}e^{-\frac{|x|^2}{2}}$, for $\mathbf{x} = (x_1, \ldots, x_k)$. If Σ is a $k\times k$ positive-definite symmetric matrix and $\mu \in \mathbb{R}^k$, then the *normal* (or *Gaussian*) *distribution* $\Phi_{\Sigma,\mu}$ *with mean* μ *and dispersion* (or *covariance*) *matrix* Σ has pdf $\varphi_{\Sigma,\mu}(x) = (2\pi)^{-\frac{k}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x-\mu)\cdot\Sigma^{-1}(x-\mu)\}$, where $\cdot$ denotes the inner (dot) product on $\mathbb{R}^k$. (a) Show that $\hat{\varphi}_{\Sigma,\mu}(\xi) = \exp\{i\xi\cdot\mu - \frac{1}{2}\xi\cdot\Sigma\xi\}$, $\xi \in \mathbb{R}^k$. (Customary abuse of notation identifies the characteristic function of the distribution with the characteristic function of the pdf). (b) If A is a $k\times k$ matrix such that $AA' = \Sigma$, show that for standard normal $\mathbf{Z}$, $A\mathbf{Z}+\mu$ has the distribution $\Phi_{\Sigma,\mu}$. (c) Prove the *inversion formula* $\varphi_{\Sigma,\mu}(x) = (2\pi)^{-k}\int_{\mathbb{R}^k}\hat{\varphi}_{\Sigma,\mu}(\xi)e^{-i\xi\cdot x}d\xi$, $x \in \mathbb{R}^k$.

(vii) Show that if $(X_1, \ldots, X_k)$ has a k-dimensional Gaussian distribution, then $\{X_1, \ldots, X_k\}$ is a collection of independent random variables if and only if they are uncorrelated.

4. Suppose that $\{P_n\}_{n=1}^{\infty}$ is a sequence of Gaussian probability distributions on $(\mathbb{R}^k, \mathcal{B}^k)$ with respective mean vectors $m^{(n)} = (m_1^{(n)}, \ldots, m_k^{(n)})$ and variance–covariance matrices $\Gamma^{(n)} = ((\gamma_{i,j}^{(n)}))_{1\le i,j\le k}$. (i) Show that if $m^{(n)} \to m$ and $\Gamma^{(n)} \to \Gamma$ (componentwise) as $n \to \infty$, then $P_n \Rightarrow P$, where P is Gaussian with mean vector m and variance–covariance matrix Γ. [*Hint*: Apply the continuity theorem for characteristic functions. Note that in the case of nonsingular Γ one may apply Scheffé's theorem, or apply Fatou's lemma to $P_n(G), G$ open.] (ii) Show that if $P_n \Rightarrow P$, then P must be Gaussian. [*Hint*: Consider the case $k = 1$, $m_n = 0, \sigma_n^2 = \int_{\mathbb{R}} x^2P_n(dx)$. Use the continuity theorem and observe that if σ_n^2 $(n \ge 1)$ is unbounded, then $\hat{P}_n(\xi) \equiv \exp\{-\frac{\sigma_n^2}{2}\xi^2\}$ does not converge to a continuous limit at $\xi = 0$.]

5. (*Change of Location/Scale/Orientation*) Let $\mathbf{X}$ be a k-dimensional random vector and compute the characteristic function of $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, where A is a $k \times k$ matrix and $\mathbf{b} \in \mathbb{R}^k$.

6. (*Fourier Transform, Fourier Series, Inversion, and Plancherel*) Suppose f is differentiable and vanishes outside a finite interval, and f' is square-integrable. Derive the inversion formula (6.48) by justifying the following steps. Define $g_N(x) := f(Nx)$, vanishing outside $(-\pi, \pi)$. Let $\sum c_{n,N} e^{inx}$, $\sum c^{(1)}_{n,N} e^{inx}$ be the Fourier series of g_N and its derivative g'_N, respectively.
   (i) Show that $c_{n,N} = \frac{1}{2N\pi}\hat{f}\left(-\frac{n}{N}\right)$.
   (ii) Show that $\sum_{n=-\infty}^{\infty} |c_{n,N}| \le \frac{1}{2\pi}\left|\int_{-\pi}^{\pi} g_N(x)\,dx\right| + A\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |g'_N(x)|^2\,dx\right)^{1/2} < \infty$, where $A = (2\sum_{n=1}^{\infty} n^{-2})^{1/2}$. [*Hint*: Split off $|c_{0,N}|$ and apply Cauchy–Schwarz inequality to $\sum_{n\neq 0} \frac{1}{|n|}(|nc_{n,N}|)$. Also note that $|c^{(1)}_{n,N}|^2 = |nc_{n,N}|^2$.]
   (iii) Show that for all sufficiently large N, the following convergence is uniform: $f(z) = g_N\left(\frac{z}{N}\right) = \sum_{n=-\infty}^{\infty} c_{n,N} e^{inz/N} = \sum_{n=-\infty}^{\infty} \frac{1}{2N\pi}\hat{f}\left(-\frac{n}{N}\right)e^{inz/N}$.
   (iv) Show that (6.48) follows by letting $N \to \infty$ in the previous step if $\hat{f} \in L^1(\mathbb{R}, dx)$.
   (v) Show that for any f that vanishes outside a finite interval and is square-integrable, hence integrable, one has, for all sufficiently large N, $\frac{1}{N}\sum_{n=-\infty}^{\infty}\left|\hat{f}\left(\frac{n}{N}\right)\right|^2 = 2\pi\int_{-\infty}^{\infty}|f(y)|^2\,dy$. [*Hint*: Check that $\frac{1}{2\pi}\int_{-\pi}^{\pi}|g_N(x)|^2\,dx = \frac{1}{2N\pi}\int_{-\infty}^{\infty}|f(y)|^2\,dy$, and $\frac{1}{2\pi}\int_{-\pi}^{\pi}|g_N(x)|^2\,dx = \sum_{n=-\infty}^{\infty}|c_{n,N}|^2 = \frac{1}{4N^2\pi^2}\sum_{n=-\infty}^{\infty}\left|\hat{f}\left(\frac{n}{N}\right)\right|^2$.]
   (vi) Show that the Plancherel identity (6.49) follows in the limit as $N \to \infty$, in (v).

7. (*(i) Inversion Formula*) Prove (6.48) assuming only that f, $\hat{f}$ are integrable. [*Hint*: *Step 1.* Continuous functions with compact support are dense in $L^1 \equiv L^1(\mathbb{R}, dx)$. *Step 2.* Show that *translation* $y \to g(\cdot + y)(\equiv g(x+y), x \in \mathbb{R})$, is continuous on $\mathbb{R}$ into L^1, for any $g \in L^1$. For this, given $\delta > 0$, find continuous h with compact support such that $\|g - h\|_1 < \delta/3$. Then find $\varepsilon > 0$ such that $\|h(\cdot + y) - h(\cdot + y')\|_1 < \delta/3$ if $|y - y'| < \varepsilon$. Then use $\|g(\cdot+y) - g(\cdot+y')\|_1 \le \|g(\cdot+y) - h(\cdot+y)\|_1 + \|h(\cdot+y) - h(\cdot+y')\|_1 + \|h(\cdot+y') - g(\cdot+y')\|_1 < \delta$, noting that the Lebesgue integral (measure) is translation invariant. *Step 3.* Use Step 2 to prove that $\mathbb{E}f(x+\varepsilon Z) \to f(x)$ in L^1 as $\varepsilon \to 0$, where Z is standard normal. *Step 4.* Use (6.51), which does not require f to be continuous, and Step 3, to show that the limit in (6.51) is equal a.e. to f.] (ii) *(Plancherel Identity).* Let $f \in L^1 \cap L^2$. Prove (6.49). [*Hint*: Let $\tilde{f}(x) := \overline{f(-x)}$, $g = f * \tilde{f}$. Then $g \in L^1$, $|g(x)| \le \|f\|_2^2$, $g(0) = \|f\|_2^2$. Also $g(x) = \langle f_x, f\rangle$, where $f_x(y) = f(x+y)$. Since $x \to f_x$ is continuous on $\mathbb{R}$ into L^2 (using arguments similar to those in Step 2 of part (i) above), and $\langle , \rangle$ is continuous on $L^2 \times L^2$ into $\mathbb{R}$, g is continuous on $\mathbb{R}$. Apply the inversion formula (in part (i)) to get $\|f\|_2^2 = g(0) = \frac{1}{2\pi}\int \hat{g}(\xi)d\xi \equiv \frac{1}{2\pi}\int |\hat{f}(\xi)|^2 d\xi$.]

8. (*Smoothing Property of Convolution*) (a) Suppose μ, ν are probabilities on $\mathbb{R}^k$, with ν absolutely continuous with pdf f; $\nu(dx) = f(x)dx$. Show that $\mu * \nu$ is absolutely continuous and calculate its pdf. (b) If $f, g \in L^1(\mathbb{R}^k, dx)$ and if g is bounded and continuous, show that $f * g$ is continuous. (c) If $f, g \in L^1(\mathbb{R}^k, dx)$, and if g and its first r derivatives $g^{(j)}$, $j = 1, \ldots, r$ are bounded and continuous, show that $f * g$ is r times continuously differentiable. [*Hint*: Use induction.]

9. Suppose $f, \hat{f}$ are integrable on $(\mathbb{R}, dx)$. Show $\hat{\hat{f}}(x) = 2\pi f(-x)$.

10. Let $Q(dx) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)dx$ be the uniform distribution on $[-1, 1]$.
    (i) Find the characteristic functions of Q and $Q^{*2} \equiv Q * Q$.
    (ii) Show that the probability with pdf $c\sin^2 x/x^2$, for appropriate normalizing constant c, has a characteristic function with compact support and compute this characteristic function. [*Hint*: Use Fourier inversion for $f = \hat{Q}^2$.]

11. Derive the multidimensional extension of the Fourier inversion formula.

12. Show that:
   (i) The Cauchy distribution with pdf $(\pi(1+x^2))^{-1}$, $x \in \mathbb{R}$, has characteristic function $e^{-|\xi|}$.
   (ii) The characteristic function of the double-sided exponential distribution $\frac{1}{2}e^{-|x|}dx$ is $(1+\xi^2)^{-1}$. [*Hint*: Use integration by parts twice to show that $\int_{-\infty}^{\infty} e^{i\xi x}(\frac{1}{2}e^{-|x|})dx \equiv \int_0^{\infty} e^{-x}\cos(\xi x)dx = (1+\xi^2)^{-1}$.]

13. (i) Give an example of a pair of *dependent* random variables X, Y such that the distribution of their sum is the convolution of their distributions. [*Hint*: Consider the Cauchy distribution with $X = Y$.] (ii) Give an example of a non-Gaussian bivariate distribution such that the marginals are Gaussian.

14. Show that if φ is the characeristic function of a probability then φ must be uniformly continuous on $\mathbb{R}$.

15. (*Symmetric Distributions*) (i) Show that the characteristic function of $\mathbf{X}$ is real-valued if and only if $\mathbf{X}$ and $-\mathbf{X}$ have the same distribution. (ii) A *symmetrization* of (the distribution of) a random variable $\mathbf{X}$ may be defined by (the distribution of) $\mathbf{X}-\mathbf{X}'$, where $\mathbf{X}'$ is an independent copy of $\mathbf{X}$, i.e., independent of $\mathbf{X}$ and having the same distribution as $\mathbf{X}$. Express symmetrization of a random variable in terms of its characteristic function.

16. (*Multidimensional Gaussian characterization*) Suppose that $\mathbf{X} = (X_1, \ldots, X_k)$ is a k-dimensional random vector having a positive pdf $f(x_1, \ldots, x_k)$ on $\mathbb{R}^k (k \geq 2)$. Assume that (a) f is differentiable, (b) $X_1, \ldots, X_k$ are independent, and (c) have an *isotropic density*, i.e. $f(x_1, \ldots, x_k)$ is a function of $\|x\|^2 = x_1^2 + \cdots + x_k^2, (x_1, \ldots, x_k) \in \mathbb{R}^k$. Show that $X_1, \ldots, X_k$ are i.i.d. normal with mean zero and common variance. [*Hint*: Let f_j denote the marginal pdf of X_j and argue that $\frac{f_j'}{2x_j f_j}$ must be a constant.]

17. (i) Show that the functions $\{e_\xi : \xi \in \mathbb{R}^k\}$ defined by $e_\xi(x) := \exp(i\boldsymbol{\xi} \cdot \mathbf{x})$, $\mathbf{x} \in \mathbb{R}^k$ constitute a measure-determining class for probabilities on $(\mathbb{R}^k, \mathcal{B}^k)$. [*Hint*: Given two probabilities P, Q for which the integrals of the indicated functions agree, construct a sequence by $P_n = P \ \forall\ n = 1, 2, \ldots$ whose characteristic functions will obviously converge to that of Q.]
   (ii) Show that the closed half-spaces of $\mathbb{R}^k$ defined by $F_a := \{x \in \mathbb{R}^k : x_j \leq a_j, 1 \leq j \leq k\}$, $a = (a_1, \ldots, a_k)$ constitute a measure-determining collection of Borel subsets of $\mathbb{R}^k$. [*Hint*: Use a trick similar to that above.]

18. Compute the distribution with characteristic function $\varphi(\xi) = \cos(\xi), \xi \in \mathbb{R}^1$.

19. (*Fourier Inversion for Lattice Random Variables*)
   (i) Let $p_j, j \in \mathbb{Z}$, be a probability mass function (pmf) of a probability distribution Q on the integer lattice $\mathbb{Z}$. Show that the Fourier transform $\hat{Q}$ is periodic with period 2π, and derive the inversion formula $p_j = (2\pi)^{-1} \int_{(-\pi,\pi]} e^{-ij\xi}\hat{Q}(\xi)d\xi$. (ii) Let Q be a *lattice distribution of span* $h > 0$ i.e., for some a_0, $Q(\{a_0 + jh : j = 0, \pm 1, \pm 2, \ldots\}) = 1$. Show that $\hat{Q}$ is periodic with period $2\pi/h$ and write down an inversion formula. (iii) Extend (i), (ii) to the multidimensional lattice distributions with $\mathbb{Z}^k$ in place of $\mathbb{Z}$.

20. (*Parseval's Relation*) (i) Let $f, g, \in L^2([-\pi, \pi))$, with Fourier coefficients $\{c_n\}, \{d_n\}$, respectively. Prove that $\sum_n c_n \overline{d}_n = \frac{1}{2\pi}\int_{(-\pi,\pi]} f(x)\overline{g}(x)dx \equiv \langle f, g\rangle$. (ii) Let $f, g \in$

$L^2(\mathbb{R}^k, dx)$ with Fourier transforms $\hat{f}, \hat{g}$. Prove that $\langle \hat{f}, \hat{g}\rangle = 2\pi\langle f, g\rangle$. [*Hint*: Use (a) the Plancherel identity and (b) the *polar identity* $4\langle f, g\rangle = \|f+g\|^2 - \|f-g\|^2$.]

21. (i) Let φ be continuous and positive-definite on $\mathbb{R}$ in the sense of Bochner, and $\varphi(0) = 1$. Show that the sequence $\{c_j \equiv \varphi(j) : j \in \mathbb{Z}\}$ is positive-definite in the sense of Herglotz (6.27). (ii) Show that there exist distinct probability measures on $\mathbb{R}$ whose characteristic functions agree at all integer points.

22. Show that the "triangular function" $\hat{f}(\xi) = (1 - |\xi|)^+$ is the characteristic function of $f(x) = \frac{1}{\pi}\frac{1-\cos(x)}{x^2}$, $x \in \mathbb{R}$.[*Hint*: Consider the characteristic function of the convolution of two uniform distributions on $[-1/2, 1/2]$ and Fourier inversion.] Also show that $1 - \cos(x) \geq x^2/4$ for $|x| \leq \pi/3$. [*Hint*: Use $\cos(y) \geq 1/2$ for $0 < y < \pi/3$.]

23. Show that if $\int_{B_\varepsilon} \mathrm{Re}\left(\frac{1}{1-\hat{Q}(\xi)}\right) d\xi = \infty$ for $\varepsilon > 0$, then the random walk with displacement distribution is neighborhood-recurrent.[1] [*Hint*: Pass to the limit as $r \to 1$ in $0 \leq \mathrm{Re}\left(\frac{1}{1-r\hat{Q}(\xi)}\right)$, using the Chung–Fuchs criterion]

24. (*Chung–Fuchs*) For the one-dimensional random walk show that if $\frac{S_n}{n} \to 0$ in probability as $n \to \infty$, i.e., WLLN holds, then 0 is neighborhood recurrent. [*Hint*: Using the lemma for the proof of Chung–Fuchs, for any positive integer m and $\delta, \varepsilon > 0$, $\sum_{n=0}^{\infty} P(S_n \in B_\varepsilon) \geq \frac{1}{2m}\sum_{n=0}^{\infty} P(S_n \in B_{m\varepsilon}) \geq \frac{1}{2m}\sum_{n=0}^{m\delta^{-1}} P(S_n \in B_{\delta\varepsilon})$, using monotonicity of $r \to P(S_n \in B_r)$. Let $m \to \infty$ to obtain for the indicated Cesàro average, using $\lim_{n\to\infty} P(S_n \in B_{\delta\varepsilon}) = 1$ from the WLLN hypothesis, that $\sum_{n=0}^{\infty} P(S_n \in B_\varepsilon) \geq \frac{1}{2\delta}$. Let $\delta \to 0$ and apply the Lemma 1.]

25. Show that 0 is neighborhood recurrent for the random walk if and only if $\sum_{n=0}^{\infty} P(S_n \in B_1) = \infty$.

26. Prove that the set of trigonometric polynomials is dense in $L^2([-\pi, \pi), \mu)$, where μ is a finite measure on $[-\pi, \pi)$.

27. (*An Equicontinuity Lemma*)
    (i) Let $\varphi_N, N \geq 1$, be a sequence of characteristic functions of probabilities μ_N. Show that if the sequence is equicontinuous at $\xi = 0$ then it is equicontinuous at all $\xi \in \mathbb{R}$. [*Hint*: Use the Cauchy–Schwarz inequality to check that $|\varphi_N(\xi) - \varphi_N(\xi + \eta)|^2 \leq 2|\varphi_N(0) - \mathrm{Re}\,\varphi_N(\eta)|$.]
    (ii) In the notation of the proof of Bochner's theorem, let μ_N be the probability on $[-2^N\pi, 2^N\pi]$ with characteristic function $\varphi_N = \hat{\mu}_N$, where $\varphi_N(\xi) = \varphi(\xi)$ for $\xi = j2^{-N}, j \in \mathbb{Z}$. (a) Show that for $h \in [-1, 1]$, $0 \leq 1 - \mathrm{Re}\,\varphi_N(h2^{-N}) \leq 1 - \mathrm{Re}\,\varphi(2^{-N})$. [*Hint*: Write the formula and simply note that $1 - cos(hx) \leq 1 - \cos(x)$ for $-\pi \leq x \leq \pi$, $0 \leq h \leq 1$.] (b) Show that φ_N is equicontinuous at 0, and hence at all points of $\mathbb{R}$ (by (i)). [*Hint*: Given $\varepsilon > 0$ find $\delta > 0, (0 < \delta < 1)$ such that $|1 - \varphi(\theta)| < \varepsilon$ for all $|\theta| < \delta$. Express each such θ as $\theta = (h_N + k_N)2^{-N}$, where $k_N = [2^N\theta]$ is the integer part of $2^N\theta$, and $h_N = 2^N\theta - [2^N\theta] \in [-1, 1]$. Use the inequality $|a + b|^2 \leq 2|a|^2 + 2|b|^2$ together with the inequality in the hint for (i) to check that

[1] That the converse is also true was independently established in Stone, C. J. (1969): On the potential operator for one-dimensional recurrent random walks, *Trans. AMS* , **136** 427–445, and Ornstein, D. (1969): Random walks, *Trans. AMS*, **138**, 1–60.

$|1-\varphi_N(\theta)|^2 = |1-\varphi_N((h_N+k_N)2^{-N})|^2 \le 2|1-\varphi(k_N 2^{-N})|^2 + 4|1-\operatorname{Re}\varphi(2^{-N})| \le 2\varepsilon^2 + 4\varepsilon$.]

28. Establish the formula $\int_{\mathbb{R}} g(x)\mu * \nu(dx) = \int_{\mathbb{R}}\int_{\mathbb{R}} g(x+y)\mu(dx)\nu(dy)$ for any bounded measurable function g, and finite measures μ, ν.

CHAPTER VII

Classical Central Limit Theorems

In view of the great importance of the *central limit theorem* (CLT) we shall give a general but self-contained version due to Lindeberg. This version is applicable to nonidentically distributed summands and provides the foundation to the following **CLT paradigm**, which permeates the sciences: "The sum of a large number of 'small' independent random terms is approximately normally distributed."

The following simple lemma is easily checked by an integration by parts left as Exercise 1.

Lemma 1 *(A Second-Order Taylor Expansion).* Let f be a real-valued function of $\mathbb{R}$ such that f, f', f'', f''' are bounded. Then for $x, h \in \mathbb{R}$,

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + h^2 \int_0^1 (1-\theta)\{f''(x+\theta h) - f''(x)\}\, d\theta.$$

Theorem 7.1 *(Lindeberg's CLT).* For each n, let $X_{n,1}, \ldots, X_{n,k_n}$ be an independent array of random variables satisfying

$$\mathbb{E}X_{n,j} = 0, \qquad \sigma_{n,j} := (\mathbb{E}X_{n,j}^2)^{1/2} < \infty, \qquad \sum_{j=1}^{k_n} \sigma_{n,j}^2 = 1, \tag{7.1}$$

and, for each $\varepsilon > 0$,

$$\text{(Lindeberg condition)} \qquad \lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}(X_{n,j}^2 \mathbf{1}_{[|X_{n,j}|>\varepsilon]}) = 0. \tag{7.2}$$

Then $\sum_{j=1}^{k_n} X_{n,j}$ converges in distribution to the standard normal law $N(0,1)$.

Proof. Let $\{Z_j : j \geq 1\}$ be a sequence of i.i.d. $N(0,1)$ random variables, independent of $\{X_{n,j} : 1 \leq j \leq k_n\}$. Write

$$Z_{n,j} := \sigma_{n,j} Z_j \qquad (1 \leq j \leq k_n), \tag{7.3}$$

so that $\mathbb{E}Z_{n,j} = 0 = \mathbb{E}X_{n,j}, \mathbb{E}Z_{n,j}^2 = \sigma_{n,j}^2 = \mathbb{E}X_{n,j}^2$. Define

$$\begin{aligned} U_{n,m} &:= \sum_{j=1}^{m} X_{n,j} + \sum_{j=m+1}^{k_n} Z_{n.j} \qquad (1 \leq m \leq k_n - 1), \\ U_{n,0} &:= \sum_{j=1}^{k_n} Z_{n,j}, \qquad U_{n,k_n} := \sum_{j=1}^{k_n} X_{n,j}, \\ V_{n,m} &:= U_{n,m} - X_{n,m} \qquad (1 \leq m \leq k_n). \end{aligned} \tag{7.4}$$

Let f be a real-valued function of $\mathbb{R}$ such that f, f', f'', f''' are bounded. Taking $x = V_{n,m}$, $h = X_{n,m}$ in the Taylor expansion Lemma 1, one has

$$\begin{aligned} \mathbb{E}f(U_{n,m}) \equiv \mathbb{E}f(V_{n,m} + X_{n,m}) &= \mathbb{E}f(V_{n,m}) + \mathbb{E}(X_{n,m} f'(V_{n,m})) \\ &\quad + \tfrac{1}{2}\mathbb{E}(X_{n,m}^2 f''(V_{n,m})) + \mathbb{E}(R_{n,m}), \end{aligned} \tag{7.5}$$

where

$$R_{n,m} := X_{n,m}^2 \int_0^1 (1-\theta)\{f''(V_{n,m} + \theta X_{n,m}) - f''(V_{n,m})\}\, d\theta. \tag{7.6}$$

Since $X_{n,m}$ and $V_{n,m}$ are independent, and $\mathbb{E}X_{n,m} = 0$, $\mathbb{E}X_{n,m}^2 = \sigma_{n,m}^2$, (7.5) reduces to

$$\mathbb{E}f(U_{n,m}) = \mathbb{E}f(V_{n,m}) + \frac{\sigma_{n,m}^2}{2}\mathbb{E}f''(V_{n,m}) + \mathbb{E}(R_{n,m}). \tag{7.7}$$

Also $U_{n,m-1} = V_{n,m} + Z_{n,m}$, and $V_{n,m}$ and $Z_{n,m}$ are independent. Therefore, exactly as above one gets, using $\mathbb{E}Z_{n,m} = 0$, $\mathbb{E}Z_{n,m}^2 = \sigma_{n,m}^2$,

$$\mathbb{E}f(U_{n,m-1}) = \mathbb{E}f(V_{n,m}) + \frac{\sigma_{n,m}^2}{2}\mathbb{E}f''(V_{n,m}) + \mathbb{E}R'_{n,m}), \tag{7.8}$$

where

$$R'_{n,m} := Z_{n,m}^2 \int_0^1 (1-\theta)\{f''(V_{n,m}+\theta Z_{n,m}) - f''(V_{n,m})\}\, d\theta. \tag{7.9}$$

Hence,

$$|\mathbb{E}f(U_{n,m}) - \mathbb{E}f(U_{n,m-1})| \le \mathbb{E}|R_{n,m}| + \mathbb{E}|R'_{n,m}| \qquad (1 \le m \le k_n). \tag{7.10}$$

Now, given an arbitrary $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{E}|R_{n,m}| &= \mathbb{E}(|R_{n,m}|\mathbf{1}_{[|X_{n,m}|>\varepsilon]}) + \mathbb{E}(|R_{n,m}|\mathbf{1}_{[X_{n,m}|\le\varepsilon]}) \\
&\le \mathbb{E}\left[X_{n,m}^2 \mathbf{1}_{[|X_{n,m}|>\varepsilon]} \int_0^1 (1-\theta) 2\|f''\|_\infty \, d\theta\right] \\
&\quad + \mathbb{E}\left[X_{n,m}^2 \mathbf{1}_{[|X_{n,m}|\le\varepsilon]} \int_0^1 (1-\theta)|X_{n,m}|\|f'''\|_\infty \, d\theta\right] \\
&\le \|f''\|_\infty \mathbb{E}(X_{n,m}^2 \mathbf{1}_{[|X_{n,m}|>\varepsilon]}) + \tfrac{1}{2}\varepsilon\sigma_{n,m}^2 \|f'''\|_\infty. \qquad (7.11)
\end{aligned}$$

We have used the notation $\|g\|_\infty := \sup\{|g(x)| : x \in \mathbb{R}\}$. By (7.1), (7.2), and (7.11),

$$\lim \sum_{m=1}^{k_n} \mathbb{E}|R_{n,m}| \le \tfrac{1}{2}\varepsilon\|f'''\|_\infty.$$

Since $\varepsilon > 0$ is arbitrary,

$$\lim \sum_{m=1}^{k_n} \mathbb{E}|R_{n,m}| = 0. \tag{7.12}$$

Also,

$$\begin{aligned}
\mathbb{E}|R'_{n,m}| &\le \mathbb{E}\left[Z_{n,m}^2 \int_0^1 (1-\theta)\|f'''\|_\infty |Z_{n,m}|\, d\theta\right] = \frac{1}{2}\|f'''\|_\infty \mathbb{E}|Z_{n,m}|^3 \\
&= \tfrac{1}{2}\|f'''\|_\infty \sigma_{n,m}^3 \mathbb{E}|Z_1|^3 \le c\sigma_{m,n}^3 \le c\left(\max_{1\le m\le k_n} \sigma_{m,n}\right)\sigma_{n,m}^2, \qquad (7.13)
\end{aligned}$$

where $c = \frac{1}{2}\|f'''\|_\infty \mathbb{E}|Z_1|^3$. Now, for each $\delta > 0$,

$$\sigma_{n,m}^2 = \mathbb{E}(X_{n,m}^2 \mathbf{1}_{[|X_{n,m}|>\delta]}) + \mathbb{E}(X_{n,m}^2 \mathbf{1}_{[X_{n,m}|\le\delta]}) \le \mathbb{E}(X_{n,m}^2 \mathbf{1}_{[|X_{n,m}|>\delta]}) + \delta^2,$$

which implies that

$$\max_{1\le m\le k_n} \sigma_{n,m}^2 \le \sum_{m=1}^{k_n} \mathbb{E}(X_{n,m}^2 \mathbf{1}_{[|X_{n,m}|>\delta]}) + \delta^2.$$

Therefore, by (7.2),

$$\max_{1\le m\le k_n} \sigma_{n,m} \to 0 \qquad \text{as } n\to\infty. \tag{7.14}$$

From (7.13) and (7.14) one gets

$$\sum_{m=1}^{k_n} \mathbb{E}|R'_{n,m}| \le c\left(\max_{1\le m\le k_n} \sigma_{n,m}\right) \to 0 \qquad \text{as } n\to\infty. \tag{7.15}$$

Combining (7.12) and (7.15), one finally gets, on telescoping the difference between (7.7) and (7.8),

$$|\mathbb{E}f(U_{n,k_n}) - \mathbb{E}f(U_{n,0})| \le \sum_{m=1}^{k_n} (\mathbb{E}|R_{n,m}| + \mathbb{E}|R'_{n,m}|) \to 0 \qquad \text{as } n\to\infty. \tag{7.16}$$

But $U_{n,0}$ is a standard normal random variable. Hence,

$$\mathbb{E}f\left(\sum_{j=1}^{k_n} X_{n,j}\right) - \int_{\mathbb{R}} f(y)(2\pi)^{-1/2}\exp\{-\tfrac{1}{2}y^2\}\,dy \to 0 \qquad \text{as } n\to\infty.$$

By Theorem 5.3, the proof is complete. ■

It has been shown by Feller[1] that in the presence of the **uniform asymptotic negligibility** condition (7.14), the Lindeberg condition is also *necessary* for the CLT to hold.

Corollary 7.2 (The Classical CLT). Let $\{X_j : j \ge 1\}$ be i.i.d. $EX_j = \mu$, $0 < \sigma^2 := \operatorname{Var} X_j < \infty$. Then $\sum_{j=1}^{n}(X_j - \mu)/(\sigma\sqrt{n})$ converges in distribution to $N(0,1)$.

Proof. Let $X_{n,j} = (X_j - \mu)/(\sigma\sqrt{n})$, $k_n = n$, and apply Theorem 7.1. ■

Remark 7.1. Note that the case $k_n = n$ corresponds to an exact **triangular array** of random variables. The general framework of the Lindeberg CLT is referred to as a triangular array as well.

[1] Billingsley, P. (1968), *Convergence of probability measures*, Wiley, NY., p. 373.

Corollary 7.3 (Lyapounov's CLT). For each n let $X_{1,n}$, $X_{2,n}, \ldots, X_{n,k_n}$ be k_n independent random variables such that

$$\sum_{j=1}^{k_n} \mathbb{E}X_{n,j} = \mu, \qquad \sum_{j=1}^{k_n} \mathrm{Var}X_{n,j} = \sigma^2 > 0,$$

$$\text{(Lyapounov condition)} \qquad \lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}|X_{n,j} - \mathbb{E}X_{n,j}|^{2+\delta} = 0 \tag{7.17}$$

for some $\delta > 0$. Then $\sum_{j=1}^{k_n} X_{n,j}$ converges in distribution to the Gaussian law with mean μ and variance σ^2.

Proof. By normalizing one may assume, without loss of generality, that

$$\mathbb{E}X_{n,j} = 0, \qquad \sum_{j=1}^{k_n} \mathbb{E}X_{n,j}^2 = 1.$$

It then remains to show that the hypothesis of the corollary implies the Lindeberg condition (7.2). This is true, since for every $\varepsilon > 0$,

$$\sum_{j=1}^{k_n} \mathbb{E}(X_{n,j}^2 \mathbf{1}[X_{n,j} > \varepsilon]) \leq \sum_{j=1}^{k_n} \mathbb{E}\frac{|X_{n,j}|^{2+\delta}}{\varepsilon^{\delta}} \to 0 \tag{7.18}$$

as $n \to \infty$, by (7.17). ■

Observe that the most crucial property of the normal distribution used in the proof of Theorem 7.1 is that the sum of independent normal random variables is normal. In fact, the normal distribution $N(0,1)$ may be realized as the distribution of the sum of independent normal random variables having zero means and variances σ_i^2 for any arbitrarily specified set of nonnegative numbers σ_i^2 adding up to 1; a form of *infinite divisibility*[2] of the normal distribution.

Definition 7.1. A probability Q on $(\mathbb{R}^k, \mathcal{B}^k)$ is said to be **infinitely divisible** if for each integer $n \geq 1$ there is a probability Q_n such that $Q = Q_n^{*n}$.

Another well-known distribution possessing infinite divisibility properties is the **Poisson distribution**, as well as the **stable laws** defined in Exercises 3 and 6.

The following multidimensional version of Corollary 7.2 was proved by the method of characteristic functions in Chapter VI.

[2] Infinitely divisible distributions are naturally associated with stochastic processes having independent increments. This connection is thoroughly developed in our companion text on stochastic processes.

***Theorem 7.4** (Multivariate Classical CLT).* Let $\{\mathbf{X}_n : n = 1, 2, \ldots\}$ be a sequence of i.i.d. random vectors with values in $\mathbb{R}^k$. Let $\mathbb{E}\mathbf{X}_1 = \mu \in \mathbb{R}^k$ (defined componentwise) and assume that the dispersion matrix (i.e., variance–covariance matrix) $\mathbf{D}$ of $\mathbf{X}_1$ is nonsingular. Then as $n \to \infty$, $n^{-1/2}(\mathbf{X}_1 + \cdots + \mathbf{X}_n - n\mu)$ converges in distribution to the Gaussian probability measure with mean zero and dispersion matrix $\mathbf{D}$.

Proof. For each $\xi \in \mathbb{R}^k \backslash \{0\}$ apply Corollary 7.2, to the sequence $\xi \cdot \mathbf{X}_n$, $n \geq 1$. Then use the Cramér–Lévy continuity theorem (Theorem 6.10) (see Exercise 11). ■

EXERCISES

Exercise Set VII

1. Give a proof of Lemma 1. [*Hint*: Use integration by parts.]

2. Define a one-dimensional normal distribution with mean μ and variance $\sigma^2 = 0$ to be δ_μ, the Dirac measure concentrated at μ. For dimensions $k > 1$, given $\mu \in \mathbb{R}^k$, and a nonnegative-definite (possibly singular) $k \times k$ matrix D, a k-dimensional normal distribution $\Phi_{D,\mu}$ is defined to be the distribution of $\mu + \sqrt{D}\mathbf{Z}$, where $\mathbf{Z}$ is k-dimensional standard normal, and $\sqrt{D}$ denotes a nonnegative definite symmetric matrix such that $\sqrt{D}\sqrt{D} = D$. Extend Corollary 7.3 and Theorem 7.4 to versions with such possible limits.

3. A nondegenerate distribution Q on $\mathbb{R}$, i.e. $Q \neq \delta_{\{c\}}$, is said to be *stable* if for every integer n there is a centering constant c_n and a scaling *index* $\alpha > 0$ such that $n^{-\frac{1}{\alpha}}(X_1 + \cdots + X_n - c_n)$ has distribution Q whenever $X_j, j \geq 1$, are i.i.d. with distribution Q. Show that the normal distribution and Cauchy distribution are both stable with respective indices $\alpha = 2$ and $\alpha = 1$.

4. Show that a stable law Q is infinitely divisible.

5. Show that if Q is a stable distribution symmetric about 0 with exponent α, then $c_n = 0$ and $0 < \alpha \leq 2$. [*Hint*: $\hat{Q}(\xi)$ must be real by symmetry, and positivity follows from the case $n = 2$ in the definition.]

6. (*One-dimensional Holtzmark Problem*) Consider $2n$ points (eg., masses or charges) $X_1, \ldots, X_{2n}$ independently and uniformly distributed within an interval $[-n/\rho, n/\rho]$ so that the density of points is the constant $\rho > 0$. Suppose that there is a fixed point (mass, charge) at the origin that exerts an *inverse rth power force* on the randomly distributed points, where $r > 1/2$. That is, the force exerted by the point at the origin on a mass at location x is $csgn(x)|x|^{-r}$ for a positive constant c. Let $F_n = c\sum_{j=1}^{2n} \frac{sgn(X_j)}{|X_j|^r}$ denote the total force exerted by the origin on the $2n$ points. (a) Calculate the characteristic function of the limit distribution Q_r of F_n as $n \to \infty$. [*Hint*: Take $c = 1$ without loss of generality. For $\xi > 0$, $\mathbb{E}e^{i\xi F_n} = \left(\mathbb{E}\cos(\frac{\xi sgn(X_1)}{|X_1|^r})\right)^{2n} = \left(1 - \frac{\rho\xi^{\frac{1}{r}}}{nr}\int_{\xi(\frac{\rho}{n})^r}^{\infty}(1-\cos(y))y^{-\frac{r+1}{r}}dy\right)^{2n}$ (after a change of variable). Use the fact that $|1 - \cos(y)| \leq 2$ to investigate integrability on $[1, \infty)$ and $\frac{1-\cos(y)}{y^2} \to \frac{1}{2}$ as $y \downarrow 0$ to investigate integrability on $(0, 1)$.] (b) Show that Q_r is a stable distribution with index $\alpha = \frac{1}{r} \in (0, 2)$.

7. (*Random Walk with Symmetric Stable Displacements: Transience/Recurrence*) Consider a one-dimensional random walk with symmetric stable displacement distribution[3] Q having characteristic function $\hat{Q}(\xi) = e^{-|\xi|^\alpha}$. Show that 0 is neighborhood recurrent for $1 \leq \alpha \leq 2$ and not neighborhood recurrent for $0 < \alpha < 1$. [*Hint*: Use the Chung–Fuchs theorem from the previous section.]

8. Suppose that $\{X_j\}_{j=1}^{\infty}, \ldots$ is a sequence of independent random variables respectively distributed uniformly on $[-j, j]$, $j \geq 1$. Show that for a suitable choice of scaling constants c_n, the rescaled sum $c_n^{-1}(X_1 + \cdots + X_n)$ is asymptotically normal with mean 0 and variance 1 as $n \to \infty$.

9. $\{X_j\}_{j=1}^{\infty}, \ldots$ is a sequence of independent random variables uniformly bounded by $M > 0$. Assume $\sigma_n^2 = \sum_{k=1}^n \mathrm{Var}(X_k) \to \infty$ as $n \to \infty$. . Show that the central limit theorem holds under suitable centering and scaling.

10. Suppose that $\{X_m\}_{m=1}^{\infty}, \ldots$ is a sequence of independent random variables respectively distributed as $P(X_m = 1) = P(X_m = -1) = p_m, P(X_m = 0) = 1 - 2p_m$, $m \geq 1$, where $\sum_{m=1}^{\infty} p_m = \infty$. Use each of the methods of (a) Lindeberg, (b) Lyapounov, and (c) characteristic functions to give a proof that for a suitable choice of scaling constants c_n, the rescaled sum $c_n^{-1}(X_1 + \cdots + X_n)$ is asymptotically normal with mean 0 and variance 1 as $n \to \infty$.

11. (*Cramér–Wold Device*) Show that a sequence of k-dimensional random vectors $\mathbf{X}_n (n \geq 1)$ converge in distribution to (the distribution of a random vector) $\mathbf{X}$ if and only if all linear functions $\mathbf{c} \cdot \mathbf{X} \equiv \sum_{j=1}^k c_j X_n^{(j)}$ converge in distribution to $\mathbf{c} \cdot \mathbf{X}$ for all $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^k$. [*Hint*: Use the Cramér- Lévy continuity theorem of Chapter VI.]

[3] This exercise begs for a word of caution, since the tails of displacement distributions in a recurrent random walk can be arbitrarily large; see Shepp, L.A. (1964): Recurrent random walks may take arbitrarily large steps, *Bull. AMS* **70**, 540–542, and/or Grey, D.R. (1989): Persistent random walks may have arbitrarily large tails, *Adv. Appld. Probab.* **21**, 229–230.

CHAPTER VIII

Laplace Transforms and Tauberian Theorem

Like the Fourier transform, the Laplace transform of a measure μ has a number of useful operational properties pertaining to moment generation, convolutions, and vague convergence. However, the main point of this chapter is to show that if μ is concentrated on a half-line, say $[0,\infty)$, then its Laplace transform can also be useful for obtaining the asymptotic behavior of $\mu[0,x]$ as $x \to \infty$.

Definition 8.1. Let μ be a measure on $[0,\infty)$. The **Laplace transform** $\hat{\mu}(\lambda)$ of μ is the real-valued function defined for $\lambda \geq c$ by

$$\hat{\mu}(\lambda) := \int_0^\infty e^{-\lambda x}\mu(dx), \quad \lambda > c, \tag{8.1}$$

where $c = \inf\{\lambda : \int_0^\infty e^{-\lambda x}\mu(dx) < \infty\}$.

Notice that by monotonicity of $e^{-\lambda x}$, $x \geq 0$, as a function of λ, the finiteness of the integral defining $\hat{\mu}(a)$ implies finiteness of $\int_0^\infty e^{-\lambda x}\mu(dx)$ for all $\lambda \geq a$. If μ is a finite measure, then $\hat{\mu}(\lambda)$ is defined at least for all $\lambda \geq 0$. On the other hand, one may also wish to view $\hat{\mu}(\lambda)$ as an extended real-valued, i.e., possibly infinite-valued, function defined for all $\lambda \in \mathbb{R}$, which is easy to do since the integrand is nonnegative. However, in general, the statement that the Laplace transform $\hat{\mu}(\lambda)$ exists is intended to mean that the defining integral is finite on some half-line.

***Remark 8.1** (Special Cases and Terminology).* In the case that μ is absolutely continuous, say $\mu(dx) = g(x)dx$, then $\hat{g}(\lambda) := \hat{\mu}(\lambda)$ is also referred to as the Laplace transform of the (Radon–Nikodym derivative) function g. Also, if $\mu = P \circ X^{-1}$ is the distribution of a nonnegative random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$, then $\hat{\mu}(\lambda)$ is also referred to as the Laplace transform of X,

$$\hat{\mu}(\lambda) = \mathbb{E}e^{-\lambda X}.$$

In the case that μ is a probability, the function $\hat{\mu}(-\lambda)$ is the **moment-generating function**.

Although the Laplace transform is an analytic tool, the theory to be developed is largely based on the probabilistic ideas already introduced in previous sections. This is made possible by the **exponential size-bias** transformation introduced in the treatment of large deviations, although in terms of the moment-generating function of a probability. Specifically, if μ is a measure on $[0, \infty)$ such that $\hat{\mu}(c) < \infty$ for some c, then one obtains a probability μ_c on $[0, \infty)$ by

$$\mu_c(dx) = \frac{1}{\hat{\mu}(c)} e^{-cx} \mu(dx). \tag{8.2}$$

Observe also that

$$\hat{\mu}_c(\lambda) = \frac{\hat{\mu}(c + \lambda)}{\hat{\mu}(c)}. \tag{8.3}$$

Just as with the Fourier transform one has the following basic operational calculus.

Proposition 8.1 ***(Moment Generation).*** If $\hat{\mu}$ exists on $(0, \infty)$, then $\hat{\mu}(\lambda)$ has derivatives of all orders $m = 1, 2, \ldots$ given by

$$\frac{d^m}{d\lambda^m}\hat{\mu}(\lambda) = (-1)^m \int_0^\infty x^m e^{-\lambda x} \mu(dx), \quad \lambda > 0.$$

In particular, μ has an mth order finite moment if and only if $\frac{d^m}{d\lambda^m}\hat{\mu}(0^+)$ exists and is finite.

Proof. For the first derivative one has for arbitrary $\lambda > 0$,

$$\lim_{h \to 0} \frac{\hat{\mu}(\lambda + h) - \hat{\mu}(\lambda)}{h} = \lim_{h \to 0} \int_0^\infty \left(\frac{e^{-hx} - 1}{h} \right) e^{-\lambda x} \mu(dx).$$

Since $|(e^{-hx} - 1)/h| \leq c(\delta)e^{\delta x}$ for some constant $c(\delta)$ if $|h| \leq \delta/2$, where $\lambda - \delta > 0$, the limit may be passed under the integral sign by the dominated convergence theorem. The remainder of the proof of the first assertion follows by induction. For the final assertion, by the monotone convergence theorem,

$$\lim_{\lambda\downarrow 0}\int_0^\infty x^m e^{-\lambda x}\mu(dx) = \int_0^\infty x^m \mu(dx). \qquad \blacksquare$$

The proof of the following property is obvious, but its statement is important enough to record.

Proposition 8.2 *(Scale Change).* Let μ be a measure on $[0,\infty)$ with Laplace transform $\hat{\mu}(\lambda)$ for $\lambda > 0$. Define $\alpha : [0,\infty) \to [0,\infty)$ by $\alpha(x) = ax$, for an $a > 0$. Then one has $\widehat{\mu\circ\alpha^{-1}}(\lambda) = \hat{\mu}(a\lambda)$.

Proposition 8.3 *(Convolution Products).* If μ and ν are measures on $[0,\infty)$ such that $\hat{\mu}(\lambda)$ and $\hat{\nu}(\lambda)$ both exist for $\lambda > 0$, then the convolution $\gamma = \mu * \nu$ has Laplace transform $\hat{\gamma}(\lambda) = \hat{\mu}(\lambda)\hat{\nu}(\lambda)$ for all $\lambda > 0$.

Proof. This is a consequence of the basic formula (Exercise 1)

$$\int_0^\infty g(x)\mu * \nu(dx) = \int_0^\infty\int_0^\infty g(x+y)\mu(dx)\nu(dy)$$

for bounded Borel-measurable functions g, using the nonnegativity and multiplicative property of the exponential function. $\blacksquare$

Theorem 8.4 *(Uniqueness & Inversion Formula).* Let μ, ν be two measures on $[0,\infty)$ such that $\int_0^\infty e^{-cx}\mu(dx) = \int_0^\infty e^{-cx}\nu(dx) < \infty$ for some c and

$$\hat{\mu}(\lambda) = \hat{\nu}(\lambda) < \infty, \qquad \forall\ \lambda \geq c.$$

Then one has $\mu = \nu$. Moreover if $\mu[0,\infty) < \infty$, then one also has the inversion formula

$$\mu[0,x] = \lim_{\lambda\to\infty}\sum_{j\leq\lambda x}\frac{(-\lambda)^j}{j!}\frac{d^j}{d\lambda^j}\hat{\mu}(\lambda)$$

at each continuity point x of the (distribution) function $x \to \mu([0,x])$.

Proof. Assume first that μ and ν are finite measures. In this case a probabilistic proof is made possible by the asserted inversion formula obtained as follows. Without loss of generality, assume that μ and ν are normalized to probabilities. For arbitrary fixed $x, z > 0$, consider the expression $\sum_{j\leq\lambda z}\frac{(-\lambda)^j}{j!}\frac{d^j}{d\lambda^j}\hat{\mu}(\lambda) = \sum_{j\leq\lambda z}\frac{(-1)^j\lambda^j}{j!}\frac{d^j}{d\lambda^j}\hat{\mu}(\lambda)$, along with the expected value

$$P(Y_{\lambda x} \le z) = \mathbb{E}h_z(Y_{\lambda x}) = \sum_{j=0}^{\infty} h_z(\frac{j}{\lambda})\frac{(\lambda x)^j}{j!}e^{-\lambda x},$$

where $Y_{\lambda x}$, $\lambda, x > 0$, is Poisson distributed on the lattice $\{0, 1/\lambda, 2/\lambda, \ldots\}$ with intensity λx, and $h_z(y) = \mathbf{1}_{[0,z]}(y), y \ge 0$. Note that $\mathbb{E}Y_{\lambda x} = x$, and $\mathrm{Var}(Y_{\lambda x}) = \frac{x}{\lambda} \to 0$ as $\lambda \to \infty$. Notice that in general, if $\{\mu_{t,a} : t \ge 0, a \in \mathbb{R}\}$ is a collection of probabilities on $\mathbb{R}$, such that $\mu_{t,a}$ has mean a and variance $\sigma^2(a) \to 0$ as $t \to \infty$, then $\mu_{t,a} \Rightarrow \delta_a$ as $t \to \infty$. In particular,

$$\lim_{\lambda\to\infty} P(Y_{\lambda x} \le z) = \begin{cases} 0, & \text{if } z < x, \\ 1 & \text{if } z > x, \end{cases} \tag{8.4}$$

Now, in view of the moment-generation formula $(-1)^j \frac{d^j}{d\lambda^j}\hat{\mu}(\lambda) = \int_0^\infty x^j e^{-\lambda x}\mu(dx)$, one has

$$\sum_{j\le\lambda z} \frac{(-\lambda)^j}{j!}\frac{d^j}{d\lambda^j}\hat{\mu}(\lambda) = \int_0^\infty P(Y_{\lambda x} \le z)\mu(dx).$$

The inversion formula and hence uniqueness follows in the limit $\lambda \to \infty$ by application of the dominated convergence theorem. The general uniqueness assertion follows by the exponential size-bias transformation. Specifically, since μ_c and ν_c are probabilities whose Laplace transforms agree, one has $\mu_c = \nu_c$. Since $\mu \ll \mu_c = \nu_c$ and $\nu_c \ll \nu$, it follows that $\mu \ll \nu$ and $\frac{d\mu}{d\nu} = \frac{d\mu}{d\mu_c}\frac{d\nu_c}{d\nu} = \frac{\hat{\mu}(c)}{e^{-cx}}\frac{e^{-cx}}{\hat{\nu}(c)} = 1$. ■

Recall from Chapter V that a sequence of measures $\mu_n (n \ge 1)$ on $[0, \infty)$ is said to **converge vaguely** to a measure μ if $\int_{[0,\infty)} g\, d\mu_n \to \int_{[0,\infty)} g\, d\mu$ for all continuous functions g vanishing at infinity, i.e., $g(x) \to 0$ as $x \to \infty$.

Theorem 8.5 (Continuity). Let $\mu_n, n \ge 1$, be a sequence of measures on $[0, \infty)$ with respective Laplace transforms $\hat{\mu}_n, n \ge 1$, defined on a common half-line $\lambda \ge c$.

- **a.** If $\mu_n, n \ge 1$, converges vaguely to μ, and if $\{\hat{\mu}_n(c) : n \ge 1\}$ is a bounded sequence of real numbers, then $\lim_n \hat{\mu}_n(\lambda) = \hat{\mu}(\lambda)$ for all $\lambda > c$. Conversely, if for a sequence of measures $\mu_n (n \ge 1)$, $\hat{\mu}_n(\lambda) \to \varphi(\lambda) > 0\ \forall\ \lambda > c$ as $n \to \infty$, then φ is the Laplace transform of a measure, μ and μ_n converges vaguely to μ.
- **b.** Suppose $c = 0$ in (a), $\varphi(0^+) = 1$, and $\mu_n, n \ge 1$, is a sequence of probabilities. Then μ is a probability and $\mu_n \Rightarrow \mu$ as $n \to \infty$.

Proof. We will prove part (b) first. For this we use the Helly selection principle (Corollary 5.6) to select a weakly convergent subsequence $\{\mu_{n_m} : m \ge 1\}$ to a measure μ with $\mu(\mathbb{R}) \le 1$ on $[0, \infty)$. Since $x \mapsto e^{-\lambda x}$ is continuous and vanishes at infinity on

$[0,\infty)$, $\hat{\mu}_{n_m}(\lambda) \to \hat{\mu}(\lambda)$ as $m \to \infty$ for each $\lambda > 0$. Thus μ is the unique measure on $[0,\infty)$ with Laplace transform φ. In particular, there can be only one (vague) limit point. Since $\varphi(0^+) = 1$ it follows that μ is a probability.

We now turn to part (a). Assume that μ_n, $n \ge 1$, converges vaguely to μ, and first suppose that $\lim_n \hat{\mu}_n(c) = m$ exists. Apply exponential size-biasing to obtain for bounded continuous functions f vanishing at infinity that

$$\lim_{n\to\infty}\int_0^\infty f(x)\frac{e^{-cx}}{\hat{\mu}_n(c)}\mu_n(dx) = \int_0^\infty f(x)\frac{e^{-cx}}{m}\mu(dx) = \int_0^\infty f(x)\mu_c(dx),$$

for some measure μ_c. For $\lambda > c$, take $f(x) = e^{-(\lambda-c)x}, x \ge 0$, to see that $\lim_n \hat{\mu}_n(\lambda) = \hat{\mu}(\lambda), \lambda > c$. Assuming only that $\{\hat{\mu}_n(c) : n \ge 1\}$ is bounded, consider any convergent subsequence $\lim_{n'} \hat{\mu}_{n'}(c) = m'$. Since the limit $\lim_n \hat{\mu}_{n'}(\lambda) = \hat{\mu}(\lambda)$ does not depend on the subsequence, $\hat{\mu}_n(\lambda) \to \hat{\mu}(\lambda)$.

For the converse part of (a) suppose that $\hat{\mu}_n(\lambda) \to \varphi(\lambda)$ for all $\lambda > c$. For any fixed $\lambda' > c$, note that $\frac{\hat{\mu}_n(\lambda+\lambda')}{\hat{\mu}_n(\lambda')}$ is the Laplace transform of the exponentially size-biased probability $\mu'_n(dx) = \frac{1}{\hat{\mu}_n(\lambda')}e^{-\lambda' x}\mu_n(dx)$. By part (b), $\mu'_n, n \ge 1$, converges vaguely to a finite measure μ', and therefore μ_n converges vaguely to $\mu(dx) = \varphi(c)e^{cx}\mu'(dx)$. ■

Definition 8.2. A function φ on $(0,\infty)$ is said to be **completely monotone** if it possesses derivatives of all orders $m = 1, 2, \dots$ on $(0,\infty)$ and $(-1)^m \frac{d^m}{d\lambda^m}\hat{\mu}(\lambda) \ge 0$ for each $\lambda > 0$.

It follows from the moment generation theorem that $\hat{\mu}(\lambda)$ is completely monotone. In fact, we will now see that this property characterizes Laplace transforms of measures on $[0,\infty)$. We preface this with two lemmas characterizing the range of generating functions (combinatorial) originally due to S. Bernstein, while the proofs here are along the lines of those given in Feller.[1]

For a given continuous function g on $[0,1]$, the **Bernstein polynomials** arise naturally in the Weierstrass approximation theorem (see Appendix B) and are defined by

$$B_n(t) = \sum_{k=0}^{n} g(\frac{k}{n})\binom{n}{k}t^k(1-t)^{n-k}, \quad 0 \le t \le 1.$$

Lemma 1 ***(Finite Differences and Bernstein Polynomials).*** The following is an equivalent representation of the Bernstein polynomials for a given continuous function

[1] See Feller, W. (1970).

g on $[0,1]$ in terms of the difference operator $\Delta_h g(t) = \frac{g(t+h)-g(t)}{h}$:

$$B_n(t) = \sum_{k=0}^{n} \binom{n}{k} (\frac{t}{n})^k \Delta_{\frac{1}{n}}^k g(0),$$

where $\Delta_h^1 = \Delta_h, \Delta_h^k = \Delta_h(\Delta_h^{k-1})$, $k \geq 1$, and Δ_h^0 is the identity operator.

Proof. Insert the binomial expansion of $(1-t)^{n-k} = \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^{n-k-j} t^{n-k-j}$ in the definition of $B_n(t)$, to obtain

$$B_n(t) = \sum_{j=0}^{n} \sum_{k=0}^{j} g\left(\frac{k}{n}\right) \binom{n}{k} \binom{n-k}{n-j} (-1)^{j-k} t^j.$$

For any finite or infinite sequence $a_0, a_1, \ldots$ of real numbers, the difference notation $\Delta_1 a_m := a_{m+1} - a_m$ is also used. For notational convenience we simply write $\Delta := \Delta_1$, i.e., $h = 1$. Upon iteration of $\Delta a_m = a_{m+1} - a_m$, one inductively arrives at

$$\Delta^k a_m = \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} a_{m+j}. \tag{8.5}$$

For another sequence $b_0, b_1, \ldots$, multiply this by $\binom{n}{k} b_k$ and sum over $k = 0, \ldots, n$. Then making a change in the order of summation, the coefficient of a_{m+j} may be read off as

$$\begin{aligned} \sum_{k=j}^{n} \binom{n}{k} \binom{k}{j} (-1)^{k-j} b_k &= (-1)^{n-j} \binom{n}{j} \sum_{l=0}^{n-j} \binom{n-j}{l} (-1)^{n-j-l} b_{l+j} \\ &= \binom{n}{j} (-1)^{n-j} \Delta^{n-j} b_j. \end{aligned}$$

The first equality is by a change of order of summation and writing $(-1)^l = (-1)^{n-j}(-1)^{n-j-l}$, and the last equality is by (8.5) applied to the sequence $b_0, b_1, \ldots$. Thus one has the so-called *general reciprocity formula* relating differences $\Delta^k a_m$ and $\Delta^k b_m$ for two arbitrary sequences $\{a_m : m = 0, 1, \ldots\}$ and $\{b_m : m = 0, 1, \ldots\}$ in a "summation by parts" form

$$\sum_{k=0}^{n} b_k \binom{n}{k} \Delta^k a_m = \sum_{j=0}^{n} a_{m+j} \binom{n}{j} (-1)^{n-j} \Delta^{n-j} b_j. \tag{8.6}$$

For $0 < t < 1$, applying (8.6) to $b_m = t^m$ using (8.5), one has $\Delta^k b_m = t^m(1-t)^k(-1)^k$. Thus, applying (8.6) yields the identity

$$\sum_{m=0}^{n} t^m \binom{n}{m} \Delta^m a_k = \sum_{j=0}^{n} a_{k+j} \binom{n}{j} t^j (1-t)^{n-j}. \tag{8.7}$$

Now fix $h = 1/n$ and consider the difference ratios $\Delta_h^m a_k \equiv h^{-m}\Delta^m a_k$ of the sequence $a_k = g(\frac{k}{n})$, $k = 0, 1, \ldots, n$. The asserted difference representation of the Bernstein polynomials for g now follows directly from (8.7). ■

The dual representation of Bernstein polynomials can be used to characterize power series with positive coefficients as follows.

Lemma 2. Let g be a function on $[0, 1)$. Then the following are equivalent: (a) $g(t) = \sum_{n=0}^{\infty} c_n t^n$, $0 \le t < 1$ with $c_n \ge 0$, $\forall n$; (b) $g^{(n)}(t) \equiv \frac{d^n}{dt^n} g(t)$ exists at each $t \in (0, 1)$ and is nonnegative for every $n = 0, 1, 2, \ldots$; (c) $\Delta_{\frac{1}{n}}^k g(0) \ge 0$, for $k = 0, 1, \ldots, n-1$, $n \ge 1$. Such functions g are said to be **absolutely monotone**.

Proof. That $(a) \Rightarrow (b)$ follows from the analyticity of Taylor series expansion and term by term differentiation (see Exercise 7, Chapter IV). Also $(b) \Rightarrow (c)$ since monotonicity of g implies $\Delta_{\frac{1}{n}} g(t) \ge 0$, and monotonicity of g' then implies monotonicity of $\Delta_{\frac{1}{n}} g(t)$, so that $\Delta_h^2 g(t) \ge 0$. Iterating this argument, one arrives at (c) from (b). In fact, $\Delta_{\frac{1}{n}}^n g(0) \ge 0$ as well. For $(c) \Rightarrow (a)$, first consider the case that g satisfies (c) for $k = 0, 1, \ldots, n$ and is continuously defined on the closed interval $[0, 1]$ with $g(1) = 1$. In view of the Weierstrass approximation theorem, the Bernstein polynomials

$$B_n(t) = \sum_{k=0}^{n} g\left(\frac{k}{n}\right) \binom{n}{k} t^k (1-t)^{n-k}, \quad 0 \le t \le 1,$$

converge uniformly to g on $[0, 1]$ as $n \to \infty$ (see Appendix B). From (c) one sees using Lemma 1 that the coefficients $p_{j,n} = \sum_{k=0}^{j} g(\frac{k}{n}) \binom{n}{k} \binom{n-k}{n-j} (-1)^{j-k}$, $j = 0, 1 \ldots, n$, are nonnegative and $\sum_{j=0}^{n} p_{j,n} = B_n(1) = 1$. Thus $B_n(e^{-\lambda})$ is the Laplace transform of the probability μ_n defined by $\{p_{j,n} : j = 0, 1, \ldots, n\}$, i.e., $\mu_n = \sum_{j=0}^{n} p_{j,n} \delta_{\{j\}}$. It follows from the Weierstrass approximation and the continuity theorem for Laplace transforms that there is a probability μ such that $\mu_n \Rightarrow \mu$, and μ has the desired Laplace transform $g(e^{-\lambda}) = \lim_{n\to\infty} B_n(e^{-\lambda})$. Take $\lambda = \log t$ to complete the proof of (a) for the case in which g continuously extends to $[0, 1]$. If $g(1^-) = \infty$, fix an arbitrary $0 < \delta < 1$ and define $g_\delta(t) = \frac{g(\delta t)}{g(\delta)}$, for $0 \le t \le 1$. Then g_δ satisfies (c) and the above proof applied to g_δ yields an expansion (in $s = \delta t$)

$$g(s) = g(\delta) \sum_{n=0}^{\infty} d_n(\delta) s^n, \quad 0 \le s < \delta.$$

By uniqueness of coefficients in a series expansion of $g(s)$ on an interval $[0,\delta)$, the coefficients $c_n = g(\delta)d_n(\delta)$ do not depend on δ, and the expansion (a) is therefore valid on $[0,\delta)$ for δ arbitrarily close to 1, i.e., valid on $[0,1)$. ■

***Theorem 8.6** (Range of Laplace Transforms).* A function φ on $(0,\infty)$ is completely monotone if and only if there is a measure μ on $[0,\infty)$ such that

$$\varphi(\lambda) = \int_0^\infty e^{-\lambda x}\mu(dx), \quad \lambda > 0.$$

In particular, μ is a probability if and only if $\varphi(0^+) = 1$.

Proof. In the case that μ is a finite measure, the necessity of complete monotonicity follows directly from the previous moment-generation formula. For general measures μ on $[0,\infty)$ for which $\hat{\mu}(\lambda)$ exists for $\lambda > 0$, it then follows from exponential size-biasing that $\frac{\hat{\mu}(\lambda+c)}{\hat{\mu}(c)}$ is completely monotone as a function of $\lambda > 0$ for any fixed $c > 0$. Thus, the necessity is proven.

Suppose that φ is a completely monotone function on $(0,\infty)$. For arbitrary fixed $h > 0$, define a measure μ_h by

$$\mu_h = \sum_{n=0}^{\infty} \frac{(-h)^n}{n!}\frac{d^n}{d\lambda^n}\varphi(h)\delta_{\{\frac{n}{h}\}}.$$

Then by linearity of the Laplace transform and the continuity theorem applied to the limit of the partial sums,

$$\hat{\mu}_h(\lambda) = \sum_{n=0}^{\infty} \frac{(-h)^n}{n!}\frac{d^n}{d\lambda^n}\varphi(h)e^{-\lambda\frac{n}{h}}.$$

Since $c_n := \frac{1}{n!}\frac{d^n}{d\lambda^n}\varphi(h(1-t))|_{t=0} = \frac{(-h)^n}{n!}\frac{d^n}{d\lambda^n}\varphi(h) \geq 0$ for each n, it follows from the preceding lemma that $\varphi(h(1-t))$ has the power series expansion

$$\varphi(h(1-t)) := \sum_{n=0}^{\infty} \frac{(-h)^n}{n!}\frac{d^n}{d\lambda^n}\varphi(h)t^n, \quad 0 \leq t < 1 \tag{8.8}$$

(also see Exercise 10). Thus $g_h(\lambda) := \varphi(h(1-e^{-\frac{\lambda}{h}}))$, $\lambda > 0$, is the Laplace transform of μ_h. Since $g_h(\lambda)$ converges to $\varphi(\lambda)$ on $(0,\infty)$ as $h \to \infty$, it follows from the continuity theorem that there exists a measure μ on $[0,\infty)$ having Laplace transform φ. ■

Already the condition that the Laplace transform $\hat{\mu}(\lambda)$ exists at some $\lambda \geq 0$, readily implies that for any bounded interval $J = (a,b)$, $\mu(J) \leq e^{\lambda b}\hat{\mu}(\lambda) < \infty$; finiteness of $\mu(J)$ for all bounded intervals J is referred to as the **Radon property**. As the following theorem illustrates, much more on the asymptotic behavior of μ may be obtained from that of its Laplace transform near zero, and vice versa. For the proofs of results it will be convenient to use the distribution function G_μ of a (Radon) measure μ on $[0, \infty)$, defined by

$$G_\mu(x) = \mu[0, x], \qquad x \geq 0.$$

Theorem 8.7 ***(Karamata Tauberian Theorem).*** Let μ be a measure on $[0, \infty)$ whose Laplace transform exists for $\lambda > 0$. Then for $\theta \geq 0$,

$$\lim_{\alpha \downarrow 0} \frac{\hat{\mu}(\alpha\lambda)}{\hat{\mu}(\alpha)} = \lambda^{-\theta} \quad \text{if and only if} \quad \lim_{a \to \infty} \frac{\mu[0, ax]}{\mu[0, a]} = x^\theta.$$

In particular, either of these implies for $\alpha \downarrow 0$, $a = \alpha^{-1} \to \infty$, that

$$\hat{\mu}(\alpha) \sim \mu[0, a]\Gamma(\theta + 1),$$

where $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$, $r > 0$, is the gamma function.

Proof. Suppose $\lim_{\alpha \downarrow 0} \frac{\hat{\mu}(\alpha\lambda)}{\hat{\mu}(\alpha)} = \lambda^{-\theta}$. Observe that the limit on the left side applies to Laplace transforms of measures μ_a obtained from μ by scale changes of the form $G_{\mu_a}(x) = \frac{G_\mu(ax)}{\hat{\mu}(\alpha)}$, where $a = \alpha^{-1}$. On the other hand, the right side is the Laplace transform $\hat{\gamma}(\lambda) = \lambda^{-\theta}$ of the measure $\gamma(dx) = \frac{1}{\Gamma(\theta)} x^{\theta-1} dx$ on $[0, \infty)$. Thus, by the continuity theorem for Laplace transforms, μ_a converges vaguely to γ as $a \to \infty$. Since γ is absolutely continuous with respect to Lebesgue measure, it follows that the (improper) distribution function converges at all points $x \geq 0$. That is,

$$G_{\mu_a}(x) \to G_\gamma(x) = \frac{1}{\Gamma(\theta + 1)} x^\theta$$

as $a \to \infty$. Take $x = 1$ to get

$$\hat{\mu}(\alpha) \sim G_\mu(a)\Gamma(\theta + 1) = \mu[0, a]\Gamma(\theta + 1).$$

With this it also follows that

$$\lim_{a \to \infty} \frac{\mu[0, ax]}{\mu[0, a]} = x^\theta.$$

For the converse, assume that $\lim_{a \to \infty} \frac{\mu[0, ax]}{\mu[0, a]} = x^\theta$. The Laplace transform of the measure μ_a with distribution function $G_{\mu_a}(x) = \frac{\mu[0, ax]}{\mu[0, a]}$ is $\frac{\hat{\mu}(\alpha\lambda)}{G_\mu(a)}$, and that of $G_\gamma(x) =$

x^θ is $\Gamma(\theta+1)\lambda^{-\theta}$. Thus, in view of the continuity theorem, if one can show that $\frac{\hat{\mu}(\alpha c)}{G_\mu(a)}$ is bounded for some $c > 0$, then it will follow that

$$\frac{\hat{\mu}(\alpha\lambda)}{G_\mu(a)} \to \Gamma(\theta+1)\lambda^{-\theta}$$

as $\alpha \to 0$, $a = \alpha^{-1}$. From here the converse assertions follow as above. So it suffices to prove the boundedness of $\frac{\hat{\mu}(\alpha c)}{G_\mu(a)}$ for some $c > 0$. For this, first observe that the assumption $\lim_{a\to\infty} \frac{\mu[0,ax]}{\mu[0,a]} = x^\theta$ implies that there is a $c > 1$ such that $G_\mu(2x) \leq 2^{\theta+1}G_\mu(x)$ for $x > c$. Thus, with $a = \alpha^{-1}$,

$$\begin{aligned}
\hat{\mu}(\alpha c) \leq \hat{\mu}(\alpha) &= \int_0^a e^{-\alpha x}\mu(dx) + \sum_{n=0}^{\infty} \int_{2^n a}^{2^{n+1}a} e^{-\alpha x}\mu(dx) \\
&\leq G_\mu(a) + \sum_{n=0}^{\infty} e^{-2^n} G_\mu(2^{n+1}a) \\
&\leq G_\mu(a) \left\{ 1 + \sum_{n=0}^{\infty} 2^{(n+1)(\theta+1)} e^{-2^n} \right\},
\end{aligned}$$

for all $a > c > 1$. In particular, this establishes a desired bound to complete the proof. ■

Definition 8.3. A function L on $[0,\infty)$ is said to be **slowly varying at infinity** if for each fixed $x > 0$, one has $\lim_{a\to\infty} \frac{L(ax)}{L(a)} = 1$.

The following corollary is essentially just a reformulation of the statement of the Tauberian theorem. The proof is left as Exercise 6.

Corollary 8.8. For L slowly varying at infinity and $0 \leq \theta < \infty$ one has

$$\hat{\mu}(\lambda) \sim \lambda^{-\theta} L\left(\frac{1}{\lambda}\right) \quad \text{as} \quad \lambda \downarrow 0$$

if and only if

$$\mu[0,x] \sim \frac{1}{\Gamma(\theta+1)} x^\theta L(x) \quad \text{as} \quad x \to \infty.$$

Remark 8.2. It is to be noted that asymptotic relations in the Tauberian theorem are also valid with the roles of α and a reversed, i.e., for $\alpha \to \infty$ and $a \to 0$.

In the case that $\mu(dx) = g(x)dx$ has a density f one may obtain a "differentiated form" of the asymptotic relation under sufficient regularity in g. One such condition[2] is the following:

Definition 8.4. A function g on $[0, \infty)$ is said to be **ultimately monotone** if it is monotone on some $[x_0, \infty)$ for some $x_0 \geq 0$.

***Lemma 3** (Monotone Density Lemma).* Suppose that $\mu(dx) = g(x)dx$ has an ultimately monotone density g. If $G_\mu(x) \sim \frac{1}{\Gamma(\theta+1)} x^\theta L(x)$ as $x \to \infty$, then $g(x) \sim \frac{x^{\theta-1}}{\Gamma(\theta)} L(x) \sim \theta G_\mu(x)/x$ as $x \to \infty$.

Proof. Assume that g is ultimately nondecreasing. Then, for arbitrary $0 < c < d < \infty$, for all x sufficiently large one may bound $\frac{G_\mu(dx) - G_\mu(cx)}{x^\theta L(x)} = \frac{\int_{cx}^{dx} g(y)dy}{x^\theta L(x)}$ above and below by

$$\frac{(d-c)xg(cx)}{x^\theta L(x)} \leq \frac{G_\mu(dx) - G_\mu(cx)}{x^\theta L(x)} \leq \frac{(d-c)xg(dx)}{x^\theta L(x)}.$$

Thus,

$$\begin{aligned}\limsup_{x\to\infty} \frac{g(cx)}{x^{\theta-1}L(x)} &\leq \limsup_{x\to\infty} \frac{G_\mu(dx) - G_\mu(cx)}{(d-c)x^\theta L(x)} \\ &= \limsup_{x\to\infty} \left\{ \frac{G_\mu(dx)}{(dx)^\theta (d-c) L(dx)} d^\theta \frac{L(dx)}{L(x)} - \frac{G_\mu(cx)}{(cx)^\theta (d-c) L(cx)} c^\theta \frac{L(cx)}{L(x)} \right\} \\ &\to \frac{(d^\theta - c^\theta)}{d-c}.\end{aligned}$$

Take $c = 1$ and let $d \downarrow 1$ to get the desired upper bound on the limsup. The same lower bound on the liminf is obtained by the same considerations applied to the other inequality. Finally, the case in which g is nonincreasing follows by the same argument but with reversed estimates for the upper and lower bounds. ■

The Tauberian theorem together with the monotone density lemma immediately yields the following consequence.

[2] A treatment of the problem with less-stringent conditions can be found in the more comprehensive monograph Bingham, N.H., C.M. Goldie, J.L. Teugels (1987).

Corollary 8.9. Suppose that $\mu(dx) = g(x)dx$ has an ultimately monotone density g. For L slowly varying at infinity and $0 \le \theta < \infty$ one has

$$\hat{\mu}(\lambda) \sim \lambda^{-\theta} L\left(\frac{1}{\lambda}\right) \quad \text{as} \quad \lambda \downarrow 0 \quad \text{if and only if} \quad g(x) \sim \frac{1}{\Gamma(\theta)} x^{\theta-1} L(x) \quad \text{as} \quad x \to \infty.$$

Finally, for discrete measures one has the following asymptotic behavior conveniently expressed in terms of *(combinatorial) generating functions*, i.e., with $t = e^{-\lambda}$.

Corollary 8.10. Let $\tilde{\mu}(t) = \sum_{n=0}^{\infty} \mu_n t^n$, $0 \le t < 1$, where $\{\mu_n\}_{n=0}^{\infty}$ is a sequence of nonnegative numbers. For L slowly varying at infinity and $0 \le \theta < \infty$ one has

$$\hat{\mu}(t) \sim (1-t)^{-\theta} L\left(\frac{1}{1-t}\right) \quad \text{as} \quad t \uparrow 1$$

if and only if

$$\sum_{j=0}^{n} \mu_j \sim \frac{1}{\Gamma(\theta)} n^{\theta} L(n) \quad \text{as} \quad n \to \infty.$$

Moreover, if the sequence $\{\mu_n\}_{n=0}^{\infty}$ is ultimately monotone and $0 < \theta < \infty$, then equivalently,

$$\mu_n \sim \frac{1}{\Gamma(\theta)} n^{\theta-1} L(n) \quad \text{as} \quad n \to \infty.$$

Proof. Let $\mu(dx) = \sum_{n=0}^{\infty} \mu_n \mathbf{1}_{[n,n+1)}(x)dx$, with (improper) distribution function G_μ. Then $G_\mu(n) = \sum_{j=0}^{n} \mu_j$. Also

$$\hat{\mu}(\lambda) = \frac{1-e^{-\lambda}}{\lambda} \sum_{n=0}^{\infty} \mu_n e^{-n\lambda} = \frac{1-e^{-\lambda}}{\lambda} \tilde{\mu}(e^{-\lambda}).$$

The assertions now follow immediately from the Tauberian theorem and previous corollary. ■

EXERCISES

Exercise Set VIII

1. Establish the formula $\int_0^\infty g(x)\mu * \nu(dx) = \int_0^\infty \int_0^\infty g(x+y)\mu(dx)\nu(dy)$ for bounded Borel-measurable functions g used in the proof of the convolution property of Laplace transforms.
2. Show that size-biasing a Gaussian distribution corresponds to a shift in the mean.

3. Show that $g(t) = \frac{1}{1-t}$, $0 \le t < 1$, is absolutely monotone and $\varphi(\lambda) = \frac{e^{\lambda}}{e^{\lambda}-1}$, $\lambda > 0$, is completely monotone. Calculate the measure μ with Laplace transform $\varphi(\lambda)$.

4. Show that if g is absolutely monotone on $[0, 1]$ with $g(1) = 1$, then $p_{j,n} = \sum_{k=0}^{j} g\left(\frac{k}{n}\right)\binom{n}{k}\binom{n-k}{n-j}(-1)^{j-k}$ is a probability.

5. Show that (i) $|\log x|^r$, $x > 0$, is slowly varying at infinity and at 0 for any exponent r; (ii) $\log\log x, x > 1$, is slowly varying at ∞; (iii) $(1 + x^{-s})^r$, $x > 0$, is slowly varying at ∞ for any exponents r and $s > 0$.

6. Complete the proofs of the corollaries to the Tauberian theorem. [*Hint*: Note that $\frac{G_\mu(ax)}{G_\mu(a)} \sim x^\theta$ as $a \to \infty$ if and only if $L(x) = \frac{G_\mu(x)}{x^\theta}$ is slowly varying at infinity, and $\frac{\hat{\mu}(\alpha\lambda)}{\hat{\mu}(\alpha)} \sim \lambda^{-\theta}$ as $\alpha \to 0$ if and only if $\lambda^\theta \hat{\mu}(\lambda)$ varies slowly at 0.]

7. (*Renewal Equation Asymptotics*) Let μ be a probability on $[0, \infty)$ not concentrated at $\{0\}$, and suppose g is a nonnegative measurable function on $[0, \infty)$. Show that $u(t) = g * \mu(t) := \int_0^t g(t-s)\mu(ds)$, $t \ge 0$, satisfies the **renewal equation** $u(t) = g(t) + \int_0^t u(t-s)\mu(ds)$, $t \ge 0$. Show that if g is integrable on $[0, \infty)$ and μ has finite first moment m, then $u(t) \sim \{\frac{1}{m}\int_0^\infty g(s)ds\}t$ as $t \to \infty$. [*Hint*: Use the Tauberian theorem.]

8. (*Branching with Geometric Offspring*) Let $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n}$ be a triangular array of i.i.d. random variables having geometric (offspring) distribution $P(Y_{n,j} = k) = qp^k, k = 0, 1, 2 \ldots$. Recursively define $X_{n+1} = \sum_{j=1}^{X_n} Y_{n,j}\mathbf{1}_{[X_n \ge 1]}$, for $n \ge 0$, with $X_0 = 1$. Then X_{n+1} may be viewed as the number of offspring in the $(n + 1)$st generation produced by ancestors in the nth generation. The geometric offspring assumption makes various explicit calculations possible that are otherwise impossible. Let $g_n(t) = \mathbb{E}t^{X_n}$, and $g_1(t) = g(t) = \mathbb{E}t^{Y_{n,j}}$ the generating function of the offspring distribution.
   (i) Show that $g_{n+1}(t) = g(g_n(t))$.
   (ii) For $p \ne q$ show that $g_n(t) = q\frac{p^n-q^n-(p^{n-1}-q^{n-1})pt}{p^{n+1}-q^{n+1}-(p^n-q^n)pt}$.
   (iii) For $p < \frac{1}{2}$, consider the *total progeny* defined by $N = \sum_{n=0}^{\infty} X_n$. Show that $P(N < \infty) = 1$. [*Hint*: Consider $P(X_n = 0) = g_n(0)$ and $[X_n = 0] \subseteq [X_{n+1} = 0]$.]
   (iv) For $p < \frac{1}{2}$, let $h(t) = \mathbb{E}t^N$ be the generating function for the total progeny. Show that $h(t) = tg(h(t)) = \frac{qh(t)}{1-ph(t)}$, $0 < t < 1$. [*Hint*: Consider the generating functions $h_n(t) = \mathbb{E}t^{\sum_{j=0}^{n} X_j}$, $n \ge 0$, in the limit as $n \to \infty$.]
   (v) For $p < \frac{1}{2}$, show that $h(t) = \frac{1-\sqrt{1-4pqt}}{2p}$, $0 < t < 1$. [*Hint*: Solve the quadratic equation implied by the preceding calculation.]
   (vi) Show that $\sum_{k=1}^{n} \frac{k}{(4pq)^k}P(N = k) \sim \frac{1}{p\sqrt{\pi}}n^{\frac{1}{2}}$ as $n \to \infty$. [*Hint*: Apply the Tauberian theorem to $h'(\frac{t}{4pq})$ and use properties of the gamma function: $\Gamma(x + 1) = x\Gamma(x), \Gamma(\frac{1}{2}) = \sqrt{\pi}$.]

9. Show that under the hypothesis of Theorem 8.5, the sequence of probabilities $\{\mu_n : n \ge 1\}$ is tight. [*Hint*: Given $\varepsilon > 0$ there exists $\lambda_\varepsilon > 0$ such that $\hat{\mu}_n(\lambda_\varepsilon) \ge 1 - \frac{\varepsilon}{2}$ for all n. Now find $M = M_\varepsilon$ such that $e^{-\lambda_\varepsilon M} < \frac{\varepsilon}{2}$. Then $\mu_n[0, M] \ge 1 - \varepsilon$ for all n.]

10. Show that the series (8.8) converges uniformly on $[0, a]$ for all $a < 1$. [*Hint*: Check that the series increases monotonically to $\varphi(h(1-t))$ and apply Dini's theorem from advanced calculus.]

11. (i) Show that under the hypothesis of part (b) of Theorem 8.5 the sequence of probabilities $\{\mu_n : n \ge 1\}$ is tight. [*Hint*: Given $\varepsilon > 0$ there exists $\lambda_\varepsilon > 0$ such that $\hat{\mu}_n(\lambda_\varepsilon) \ge 1 - \frac{\varepsilon}{2}$

for all n. Now find $M = M_\varepsilon$ such that $e^{-\lambda_\varepsilon M} < \varepsilon/2$. Then $\mu_n[0, M] \geq 1 - \varepsilon$ for all n.]
(ii) Give an example to show that the boundedness of $\hat{\mu}_n(c)$ is necessary in part (a). [*Hint*: Consider point-mass measures μ_n at positive integers n.]

CHAPTER IX

Random Series of Independent Summands

The convergence of an infinite series $\sum_{n=1}^{\infty} X_n$ is a tail event. Thus, if $X_1, X_2, \ldots$ is a sequence of independent random variables, the convergence takes place with probability one or zero. For a concrete example, consider the so-called random signs question for the divergent series $\sum_{n=1}^{\infty} \frac{1}{n}$. Namely, while $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$ is convergent, one might ask what happens if the signs are assigned by i.i.d. tosses of a balanced coin (see Exercise 1)?

To answer questions about almost-sure convergence of a random series, one often proceeds with an effort to show that the sequence $\{S_n = X_1 + \cdots + X_n : n \geq 1\}$ of partial sums is *not* Cauchy with probability zero. A "non-Cauchy with probability zero" statement may be formulated by first observing that the event that $\{S_n\}_{n=1}^{\infty}$ is not Cauchy implies $\cup_{\varepsilon>0}^{*}[\sup_{j,k\geq n} |S_j - S_k| \geq \varepsilon \ \forall\ n] = \cup_{\varepsilon>0}^{*} \cap_{n=1}^{\infty} [\sup_{j,k\geq n} |S_j - S_k| \geq \varepsilon]$, where $\cup_{\varepsilon>0}^{*}$ denotes a countable union over rational $\varepsilon > 0$. Moreover, by continuity of the probability P from above, $P(\cap_{n=1}^{\infty}[\sup_{j,k\geq n} |S_j - S_k| \geq \varepsilon]) = \lim_{n\to\infty} P(\sup_{j,k\geq n} |S_j - S_k| \geq \varepsilon)$. Now, since $\sup_{j,k\geq n} |S_j - S_k| = \sup_{j,k\geq 0} |S_{n+j} - S_{n+k}| \leq \sup_{j\geq 0} |S_{n+j} - S_n| + \sup_{k\geq 0} |S_{n+k} - S_n| = 2\sup_{m\geq 1} |S_{n+m} - S_n|$, one has

$$
\begin{aligned}
P\left(\sup_{j,k\geq n} |S_j - S_k| \geq \varepsilon\right) &\leq 2P\left(\sup_{m\geq 1} |S_{m+n} - S_n| \geq \frac{\varepsilon}{2}\right) \\
&= 2\lim_{N\to\infty} P\left(\max_{1\leq m\leq N} |S_{n+m} - S_n| \geq \frac{\varepsilon}{2}\right). \qquad (9.1)
\end{aligned}
$$

Thus, to prove non-Cauchy with probability zero it is sufficient to show that

$$\lim_{n,N\to\infty} P\left(\max_{1\le m\le N} |S_{n+m} - S_n| \ge \varepsilon\right) = 0. \tag{9.2}$$

This approach is facilitated by the use of maximal inequalities of the type found previously for martingales. At the cost of some redundancy, here is another statement and derivation of *Kolmogorov's maximal inequality* for sums of independent random variables.

Theorem 9.1 *(Kolmogorov's Maximal Inequality).* Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}X_j = 0$, $\operatorname{Var} X_j < \infty$, for $j = 1, \ldots, n$. Let $S_k = \sum_{j=1}^k X_j$. For $\delta > 0$ one has $P(\max_{1\le k\le n} |S_k| \ge \delta) \le \frac{\operatorname{Var} S_n}{\delta^2}$.

Proof. Let $\tau = \min\{k \le n : |S_k| > \delta\}$, with $\tau = \infty$ if $|S_k| \le \delta$ for all $k \le n$. Then

$$\begin{aligned}
\mathbb{E}\big(S_n^2\big) &\ge \sum_{k=1}^n \mathbb{E}\big(S_n^2 \mathbf{1}_{[\tau=k]}\big) \\
&= \sum_{k=1}^n \mathbb{E}(\{S_k + (S_n - S_k)\}^2 \mathbf{1}_{[\tau=k]}) \qquad (9.3)\\
&= \sum_{k=1}^n \mathbb{E}\big(\{S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2\}\mathbf{1}_{[\tau=k]}\big) \\
&\ge \sum_{k=1}^n \mathbb{E}\{S_k^2 + 2S_k(S_n - S_k)\}\mathbf{1}_{[\tau=k]}. \qquad (9.4)
\end{aligned}$$

Now observe that $[\tau = k] \in \sigma(X_1, \ldots, X_k)$ and S_k is $\sigma(X_1, \ldots, X_k)$-measurable. Thus $\mathbf{1}_{[\tau=k]}S_k$ and $S_n - S_k$ are independent. Since the latter has mean zero, the expected value of their product is zero, and the above bound reduces to

$$\mathbb{E}S_n^2 \ge \sum_{k=1}^n \mathbb{E}\{S_k^2 \mathbf{1}_{[\tau=k]}\} \ge \sum_{k=1}^n \delta^2 P(\tau = k).$$

Noting that $\sum_{k=1}^n P(\tau = k) = P(\max_{1\le k\le n} |S_k| \ge \delta)$ completes the proof. ■

Theorem 9.2 *(Mean-Square-Summability Criterion).* Let $X_1, X_2, \ldots$ be independent random variables with mean zero. If $\sum_{n=1}^\infty \operatorname{Var}(X_n) < \infty$ then $\sum_{n=1}^\infty X_n$ converges a.s.

Proof. Applying Kolmogorov's maximal inequality to the sum of $X_{n+1}, \ldots, X_{n+m}$ yields for arbitrary $\varepsilon > 0$,

$$P\left(\max_{1\le k\le m} |S_{n+k} - S_n| > \varepsilon\right) \le \frac{1}{\varepsilon^2}\sum_{k=1}^{m} \mathrm{Var}(X_{n+k}) \le \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty} \mathrm{Var}(X_{n+k}).$$

Using continuity of the probability P, it follows that $P(\sup_{k\ge1} |S_{n+k} - S_n| > \varepsilon) = \lim_{m\to\infty} P\left(\max_{1\le k\le m} |S_{n+k} - S_n| > \varepsilon\right) \le \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty} \mathrm{Var}(X_{n+k})$. Since the bound is by the tail of a convergent series, one has, letting $n \to \infty$, that

$$\lim_{n\to\infty} P(\sup_{k\ge1} |S_{n+k} - S_n| > \varepsilon) = 0.$$

It follows by the method leading up to (9.2) that the event $[\{S_n\}_{n=1}^{\infty}$ is not a Cauchy sequence] has probability zero. ■

As a quick application of the mean-square-summability criterion one may obtain a strong law of large numbers for sums of independent centered random variables whose variances do not grow too rapidly; see Exercise 3.

We will see below that it can also be employed in a proof of strong laws for rescaled averages of i.i.d. sequences under suitable moment conditions. This will use truncation arguments stemming from the following further consequence; also see Exercises 5, 6.

***Corollary 9.3** (Kolmogorov's Three-Series Criteria: Sufficiency Part).* Let X_1, X_2, ... be independent random variables. Suppose that there is a (truncation level) number $a > 0$ such that the following three series converge: (i) $\sum_{n=1}^{\infty} P(|X_n| > a)$; (ii)$\sum_{n=1}^{\infty} \mathbb{E}(X_n\mathbf{1}_{[|X_n|\le a]})$; (iii)$\sum_{n=1}^{\infty} \mathrm{Var}(X_n\mathbf{1}_{[|X_n|\le a]})$. Then $\sum_{n=1}^{\infty} X_n$ converges with probability one.

Proof. Convergence of (i) implies that the truncated and nontruncated series converge and diverge together, since by Borel–Cantelli I, they differ by at most finitely many terms with probability one. In view of the mean-square-summability criterion, part (iii) gives a.s. convergence of the centered truncated sum, and adding (ii) gives the convergence of the uncentered truncated sum. ■

As an application of the CLT one may also establish the necessity of Kolmogorov's three series criteria as follows.

***Corollary 9.4** (Kolmogorov's Three-Series Criteria: Necessary Part).* Let X_1, X_2, ... be independent random variables. If $\sum_{n=1}^{\infty} X_n$ converges with probability one, then for any (truncation level) number $a > 0$ the following three series converge: (i) $\sum_{n=1}^{\infty} P(|X_n| > a)$; (ii) $\sum_{n=1}^{\infty} \mathbb{E}[X_n\mathbf{1}_{[|X_n|\le a]}]$; (iii) $\sum_{n=1}^{\infty} \mathrm{Var}(X_n\mathbf{1}_{[|X_n|\le a]})$.

Proof. Assume that $\sum_{n=1}^{\infty} X_n$ converges a.s. and let $a > 0$. Necessity of condition (i) follows from Borel–Cantelli II. Let $S_n^{(a)} = \sum_{k=1}^{n} X_k\mathbf{1}_{[|X_k|\le a]}$, and $\sigma_n^2(a) =$

$\mathrm{Var}(S_n^{(a)}), \mu_n(a) = \mathbb{E}S_n^{(a)}$. Suppose for the sake of contradiction of (iii) that $\sigma_n(a) \to \infty$. Then $S_n^{(a)}/\sigma_n(a) \to 0$ a.s. as $n \to \infty$, and hence in probability as well. However, since the terms $X_k\mathbf{1}_{[|X_k|\le a]} - \mathbb{E}\{X_k\mathbf{1}_{[|X_k|\le a]}\}$ are uniformly bounded, one may use the central limit theorem to compute for an arbitrary interval $J = (c, d]$, $c < d$, that

$$P\left(\frac{S_n^{(a)} - \mu_n(a)}{\sigma_n(a)} \in J, \frac{|S_n^{(a)}|}{\sigma_n(a)} < 1\right) \ge P\left(\frac{S_n^{(a)} - \mu_n(a)}{\sigma_n(a)} \in J\right) - P\left(\frac{|S_n^{(a)}|}{\sigma_n(a)} \ge 1\right)$$

is bounded away from zero for all sufficiently large n. This is a contradiction since it implies that for sufficiently large n, the numbers $-\mu_n(a)/\sigma_n(a)$ are between $c-1$ and $d+1$ for two distinct choices of intervals J more than 2 units apart. Thus condition (iii) holds. The necessity of condition (ii) now follows by applying the mean-square-summability criterion, Theorem 9.2, to see that $\sum_{n=1}^{\infty}\{X_n\mathbf{1}_{[|X_n|\le a]} - \mu_n(a)\}$ is a.s. convergent by (ii). Thus $\sum_{n=1}^{\infty}\mu_n(a)$ must converge. ■

In preparation for an extension[1] of the strong law of large numbers, we record here two very basic facts pertaining to the ordinary "calculus of averages"; their proofs are left as Exercise 4.

Lemma 1. Let $\{c_n\}_{n=1}^{\infty}$ be a sequence of positive real numbers such that $c_n \uparrow \infty$ as $n \to \infty$. Let $\{a_n\}_{n=1}^{\infty}$ be an arbitrary sequence of real numbers. (a) If $a_n \to a$ as $n \to \infty$, then defining $c_0 = 0$,

$$[\text{Cesàro}] \qquad \lim_{n\to\infty} \frac{1}{c_n}\sum_{j=1}^{n}(c_j - c_{j-1})a_j = a. \tag{9.5}$$

(b) If $\sum_{j=1}^{\infty}\frac{a_j}{c_j}$ converges then

$$[\mathit{Kronecker}] \qquad \lim_{n\to\infty} \frac{1}{c_n}\sum_{j=1}^{n} a_j = 0. \tag{9.6}$$

Theorem 9.5 ***(Strong Law of Large Numbers).*** Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables, and let $0 < \theta < 2$. Then $n^{-\frac{1}{\theta}}\sum_{j=1}^{n} X_j$ converges a.s. if and only if $\mathbb{E}|X_1|^{\theta} < \infty$ and either (i) $\theta \le 1$ or (ii) $\theta > 1$ and $\mathbb{E}X_1 = 0$. When the limit exists it is $\mathbb{E}X_1$ in the case $\theta = 1$, and is otherwise zero for all other cases of $\theta \in (0, 2)$, $\theta \ne 1$.

[1] Theorem 9.5 is a stronger statement than Kolmogorov's classical strong law due to Marcinkiewicz and Zygmund (1937): Sur les fonctions indépendentes, *Fund. Math.* **29**, 60–90., but clearly contains the classical law as a special case.

Proof. The case $\theta = 1$ was considered in Chapter IV (also see Exercise 6). Fix $\theta \in (0,2)$, $\theta \neq 1$, and assume $\mathbb{E}|X_1|^\theta < \infty$. For the cases in which $\theta > 1$, one has $\mathbb{E}|X_1| < \infty$, but assume for the *sufficiency* part that $\mathbb{E}X_1 = 0$ in such cases. We will show that $n^{-\frac{1}{\theta}} \sum_{j=1}^n X_j \to 0$ a.s. as $n \to \infty$.

The basic idea for the proof is to use "truncation methods" as follows. Let $Y_n = X_n \mathbf{1}_{\left[|X_n| \leq n^{\frac{1}{\theta}}\right]}, n \geq 1$. Then it follows from the identical distribution and moment hypothesis, using Borel-Cantelli lemma I, that $P(Y_n \neq X_n \ i.o) = 0$ since

$$\sum_{n=1}^\infty P(Y_n \neq X_n) = \sum_{n=1}^\infty P(|X_1|^\theta > n) \leq \int_0^\infty P(|X_1|^\theta > x)dx = \mathbb{E}|X_1|^\theta < \infty.$$

Thus, it is sufficient to show that $n^{-\frac{1}{\theta}} \sum_{j=1}^n Y_j$ a.s. converges to zero. In view of Kronecker's lemma, for this one needs only to show that $\sum_{n=1}^\infty \frac{Y_n}{n^{\frac{1}{\theta}}}$ is a.s. convergent. If $\theta < 1$, then this follows by the direct calculation that

$$\begin{aligned}\mathbb{E}\sum_{n=1}^\infty \frac{1}{n^{\frac{1}{\theta}}}|Y_n| &= \sum_{n=1}^\infty n^{-\frac{1}{\theta}}\mathbb{E}|X_n|\mathbf{1}_{[|X_n| \leq n^{\frac{1}{\theta}}]} \\ &\leq \int_0^\infty x^{-\frac{1}{\theta}}\mathbb{E}|X_1|\mathbf{1}_{[|X_1| \leq x^{\frac{1}{\theta}}]}dx \\ &= \mathbb{E}\left\{|X_1| \int_{|X_1|^\theta}^\infty x^{-\frac{1}{\theta}}dx\right\} \leq c\mathbb{E}|X_1|^\theta < \infty,\end{aligned}$$

for a positive constant c. Thus $n^{-\frac{1}{\theta}} \sum_{j=1}^n Y_j$ is a.s. absolutely convergent for $\theta < 1$. For $\theta > 1$, using the three-series theorem, it suffices to check that $\sum_{n=1}^\infty \mathbb{E}\frac{Y_n}{n^{\frac{1}{\theta}}}$ is convergent, and $\sum_{n=1}^\infty \frac{\mathrm{Var}(Y_n)}{n^{\frac{2}{\theta}}} < \infty$. For the first of these, noting that $\mathbb{E}Y_n = -\mathbb{E}X_n\mathbf{1}_{[|X_n| > n^{\frac{1}{\theta}}]}$, one has

$$\begin{aligned}\sum_{n=1}^\infty n^{-\frac{1}{\theta}}|\mathbb{E}Y_n| &\leq \sum_{n=1}^\infty n^{-\frac{1}{\theta}}\mathbb{E}|X_n|\mathbf{1}_{[|X_n| > n^{\frac{1}{\theta}}]} \\ &\leq \int_0^\infty x^{-\frac{1}{\theta}}\mathbb{E}|X_1|\mathbf{1}_{[|X_1| > x^{\frac{1}{\theta}}]}dx \\ &= \mathbb{E}\left\{|X_1| \int_0^{|X_1|^\theta} x^{-\frac{1}{\theta}}dx\right\} \leq c'\mathbb{E}|X_1|^\theta < \infty,\end{aligned}$$

for some constant c'. Similarly, for the second one has

$$\begin{aligned}\sum_{n=1}^{\infty} n^{-\frac{2}{\theta}} \operatorname{Var}(|Y_n|) &\le \sum_{n=1}^{\infty} n^{-\frac{2}{\theta}} \mathbb{E}|Y_n|^2 \\ &= \sum_{n=1}^{\infty} n^{-\frac{2}{\theta}} \mathbb{E}|X_n|^2 \mathbf{1}_{[|X_n| \le n^{\frac{1}{\theta}}]} \\ &\le \int_0^{\infty} x^{-\frac{2}{\theta}} \mathbb{E}|X_1|^2 \mathbf{1}_{[|X_1| \le x^{\frac{1}{\theta}}]} dx \\ &= \mathbb{E}\left\{ |X_1|^2 \int_{|X_1|^{\theta}}^{\infty} x^{-\frac{2}{\theta}} dx \right\} \le c'\mathbb{E}|X_1|^{\theta} < \infty.\end{aligned}$$

For the converse, suppose that $n^{-\frac{1}{\theta}} \sum_{j=1}^{n} X_j$ is a.s. convergent. Let $S_n := \sum_{j=1}^{n} X_j$. Since a.s.

$$\frac{X_n}{n^{\frac{1}{\theta}}} = \frac{S_n}{n^{\frac{1}{\theta}}} - \left(\frac{n-1}{n}\right)^{\frac{1}{\theta}} \frac{S_{n-1}}{(n-1)^{\frac{1}{\theta}}} \to 0$$

as $n \to \infty$, it follows from Boret Cantelli II that

$$\mathbb{E}|X_1|^{\theta} = \int_0^{\infty} P(|X_1|^{\theta} > x) dx \le 1 + \sum_{n=1}^{\infty} P(|X_1|^{\theta} > n) < \infty.$$

In the case that $\theta > 1$, one may further conclude that $\mathbb{E}X_1 = 0$ in view of the scaling. That is, knowing that $\mathbb{E}|X_1|^{\frac{1}{\theta}} < \infty$, one may apply the direct half to conclude that a.s. $n^{-\frac{1}{\theta}}(S_n - n\mathbb{E}X_1) \to 0$. Since $n^{-\frac{1}{\theta}} S_n$ is assumed to converge a.s., so must $n^{1-\frac{1}{\theta}} \mathbb{E}X_1$ and hence $\mathbb{E}X_1 = 0$ in this case as asserted. ■

***Proposition 9.6** (Almost-Sure & Convergence in Probability for Series of Independent Terms).* Let $X_1, X_2, \ldots$ be independent random variables. Then $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if $\sum_{n=1}^{\infty} X_n$ converges in probability.

Proof. One part is obvious since almost-sure convergence always implies convergence in probability. For the converse suppose, for contradiction, that $\lim_n \sum_{j=1}^{n} X_j$ exists in probability, but with positive probability is divergent. Then there is an $\varepsilon > 0$ and a $\gamma > 0$ such that for any fixed k, $P(\sup_{n>k} |S_n - S_k| > \varepsilon) > \gamma$. Use *Skorokhod's maximal inequality* in Exercise 2 to bound $P(\max_{k<n\le m} |S_n - S_k| > \varepsilon)$ for fixed k, m. Then note that $p = p_{k,m} := \max_{k<n\le m} P(|S_m - S_{n-1}| > \varepsilon/2) \to 0$ as $k, m \to \infty$, since $|S_m - S_k| \to 0$ in probability as $k, m \to \infty$. This indicates a contradiction. ■

***Proposition 9.7** (Almost-Sure & Convergence in Distribution for Series of Independent Summands).* Let $\{X_n : n \geq 1\}$ be a sequence of independent real-valued random variables. Then $\sum_{k=1}^{n} X_k$ converges a.s. as $n \to \infty$ if and only if it converges in distribution.

Proof. One way follows from the dominated convergence theorem using characteristic functions. For the other, suppose $\sum_{k=1}^{n} X_k \to Y$ a.s. Then letting $\varphi_k(\xi) = \mathbb{E}e^{i\xi X_k}$, $\varphi(\xi) = \mathbb{E}e^{i\xi Y}$, one has $\prod_{k=1}^{n} \varphi_k(\xi) \to \varphi(\xi)$ as $n \to \infty$. Thus $\prod_{k=m}^{n} \varphi_k(\xi) \to 1$ for all ξ as $m, n \to \infty$. By (6.57), for every $\varepsilon > 0$, one has $P(|\sum_{k=m}^{n} X_k| > 2\varepsilon) \leq \varepsilon \int_{[-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}]} (1 - \prod_{k=m}^{n} \varphi_k(\xi)) d\xi \to 0$ as $m, n \to \infty$. Now use Proposition 9.6 to complete the proof. ■

EXERCISES

Exercise Set IX

1. (*Random Signs Problem*) Suppose that $a_1, a_2, \ldots$ is a sequence of real numbers, and $X_1, X_2, \ldots$ an i.i.d. sequence of symmetrically distributed Bernoulli ±1-valued random variables. Show that $\sum_{n=1}^{\infty} X_n a_n$ converges with probability one if and only if $\sum_{n=1}^{\infty} a_n^2 < \infty$. [*Hint*: Use mean-square-summability in one direction and a Kolmogorov's three-series theorem for the other.]

2. (*Skorokhod's Maximal Inequality*) Let $X_1, \ldots, X_n$ be independent random variables, $S_k = \sum_{j=1}^{k} X_j$. For a given $\delta > 0$, let $p = \max_{k \leq n} P(|S_n - S_k| > \delta) < 1$. Show that $P(\max_{k \leq n} |S_k| > 2\delta) \leq \frac{1}{q} P(|S_n| > \delta)$, where $q = 1 - p = \min_{k \leq n} P(|S_n - S_k| \leq \delta) > 0$. [*Hint*: Proceed along the lines of the proof of Kolmogorov's maximal inequality by using values of $\tau := \inf\{k : |S_k| > 2\delta\}$ to decompose the event $[|S_n| > \delta]$, and noting that $P(|S_n| > \delta, \tau = k) \geq P(|S_n - S_k| \leq \delta, \tau = k)$. The latter probability factors by independence.]

3. (*A Strong Law of Large Numbers*) Use the mean-square-summability criterion to formulate and prove a strong law of large numbers for a sequence of independent random variables $X_1, X_2, \ldots$ such that $\mathbb{E}X_n = 0$ for each $n \geq 1$, and $\sum_{n=1}^{\infty} \frac{\mathbb{E}X_n^2}{n^2} < \infty$. For what values of θ does one have this strong law with $\text{Var}(X_n) = n^\theta$?

4. (*Cesàro Limits and Kronecker's Lemma*) Give a proof of the Cesàro and Kronecker lemmas. [*Hint*: For the Cesàro limit, let $\varepsilon > 0$ and choose N sufficiently large that $a + \varepsilon > a_j > a - \varepsilon$ for all $j \geq N$. Consider lim sup and lim inf in the indicated average. For Kronecker's lemma make a "summation by parts" to the indicated sum, and apply the Cesàro limit result. Use $b_n := -\sum_{j=n}^{\infty} \frac{a_j}{c_j}$, $\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} (b_{j+1} - b_j) c_j = a_1 + b_n c_n - b_1 c_1 - \sum_{j=1}^{n-1} b_{j+1}(c_{j+1} - c_j)$.]

5. (*Kolmogorov's Truncation Method*) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with $\mathbb{E}|X_1| < \infty$. Define $Y_n = X_n \mathbf{1}_{[|X_n| \leq n]}$, for $n \geq 1$. Show that in the limit as $n \to \infty$, (a) $\mathbb{E}Y_n \to \mathbb{E}X_1$; (b) $P(Y_n \neq X_n i.o.) = 0$; and (c) $\sum_{n=1}^{\infty} \frac{\text{Var}(Y_n)}{n^2} < \infty$. [*Hint*: For (a), Lebesgue's dominated convergence; for (b), Borel–Cantelli I; for (c), $\text{Var}(Y_n) \leq \mathbb{E}Y_n^2 = \mathbb{E}\{X_1^2 \mathbf{1}[|X_1| \leq n]\}$, and $\sum_{n=1}^{\infty} \frac{1}{n^2} \mathbb{E}X_1^2 \mathbf{1}_{[|X_1| \leq n]} \leq \mathbb{E}X_1^2 \int_{|X_1|}^{\infty} x^{-2} dx = \mathbb{E}|X_1|$.]

6. (*A strong law of large numbers*) Use Kolmogorov's truncation method from the previous exercise together with Exercise 3 to prove the classic strong law for i.i.d. sequences having finite first moment.

CHAPTER X

Kolmogorov's Extension Theorem and Brownian Motion

Suppose a probability measure Q is given on a product space $\Omega = \prod_{t\in\Lambda} S_t$ with the product σ-field $\mathcal{F} = \otimes_{t\in\Lambda}\mathcal{S}_t$. Let $\mathcal{C}$ denote the class of all **finite-dimensional cylinders** C of the form

$$C = \left\{\omega = (x_t, t\in\Lambda) \in \prod_{t\in\Lambda} S_t : (x_{t_1}, x_{t_2}, \dots, x_{t_n}) \in B\right\}, \tag{10.1}$$

for $n \geq 1$, $B \in \mathcal{S}_{t_1} \otimes \cdots \otimes \mathcal{S}_{t_n}$, and $(t_1, t_2, \dots, t_n)$ an arbitrary n-tuple of distinct elements of Λ. Since $\otimes_{t\in\Lambda}\mathcal{S}_t$ is the smallest σ-field containing $\mathcal{C}$, it is simple to check from the $\pi - \lambda$ theorem that Q is determined by its values on $\mathcal{C}$. Write $\mu_{t_1,t_2,\dots,t_n}$ for the probability measure on the product space $(S_{t_1} \times S_{t_2} \times \cdots \times S_{t_n}, \mathcal{S}_{t_2} \otimes \mathcal{S}_{t_2} \otimes \cdots \otimes \mathcal{S}_{t_n})$ given by

$$\mu_{t_1,t_2,\dots,t_n}(B) := Q(C) \qquad (B \in \mathcal{S}_{t_1} \otimes \cdots \otimes \mathcal{S}_{t_n}), \tag{10.2}$$

where $C \in \mathcal{C}$ is of the form (10.1) for a given n-tuple $(t_1, t_2, \dots, t_n)$ of distinct elements in Λ. Note that this collection of **finite-dimensional distributions** $\mathcal{P}_f := \{\mu_{t_1,t_2,\dots,t_n} : t_i \in \Lambda, t_i \neq t_j \text{ for } i \neq j, n \geq 1\}$ satisfies the following so-called **consistency properties**:

(a) For any n-tuple of distinct elements $(t_1, t_2, \dots, t_n)$, $n \geq 1$, and all permutations $(t'_1, t'_2, \dots, t'_n) = (t_{\pi(1)}, \dots, t_{\pi(n)})$ of $(t_1, t_2, \dots, t_n)$, $(n \geq 1)$, one has $\mu_{t'_1,t'_2,\dots,t'_n} = \mu_{t_1,t_2,\dots,t_n} \circ T^{-1}$ under the permutation of coordinates $T : S_{t_1} \times \cdots \times S_{t_n} \to$

$S_{t'_1} \times \cdots \times S_{t'_n}$ given by $T(x_{t_1}, x_{t_2}, \ldots, x_{t_n}) := (x_{t'_1}, x_{t'_2}, \ldots, x_{t'_n})$, i.e., for finite-dimensional rectangles

$$\mu_{t'_1, t'_2, \ldots, t'_n}(B_{t'_1} \times \cdots \times B_{t'_n}) = \mu_{t_1, t_2, \ldots, t_n}(B_{t_1} \times \cdots \times B_{t_n}), \quad B_{t_i} \in \mathcal{S}_{t_i} (1 \leq i \leq n). \tag{10.3}$$

(b) If $(t_1, t_2, \ldots, t_n, t_{n+1})$ is an $(n+1)$-tuple of distinct elements of Λ, then $\mu_{t_1, t_2, \ldots, t_n}$ is the image of $\mu_{t_1, t_2, \ldots, t_{n+1}}$ under the projection map $\pi : S_{t_1} \times \cdots \times S_{t_{n+1}} \to S_{t_1} \times \cdots \times S_{t_n}$ given by $\pi(x_{t_1}, x_{t_2}, \ldots, x_{t_n}, x_{t_{n+1}}) := (x_{t_1}, x_{t_2}, \ldots, x_{t_n})$,

i.e., for finite-dimensional cylinders,

$$\mu_{t_1, t_2, \ldots, t_n}(B) = \mu_{t_1, t_2, \ldots, t_n, t_{n+1}}(B \times S_{t_{n+1}}) \;\forall\; B \in \mathcal{S}_{t_1} \otimes \cdots \otimes \mathcal{S}_{t_n}. \tag{10.4}$$

The theorem below, variously referred to by other names such as *Kolmogorov's existence theorem* or *Kolmogorov's consistency theorem*, says, conversely, that given a family $\mathcal{P}_f$ of consistent finite-dimensional probabilities, there exists a Q on $(S, \mathcal{S})$ with these as the finite-dimensional distributions.

***Theorem 10.1** (Kolmogorov's Existence Theorem).* Suppose S_t, $t \in \Lambda$, are Polish spaces and $\mathcal{S}_t = \mathcal{B}(S_t)$ $\forall\, t \in \Lambda$. Then given any family $\mathcal{P}_f$ of finite-dimensional probabilities, $\mathcal{P}_f = \{\mu_{t_1, \ldots, t_n} : t_i \in \Lambda, t_i \neq t_j \text{for } i \neq j, n \geq 1\}$ satisfying the consistency properties (a) and (b), there exists a unique probability Q on the product space $(\Omega = \prod_{t \in \Lambda} S_t, \mathcal{F} = \bigotimes_{t \in \Lambda} \mathcal{S}_t)$ satisfying (10.2) for all $n \geq 1$ and every $(t_1, \ldots, t_n)$ (n-tuple of distinct elements of Λ), $\mu_{t_1, t_2, \ldots, t_n}$. Moreover, the stochastic process $\mathbf{X} = (X_t : t \in \Lambda)$ defined on Ω by the coordinate projections $X_t(\omega) = x_t, \omega = (x_t, t \in \Lambda) \in \Omega$ has distribution Q.

Proof. A complete proof is sketched in Exercises 6, 7 with broad hints. We give a proof[1] here in the case that each image space S_t, $t \in \Lambda$, is assumed a compact metric space. The more general statement can be proved using an embedding into a compact metric space (see Exercise 8). Assuming compactness of the image spaces makes $\Omega = \prod_{t \in \Lambda} S_t$ compact for the product topology by Tychonov's[2] Theorem. On the Banach space $C(\Omega)$ of continuous functions on Ω with the uniform norm $\|f\| := \max_{\mathbf{x} \in \Omega} |f(\mathbf{x})|$, define a bounded linear functional h as follows: For a function $f \in C(S)$ that depends on only finitely many coordinates, say

$$f(\mathbf{x}) = \overline{f}(x_{t_1}, \ldots, x_{t_n}),$$

[1] This proof is due to Edward Nelsen (1959), *Regular Probability Measures on Function Spaces,* Ann. of Math. **69**, 630–643.

[2] See Appendix B for a proof of Tychonov's theorem for the case of countable Λ. For uncountable Λ, see Folland, G.B. (1984).

for some $n \geq 1$, distinct $t_1, \ldots, t_n$, and $\overline{f} : S_{t_1} \times \cdots \times S_{t_n} \to \mathbb{R}$, define

$$h(f) = \int_{S_{t_1} \times \cdots \times S_{t_n}} \overline{f} \, d\mu_{t_1,\ldots,t_n}.$$

One may check from the consistency properties that h is well-defined. By the Stone–Weierstrass theorem from real analysis, see Appendix B, the class of functions in $C(\Omega)$ depending on finitely many coordinates is dense in $C(\Omega)$. Thus h uniquely extends to a bounded (i.e., continuous) linear functional defined on all of $C(\Omega)$. One may then apply the Riesz representation theorem[3] to obtain the desired probability Q. In particular, since $C(S_{t_1} \times \cdots \times S_{t_n})$ is a measure-determining class of functions, it follows that $Q \circ \pi^{-1}_{t_1,\ldots,t_n} = \mu_{t_1 \ldots t_n}$, where $\pi_{t_1 \ldots t_n}(\omega) = (x_{t_1}, \ldots, x_{t_n})$, for $\omega = (x_t : t \in \Lambda)$. ■

Remark 10.1. In the full generality of the specification of finite-dimensional distributions for *Kolmogorov's extension theorem*, topological assumptions (for compactness) are used to prove countable additivity of Q. However, for constructing an **infinite product probability measure**, or even the distribution of a discrete parameter **Markov process** with arbitrary measurable state spaces $(S_t, \mathcal{S}_t), t \in \Lambda$, from specified transition probabilities and initial distribution, consistency is sufficient to prove that Q is a probability. The trade-off is that one is assuming more on the type of dependence structure for the finite-dimensional distributions. The extension theorem is referred to as *Tulcea's extension theorem.* The precise statement is as follows in the case $\Lambda = \{0, 1, \ldots\}$.

***Theorem 10.2** (Tulcea's Extension Theorem).* Let $(S_m, \mathcal{S}_m)$, $m = 0, 1, 2, \ldots$, be an arbitrary sequence of measurable spaces, and let $\Omega = \prod_{m=0}^{\infty} S_m$, $\mathcal{F} = \otimes_{m=0}^{\infty} \mathcal{S}_m$ denote the corresponding product space and product σ-field. Let μ_0 be a probability on $\mathcal{S}_0$ and suppose that (a) for each $n \geq 1$ and $(x_0, x_1, \ldots, x_n) \in S_1 \times \cdots \times S_n$, $B \to \mu_n(x_0, x_1, \ldots, x_{n-1}, B)$, $B \in \mathcal{S}_n$, is a probability on $\mathcal{S}_n$ and (b) for each $n \geq 1$, $B \in \mathcal{S}_n$, the map $(x_0, x_1, \ldots, x_{n-1}) \mapsto \mu_n(x_0, x_1, \ldots, x_{n-1}, B)$ is a Borel-measurable map from $\prod_{m=0}^{n-1} S_m$ into $[0, 1]$. Then there is a probability Q on $(\Omega, \mathcal{F})$ such that for each finite-dimensional cylinder set $C = B \times S_n \times S_{n+1} \times \cdots \in \mathcal{F}$, $B \in \otimes_{m=0}^{n-1} \mathcal{S}_m$ $(n \geq 1)$,

$$Q(C) = \int_{S_0} \cdots \int_{S_{n-1}} \mathbf{1}_B(x_0, \ldots, x_{n-1}) \mu_{n-1}(x_0, \ldots, x_{n-2}, dx_{n-1}) \cdots \mu_1(x_0, dx_1) \mu_0(dx_0).$$

In the case that the spaces are Polish spaces this is a consistent specification and the theorem is a special case of Kolmogorov's extension theorem. However, in the absence

[3] See Appendix A for a proof of the Riesz representation theorem for compact metric spaces S. For general locally compact Hausdorff spaces see Folland, G.B. (1984), or Royden, H.L. (1988).

of topology, it stands alone. The proof[4] is essentially a matter of checking countable additivity so that the Carathéodory extension theorem may be applied.

Remark 10.2. For the case of a product probability measure $\prod_{t\in\Lambda}\mu_t$ on $(\times_{t\in\Lambda}S_t, \otimes_{t\in\Lambda}\mathcal{S}_t)$ the component probability spaces $(S_t, \mathcal{S}_t, \mu_t)$, $t \in \Lambda$, may be arbitrary measure spaces, and Λ may be uncountable.[5] On such a space the coordinate projections $X_s(\omega) = \omega_s$, $\omega = (\omega_t : t \in \Lambda)$, define a family of independent random variables with marginal distributions μ_s $(s \in \Lambda)$.

The following example is a recasting of the content of Tulcea's theorem in the language of Markov processes whose transition probabilities are assumed to have densities.

Example 1 *(Discrete-Parameter Markov Process.)*. Let $(S, \mathcal{S})$ be a measurable space, ν a σ-finite measure on $(S, \mathcal{S})$. Let $p(x, y)$ be a nonnegative measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ such that $\int_S p(x, y)\nu(dy) = 1 \ \forall\ x \in S$. The function $p(x, y)$ is the (one-step) **transition probability density** of a Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ constructed here on the infinite product space $(S^\infty, \mathcal{S}^{\otimes\infty})$ of all sequences $\mathbf{x} := (x_0, x_1, x_2, \ldots)$ in S. Here, as usual, $\mathcal{S}^{\otimes\infty}$ is the product σ-field on S^∞ generated by the class of all finite-dimensional rectangles of the form

$$C = \{x = (x_0, x_1, \ldots) \in S^\infty : x_i \in B_i \text{ for } i = 0, 1, 2, \ldots, m\}, \tag{10.5}$$

for $m \geq 1$, $B_i \in \mathcal{S}$, $i = 0, 1, \ldots, m$. For this construction, fix a probability measure μ_0 on $(S, \mathcal{S})$ and define for $B_i \in \mathcal{S}$, $i = 0, 1, \ldots, n$,

$$\begin{aligned} &\mu_{0,1,2,\ldots,n}(B_0 \times B_1 \times \cdots \times B_n) \qquad\qquad (10.6)\\ &= \int_{B_0}\int_{B_1}\cdots\int_{B_n} p(x_0, x_1)p(x_1, x_2)\cdots p(x_{n-1}, x_n)\nu(dx_n)\nu(dx_{n-1})\cdots\nu(dx_1)\mu_0(dx_0). \end{aligned}$$

More generally, $\mu_{0,1,2,\ldots n}(B)$ is defined $\forall\ B \in \mathcal{S}^{\otimes(n+1)}$ by integration of the function $p(x_0, x_1)\cdots p(x_{n-1}, x_n)$ over B with respect to the product measure $\mu_0 \times \nu \times \cdots \times \nu$. Since $\int_S p(x_n, x_{n+1})\nu(dx_{n+1}) = 1$, the condition (b) for *consistency* of $\mu_{0,1,\ldots,n}, n \geq 0$, required by Theorem 10.1 is easily checked. For integers $0 \leq m_0 < m_1 < \cdots < m_n (n \geq 0)$, the finite-dimensional probability $\mu_{m_0,\ldots,m_n}$ can then be consistently *defined* by $\mu_{m_0,\ldots,m_n} = \mu_{0,1,\ldots,m_n} \circ \pi^{-1}_{m_0,\ldots,m_n}$, where $\pi_{m_0,\ldots,m_n}(x_0, \ldots, x_{m_n}) := (x_{m_0}, \ldots, x_{m_n})$, for $(x_0, \ldots, x_{m_n}) \in S^{m_n+1}$. If one also defines $\mu_{\tau(m_0),\tau(m_1),\ldots,\tau(m_n)}$ for any given permutation τ of $(0, 1, 2, \ldots, n)$ as the induced-image measure on $(S^{m_n+1}, \mathcal{S}^{\otimes(m_n+1)})$ under the map $(x_0, x_1, \ldots, x_{m_n}) \mapsto (x_{m_{\tau(0)}}, x_{m_{\tau(1)}}, \ldots, x_{m_{\tau(n)}})$ on $(S^{m_n+1}, \mathcal{S}^{\otimes(m_n+1)}, \mu_{0,1,2,\ldots,m_n})$ into $(S^{m_n+1}, \mathcal{S}^{\otimes(m_n+1)})$, then the family $\mathcal{P}_f$ of Theorem 10.1 is obtained, and it automatically satisfies (a) as well as (b). As noted

[4] For a proof of Tulcea's theorem see S.N. Ethier and T. Kurtz (1986), or J. Neveu (1965).

[5] See Neveu (1965).

above, according to Tulcea's proof the conclusion of Theorem 10.1 holds without any topological conditions on $(S, \mathcal{S})$. The coordinate process $\{X_n : n = 0, 1, 2, \ldots\}$ defined by $X_n(\omega) = x_n \; \forall \; \omega = (x_0, x_1, \ldots, x_n, \ldots) \in S^\infty \; n = 0, 1, 2, \ldots$ on $(S^\infty, \mathcal{S}^{\otimes\infty}, Q)$ is a **Markov process** in the sense of Theorem 2.9: The **(regular) conditional distribution of** $X_m^+ := (X_m, X_{m+1}, \ldots)$ **given** $\mathcal{F}_m := \sigma(X_0, X_1, \ldots, X_m)$ is $(Q_y)_{y=X_m} \equiv Q_{X_m}$, where $Q_y = Q$ with the **initial distribution** μ_0 taken to be the Dirac delta measure δ_y (i.e., $\mu_0(\{y\}) = 1, \; \mu_0(S \backslash \{y\}) = 0$) (Exercise 2).

Remark 10.3. As illustrated by this example, in the discrete parameter case in which $\Lambda = \{0, 1, 2, \ldots\}$, it is enough to consistently specify $\mu_{0,1,\ldots,n}$ for $n = 0, 1, 2 \ldots$, subject to condition (b) and then consistently *define* the other finite-dimensional probabilities as being induced by the coordinate projections and permutation maps. More generally, the condition (a) on permutation consistency can always be built into the specification of finite-dimensional probabilities when Λ is *linearly ordered.* This is accomplished by specifying $\mu_{t_1,t_2,\ldots,t_n}$ for $t_1 < t_2 < \cdots < t_n$ and then defining $\mu_{\tau(1),\tau(2),\ldots,\tau(n)}$ as the image (measure) of $\mu_{t_1,t_2,\ldots,t_n}$ under the permutation map $(x_{t_1}, x_{t_2}, \ldots, x_{t_n}) \to (x_{\tau(1)}, x_{\tau(2)}, \ldots, x_{\tau(n)})$. Thus one needs only to check the consistency property (b) to hold for ordered n-tuples $(t_1, t_2, \ldots, t_n)$ with $t_1 < t_2 < \cdots < t_n$.

Remark 10.4. On an arbitrary measurable space $(S, \mathcal{S})$ one defines a **transition probability** $p(x, B) : S \times \mathcal{S} \to [0, 1]$ requiring only that (i) $x \mapsto p(x, B)$ be *measurable* for each $B \in \mathcal{S}$, and that (ii) for each $x \in S$, $B \mapsto p(x, B)$ is a *probability* on $\mathcal{S}$. The construction of a **Markov process** with a given transition probability $p(\cdot, \cdot)$ and a given **initial distribution** μ_0 is now defined by the *successive iterated integration*, generalizing (10.7), beginning with the integral of $p(x_{n-1}, B_n)$ with respect to the measure $p(x_{n-2}, dx_{n-1})$ to get $\int_{B_{n-1}} p(x_{n-1}, B_n) p(x_{n-2}, dx_{n-1}) = g_{n-2}(x_{n-2})$, say. Then integrate this with respect to $p(x_{n-3}, dx_{n-2})$ to get $\int_{B_{n-2}} g_{n-2}(x_{n-2}) p(x_{n-3}, dx_{n-2}) = g_{n-3}(x_{n-3})$, say, and so on. In this manner one has $\mu_{0,1,2,\ldots,n}(B_0 \times B_1 \times \cdots \times B_n) = \int_{B_0} g_0(x_0)\mu_0(dx_0)$, $g_0(x_0) = \int_{B_1} g_1(x_1) p(x_0, dx_1)$, $g_1(x_1) = \int_{B_2} g_1(x_2) p(x_1, dx_2)$, $\ldots, g_{n-2}(x_{n-2}) = \int_{B_{n-1}} g_{n-1}(x_{n-1}) p(x_{n-2}, dx_{n-1})$, $g_{n-1}(x_{n-1}) = p(x_{n-1}, B_n) \equiv \int_{B_n} p(x_{n-1}, dx_n)$, beginning with the last term and moving successively backward.

Example 2 (Gaussian Process/Random Field). Here we take $S_t = \mathbb{R}$, $\mathcal{S}_t = \mathcal{B}(\mathbb{R}), t \in \Lambda$. The Kolmogorov extension theorem may be used to construct a probability space $(\Omega, \mathcal{F}, P)$ on which a family of Gaussian, or normal, random variables $\{X_t : t \in \Lambda\}$ are defined with arbitrarily specified (i) means $m_t = \mathbb{E}(X_t)$, $t \in \Lambda$, and (ii) covariances $\sigma_{t,t'} = \mathrm{Cov}(\mathrm{X_t}, \mathrm{X_{t'}})$, t and $t' \in \Lambda$, with the property that for every n-tuple $(t_1, t_2, \ldots, t_n)$ of distinct indices $(n \geq 1)$, the matrix $((\sigma_{t_i,t_j}))_{1 \leq i, \; j \leq n}$ is symmetric and nonnegative definite. In this case, using the notation above, $\mu_{t_1,t_2,\ldots,t_n}$ is the Gaussian probability distribution parameterized by a (mean) vector $(m_{t_1}, m_{t_2}, \ldots, m_{t_n})^t \in \mathbb{R}^n$ and symmetric, nonnegative definite (covariance) matrix $((\sigma_{t_i,t_j}))_{1 \leq i, \; j \leq n}$. More specifically, $\mu_{t_1,t_2,\ldots,t_n}$ is defined as the distribution of

$Y = AZ + m$, where $Z = (Z_1, \ldots, Z_n)^t$ is n-dimensional standard normal with pdf $\varphi(z_1, \ldots, z_n) = (2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{j=1}^n z_j^2\}$, $m = (m_{t_1}, \ldots, m_{t_n})$ and $A^t A = \Gamma := ((\sigma_{t_i,t_j}))_{1\le i,\ j\le n}$. Consistency properties (a), (b) are easily checked. Hence there exists a probability measure Q on the product space $(\Omega = \mathbb{R}^\Lambda, \mathcal{F} = \mathcal{B}(\mathbb{R})^{\otimes\Lambda})$ such that the coordinate process $\{X_t : t \in \Lambda\}$ is the desired Gaussian process. Here $X_t(\omega) = x_t$ for $\omega = (x_{t'}, t' \in \Lambda) \in \Omega \equiv \mathbb{R}^\Lambda$ $(t \in \Lambda)$. The indexing set Λ is general and includes examples such as $\Lambda = [0,\infty), [0,1]$, or in a construction, for example, of **Gaussian random fields** where $\Lambda = \mathbb{R}^k$.

As a special case, let $\Lambda = [0,\infty)$ (or $\Lambda = [0,1]$), $m_t = 0\ \forall\ t$, and $\sigma_{t,t'} = \min\{t,t'\}$ $(t,t' \in \Lambda)$. The check that $((\sigma_{t_i,t_j}))_{1\le i,j\le n}$ is nonnegative-definite for all n-tuples of distinct indices is outlined in Exercise 1. The process so constructed on $(\Omega = \mathbb{R}^{[0,\infty)}$ or $\mathbb{R}^{[0,1]})$ defines a **Brownian motion process on the Kolmogorov σ-field** $\mathcal{B}(\mathbb{R})^{\otimes[0,\infty)}$ (or $\mathcal{B}(\mathbb{R})^{\otimes[0,1]}$), i.e., on the product σ-field for Ω generated by finite-dimensional cylinders of the form $C = \{\omega = (x_t, t \in \Lambda) : (x_{t_1}, x_{t_2}, \ldots, x_{t_n}) \in B\}$ for arbitrary $n \ge 1,\ t_1 < t_2 < \cdots < t_n,\ B \in \mathcal{B}(\mathbb{R}^n)$. Unfortunately, the Kolmogorov σ-field does not include the set of (all) continuous functions $C[0,\infty)$ (or $C[0,1]$). The reason for this is that the product σ-field consists only of sets determined by countably many coordinates, rendering this model mathematically inadequate for computing probabilities of many "events"of interest due to nonmeasurability (Exercise 4). The first resolution of this situation was obtained by the seminal construction of Norbert Wiener. This led to the following definition of Brownian motion.

Definition 10.1. A stochastic process $B = \{B_t : t \ge 0\}$, $B_0 = 0$, defined on a probability space $(\Omega, \mathcal{F}, P)$ a.s. having continuous sample paths $t \to B_t, t \ge 0$, and such that for any $0 < t_1 < t_2 < \cdots < t_k$, $k \ge 1$ $(B_{t_1}, \ldots, B_{t_k})$ has a k-dimensional Gaussian distribution with zero mean and variance-covariance matrix $((t_i \wedge t_j))_{1\le i,j\le k}$ is referred to as **one-dimensional standard Brownian motion** started at $B_0 = 0$. The distribution $P \circ B^{-1}$ of the process B is a probability measure concentrated on the Borel σ-field of $C[0,\infty)$, referred to as **Wiener measure**.

Since Wiener's construction, a number of alternative approaches have become known, several of which will arise in the main course of the companion textbook on stochastic processes. For the present, however, a resolution of this problem is given below by a so-called wavelet construction in close resemblance to the classic "Fourier construction" of Wiener, but technically much simpler. Specifically, a construction is made of a probability space $(\Omega, \mathcal{F}, P)$ and stochastic process $B = \{B_t : t \in [0,\infty)\}$ such that, as above, for each $0 \le t_1 < t_2 < \cdots < t_k$ $(k \ge 1)$, $(B_{t_1}, \ldots, B_{t_k})$ is Gaussian with mean $\mathbf{0}$ and covariance matrix $((\min\{t_i, t_j\}))_{1\le i,j\le k}$. Equivalently, the increments $B_{t_j} - B_{t_{j-1}}$, $1 \le j \le k$, are independent Gaussian random variables with zero mean and variance $t_j - t_{j-1}$, respectively; cf. Exercise 1. Moreover, for such a model of Brownian motion, the subset $[B \in C[0,\infty)] \in \mathcal{F}$ is a measurable event (and has probability one).

10.1 A Wavelet Construction of Brownian Motion: The Lévy–Ciesielski Construction

A construction[6] of Brownian motion based on a.s. uniform and absolute convergence on the time interval [0,1] of a random series expansion in terms of the integrated Haar wavelet basis, referred to as the Schauder basis, of $L^2[0,1]$ may be obtained as a consequence of the following sequence of lemmas. First, though, recursively define **Haar wavelet functions** $H_{0,0}, H_{n,k}$ $n = 0,1,2,\ldots$, $2^n \leq k < 2^{n+1}$, on $0 \leq t \leq 1$, $H_{0,0}(t) \equiv 1$; $H_{0,1}(t) := \mathbf{1}_{[0,1/2]}(t) - \mathbf{1}_{(1/2,1]}(t)$; and $H_{n,k}(t) := 2^{\frac{n}{2}}\mathbf{1}_{[k2^{-n}-1,k2^{-n}-1+2^{-n-1}]}(t) - 2^{\frac{n}{2}}\mathbf{1}_{(k2^{-n}-1+2^{-n-1},k2^{-n}-1+2^{-n}]}(t), 0 \leq t \leq 1$. Recall the definition of a complete orthonormal basis in a Hilbert space (see Appendix C).

Lemma 1. The collection of Haar wavelet functions $\{H_{n,k}\}$ is a complete orthonormal basis for $L^2[0,1]$. In particular, $\langle f, g\rangle = \sum_{(n,k)}\langle f, H_{n,k}\rangle\langle g, H_{n,k}\rangle$ holds for $f, g \in L^2[0,1]$.

Proof. Orthonormality follows by a direct calculation. To prove completeness one needs to show that if $f \in L^2[0,1]$ and $\langle f, H_{n,k}\rangle = 0$ for all n, k, then $f = 0$ almost everywhere with respect to Lebesgue measure on $[0,1]$. Define

$$I_f(t) = \int_0^t f(s)ds, \qquad 0 \leq t \leq 1.$$

Then I_f is continuous with $I_f(0) = 0$. Moreover, orthogonality with respect to $H_{0,0}$ implies that $I_f(1) = 0$. Next $I_f(\frac{1}{2}) = 0$ in view of $I_f(0) = I_f(1) = 0$ and orthogonality of f to $H_{0,1}$. Using the orthogonality of f to $H_{1,2}$, one shows that $I_f(\frac{1}{4}) - (I_f(\frac{1}{2}) - I_f(\frac{1}{4})) = 0$, so that $I_f(\frac{1}{4}) = 0$. Orthogonality with $H_{1,3}$ means that $I_f(\frac{3}{4}) - I_f(\frac{1}{2}) - (I_f(1) - I_f(\frac{3}{4})) = 0$, implying $I_f(\frac{3}{4}) = 0$. Continuing by induction one finds that $I_f(k2^{-n}) = 0$ for all dyadic rationals $k2^{-n} \in [0,1]$. By continuity it now follows that $I_f(t) = 0$ for all $t \in [0,1]$ and hence $f = 0$ a.e., as asserted. The last equality is then simply Parseval's relation which holds for any complete orthonormal system (see Appendix C). ■

Definition 10.2. The functions defined by $S_{n,k}(t) := \int_0^t H_{n,k}(s)ds$, $0 \leq t \leq 1$, are called the **Schauder functions**.

Lemma 2. The Schauder functions $S_{n,k}$ on $[0,1]$ are continuous, nonnegative, and attain a maximum value $2^{-(\frac{n}{2}+1)}$. Moreover, for fixed n, the functions $S_{n,k}, k = 2^n, \ldots, 2^{n+1} - 1$, have disjoint supports.

[6] This construction originated in Ciesielski, Z. (1961): Hölder condition for realization of Gaussian processes, *Trans. Amer. Math. Soc.* **99** 403–413, based on a general approach of Lévy, P. (1948): *Processes stochastique et mouvement Brownian*, Gauthier-Villars, Paris.

Proof. Continuity is obvious. The assertions are also clearly true for $S_{0,0}$ and $S_{0,1}$. Since $H_{n,k}$ is positive, with constant value $2^{\frac{n}{2}}$ on the interval $[k2^{-n}-1, k2^{-n}-1+2^{-n-1}]$ to the left of $(k2^{-n}-1, k2^{-n}-1+2^{-n-1}]$, where it takes negative constant value $-2^{\frac{n}{2}}$, and it has the value 0 off these two intervals, $S_{n,k}$ is positive and increasing on the first interval with a maximum value $S_{n,k}(t_M) = 2^{\frac{n}{2}}(k2^{-n}-1+2^{-n-1}-k2^{-n}+1) = 2^{-(\frac{n}{2}+1)}$ at the endpoint $t_M = k2^{-n}-1+2^{-n-1}$. Moreover, it attains a minimum value $S_{n,k}(t_m) = 0$ at the rightmost endpoint $t_m = k2^{-n}-1+2^{-n-1}$. Thus $S_{n,k}$ is nonnegative with disjoint supports $[k2^{-n}-1, k2^{-n}-1+2^{-n-1}]$ for $k = 2^n, \ldots, 2^{n+1}-1$. ■

Lemma 3. For $0 \le s \le t \le 1$,

$$\sum_{n,k} S_{n,k}(s)S_{n,k}(t) = \min\{s,t\} = s.$$

Proof. By definition of the Schauder functions one has $S_{n,k}(t) = \langle \mathbf{1}_{[0,t]}, H_{n,k}\rangle$ for fixed $t \in [0,1]$. Thus one may apply Parseval's equation to obtain for $s \le t$, $\sum_{n,k} S_{n,k}(s)S_{n,k}(t) = \sum_{n,k}\langle \mathbf{1}_{[0,s]}, H_{n,k}\rangle\langle \mathbf{1}_{[0,t]}, H_{n,k}\rangle = \langle \mathbf{1}_{[0,s]}, \mathbf{1}_{[0,t]}\rangle = s$, since $\mathbf{1}_{[0,s]}\mathbf{1}_{[0,t]} = \mathbf{1}_{[0,s]}$. ■

Since the maximum of the Schauder functions are decaying exponentially, there is some room for growth in the coefficients of a series expansion in these functions as furnished by the next lemma.

Lemma 4. If $\max_{2^n \le k < 2^{n+1}} |a_{n,k}| = O(2^{n\varepsilon})$, for some $0 < \varepsilon < 1/2$, then $\sum_{n,k} a_{n,k}S_{n,k}$ on $[0,1]$ converges uniformly and absolutely to a continuous function.

Proof. The key is to observe that since for given n, the Schauder functions have disjoint supports for $2^n \le k < 2^{n+1}$, the maximum value of $|\sum_{k=2^n}^{2^{n+1}-1} a_{n,k}S_{n,k}|$ on $[0,1]$ is $(\max_{2^n \le k < 2^{n+1}} |a_{n,k}|)2^{-(\frac{n}{2}+1)}$. Thus for some $c > 0$,

$$\sum_{n \ge m}\left|\sum_{k=2^n}^{2^{n+1}-1} a_{n,k}S_{n,k}(t)\right| \le \sum_{n \ge m} c2^{n\varepsilon}2^{-(n/2+1)}$$

is the tail of a convergent geometric series. In particular, the partial sums are uniformly Cauchy. The assertion follows since the uniform limit of continuous functions on $[0,1]$ is continuous. ■

Lemma 5 ***(Feller's Tail Probability Estimates).*** For a standard normal random variable Z, $(z^{-1} - z^{-3})\sqrt{\frac{2}{\pi}}\exp\{-z^2/2\} \le P(|Z| \ge z) \le \sqrt{\frac{2}{\pi z^2}}\exp\{-z^2/2\}, z > 1$.

In particular,

$$\lim_{z\to\infty} \frac{P(|Z| > z)}{\sqrt{\frac{2}{\pi z^2}}\exp\{-z^2/2\}} = 1.$$

Proof. One may obtain simple upper and lower bounds on the integrand by perfect derivatives as follows: $\int_z^\infty e^{-\frac{x^2}{2}}dx \le \int_z^\infty \frac{x}{z}e^{-\frac{x^2}{2}}dx$, and for the other direction $-\frac{d}{dx}\{(\frac{1}{x} - \frac{1}{x^3})e^{-\frac{x^2}{2}}\} = (1 - \frac{3}{x^4})e^{-\frac{x^2}{2}} \le e^{-\frac{x^2}{2}}$. ■

Lemma 6. There is an i.i.d. sequence $a_{n,k} = Z_{n,k}$ of standard normal random variables on a probability space $(\Omega, \mathcal{F}, P)$. Moreover, $\sum_{n,k} Z_{n,k}S_{n,k}$ is uniformly and absolutely convergent on $[0,1]$ with probability one.

Proof. The existence of the i.i.d. sequence follows from Kolmogorov's extension theorem. From here apply Borel–Cantelli I and the Feller's tail probability estimates to obtain by the preceding lemma that with probability one, $\sum_{n,k} Z_{n,k}S_{n,k}$ is uniformly and absolutely convergent on $[0,1]$. Specifically, for some $c' > 0$, $\sum_{n=1}^\infty P(\max_{2^n \le k < 2^{n+1}} |Z_{n,k}| > 2^{n\varepsilon}) \le c' \sum_{n=1}^\infty 2^n 2^{-\frac{n\varepsilon}{2}} e^{-\frac{1}{2}2^{2\varepsilon n}} < \infty$. Thus $\max_{2^n \le k < 2^{n+1}} |Z_{n,k}|$ is a.s. $O(2^{n\varepsilon})$ for any choice of $0 < \varepsilon < 1/2$. ■

Lemma 7. Define $B_t := \sum_{n,k} Z_{n,k}S_{n,k}(t)$, $0 \le t \le 1$. Then with probability one, $\{B_t : 0 \le t \le 1\}$ has continuous sample paths, $B_0 = 0$, and for any $0 = t_0 < t_1 < \cdots < t_m \le 1$, $m \ge 1$, the increments $B_{t_j} - B_{t_{j-1}}$, $j = 1, \ldots, m$, are distributed as independent normal random variables with zero mean and respective variances $t_j - t_{j-1}$, $j = 1, \ldots, m$.

Proof. Observe that using the Parseval's relation as in Lemma 3,

$$\begin{aligned}
\mathbb{E}e^{i\xi B_t} &= \prod_{(n,k)} \mathbb{E}e^{i\xi Z_{n,k}S_{n,k}(t)} \\
&= \prod_{(n,k)} e^{-\frac{1}{2}\xi^2 S_{n,k}^2(t)} \\
&= \exp\left\{-\frac{1}{2}\xi^2 \sum_{(n,k)} S_{n,k}^2(t)\right\} = e^{-\frac{1}{2}\xi^2 t}.
\end{aligned} \tag{10.7}$$

Proceed inductively on m, similarly using Parseval's relation, to check that the increments $B_{t_j} - B_{t_{j-1}}$, $j = 1, \ldots, m$, have the multivariate characteristic function $\mathbb{E}\exp\{i\sum_{j=1}^m \xi_j(B_{t_j} - B_{t_{j-1}})\} = \prod_{j=1}^m \exp(-\frac{1}{2}(t_j - t_{j-1})\xi_j^2)$. ■

Theorem 10.3. There is a stochastic process $B = \{B_t : t \geq 0\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ with continuous sample paths and having stationary independent Gaussian increments $B_t - B_s$ with mean zero and variance $t - s$ for each $0 \leq s < t$.

Proof. First use Lemma 7 to construct a Brownian motion on its image space $C[0,1]$. By the Kolmogorov extension theorem one may construct a sequence $B_t^{(r)}, 0 \leq t \leq 1, r = 1, 2, \ldots$ of independent standard Brownian motions on $[0,1]$ each starting at 0. Inductively extend $B_t := B_t^{(1)}$, $0 \leq t \leq 1$, by $B_t := B_{t-r+1}^{(r)} + B_{r-1}$, $r - 1 \leq t \leq r$, $r = 1, 2, \ldots$. Then it is simple to check that the stochastic process $\{B_t : t \geq 0\}$ satisfies all the properties that define a standard Brownian motion on $[0, \infty)$ starting at 0. ■

Definition 10.3. A **k-dimensional standard Brownian motion** is a stochastic process $\{\mathbf{B}_t = (B_t^{(1)}, \ldots, B_t^{(k)}) : t \geq 0\}$ such that $\{B_t^{(j)} : t \geq 0\}$, $j = 1, \ldots, k$, are k independent one-dimensional standard Brownian motions.

In the next chapter some fine-scale properties of Brownian motion paths are presented. In particular, see Exercise 5 of Chapter XI for a simple application of the wavelet construction in this connection.

The idea of a process $\{B_t : t \geq 0\}$ that has independent Gaussian increments derives from a central limit theorem (CLT) governing the distribution of sums of large numbers of "independent small displacements".

Many classical limit theorems for sums of independent random variables in fact arise as *consequences* of much more general theories that lead to the existence of Brownian motion with a.s. continuous sample paths. This point is explored in Chapter XII via a beautiful proof of the weak convergence of suitably scaled random walks to Brownian motion by the so-called **Skorokhod embedding**.

EXERCISES

Exercise Set X

1. (i) Suppose that $Y_1, Y_2, \ldots$ is a sequence of real-valued random variables in $L^2(\Omega, \mathcal{F}, P)$ and let $\Gamma := ((\mathrm{Cov}(Y_i, Y_j))_{1 \leq i,j \leq n}$. Show that Γ is nonnegative-definite. [*Hint*: Expand $0 \leq \mathbb{E}|\sum_{i=}^{n} c_i Y_i|^2$ for real numbers $c_1, \ldots, c_n$.]

   (ii) Show that $((\sigma_{t_i,t_j}))_{1 \leq i,j \leq n} := ((\min\{t_i, t_j\}))_{1 \leq i,j \leq n}$ is nonnegative-definite for all n-tuples of distinct indices $t_1, \ldots, t_j$. [*Hint*: Take $0 \leq t_1 < t_2 < \cdots < t_n$, and let $Z_1, Z_2, \ldots, Z_n$ be independent mean-zero Gaussian random variables (for example defined on a finite product space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\mathrm{Var}(Z_1) = t_1$, $\mathrm{Var}(Z_j) = t_j - t_{j-1} (j = 2, \ldots, n)$. Consider $Y_1 = Z_1$, $Y_2 = Z_1 + Z_2, \ldots, Y_n = Z_1 + Z_2 + \cdots + Z_n$. Compute the covariance matrix of $Y_1, Y_2, \ldots, Y_n$.]

2. Prove the assertion in Example 1 that the (regular) conditional distribution of $X_m^+ := (X_m, X_{m+1}, \ldots)$, given $\mathcal{F}_m := \sigma(X_0, X_1, \ldots, X_m)$, is $(Q_y)_{y = X_m} \equiv Q_{X_m}$, where

$Q_y = Q$ with the *initial distribution* μ_0 taken to be the Delta measure δ_y (i.e., $\mu_0(\{y\}) = 1$, $\mu_0(S\backslash\{y\}) = 0$). [*Hint*: First consider a cylinder set C as in (10.5) and show that $Q_{X_m}(C)$, as given by the right side of (10.6) with $\mu_0 = \delta_{X_m}$ equals $P([X_m^+ \in C]|\sigma(X_0, X_1, \ldots, X_m)) \equiv P([(X_m, X_{m+1}, \ldots, X_{m+n}) \in B_0 \times B_1 \times \cdots \times B_n]|\sigma(X_0, X_1, \ldots, X_m)) \equiv \mathbb{E}(\mathbf{1}_{[(X_m, X_{m+1}, \ldots, X_{m+n}) \in B_0 \times \cdots \times B_n]}|\sigma(X_0, X_1, \ldots, X_m))$. For this, first check with $n = 0$ and then $n = 1$. For the latter, let $g(X_0, X_1, \ldots, X_m)$ be nonnegative, bounded measurable and calculate $\mathbb{E}(\mathbf{1}_{[X_m \in B_0, X_{m+1} \in B_1]} g(X_0, \ldots, X_m))$ using (10.6). Finally, use induction and properties of conditional expectation to calculate $\mathbb{E}(\mathbf{1}_{[X_m \in B_0, \ldots, X_{n+m} \in B_n]} g(X_0, X_1, \ldots, X_m))$.]

3. Let $S_n = \{0, 1\}$, with the power set σ-field $\mathcal{S}_n = 2^{S_n}$, $n = 1, 2, \ldots$. Suppose that $p_n : S_1 \times \cdots \times S_n \to [0, 1], n \geq 1$, are probability mass functions, i.e., $\sum_{(s_1, \ldots, s_n) \in \{0,1\}^n} p_n(s_1, \ldots, s_n) = 1$, for each n. Assume the following consistency condition: $p_n(s_1, \ldots, s_n) = p_{n+1}(s_1, \ldots, s_n, 0) + p_{n+1}(s_1, \ldots, s_n, 1), s_i \in \{0, 1\}, 1 \leq i \leq n$. Give a direct proof of the existence of a probability space $(\Omega, \mathcal{F}, P)$ and a sequence of random variables $X_1, X_2, \ldots$ such that $P(X_1 = s_1, \ldots, X_n = s_n) = p_n(s_1, \ldots, s_n)$, $s_i \in \{0, 1\}, 1 \leq i \leq n, n \geq 1$. [*Hint*: $\prod_{n \in \mathbf{N}} S_n$ may be viewed as a compact space, with Borel σ-field $\mathcal{B} \equiv \otimes_{n \in \mathbf{N}} \mathcal{S}_n$ and such that the finite-dimensional cylinders are both open and closed. Define a set function on the field of finite-dimensional cylinders and use the Heine–Borel compactness property to prove countable additivity on this field. The rest follows by Carathéodory extension theory.]

4. For $\Omega = \mathbb{R}^{[0,1]}$, we write $\omega = (x_t, 0 \leq t \leq 1) \in \Omega$ to denote a real-valued function on $[0, 1]$. Also $\mathcal{B}^{\otimes[0,1]} = \sigma(\mathcal{C})$, where $C \in \mathcal{C}$ if and only if $C \equiv C(T, B_1, B_2, \ldots) := \{\omega = (x_t, 0 \leq t \leq 1) \in \Omega : x_{t_1} \in B_1, \ldots, x_{t_n} \in B_n, \ldots\}$ for some countable set $T = \{t_1, t_2, \ldots\} \subseteq [0, 1]$, and Borel sets $B_1, B_2, \ldots$. Let $\mathcal{T}$ denote the collection of countable subsets of $[0, 1]$. For fixed $T \in \mathcal{T}$, let $\mathcal{C}_T$ be the collection of all subsets of Ω of the form $C(T, B_1, B_2, \ldots)$.
   (i) Show that $\mathcal{B}^{\otimes[0,1]} = \cup_{T \in \mathcal{T}} \sigma(\mathcal{C}_T)$.
   (ii) For fixed $T \in \mathcal{T}$, let $\varphi_T : \mathbb{R}^{[0,1]} \to \mathbb{R}^\infty$ by $\varphi_T(\omega) = (x_{t_1}, x_{t_2}, \ldots), \omega = (x_t, 0 \leq t \leq 1)$. Show that $\sigma(\mathcal{C}_T) = \sigma(\varphi_T)$ is the smallest σ-field on $\mathbb{R}^{[0,1]}$ that makes φ_T measurable for the product σ-field $\mathcal{B}^\infty$ on $\mathbb{R}^\infty$.
   (iii) Show that if $A \in \sigma(\mathcal{C}_T)$ and $\omega \in A$, then $\omega' \in \Omega$ with $\varphi_T(\omega) = \varphi_T(\omega')$ implies $\omega' \in A$. [*Hint*: $\sigma(\varphi_T) = \{\varphi_T^{-1}(F) : F \in \mathcal{B}^\infty\}$.]
   (iv) Show that $C[0, 1]$ is not measurable for the Kolmogorov product σ-field $\mathcal{B}^{\otimes[0,1]}$.
   (v) Show that $\{x \in \mathbb{R}^{[0,1]} : \sup_{0 \leq t \leq 1} x(t) \leq 1\}$ is not a measurable set for the Kolmogorov product σ-field.

5. Suppose that S is a Polish space with Borel σ-field $\mathcal{S} = \mathcal{B}(S)$. (i) Let $\mu_n(x_0, \ldots, x_{n-1}, B) = p(x_{n-1}, B)$ in Tulcea's theorem 10.2, with $p(x, dy)$ a probability on $(S, \mathcal{S})$ for each $x \in S$, and such that $x \to p(x, B)$ is Borel measurable for each fixed $B \in \mathcal{S}$. Show that the existence of the probability Q asserted in Tulcea's extension theorem follows from Kolmogorov's extension theorem. [*Hint*: Show that the specification of finite-dimensional distributions by the initial distribution μ and the transition probabilities $p(x, dy)$, $x \in S$, satisfy the Kolmogorov consistency condition (b).] (ii) Following Remark 10.4, extend the specification (10.3) to define $\mu_{0,1,\ldots,n}(B)$ to all $B \in \mathcal{S}^{\otimes(n+1)}$. [*Hint*: Successively define, (i)$g_{n-1}(x_{n-1}; x_0, \ldots, x_{n-2}) = p(x_{n-1}, B_{x_0, x_1, \ldots, x_{n-1}})$, where $B_{x_0, x_1, \ldots, x_{n-1}} = \{y \in S : (x_0, x_1, \ldots, x_n) \in B\}$ is the $(x_0, x_1, \ldots, x_{n-1})$-section of B. Then define $g_{n-2}(x_{n-2}; x_0, \ldots, x_{n-3}) = \int g_{n-1}(x_{n-1}; x_0, \ldots, x_{n-2}) p(x_{n-2}, dx_{n-1})$, and so on.] (iii) Show that the **canonical process** given by coordinate projections $\mathbf{x} \to x_n$, ($\mathbf{x} = (x_0, x_1, \ldots \in S^\infty$), say $X_n (n \geq 0)$, on $(S^\infty, \mathcal{S}^{\otimes\infty}, Q)$, has the **Markov property**:

the conditional distribution of X_{m+1} given $\mathcal{F}_m = \sigma(X_j : 0 \leq j \leq m)$ is $p(X_m, dy)$, and it is a Markov process as defined in Example 1.

6. (*Kolmogorov's Existence Theorem*) Let $S_{t_j}(j = 1, 2, \ldots)$ be Polish spaces and let $\mu_{t_1,t_2,\ldots,t_n}$ be a consistent sequence of probability measures on $(S_{t_1} \times \cdots \times S_{t_n}, \mathcal{S}_{t_1} \otimes \cdots \otimes \mathcal{S}_{t_n})$ $(n \geq 1)$. Define a sequence of probabilities on $S = \times_{j=1}^{\infty} S_{t_j}$, with the product σ-field $\mathcal{S}$, as follows. Fix $\mathbf{x} = (x_{t_1}, x_{t_2}, \ldots) \in S$. Define $P_n(B) := \mu_{t_1,\ldots,t_n}(B_{\mathbf{x}_{n+}})$, where $\mathbf{x}_{n+} = (x_{t_{n+1}}, x_{t_{n+2}}, \ldots) \in S$, and $B_{\mathbf{x}_{n+1}} = \{\mathbf{y} \in B : y_{t_{n+j}} = x_{t_{n+j}} : j = 1, 2, \ldots\}$ $(B \in \mathcal{S})$. (i) Show that $\{P_n : n \geq 1\}$ is tight. [*Hint*: Fix $\varepsilon > 0$. Find a compact set $K_{t_n} \subseteq S_{t_n}$ such that $x_{t_n} \in K_{t_n}$ and $\mu(K_{t_n}) > 1 - \frac{\varepsilon}{2^n}$ (use the fact that each probability on a Polish space is tight). Then $P_n(\times_{j=1}^{\infty} K_{t_j}) > 1 - \varepsilon$.] (ii) Show that if $P_{n'} \Rightarrow Q$ for some sequence $n'(n \geq 1)$, then Q is the desired probability.

7. (*Kolmogorov's Existence Theorem*) Assume the hypothesis of Theorem 10.1 with Λ uncountable. On the field $\mathcal{C}$ of all finite-dimensional cylinders (see (10.1)) define the set function Q as in (10.2). (i) Show that Q is a measure on $\mathcal{C}$. [*Hint*: If $\{C_n : n = 0, 1, \ldots\}$ is a disjoint collection in $\mathcal{C}$ whose union $C = \cup_{n=1}^{\infty} C_n \in \mathcal{C}$, there exists a countable set $T = \{t_j : j = 1, 2, \ldots\}$ such that C_n, $C(n \geq 1)$ belong to the σ-field $\mathcal{F}_T$ on $\Omega = \times_{t \in \Lambda} S_t$ generated by the coordinate projections $\mathbf{x} \mapsto x_{t_j}$, $t_j \in T$. By Exercise 6, there is a unique extension of Q to $\mathcal{F}_T$ that is countably additive.] (ii) Show that there is a unique extension of Q from $\mathcal{C}$ to the product σ-field $\otimes_{t \in \Lambda} \mathcal{S}_t$. [*Hint*: Use the Carathéodory extension theorem.]

8. (i) Show that every Polish space S_t $(t \in \Lambda)$ has a homeomorphic image $h_t(S_t)$ in a compact metric space K_t. [*Hint*: See Lemma 1, Chapter V.] (ii) Show that the construction of the product probability given in the text holds on $\times_{t \in \Lambda} \overline{h_t(S_t)}$.

CHAPTER XI

Brownian Motion: The LIL and Some Fine-Scale Properties

In this chapter we analyze the *growth* of the Brownian paths $t \mapsto B_t$ as $t \to \infty$. We will see by a property of "time inversion" of Brownian motion that this leads to small-scale properties as well. First, however, let us record some basic properties of the Brownian motion that follow somewhat directly from its definition.

Theorem 11.1. Let $B = \{B_t : t \geq 0\}$ be a standard one-dimensional Brownian motion starting at 0. Then

1. *(Symmetry)* $W_t := -B_t$, $t \geq 0$, is a standard Brownian motion starting at 0.
2. *(Homogeneity and Independent Increments)* $\{B_{t+s} - B_s : t \geq 0\}$ is a standard Brownian motion independent of $\{B_u : 0 \leq u \leq s\}$, for every $s \geq 0$.
3. *(Scale-Change Invariance).* For every $\lambda > 0$, $\{B_t^{(\lambda)} := \lambda^{-\frac{1}{2}} B_{\lambda t} : t \geq 0\}$ is a standard Brownian motion starting at 0.
4. *(Time-Inversion Invariance)* $W_t := tB_{1/t}$, $t > 0$, $W_0 = 0$, is a standard Brownian motion starting at 0.

Proof. Each of these is obtained by showing that the conditions defining a Brownian motion are satisfied. In the case of the time-inversion property one may apply the strong law of large numbers to obtain continuity at $t = 0$. That is, if $0 < t_n \to 0$ then write $s_n = 1/t_n \to \infty$ and $N_n := [s_n]$, where $[\cdot]$ denotes the greatest integer function, so that by the strong law of large numbers, with probability one

$$W_{t_n} = \frac{1}{s_n} B_{s_n} = \frac{N_n}{s_n} \frac{1}{N_n} \sum_{j=1}^{N_n} (B_i - B_{i-1}) + \frac{1}{s_n}(B_{s_n} - B_{N_n}) \to 0,$$

since $B_i - B_{i-1}$, $i \geq 1$, is an i.i.d. mean-zero sequence, $N_n/s_n \to 1$, and $(B_{s_n} - B_{N_n})/s_n \to 0$ a.s. as $n \to \infty$ (see Exercise 1). ■

In order to prove our main result of this section, we will make use of the following important inequality due to Paul Lévy.

Proposition 11.2 (Lévy's Inequality). Let $X_j, j = 1, \ldots, N$, be independent and symmetrically distributed (about zero) random variables. Write $S_j = \sum_{i=1}^{j} X_i, 1 \leq j \leq N$. Then, for every $y > 0$,

$$P\left(\max_{1\leq j\leq N} S_j \geq y\right) \leq 2P(S_N \geq y) - P(S_N = y) \leq 2P(S_N \geq y).$$

Proof. Write $A_j = [S_1 < y, \ldots, S_{j-1} < y,\, S_j \geq y]$, for $1 \leq j \leq N$. The events $[S_N - S_j < 0]$ and $[S_N - S_j > 0]$ have the same probability and are independent of A_j. Therefore

$$\begin{aligned}
P\left(\max_{1\leq j\leq N} S_j \geq y\right) &= P(S_N \geq y) + \sum_{j=1}^{N-1} P(A_j \cap [S_N < y]) \\
&\leq P(S_N \geq y) + \sum_{j=1}^{N-1} P(A_j \cap [S_N - S_j < 0]) \\
&= P(S_N \geq y) + \sum_{j=1}^{N-1} P(A_j)P([S_N - S_j < 0]) \\
&= P(S_N \geq y) + \sum_{j=1}^{N-1} P(A_j \cap [S_N - S_j > 0]) \\
&\leq P(S_N \geq y) + \sum_{j=1}^{N-1} P(A_j \cap [S_N > y]) \\
&\leq P(S_N \geq y) + P(S_N > y) \\
&= 2P(S_N \geq y) - P(S_N = y). \qquad (11.1)
\end{aligned}$$

This establishes the basic inequality. ■

Corollary 11.3. For every $y > 0$ one has for any $t > 0$,

$$P\left(\max_{0\leq s\leq t} B_s \geq y\right) \leq 2P(B_t \geq y).$$

Proof. Partition $[0,t]$ by equidistant points $0 < u_1 < u_2 < \cdots < u_N = t$, and let $X_1 = B_{u_1}, X_{j+1} = B_{u_{j+1}} - B_{u_j}$, $1 \le j \le N-1$, in the proposition. Now let $N \to \infty$, and use the continuity of Brownian motion. ■

Remark 11.1. It is shown in the text on stochastic processes that $P(\max_{0\le s\le t} B_s \ge y) = 2P(B_t \ge y)$. Thus Lévy's inequality is sharp in its stated generality. The following proposition concerns the **simple symmetric random walk** defined by $S_0 = 0, S_j = X_1 + \cdots + X_j$, $j \ge 1$, with $X_1, X_2, \ldots$ i.i.d. ±1-valued with equal probabilities. It demonstrates the remarkable strength of the **reflection method** used in the proof of the lemma, allowing one in particular to compute the distribution of the maximum of a random walk over a finite time.

Proposition 11.4. For the simple symmetric random walk one has for every positive integer y,

$$P\left(\max_{0\le j\le N} S_j \ge y\right) = 2P(S_N \ge y) - P(S_N = y).$$

Proof. In the notation of Lévy's inequality given in Proposition 11.2 one has, for the present case of the random walk moving by ±1 units at a time, that $A_j = [S_1 < y, \ldots, S_{j-1} = y]$, $1 \le j \le N$. Then in (11.1) the probability inequalities are all equalities for this special case. ■

Theorem 11.5 ***(Law of the Iterated Logarithm (LIL) for Brownian Motion).*** Each of the following holds with probability one:

$$\overline{\lim}_{t\to\infty} \frac{B_t}{\sqrt{2t\log\log t}} = 1, \qquad \underline{\lim}_{t\to\infty} \frac{B_t}{\sqrt{2t\log\log t}} = -1.$$

Proof. Let $\varphi(t) := \sqrt{2t\log\log t}, t > 0$. Let us first show that for any $0 < \delta < 1$, one has with probability one that

$$\overline{\lim}_{t\to\infty} \frac{B_t}{\varphi(t)} \le 1 + \delta. \tag{11.2}$$

For arbitrary $\alpha > 1$, partition the time interval $[0,\infty)$ into subintervals of exponentially growing lengths $t_{n+1} - t_n$, where $t_n = \alpha^n$, and consider the event

$$E_n := \left[\max_{t_n\le t\le t_{n+1}} \frac{B_t}{(1+\delta)\varphi(t)} > 1\right].$$

Since $\varphi(t)$ is a nondecreasing function, one has, using Corollary 11.3, a scaling property, and Lemma 5 from Chapter X, that

$$\begin{aligned} P(E_n) &\le P\left(\max_{0\le t\le t_{n+1}} B_t > (1+\delta)\varphi(t_n)\right) \\ &\le 2P\left(B_1 > \frac{(1+\delta)\varphi(t_n)}{\sqrt{t_{n+1}}}\right) \\ &\le \sqrt{\frac{2}{\pi}}\frac{\sqrt{t_{n+1}}}{(1+\delta)\varphi(t_n)}e^{-\frac{(1+\delta)^2\varphi^2(t_n)}{2t_{n+1}}} \le c\frac{1}{n^{(1+\delta)^2/\alpha}} \end{aligned} \tag{11.3}$$

for a constant $c > 0$ and all $n \ge (\log\alpha)^{-1}$. For a given $\delta > 0$ one may select $1 < \alpha < (1+\delta)^2$ to obtain $P(E_n\ i.o.) = 0$ from the Borel–Cantelli lemma (Part I). Thus we have (11.2). Since $\delta > 0$ is arbitrary we have with probability one that

$$\overline{\lim}_{t\to\infty}\frac{B_t}{\varphi(t)} \le 1. \tag{11.4}$$

Next let us show that with probability one,

$$\overline{\lim}_{t\to\infty}\frac{B_t}{\varphi(t)} \ge 1. \tag{11.5}$$

For this consider the independent increments $B_{t_{n+1}} - B_{t_n}$, $n \ge 1$. For $\theta = \frac{t_{n+1}-t_n}{t_{n+1}} = \frac{\alpha-1}{\alpha} < 1$, using Feller's tail probability estimate (Lemma 5, Chapter X) and Brownian scale change,

$$\begin{aligned} P\left(B_{t_{n+1}} - B_{t_n} > \theta\varphi(t_{n+1})\right) &= P\left(B_1 > \sqrt{\frac{\theta}{t_{n+1}}}\varphi(t_{n+1})\right) \\ &\ge c'e^{-\theta\log\log t_{n+1}} \\ &\ge cn^{-\theta} \end{aligned} \tag{11.6}$$

for suitable constants c, c' depending on α and for all sufficiently large n. It follows from the Borel–Cantelli Lemma (Part II) that with probability one,

$$B_{t_{n+1}} - B_{t_n} > \theta\varphi(t_{n+1})\ i.o. \tag{11.7}$$

Also, by (11.4) and replacing $\{B_t : t \ge 0\}$ by the standard Brownian motion $\{-B_t : t \ge 0\}$,

$$\underline{\lim}_{t\to\infty}\frac{B_t}{\varphi(t)} \ge -1,\ a.s. \tag{11.8}$$

Since $t_{n+1} = \alpha t_n > t_n$, we have

$$\frac{B_{t_{n+1}}}{\sqrt{2t_{n+1}\log\log t_{n+1}}} = \frac{B_{t_{n+1}} - B_{t_n}}{\sqrt{2t_{n+1}\log\log t_{n+1}}} + \frac{1}{\sqrt{\alpha}}\frac{B_{t_n}}{\sqrt{2t_n(\log\log t_n + \log\log\alpha)}}. \tag{11.9}$$

Now, using (11.7) and (11.8), it follows that with probability one,

$$\overline{\lim}_{n\to\infty}\frac{B_{t_{n+1}}}{\varphi(t_{n+1})} \geq \theta - \frac{1}{\sqrt{\alpha}} = \frac{\alpha-1}{\alpha} - \frac{1}{\sqrt{\alpha}}. \tag{11.10}$$

Since $\alpha > 1$ may be selected arbitrarily large, one has with probability one that

$$\overline{\lim}_{t\to\infty}\frac{B_t}{\varphi(t)} \geq \overline{\lim}_{n\to\infty}\frac{B_{t_{n+1}}}{\varphi(t_{n+1})} \geq 1. \tag{11.11}$$

This completes the computation of the limit superior. To get the limit inferior simply replace $\{B_t : t \geq 0\}$ by $\{-B_t : t \geq 0\}$. ■

The time inversion property for Brownian motion turns the law of the iterated logarithm (LIL) into a statement concerning the degree (or lack) of *local smoothness.* (Also see Exercise 5).

Corollary 11.6. Each of the following holds with probability one:

$$\overline{\lim}_{t\to 0}\frac{B_t}{\sqrt{2t\log\log\frac{1}{t}}} = 1, \qquad \underline{\lim}_{t\to 0}\frac{B_t}{\sqrt{2t\log\log\frac{1}{t}}} = -1.$$

EXERCISES

Exercise Set XI

1. Use Feller's tail estimate (Lemma 5, Chapter X). to prove that $\max\{|B_i - B_{i-1}| : i = 1, 2, \ldots, N+1\}/N \to 0$ a.s. as $N \to \infty$.

2. Show that with probability one, standard Brownian motion has arbitrarily large zeros. [*Hint*: Apply the LIL.]

3. Fix $t \geq 0$ and use the law of the iterated logarithm to show that $\lim_{h\to 0}\frac{B_{t+h}-B_t}{h}$ exists only with probability zero. [*Hint*: Check that $Y_h := B_{t+h} - B_t, h \geq 0$, is distributed as standard Brownian motion starting at 0. Consider $\frac{1}{h}Y_h = \frac{Y_h}{\sqrt{2h\log\log(1/h)}}\frac{\sqrt{2h\log\log(1/h)}}{h}$.]

4. For the simple symmetric random walk, find the distributions of the extremes: (a) $M_N = \max\{S_j : j = 0, \ldots, N\}$, and (b) $m_N = \min\{S_j : 0 \leq j \leq N\}$.

5. (*Lévy Modulus of Continuity*[1]) Use the wavelet construction $B_t := \sum_{n,k} Z_{n,k} S_{n,k}(t)$, $0 \le t \le 1$, of standard Brownian motion to establish the following fine-scale properties.
   (i) Let $0 < \delta < \frac{1}{2}$. With probability one there is a random constant K such that if $|t-s| \le \delta$ then $|B_t - B_s| \le K\sqrt{\delta \log \frac{1}{\delta}}$. [*Hint*: Fix N and write the increment as a sum of three terms: $B_t - B_s = Z_{00}(t-s) + \sum_{n=0}^{N} \sum_{k=2^n}^{2^{n+1}-1} Z_{n,k} \int_s^t H_{n,k}(u)du + \sum_{n=N+1}^{\infty} \sum_{k=2^n}^{2^{n+1}-1} Z_{n,k} \int_s^t H_{n,k}(u)du = a + b + c$. Check that for a suitable (random) constant K' one has $|b| \le |t-s|K' \sum_{n=0}^{N} n^{\frac{1}{2}} 2^{\frac{n}{2}} \le |t-s|K' \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{N} 2^{\frac{N}{2}}$, and $|c| \le K' \sum_{n=N+1}^{\infty} n^{\frac{1}{2}} 2^{-\frac{n}{2}} \le K' \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{N} 2^{-\frac{N}{2}}$. Use these estimates, taking $N = [-\log_2(\delta)]$ such that $\delta 2^N \sim 1$, to obtain the bound $|B_t - B_s| \le |Z_{00}|\delta + 2K'\sqrt{-\delta \log_2(\delta)}$. This is sufficient since $\delta < \sqrt{\delta}$.]
   (ii) The modulus of continuity is sharp in the sense that with probability one, there is a sequence of intervals $(s_n, t_n), n \ge 1$, of respective lengths $t_n - s_n \to 0$ as $n \to \infty$ such that the ratio $\frac{B_{t_n} - B_{s_n}}{\sqrt{-(t_n - s_n)\log(t_n - s_n)}}$ is bounded below by a positive constant. [*Hint*: Use Borel–Cantelli I together with Feller's tail probability estimate for the Gaussian distribution to show that $P(A_n\ i.o.) = 0$, where $A_n := [|B_{k2^{-n}} - B_{(k-1)2^{-n}}| \le c\sqrt{n2^{-n}}, k = 1, \ldots, 2^n]$ and c is fixed in $(0, \sqrt{2\log 2})$. Interpret this in terms of the certain occurrence of the complimentary event $[A_n\ i.o.]^c$.]
   (iii) The paths of Brownian motion are a.s. nowhere differentiable.

[1] The calculation of the modulus of continuity for Brownian motion is due to Lévy, P. (1937), *Théorie de l'addition des variables aléatores*, Gauthier-Villars, Paris. However this exercise follows Pinsky, M. (1999): Brownian continuity modulus via series expansions, *J. Theor. Probab.* **14** (1), 261–266.

CHAPTER XII

Skorokhod Embedding and Donsker's Invariance Principle

This chapter ties together a number of the topics introduced in the text via applications to the further analysis of Brownian motion, a fundamentally important stochastic process whose existence was established in Chapter X.

The discrete-parameter random walk was introduced in Chapter II, where it was shown to have the Markov property. Markov processes on a general state space S with a given transition probability $p(x, dy)$ were introduced in Chapter X (see Example 1 and Remark 10.4 in Chapter X). Generalizing from this example, a sequence of random variables $\{X_n : n \geq 0\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ with values in a measurable space $(S, \mathcal{S})$ has the **Markov property** if for every $m \geq 0$, the conditional distribution of X_{m+1} given $\mathcal{F}_m := \sigma(X_j, 0 \leq j \leq m)$ is the same as its conditional distribution given $\sigma(X_m)$. In particular, the conditional distribution is a function of X_m, denoted by $p_m(X_m, dy)$, where $p_m(x, dy)$, $x \in S$ is referred to as the (one-step) **transition probability** at time m and satisfies the following:

1. For $x \in S$, $p_m(x, dy)$ is a probability on $(S, \mathcal{S})$.
2. For $B \in \mathcal{S}$, the function $x \to p_m(x, B)$ is a real-valued measurable function on S.

In the special case that $p_m(x, dy) = p(x, dy)$, for every $m \geq 0$, the transition probabilities are said to be **homogeneous** or **stationary**.

With the random walk example as background, let us recall some basic definitions. Let P_z denote the **distribution** of a discrete-parameter stochastic process $X = \{X_n : n \geq 0\}$, i.e., a probability on the product space $(S^\infty, \mathcal{S}^{\otimes\infty})$, with transition probability $p(x, dy)$ and initial distribution $P(X_0 = z) = 1$. The notation $\mathbb{E}_z$ is used to denote expectations with respect to the probability P_z.

Definition 12.1. Fix $m \geq 0$. The **after-m** (future) process is defined by $X_m^+ := \{X_{n+m} : n \geq 0\}$.

It follows from the definition of a Markov process $\{X_n : n = 0, 1, 2, \ldots\}$ with a stationary transition probability given above that for every $n \geq 0$ the conditional distribution of $(X_m, X_{m+1}, \ldots, X_{m+n})$, given $\sigma(X_0, \ldots, X_m)$ is the same as the P_x-distribution of $(X_0, \ldots, X_n)$, evaluated at $x = X_m$. To see this, let f be a bounded measurable function on $(S^{n+1}, \mathcal{S}^{\otimes(n+1)})$. Then the claim is that

$$\mathbb{E}\big(f(X_m, X_{m+1}, \ldots, X_{m+n})|\sigma(X_0, \ldots, X_m)\big) = g_0(X_m), \tag{12.1}$$

where given $X_0 = x$,

$$g_0(x) := \mathbb{E}_x f(X_0, X_1, \ldots, X_n). \tag{12.2}$$

For $n = 0$ this is trivial. For $n \geq 1$, first take the conditional expectation of $f(X_m, X_{m+1}, \ldots, X_{m+n})$, given $\sigma(X_0, \ldots, X_m, \ldots, X_{m+n-1})$ to get, by the Markov property, that

$$\begin{aligned}
&\mathbb{E}\big(f(X_m, X_{m+1}, \ldots, X_{m+n})\ |\sigma(X_0, \ldots, X_m, \ldots, X_{m+n-1})\big)\\
&= \int_S f(X_m, \ldots, X_{m+n-1}, x_{m+n}) p(X_{m+n-1}, dx_{m+n})\\
&= g_{n-1}(X_m, \ldots, X_{m+n-1}), \quad \text{say.}
\end{aligned} \tag{12.3}$$

Next take the conditional expectation of the above with respect to $\sigma(X_0, \ldots, X_{m+n-2})$ to get

$$\begin{aligned}
&\mathbb{E}\big(f(X_m, X_{m+1}, \ldots, X_{m+n})\ |\sigma(X_0, \ldots, X_m, \ldots, X_{m+n-2})\big)\\
&= \mathbb{E}\big(g_{n-1}(X_m, \ldots, X_{m+n-1})|\sigma(X_0, \ldots, X_{m+n-2})\big)\\
&= \mathbb{E} \int_S g_{n-1}(X_m, \ldots, X_{m+n-2}, x_{m+n-1}) p(X_{m+n-2}, dx_{m+n-1})\\
&= g_{n-2}(X_m, \ldots, X_{m+n-2}), \quad \text{say.}
\end{aligned} \tag{12.4}$$

Continuing in this manner one finally arrives at

$$\begin{aligned}
&\mathbb{E}\big(f(X_m, X_{m+1}, \ldots, X_{m+n})\ |\sigma(X_0, \ldots, X_m, \ldots, X_m)\big)\\
&= \mathbb{E}\big(g_1(X_m, X_{m+1})|\sigma(X_0, \ldots, X_m, \ldots, X_m)\big)\\
&= \int_S g_1(X_m, x_{m+1}) p(X_m, dx_{m+1}) = g_0(X_m), \quad \text{say.}
\end{aligned} \tag{12.5}$$

Now, on the other hand, let us compute $\mathbb{E}_x f(X_0, X_1, \ldots, X_n)$. For this, one follows the same steps as above, but with $m = 0$. That is, first take the conditional expectation of $f(X_0, X_1, \ldots, X_n)$, given $\sigma(X_0, X_1, \ldots, X_{n-1})$, arriving at $g_{n-1}(X_0, X_1, \ldots, X_{n-1})$. Then take the conditional expectation of this given $\sigma(X_0, X_1, \ldots, X_{n-2})$, arriving at $g_{n-2}(X_0, \ldots, X_{n-2})$, and so on. In this way one again arrives at $g_0(X_0)$, which is (12.1) with $m = 0$, or (12.2) with $x = X_m$.

Since finite-dimensional cylinders $C = B \times S^\infty$, $B \in \mathcal{S}^{\otimes(n+1)}$ $(n = 0, 1, 2, \ldots)$ constitute a π-system, and taking $f = \mathbf{1}_B$ in (12.1), (12.2), one has, for every $A \in \sigma(X_0, \ldots, X_m)$,

$$\mathbb{E}\big(\mathbf{1}_A \mathbf{1}_{[X_m^+ \in C]}\big) = \mathbb{E}\big(\mathbf{1}_A \mathbf{1}_{[(X_m, X_{m+1}, \ldots, X_{m+n}) \in B]}\big) = \mathbb{E}\big(\mathbf{1}_A P_x(C)|_{x=X_m}\big). \tag{12.6}$$

It follows from the π-λ theorem that

$$\mathbb{E}\big(\mathbf{1}_A \mathbf{1}_{[X_m^+ \in C]}\big) = \mathbb{E}\big(\mathbf{1}_A P_x(C)|_{x=X_m}\big), \tag{12.7}$$

for all $C \in \mathcal{S}^\infty$; here $P_x(C)|_{x=X_m}$ denotes the (composite) evaluation of the function $x \mapsto P_x(C)$ at $x = X_m$. Thus, we have arrived at the following equivalent, but seemingly stronger, definition of the Markov property.

***Definition 12.2** (Markov Property).* We say that $X = \{X_n : n \geq 0\}$ has the **(homogeneous) Markov Property** if for every $m \geq 0$, the conditional distribution of X_m^+, given the σ-field $\mathcal{F}_m$, is P_{X_m}, i.e., equals P_y on the set $[X_m = y]$.

This notion may be significantly strengthened by considering the future evolution given its history up to and including a random stopping time. Let us recall that given a stopping time τ, the **pre-τ σ-field** $\mathcal{F}_\tau$ is defined by

$$\mathcal{F}_\tau = \{A \in \mathcal{F} : A \cap [\tau = m] \in \mathcal{F}_m, \forall m \geq 0\}. \tag{12.8}$$

Definition 12.3. The **after-τ process** $X_\tau^+ = \{X_\tau, X_{\tau+1}, X_{\tau+2}, \ldots\}$ is well defined on the set $[\tau < \infty]$ by $X_\tau^+ = X_m^+$ on $[\tau = m]$.

The following theorem shows that for discrete-parameter Markov processes, this stronger (Markov) property that "conditionally given the past and the present the future starts afresh at the present state" holds more generally for a stopping time τ in place of a constant "present time" m.

***Theorem 12.1** (Strong Markov Property).* Let τ be a stopping time for the process $\{X_n : n \geq 0\}$. If this process has the Markov property of Definition 12.2, then on $[\tau < \infty]$ the conditional distribution of the after-τ process X_τ^+, given the pre-τ σ-field $\mathcal{F}_\tau$, is P_{X_τ}.

Proof. Let f be a real-valued bounded measurable function on $(S^\infty, \mathcal{S}^{\otimes\infty})$, and let $A \in \mathcal{F}_\tau$. Then

$$\begin{aligned}\mathbb{E}(\mathbf{1}_{[\tau<\infty]}\mathbf{1}_A f(X_\tau^+)) &= \sum_{m=0}^{\infty} \mathbb{E}(\mathbf{1}_{[\tau=m]}\mathbf{1}_A f(X_m^+)) \\ &= \sum_{m=0}^{\infty} \mathbb{E}(\mathbf{1}_{[\tau=m]\cap A}\mathbb{E}_{X_m} f) \\ &= \sum_{m=0}^{\infty} \mathbb{E}(\mathbf{1}_{[\tau=m]\cap A}\mathbb{E}_{X_\tau} f) = \mathbb{E}(\mathbf{1}_{[\tau<\infty]}\mathbf{1}_A \mathbb{E}_{X_\tau} f). \qquad (12.9)\end{aligned}$$

The second equality follows from the Markov property in Definition 12.2 since $A \cap [\tau = m] \in \mathcal{F}_m$. ■

Let us now consider the continuous-parameter Brownian motion process along similar lines. It is technically convenient to consider the canonical model of standard Brownian motion $\{B_t : t \geq 0\}$ started at 0, on $\Omega = C[0,\infty)$ with $\mathcal{B}$ the Borel σ-field on $C[0,\infty)$, P_0, referred to as Wiener measure, and $B_t(\omega) := \omega(t)$, $t \geq 0, \omega \in \Omega$, the coordinate projections. However, for continuous-parameter processes it is often useful to make sure that all events that have probability zero are included in the σ-field for Ω. For example, in the analysis of fine-scale structure of Brownian motion certain sets D may arise that *imply* events $E \in \mathcal{B}$ for which one is able to compute $P(E) = 0$. In particular, then, one would want to conclude that D is measurable (and hence assigned $P(D) = 0$ too). For this it may be necessary to replace $\mathcal{B}$ by its σ-field completion $\mathcal{F} = \overline{\mathcal{B}}$. We have seen that this can always be achieved, and there is no loss in generality in assuming that the underlying probability space $(\Omega, \mathcal{F}, P)$ is **complete** from the outset (see Appendix A).

Although the focus is on Brownian motion, just as for the above discussion of random walk, some of the definitions apply more generally and will be so stated in terms of a generic continuous-parameter stochastic process $\{Z_t : t \geq 0\}$, having continuous sample paths (outside a P-null set).

Definition 12.4. For fixed $s > 0$ the **after-s** process is defined by $Z_s^+ := \{Z_{s+t} : t \geq 0\}$.

Definition 12.5. A continuous-parameter stochastic process $\{Z_t : t \geq 0\}$, with a.s. continuous sample paths, such that for each $s > 0$, the conditional distribution of the after-s process Z_s^+ given $\sigma(Z_t, t \leq s)$ coincides with its conditional distribution given $\sigma(Z_s)$ is said to have the **Markov property**.

As will become evident from the calculations in the proof below, the Markov property of a Brownian motion $\{B_t : t \geq 0\}$ follows from the fact that it has independent increments.

Proposition 12.2 (Markov Property of Brownian Motion). Let P_x denote the distribution on $C[0,\infty)$ of standard Brownian motion $B^x = \{x + B_t : t \geq 0\}$ started at x. For every $s \geq 0$, the conditional distribution of $(B^x_s)^+ := \{B^x_{s+t} : t \geq 0\}$ given $\sigma(B^x_u : 0 \leq u \leq s)$ is $P_{B^x_s}$.

Proof. Write $\mathcal{G} := \sigma(B^x_u : 0 \leq u \leq s)$. Let f be a real-valued bounded measurable function on $C[0,\infty)$. Then $Ef\big((B^x_s)^+|\mathcal{G}\big) = E\big(\psi(U,V)|\mathcal{G}\big)$, where $U = B^x_s$, $V = \{B^x_{s+t} - B^x_s : t \geq 0\}$, $\psi(y,\omega) := f(\omega^y), y \in \mathbb{R}$, $\omega \in C[0,\infty)$, and $\omega^y \in C[0,\infty)$ by $\omega^y(t) = \omega(t) + y$. By the substitution property for conditional expectation (Theorem 2.7), since U is $\mathcal{G}$-measurable and V is independent of $\mathcal{G}$, one has

$$\mathbb{E}\big(\psi(U,V)|\mathcal{G}\big) = h(U) = h(B^x_s),$$

where, simplifying notation by writing $B_t = B^0_t$ and, in turn, $\{B_t : t \geq 0\}$ for a standard Brownian motion starting at 0,

$$h(y) = \mathbb{E}\psi(y,V) = \mathbb{E}\psi(y,\{B_t : t \geq 0\}) = \mathbb{E}f(B^y) = \int_{C[0,\infty)} f\,dP_y. \qquad \blacksquare$$

It is sometimes useful to extend the definition of standard Brownian motion as follows.

Definition 12.6. Let $(\Omega,\mathcal{F},P)$ be a probability space and $\mathcal{F}_t, t \geq 0$, a filtration. The **k-dimensional standard Brownian motion with respect to this filtration** is a stochastic process $\{B_t : t \geq 0\}$ on $(\Omega,\mathcal{F},P)$ having (i) stationary, independent Gaussian increments $B_{t+s} - B_s$ with mean zero and covariance matrix $(t-s)I_k$; (ii) a.s. continuous sample paths $t \mapsto B_t$ on $[0,\infty) \to \mathbb{R}^k$; and (iii) for each $t \geq 0, B_t$ is $\mathcal{F}_t$-measurable and $B_t - B_s$ is independent of $\mathcal{F}_s$, $0 \leq s < t$. Taking $B_0 = 0$ a.s., then $B^x := \{x + B_t : t \geq 0\}$, is referred to as the **standard Brownian motion started at** $x \in \mathbb{R}^k$ (with respect to the given filtration).

For example, one may take the completion $\mathcal{F}_t = \overline{\sigma}(B_s : s \leq t)$, $t \geq 0$, of the σ-field generated by the coordinate projections $t \mapsto \omega(t)$, $\omega \in C[0,\infty)$. Alternatively, one may have occasion to use $\mathcal{F}_t = \sigma(B_s, s \leq t) \vee \mathcal{G}$, where $\mathcal{G}$ is some σ-field independent of $\mathcal{F}$. The definition of the Markov property can be modified accordingly as follows.

Proposition 12.3. The Markov property of Brownian motions B^x on $\mathbb{R}^k$ defined on $(\Omega,\mathcal{F},P)$ holds with respect to (i) the right-continuous filtration defined by

$$\mathcal{F}_{t+} := \bigcap_{\varepsilon>0} \mathcal{F}_{t+\varepsilon} \qquad (t \geq 0), \tag{12.10}$$

where $\mathcal{F}_t = \mathcal{G}_t := \sigma(B_u : 0 \leq u \leq t)$, or (ii) $\mathcal{F}_t$ is the P-completion of $\mathcal{G}_t$, or (iii) $\mathcal{F}_t = \mathcal{G}_t \vee \mathcal{G}$ $(t \geq 0)$, where $\mathcal{G}$ is independent of $\mathcal{F}$.

Proof. (i) It is enough to prove that $B_{t+s} - B_s$ is independent of $\mathcal{F}_{s+}$ for every $t > 0$. Let $G \in \mathcal{F}_{s+}$ and $t > 0$. For each $\varepsilon > 0$ such that $t > \varepsilon$, $G \in \mathcal{F}_{s+\varepsilon}$, so that if $f \in C_b(\mathbb{R}^k)$, one has

$$\mathbb{E}(\mathbf{1}_G f(B_{t+s} - B_{s+\varepsilon})) = P(G) \cdot \mathbb{E}f(B_{t+s} - B_{s+\varepsilon}).$$

Letting $\varepsilon \downarrow 0$ on both sides,

$$\mathbb{E}(\mathbf{1}_G f(B_{t+s} - B_s)) = P(G)\mathbb{E}f(B_{t+s} - B_s).$$

Since the indicator of every closed subset of $\mathbb{R}^k$ is a decreasing limit of continuous functions bounded by 1 (see the proof of Alexandrov's theorem in Chapter V), the last equality also holds for indicator functions f of closed sets. Since the class of closed sets is a π-system, and the class of Borel sets whose indicator functions f satisfy the equality is a σ-field, one can use the π-λ theorem to obtain the equality for all $B \in \mathcal{B}(\mathbb{R}^k)$. The proofs of (ii) and (iii) are left to Exercise 2 . ■

One may define the σ-field governing the "past up to time τ" as the σ-field of events $\mathcal{F}_\tau$ given by

$$\mathcal{F}_\tau := \sigma(Z_{t\wedge\tau} : t \geq 0). \tag{12.11}$$

The stochastic process $\{\tilde{Z}_t : t \geq 0\} := \{Z_{t\wedge\tau} : t \geq 0\}$ is referred to as the **process stopped at** τ. Events in $\mathcal{F}_\tau$ depend only on the process stopped at τ. The stopped process contains no further information about the process $\{Z_t : t \geq 0\}$ beyond the time τ. Alternatively, in analogy with the discrete-parameter case, a description of the past up to time τ that is often more useful for checking whether a particular event belongs to it may be formulated as follows.

Definition 12.7. Let τ be a stopping time with respect to a filtration $\mathcal{F}_t$, $t \geq 0$. The **pre-τ σ-field** is

$$\mathcal{F}_\tau = \{F \in \mathcal{F} : F \cap [\tau \leq t] \in \mathcal{F}_t \text{ for all } t \geq 0\}.$$

For example, using this definition it is simple to check that

$$[\tau \leq t] \in \mathcal{F}_\tau, \forall t \geq 0, \qquad [\tau < \infty] \in \mathcal{F}_\tau. \tag{12.12}$$

Remark 12.1. We will always use[1] Definition 12.7, and not (12.11). Note, however, that $t \wedge \tau \leq t$ for all t, so that $\sigma(X_{t\wedge\tau} : t \geq 0\}$ is contained in $\mathcal{F}_\tau$ (see Exercise 1).

[1] The proof of the equivalence of (12.11) and that of Definition 12.7 for processes with continuous sample paths may be found in Stroock and Varadhan (1980, p. 33).

The future relative to τ is the **after-τ process** $Z_\tau^+ = \{(Z_\tau^+)_t : t \geq 0\}$ obtained by viewing $\{Z_t : t \geq 0\}$ from time $t = \tau$ onwards, for $\tau < \infty$. This is

$$(Z_\tau^+)_t(\omega) = Z_{\tau(\omega)+t}(\omega), \qquad t \geq 0, \qquad \text{on } [\tau < \infty]. \tag{12.13}$$

Theorem 12.4 *(Strong Markov Property for Brownian Motion)*. Let $\{B_t : t \geq 0\}$ be a k-dimensional Brownian motion with respect to a filtration $\{\mathcal{F}_t : t \geq 0\}$ starting at 0 and let P_0 denote its distribution (Wiener measure) on $C[0, \infty)$. For $x \in \mathbb{R}^k$ let P_x denote the distribution of the Brownian motion process $B_t^x := x + B_t$, $t \geq 0$, started at x. Let τ be a stopping time. On $[\tau < \infty]$, the conditional distribution of B_τ^+ given $\mathcal{F}_\tau$ is the same as the distribution of $\{B_t^y : t \geq 0\}$ starting at $y = B_\tau$. In other words, this conditional distribution is P_{B_τ} on $[\tau < \infty]$.

Proof. First assume that τ has countably many values ordered as $0 \leq s_1 < s_2 < \cdots$. Consider a finite-dimensional function of the after-τ process of the form

$$h(B_{\tau+t_1'}, B_{\tau+t_2'}, \ldots, B_{\tau+t_r'}), \qquad [\tau < \infty], \tag{12.14}$$

where h is a bounded continuous real-valued function on $(\mathbb{R}^k)^r$ and $0 \leq t_1' < t_2' < \cdots < t_r'$. It is enough to prove

$$\mathbb{E}\left[h(B_{\tau+t_1'}, \ldots, B_{\tau+t_r'})\mathbf{1}_{[\tau<\infty]} \mid \mathcal{F}_\tau\right] = [\mathbb{E}h(B_{t_1'}^y, \ldots, B_{t_r'}^y)]_{y=B_\tau}\mathbf{1}_{[\tau<\infty]}. \tag{12.15}$$

That is, for every $A \in \mathcal{F}_\tau$ we need to show that

$$\mathbb{E}(\mathbf{1}_A h(B_{\tau+t_1'}, \ldots, B_{\tau+t_r'})\mathbf{1}_{[\tau<\infty]}) = \mathbb{E}\left(\mathbf{1}_A \left[\mathbb{E}h(B_{t_1'}^y, \ldots, B_{t_r'}^y)\right]_{y=B_\tau} \mathbf{1}_{[\tau<\infty]}\right). \tag{12.16}$$

Now

$$[\tau = s_j] = [\tau \leq s_j] \cap [\tau \leq s_{j-1}]^c \in \mathcal{F}_{s_j},$$

so that $A \cap [\tau = s_j] \in \mathcal{F}_{s_j}$. Express the left side of (12.16) as

$$\sum_{j=1}^{\infty} \mathbb{E}\big(\mathbf{1}_{A\cap[\tau=s_j]} h(B_{s_j+t_1'}, \ldots, B_{s_j+t_r'})\big). \tag{12.17}$$

By the Markov property, the jth summand in (12.17) equals

$$\mathbb{E}(\mathbf{1}_A \mathbf{1}_{[\tau=s_j]}[\mathbb{E}h(B_{t_1'}^y, \ldots, B_{t_r'}^y)]_{y=B_{s_j}}) = \mathbb{E}(\mathbf{1}_A \mathbf{1}_{[\tau=s_j]}[\mathbb{E}h(B_{t_1'}^y, \ldots, B_{t_r'}^y)]_{y=B_\tau}). \tag{12.18}$$

Summing this over j, one obtains the desired relation (12.16). This completes the proof in the case that τ has countably many values $0 \leq s_1 < s_2 < \cdots$.

The case of more general τ may be dealt with by approximating it by stopping times assuming countably many values. Specifically, for each positive integer n define

$$\tau_n = \begin{cases} \dfrac{j}{2n} & \text{if } \dfrac{j-1}{2^n} < \tau \leq \dfrac{j}{2^n}, \quad j = 0, 1, 2, \ldots \\ \infty & \text{if } \tau = \infty. \end{cases} \tag{12.19}$$

Since

$$\left[\tau_n = \frac{j}{2^n}\right] = \left[\frac{j-1}{2^n} < \tau \leq \frac{j}{2^n}\right] = \left[\tau \leq \frac{j}{2^n}\right] \setminus \left[\tau \leq \frac{j-1}{2^n}\right] \in \mathcal{F}_{j/2^n}, \tag{12.20}$$

it follows that

$$[\tau_n \leq t] = \bigcup_{j : j/2^n \leq t} \left[\tau_n = \frac{j}{2^n}\right] \in \mathcal{F}_t \qquad \text{for all } t \geq 0. \tag{12.21}$$

Therefore, τ_n is a stopping time for each n and $\tau_n(\omega) \downarrow \tau(\omega)$ as $n \uparrow \infty$ for each $\omega \in \Omega$. Also one may easily check that $\mathcal{F}_\tau \subseteq \mathcal{F}_{\tau_n}$ from the definition (see Exercise 1). Let h be a bounded continuous function on $(\mathbb{R}^k)^r$. Define

$$\varphi(y) \equiv \mathbb{E}h(B^y_{t'_1}, \ldots, B^y_{t'_r}). \tag{12.22}$$

One may also check that φ is continuous using the continuity of $y \rightarrow (B^y_{t'_1}, \ldots, B^y_{t'_r})$. Let $A \in \mathcal{F}_\tau (\subseteq \mathcal{F}_{\tau_n})$. Applying (12.16) to $\tau = \tau_n$ one has

$$\mathbb{E}(\mathbf{1}_A h(B_{\tau_n + t'_1}, \ldots, B_{\tau_n + t'_r}) \mathbf{1}_{[\tau_n < \infty]}) = \mathbb{E}(\mathbf{1}_A \varphi(B_{\tau_n}) \mathbf{1}_{[\tau_n < \infty]}). \tag{12.23}$$

Since h, φ are continuous, $\{B_t : t \geq 0\}$ has continuous sample paths, and $\tau_n \downarrow \tau$ as $n \rightarrow \infty$, Lebesgue's dominated convergence theorem may be used on both sides of (12.23) to get

$$\mathbb{E}(\mathbf{1}_A h(B_{\tau + t'_1}, \ldots, B_{\tau + t'_r}) \mathbf{1}_{[\tau < \infty]}) = \mathbb{E}(\mathbf{1}_A \varphi(B_\tau) \mathbf{1}_{[\tau < \infty]}). \tag{12.24}$$

This establishes (12.16). Since finite-dimensional distributions determine a probability on $C[0, \infty)$, the proof is complete. ■

Remark 12.2. Note that the proofs of the Markov property (Proposition 12.3 and the strong Markov property (Theorem 12.1) hold for $\mathbb{R}^k$-valued Brownian motions on $\mathbb{R}^k$ with arbitrary drift and positive definite diffusion matrix (Exercise 2).

The examples below illustrate the usefulness of Theorem 12.4 in typical computations. In examples 2–4, $B = \{B_t : t \geq 0\}$ is a one-dimensional standard Brownian motion starting at zero. For $\omega \in C([0,\infty) : \mathbb{R})$ define, for every $a \in \mathbb{R}$,

$$\overline{\tau}_a^{(1)}(\omega) \equiv \overline{\tau}_a(\omega) := \inf\{t \geq 0 : \omega(t) = a\}, \tag{12.25}$$

and, recursively,

$$\overline{\tau}_a^{(r+1)}(\omega) := \inf\{t > \overline{\tau}_a^{(r)} : \omega(t) = a\}, \quad r \geq 1, \tag{12.26}$$

with the usual convention that the infimum of an empty set of numbers is ∞.

Similarly, in the context of the simple random walk, put $\Omega = \mathbb{Z}^\infty = \{\omega = (\omega_0, \omega_1, \ldots) : \omega_n \in \mathbb{Z}, \forall n \geq 1\}$, and define

$$\overline{\tau}_a^{(1)}(\omega) \equiv \overline{\tau}_a(\omega) := \inf\{n \geq 0 : \omega_n = a\}, \tag{12.27}$$

and, recursively,

$$\overline{\tau}_a^{(r+1)}(\omega) := \inf\{n > \overline{\tau}_a^{(r)} : \omega_n = a\}, \quad r \geq 1. \tag{12.28}$$

Example 1 *(Recurrence of Simple Symmetric Random Walk)***.** Consider the simple symmetric random walk $S^x := \{S_n^x = x + S_n^0 : n \geq 0\}$ on $\mathbb{Z}$ started at x. Suppose one wishes to prove that $P_x(\overline{\tau}_y < \infty) = 1$ for $y \in \mathbb{Z}$. This may be obtained from the (ordinary) Markov property applied to $\varphi(x) := P_x(\overline{\tau}_y < \overline{\tau}_a)$, $a \leq x \leq y$. For $a < x < y$, conditioning on S_1^x, and writing $S_1^{x+} = \{S_{1+n}^x : n \geq 0\}$, we have

$$\begin{aligned}
\varphi(x) &= P_x(\overline{\tau}_y < \overline{\tau}_a) = P(\overline{\tau}_y \circ S^x < \overline{\tau}_a \circ S^x) \\
&= P(\overline{\tau}_y \circ S_1^{x+} < \overline{\tau}_a \circ S_1^{x+}) \\
&= \mathbb{E}_x P_{S_1^x}(\overline{\tau}_y < \overline{\tau}_a) = \mathbb{E}\varphi(S_1^x) \\
&= \mathbb{E}(\mathbf{1}_{[S_1^x = x+1]}\varphi(x+1) + \mathbf{1}_{[S_1^x = x-1]}\varphi(x-1)) \\
&= \frac{1}{2}\varphi(x+1) + \frac{1}{2}\varphi(x-1),
\end{aligned} \tag{12.29}$$

with boundary values $\varphi(y) = 1$, $\varphi(a) = 0$. Solving, one obtains $\varphi(x) = (x-a)/(y-a)$. Thus $P_x(\tau_y < \infty) = 1$ follows by letting $a \to -\infty$ using basic "continuity properties" of probability measures. Similarly, letting $y \to \infty$, one gets $P_x(\overline{\tau}_a < \infty) = 1$. Write $\overline{\eta}_a := \inf\{n \geq 1 : \omega_n = a\}$ for the **first return time to** a. Then $\overline{\eta}_a = \overline{\tau}_a$ on $\{\omega : \omega_0 \neq a\}$, and $\overline{\eta}_a > \overline{\tau}_a = 0$ on $\{\omega : \omega_0 = a\}$. By conditioning on S_1^x again, one has $P_x(\overline{\eta}_x < \infty) = \frac{1}{2}P_{x-1}(\overline{\tau}_x < \infty) + P_{x+1}P(\overline{\tau}_x < \infty) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1$. While this calculation required only 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 under P_x, the rth passage time to y, $\overline{\tau}_y^{(r)}$, is a.s. finite for every $r = 1, 2, \ldots$. First note that by the (ordinary) Markov property, $P_x(\tau_y < \infty) = 1\ \forall x$. To simplify notation, write $\tau_y^{(r)} = \overline{\tau}_y^{(r)} \circ S^x$, and $S^{x+}_{\tau_y^{(r)}} = \{S^{x+}_{\tau_y^{(r)}+n} : n \geq 0\}$ is then the after-$\tau_y^{(r)}$ process (for the random walk S^x). Applying the strong Markov property with respect to the stopping time $\tau_y^{(r)}$ one has, remembering that $S^x_{\tau_y^{(r)}} = y$,

$$\begin{aligned}
P_x(\overline{\tau}_y^{(r+1)} < \infty) &= P(\tau_y^{(r)} < \infty, \overline{\eta}_y \circ S^{x+}_{\tau_y^{(r)}} < \infty) \\
&= \mathbb{E}\big(\mathbf{1}_{[\tau_y^{(r)} < \infty]} P_y(\overline{\eta}_y < \infty)\big) \\
&= \mathbb{E}\big(\mathbf{1}_{[\tau_y^{(r)} < \infty]}\big) \cdot 1 \\
&= P_x(\overline{\tau}_y^{(r)} < \infty) = 1 \quad (r = 1, 2, \ldots),
\end{aligned} \tag{12.30}$$

by induction on r. If $x = y$, then $\overline{\tau}_x^{(1)}$ is replaced by $\overline{\eta}_x$. Otherwise, the proof remains the same. This is equivalent to the **recurrence** of the state y in the sense that

$$P(S_n^x = y \text{ for infinitely many } n) = P(\cap_{r=1}^{\infty}[\tau_y^{(r)} < \infty]) = 1. \tag{12.31}$$

Example 2 *(Boundary Value Distribution of Brownian Motion).* Let $B^x = \{B_t^x := x + B_t : t \geq 0\}$ be a one-dimensional standard Brownian motion started at $x \in [c, d]$ for $c < d$, and let $\tau_y = \overline{\tau}_y \circ B^x$. The stopping time $\tau_c \wedge \tau_d$ denotes the first time for B^x to reach the "boundary" states $\{c, d\}$, referred to as a **hitting time** for B^x. Define

$$\psi(x) := P(B^x_{\tau_c \wedge \tau_d} = c) \equiv P(\{B_t^x : t \geq 0\} \text{ reaches } c \text{ before } d), \qquad (c \leq x \leq d). \tag{12.32}$$

Fix $x \in (c, d)$ and $h > 0$ such that $[x - h, x + h] \subset (c, d)$. In contrast to the discrete-parameter case there is no "first step" to consider. It will be convenient to consider $\tau = \tau_{x-h} \wedge \tau_{x+h}$, i.e., τ is the first time $\{B_t^x : t \geq 0\}$ reaches $x - h$ or $x + h$. Then $P(\tau < \infty) = 1$, by the law of the iterated logarithm (see Exercise 5 for an alternative argument). Now,

$$\begin{aligned}
\psi(x) &= P(\{B_t^x : t \geq 0\} \text{ reaches } c \text{ before } d) = P(\{(B_\tau^{x+})_t : t \geq 0\} \text{ reaches } c \text{ before } d) \\
&= \mathbb{E}(P(\{(B_\tau^{x+})_t : t \geq 0\} \text{ reaches } c \text{ before } d \mid \mathcal{F}_\tau)).
\end{aligned} \tag{12.33}$$

The strong Markov property (Theorem 12.4) now gives that

$$\psi(x) = \mathbb{E}(\psi(B_\tau^x)), \tag{12.34}$$

so that by symmetry of Brownian motion, i.e., B^0 and $-B^0$ have the same distribution,

$$\begin{aligned}\psi(x) &= \psi(x-h)P(B_\tau^x = x-h) + \psi(x+h)P(B_\tau^x = x+h)\\ &= \psi(x-h)\frac{1}{2} + \psi(x+h)\frac{1}{2},\end{aligned} \tag{12.35}$$

where, by (12.32), $\psi(x)$ satisfies the boundary conditions $\psi(c) = 1$, $\psi(d) = 0$. Therefore,

$$\psi(x) = \frac{d-x}{d-c}. \tag{12.36}$$

Now, by (12.36) (see also Exercise 5),

$$P(\{B_t^x : t \geq 0\} \text{ reaches } d \text{ before } c) = 1 - \psi(x) = \frac{x-c}{d-c} \tag{12.37}$$

for $c \leq x \leq d$. It follows, on letting $d \uparrow \infty$ in (12.36), and $c \downarrow -\infty$ in (12.37) that

$$P_x(\overline{\tau}_y < \infty) = 1 \qquad \text{for all } x, y. \tag{12.38}$$

As another illustrative application of the strong Markov property one may derive a Cantor-like structure of the random set of zeros of Brownian motion as follows.

Example 3.

Proposition 12.5. With probability one, the set $\mathcal{Z} := \{t \geq 0 : B_t = 0\}$ of zeros of the sample path of a one dimensional standard Brownian motion, starting at 0, is uncountable, closed, unbounded, and has no isolated point. Moreover, $\mathcal{Z}$ a.s. has Lebesgue measure zero.

Proof. The law of iterated logarithm (LIL) may be applied as $t \downarrow 0$ to show that with probability one, $B_t = 0$ for infinitely many t in every interval $[0, \varepsilon]$. Since $t \mapsto B_t(\omega)$ is continuous, $\mathcal{Z}(\omega)$ is closed. Applying the LIL as $t \uparrow \infty$, it follows that $\mathcal{Z}(\omega)$ is unbounded a.s.

We will now show that for $0 < c < d$, the probability is zero of the event $A(c, d)$, say, that B has a single zero in $[c, d]$. For this consider the stopping time $\tau := \inf\{t \geq c : B_t = 0\}$. By the strong Markov property, B_τ^+ is a standard Brownian motion, starting at zero. In particular, τ is a point of accumulation of zeros from the right (a.s.). Also, $P(B_d = 0) = 0$. This implies $P(A(c, d)) = 0$. Considering all pairs of rationals c, d with $c < d$, it follows that $\mathcal{Z}$ has no isolated point outside a set of probability zero (see Exercise 4 for an alternate argument).

Finally, for each $T > 0$ let $H_T = \{(t, \omega) : 0 \le t \le T,\ B_t(\omega) = 0\} \subset [0, T] \times \Omega$. By the Fubini–Tonelli theorem, denoting the Lebesgue measure on $[0, \infty)$ by m, one has

$$(m \times P)(H_T) = \int_0^T \left\{ \int_\Omega \mathbf{1}_{\{\omega : B_t(\omega)=0\}} P(d\omega) \right\} dt = \int_0^T P(B_t = 0) dt = 0, \quad (12.39)$$

so that $m(\{t \in [0, T] : B_t(\omega) = 0\}) = 0$ for P-almost all ω. ■

The following general consequence of the Markov property can also be useful in the analysis of the (infinitesimal) fine-scale structure of Brownian motion and may be viewed as a corollary to Proposition 12.3. As a consequence, for example, one sees that for any given function $\varphi(t)$, $t > 0$, the event

$$D_\varphi := [B_t < \varphi(t) \text{ for all sufficiently small } t] \quad (12.40)$$

will certainly occur or is certain not to occur. Functions φ for which $P(D_\varphi) = 1$ are said to belong to the **upper class**. Thus $\varphi(t) = \sqrt{2t \log \log t}$ belongs to the upper class by the law of the iterated logarithm for Brownian motion (Theorem 11.5).

Proposition 12.6 ***(Blumenthal's Zero–One Law).*** With the notation of Proposition 12.3,

$$P(A) = 0 \text{ or } 1 \qquad \forall\, A \in \mathcal{F}_{0+}. \quad (12.41)$$

Proof. It follows from (the proof of) Proposition 12.3 that $\mathcal{F}_{s+}$ is independent of $\sigma\{B_{t+s} - B_s : t \ge 0\}$ $\forall\, s \ge 0$. Set $s = 0$ to conclude that $\mathcal{F}_{0+}$ is independent of $\sigma(B_t : t \ge 0) \supseteq \mathcal{F}_{0+}$. Thus $\mathcal{F}_{0+}$ is independent of $\mathcal{F}_{0+}$, so that $\forall\ A \in \mathcal{F}_{0+}$ one has $P(A) \equiv P(A \cap A) = P(A) \cdot P(A)$. ■

In addition to the strong Markov property, another powerful tool for the analysis of Brownian motion is made available by observing that both the processes $\{B_t : t \ge 0\}$ and $\{B_t^2 - t : t \ge 0\}$ are martingales. Thus one has available the optional sampling theory (Theorem 3.6).

Example 4 ***(Hitting by BM of a Two-Point Boundary).*** Let $\{B_t^x : t \ge 0\}$ be a one-dimensional standard Brownian motion starting at x, and let $0 < x < d$. Let τ denote the stopping time, $\tau = \inf\{t \ge 0 : B_t^x = c \text{ or } d\}$. Then writing $\psi(x) := P(\{B_t^x\}_{t \ge 0}$ reaches d before $c)$, one has (see (12.36))

$$\psi(x) = \frac{x - c}{d - c} \qquad c < x < d. \quad (12.42)$$

Applying the optional sampling theorem to the martingale $X_t := (B_t^x - x)^2 - t$, one gets $\mathbb{E}X_\tau = 0$, or $(d - x)^2 \psi(x) + (x - c)^2 (1 - \psi(x)) = \mathbb{E}\tau$, so that $\mathbb{E}\tau = [(d - x)^2 -$

$(x-c)^2]\psi(x) + (x-c)^2$, or

$$\mathbb{E}\tau = (d-x)(x-c). \tag{12.43}$$

Consider now a Brownian motion $\{Y_t^x : t \geq 0\}$ with nonzero drift μ and diffusion coefficient $\sigma^2 > 0$, starting at x. Then $\{Y_t^x - t\mu : t \geq 0\}$ is a martingale, so that (see Exercise 5) $\mathbb{E}(Y_\tau^x - \mu\tau) = x$, i.e., $d\psi_1(x) + c(1-\psi_1(x)) - \mu\mathbb{E}\tau = x$, or

$$(d-c)\psi_1(x) - \mu\mathbb{E}\tau = x - c, \tag{12.44}$$

where $\psi_1(x) = P(Y_\tau^x = d)$, i.e., the probability that $\{Y_t^x : t \geq 0\}$ reaches d before c. There are two unknowns, ψ_1 and $\mathbb{E}\tau$ in (12.44), so we need one more relation to solve for them. Consider the exponential martingale $Z_t := \exp\left\{\xi(Y_t^x - t\mu) - \frac{\xi^2\sigma^2}{2}\,t\right\}$ $(t \geq 1)$. Then $Z_0 = e^{\xi x}$, so that $e^{\xi x} = \mathbb{E}Z_\tau = \mathbb{E}\exp\{\xi(d-\tau\mu) - \xi^2\sigma^2\,\tau/2\}\mathbf{1}_{[Y_\tau^x=d]} + \mathbb{E}[\exp\{\xi(c-\tau\mu) - \xi^2\sigma^2\,\tau/2\}\mathbf{1}_{[Y_\tau^x=c]}]$. Take $\xi \neq 0$ such that the coefficient of τ in the exponent is zero, i.e., $\xi\mu + \xi^2\,\sigma^2/2 = 0$, or $\xi = -2\mu/\sigma^2$. Then optional stopping yields

$$\begin{aligned} e^{-2\mu x/\sigma^2} &= \exp\{\xi d\}\psi_1(x) + \exp\{\xi c\}(1-\psi_1(x)), \\ &= \psi_1(x)\left[\exp\left\{-\frac{2\mu d}{\sigma^2}\right\} - \exp\left\{-\frac{2\mu c}{\sigma^2}\right\}\right] + \exp\left\{-\frac{2\mu c}{\sigma^2}\right\}, \end{aligned}$$

or

$$\psi_1(x) = \frac{\exp\{-2\mu x/\sigma^2\} - \exp\{-2\mu c/\sigma^2\}}{\exp\{-\frac{2\mu d}{\sigma^2}\} - \exp\{-\frac{2\mu c}{\sigma^2}\}}. \tag{12.45}$$

One may use this to compute $\mathbb{E}\tau$:

$$\mathbb{E}\tau = \frac{(d-c)\psi_1(x) - (x-c)}{\mu}. \tag{12.46}$$

Checking the hypothesis of the optional sampling theorem for the validity of the relations (12.42)–(12.46) is left to Exercise 5.

Our main goal for this chapter is to derive a beautiful result of Skorokhod (1965) representing a general random walk (partial sum process) as values of a Brownian motion at a sequence of successive stopping times (with respect to an enlarged filtration). This will be followed by a proof of the functional central limit theorem (invariance principle) based on the Skorokhod embedding representation. Recall that for $c < x < d$,

$$P(\tau_d^x < \tau_c^x) = \frac{x-c}{d-c}, \tag{12.47}$$

where $\tau_a^x := \overline{\tau}_a(B^x) \equiv \inf\{t \geq 0 : B_t^x = a\}$. Also,

$$\mathbb{E}(\tau_c^x \wedge \tau_d^x) = (d - x)(x - c). \tag{12.48}$$

Write $\tau_a = \tau_a^0$, $B^0 = B = \{B_t : t \geq 0\}$. Consider now a two-point distribution $F_{u,v}$ with support $\{u, v\}$, $u < 0 < v$, having mean zero. That is, $F_{u,v}(\{u\}) = v/(v-u)$ and $F_{u,v}(\{v\}) = -u/(v-u)$. It follows from (12.47) that with $\tau_{u,v} = \tau_u \wedge \tau_v$, $B_{\tau_{u,v}}$ has distribution $F_{u,v}$ and, in view of (12.48),

$$\mathbb{E}\tau_{u,v} = -uv = |uv|. \tag{12.49}$$

In particular, the random variable $Z := B_{\tau_{u,v}}$ with distribution $F_{u,v}$ is naturally **embedded** in the Brownian motion. We will see by the theorem below that any given nondegenerate distribution F with mean zero may be similarly embedded by randomizing over such pairs (u, v) to get a random pair (U, V) such that $B_{\tau_{U,V}}$ has distribution F, and $\mathbb{E}\tau_{U,V} = \int_{(-\infty,\infty)} x^2 F(dx)$, the variance of F. Indeed, this is achieved by the distribution γ of (U, V) on $(-\infty, 0) \times (0, \infty)$ given by

$$\gamma(du\,dv) = \theta(v - u)F_-(du)F_+(dv), \tag{12.50}$$

where F_+ and F_- are the restrictions of F to $(0, \infty)$ and $(-\infty, 0)$, respectively. Here θ is the normalizing constant given by

$$1 = \theta\left[\left(\int_{(0,\infty)} vF_+(dv)\right)F_-((-\infty,0)) + \left(\int_{(-\infty,0)} (-u)F_-(du)\right)F_+(0,\infty)\right],$$

or, noting that the two integrals are each equal to $\frac{1}{2}\int_{-\infty}^{\infty} |x|F(dx)$ since the mean of F is zero, one has

$$1/\theta = \left(\frac{1}{2}\int_{-\infty}^{\infty} |x|F(dx)\right)[1 - F(\{0\})]. \tag{12.51}$$

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined (1) a standard Brownian motion $B \equiv B^0 = \{B_t : t \geq 0\}$, and (2) a sequence of i.i.d. pairs (U_i, V_i) independent of B, with the common distribution γ above. Let $\mathcal{F}_t := \sigma\{B_s : 0 \leq s \leq t\} \vee \sigma\{(U_i, V_i) : i \geq 1\}$, $t \geq 0$. Define the $\{\mathcal{F}_t : t \geq 0\}$-stopping times (Exercise 12)

$$\begin{aligned} &T_0 \equiv 0, \quad T_1 := \inf\{t \geq 0 : B_t = U_1 \text{ or } V_1\}, \\ &\qquad T_{i+1} := \inf\{t > T_i : B_t = B_{T_i} + U_{i+1} \text{ or } B_{T_i} + V_{i+1}\}\,(i \geq 1). \end{aligned} \tag{12.52}$$

Theorem 12.7 *(Skorokhod Embedding).* Assume that F has mean zero and finite variance. Then (a) B_{T_1} has distribution F, and $B_{T_{i+1}} - B_{T_i}$ $(i \geq 0)$ are i.i.d. with common distribution F, and (b) $T_{i+1} - T_i$ $(i \geq 0)$ are i.i.d. with

$$\mathbb{E}\left(T_{i+1} - T_i\right) = \int_{(-\infty,\infty)} x^2 F(dx). \tag{12.53}$$

Proof. (a) Given (U_1, V_1), the conditional probability that $B_{T_1} = V_1$ is $\frac{-U_1}{V_1 - U_1}$. Therefore, for all $x > 0$,

$$\begin{aligned} P\left(B_{T_1} > x\right) &= \theta \int_{\{v>x\}} \int_{(-\infty,0)} \frac{-u}{v-u} \cdot (v-u) F_-(du) F_+(dv) \\ &= \theta \int_{\{v>x\}} \left\{ \int_{(-\infty,0)} (-u) F_-(du) \right\} F_+(dv) = \int_{\{v>x\}} F_+(dv), \end{aligned} \tag{12.54}$$

since $\int_{(-\infty,0)}(-u)F_-(du) = \frac{1}{2}\int |x| F(dx) = 1/\theta$. Thus the restriction of the distribution of B_{T_1} on $(0,\infty)$ is F_+. Similarly, the restriction of the distribution of B_{T_1} on $(-\infty, 0)$ is F_-. It follows that $P(B_{T_1} = 0) = F(\{0\})$. This shows that B_{T_1} has distribution F. Next, by the strong Markov property, the conditional distribution of $B^+_{T_i} \equiv \{B_{T_i+t} : t \geq 0\}$, given $\mathcal{F}_{T_i}$, is $P_{B_{T_i}}$ (where P_x is the distribution of B^x). Therefore, the conditional distribution of $B^+_{T_i} - B_{T_i} \equiv \{B_{T_i+t} - B_{T_i};\, t \geq 0\}$, given $\mathcal{F}_{T_i}$, is P_0. In particular, $Y_i := \{(T_j, B_{T_j}) : 1 \leq j \leq i\}$ and $X^i := B^+_{T_i} - B_{T_i}$ are independent. Since Y_i and X^i are functions of $B \equiv \{B_t : t \geq 0\}$ and $\{(U_j, V_j);\, 1 \leq j \leq i\}$, they are both independent of (U_{i+1}, V_{i+1}). Since $\tau^{(i+1)} := T_{i+1} - T_i$ is the first hitting time of $\{U_{i+1}, V_{i+1}\}$ by X^i, it now follows that (1) $(T_{i+1} - T_i \equiv \tau^{(i+1)}, B_{T_{i+1}} - B_{T_i} \equiv X^i_{\tau^{(i+1)}})$ is independent of $\{(T_j, B_{T_j}) : 1 \leq j \leq i\}$, and (2) $(T_{i+1} - T_i, B_{T_{i+1}} - B_{T_i})$ has the same distribution as (T_1, B_{T_1}).

(b) It remains to prove (12.53). But this follows from (12.49):

$$\begin{aligned} \mathbb{E}T_1 &= \theta \int_{(-\infty,0)} \int_{(0,\infty)} (-uv)(v-u) F_-(du) F_+(dv) \\ &= \theta \left[\int_{(0,\infty)} v^2 F_+(dv) \cdot \int_{(-\infty,0)} (-u) F_-(du) + \int_{(-\infty,0)} u^2 F_-(du) \cdot \int_{(0,\infty)} v F_+(dv) \right] \\ &= \int_{(0,\infty)} v^2 F_+(dv) + \int_{(-\infty,0)} u^2 F_-(du) = \int_{(-\infty,\infty)} x^2 F(dx). \qquad \blacksquare \end{aligned}$$

We now present an elegant proof of **Donsker's invariance principle**, or **functional central limit theorem**, using Theorem 12.7. Consider a sequence of i.i.d. random variables Z_i $(i \geq 1)$ with common distribution having mean zero and variance 1. Let $S_k = Z_1 + \cdots + Z_k$ $(k \geq 1)$, $S_0 = 0$, and define the polygonal random function $S^{(n)}$ on $[0,1]$ as follows:

$$S^{(n)}_t := \frac{S_{k-1}}{\sqrt{n}} + n\left(t - \frac{k-1}{n}\right) \frac{S_k - S_{k-1}}{\sqrt{n}}$$

$$\text{for } t \in \left[\tfrac{k-1}{n}, \tfrac{k}{n}\right], 1 \leq k \leq n. \tag{12.55}$$

That is, $S_t^{(n)} = \frac{S_k}{\sqrt{n}}$ at points $t = \frac{k}{n}$ $(0 \leq k \leq n)$, and $t \mapsto S_t^{(n)}$ is linearly interpolated between the endpoints of each interval $\left[\frac{k-1}{n}, \frac{k}{n}\right]$.

***Theorem 12.8** (Invariance Principle).* $S^{(n)}$ converges in distribution to the standard Brownian motion, as $n \to \infty$.

Proof. Let T_k, $k \geq 1$, be as in Theorem 12.7, defined with respect to a standard Brownian motion $\{B_t : t \geq 0\}$. Then the random walk $\{S_k : k = 0, 1, 2, \ldots\}$ has the same distribution as $\{\widetilde{S}_k := B_{T_k} : k = 0, 1, 2, \ldots\}$, and therefore, $S^{(n)}$ has the same distribution as $\widetilde{S}^{(n)}$ defined by $\widetilde{S}_{k/n}^{(n)} := n^{-\frac{1}{2}} B_{T_K}$ $(k = 0, 1, \ldots, n)$ and with linear interpolation between k/n and $(k+1)/n$ $(k = 0, 1, \ldots, n-1)$. Also, define, for each $n = 1, 2, \ldots$, the standard Brownian motion $\widetilde{B}_t^{(n)} := n^{-\frac{1}{2}} B_{nt}$, $t \geq 0$. We will show that

$$\max_{0 \leq t \leq 1} \left| \widetilde{S}_t^{(n)} - \widetilde{B}_t^{(n)} \right| \longrightarrow 0 \quad \text{in probability as } n \to \infty, \tag{12.56}$$

which implies the desired weak convergence. Now

$$\begin{aligned} \max_{0 \leq t \leq 1} \left| \widetilde{S}_t^{(n)} - \widetilde{B}_t^{(n)} \right| &\leq n^{-\frac{1}{2}} \max_{1 \leq k \leq n} |B_{T_k} - B_k| \\ &\quad + \max_{0 \leq k \leq n-1} \left\{ \max_{\frac{k}{n} \leq t \leq \frac{k+1}{n}} \left| \widetilde{S}_t^{(n)} - \widetilde{S}_{k/n}^{(n)} \right| + n^{-\frac{1}{2}} \max_{k \leq t \leq k+1} |B_t - B_k| \right\} \\ &= I_n^{(1)} + I_n^{(2)} + I_n^{(3)}, \quad \text{say.} \end{aligned} \tag{12.57}$$

Now, writing $\tilde{Z}_k = \tilde{S}_k - \tilde{S}_{k-1}$, it is simple to check (Exercise 13) that as $n \to \infty$,

$$\begin{aligned} &I_n^{(2)} \leq n^{-\frac{1}{2}} \max\{|\tilde{Z}_k| : 1 \leq k \leq n\} \to 0 \quad \text{in probability,} \\ &I_n^{(3)} \leq n^{-\frac{1}{2}} \max_{0 \leq k \leq n-1} \max\{|B_t - B_k| : k \leq t \leq k+1\} \to 0 \quad \text{in probability.} \end{aligned}$$

Hence we need to prove, as $n \to \infty$,

$$I_n^{(1)} := n^{-\frac{1}{2}} \max_{1 \leq k \leq n} |B_{T_k} - B_k| \longrightarrow 0 \quad \text{in probability.} \tag{12.58}$$

Since $T_n/n \to 1$ a.s., by SLLN, it follows that (Exercise 13)

$$\varepsilon_n := \max_{1 \leq k \leq n} \left| \frac{T_k}{n} - \frac{k}{n} \right| \longrightarrow 0 \quad \text{as } n \to \infty \text{ (almost surely).} \tag{12.59}$$

In view of (12.59), there exists for each $\varepsilon > 0$ an integer n_ε such that $P(\varepsilon_n < \varepsilon) > 1 - \varepsilon$ for all $n \geq n_\varepsilon$. Hence with probability greater than $1 - \varepsilon$ one has for all $n \geq n_\varepsilon$ the estimate (writing $\overset{d}{=}$ for equality in distribution)

$$\begin{aligned} I_n^{(1)} &\leq \max_{\substack{|s-t| \leq n\varepsilon, \\ 0 \leq s,t \leq n+n\varepsilon}} n^{-\frac{1}{2}} |B_s - B_t| = \max_{\substack{|s-t| \leq n\varepsilon, \\ 0 \leq s,t \leq n(1+\varepsilon)}} \left| \widetilde{B}^{(n)}_{s/n} - \widetilde{B}^{(n)}_{t/n} \right| \\ &= \max_{\substack{|s'-t'| \leq \varepsilon, \\ 0 \leq s',t' \leq 1+\varepsilon}} \left| \widetilde{B}^{(n)}_{s'} - \widetilde{B}^{(n)}_{t'} \right| \overset{d}{=} \max_{\substack{|s'-t'| \leq \varepsilon, \\ 0 \leq s',t' \leq 1+\varepsilon}} |B_{s'} - B_{t'}| \\ &\longrightarrow 0 \quad \text{as } \varepsilon \downarrow 0, \end{aligned}$$

by the continuity of $t \to B_t$. Given $\delta > 0$ one may then choose $\varepsilon = \varepsilon_\delta$ such that for all $n \geq n(\delta) := n_{\varepsilon_\delta}$, $P(I_n^{(1)} > \delta) < \delta$. Hence $I_n^{(1)} \to 0$ in probability. ■

For another application of Skorokhod embedding let us see how to obtain a **law of the iterated logarithm** (LIL) for sums of i.i.d. random variables using the LIL for Brownian motion.

Theorem 12.9 *(Law of the Iterated Logarithm).* Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with $\mathbb{E}X_1 = 0$, $0 < \sigma^2 := \mathbb{E}X_1^2 < \infty$, and let $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Then with probability one,

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = 1.$$

Proof. By rescaling if necessary, one may take $\sigma^2 = 1$ without loss of generality. In view of Skorokhod embedding one may replace the sequence $\{S_n : n \geq 0\}$ by the embedded random walk $\{\tilde{S}_n = B_{T_n} : n \geq 0\}$. By the SLLN one also has $\frac{T_n}{n} \to 1$ a.s. as $n \to \infty$. In view of the law of the iterated logarithm for Brownian motion, it is then sufficient to check that $\frac{\tilde{S}_{[t]} - B_t}{\sqrt{t \log\log t}} \to 0$ a.s. as $t \to \infty$. From $\frac{T_n}{n} \to 1$ a.s., it follows for given $\varepsilon > 0$ that with probability one, $\frac{1}{1+\varepsilon} < \frac{T_{[t]}}{t} < 1 + \varepsilon$ for all t sufficiently large. Let $t_n = (1+\varepsilon)^n$, $n = 1, 2, \ldots$. Then for $t_n \leq t \leq t_{n+1}$, for some $n \geq 1$, one has

$$\begin{aligned} M_t &:= \max\left\{ |B_s - B_t| : \frac{t}{1+\varepsilon} \leq s \leq t(1+\varepsilon) \right\} \\ &\leq \max\left\{ |B_s - B_t| : \frac{t}{1+\varepsilon} \leq s \leq t \right\} + \max\left\{ |B_s - B_t| : t \leq s \leq t(1+\varepsilon) \right\} \\ &\leq \max\left\{ |B_s - B_{t_n}| : \frac{t_n}{1+\varepsilon} \leq s \leq t_{n+1} \right\} + \max\left\{ |B_s - B_{t_n}| : t_n \leq s \leq t_{n+1} \right\} \\ &\leq 2M_{t_n} + 2M_{t_{n+1}}. \end{aligned}$$

Since $t_{n+1} - t_{n-1} = \gamma t_{n-1} = \frac{\gamma}{1+\varepsilon} t_n$, where $\gamma = (1+\varepsilon)^2 - 1$, it follows from the scaling property of Brownian motion, using Lévy's Inequality and Feller's tail probability estimate, that

$$\begin{aligned} P\left(M_{t_n} > \sqrt{3\frac{\gamma}{1+\varepsilon} t_n \log\log t_n}\right) &= P\left(\max_{0\le u\le 1} |B_u| > \sqrt{3\log\log t_n}\right) \\ &\le 4P\left(B_1 \ge \sqrt{3\log\log(t_n)}\right) \\ &\le \frac{4}{\sqrt{3\log\log t_n}} \exp\left(-\frac{3}{2}\log\log t_n\right) \\ &\le cn^{-\frac{3}{2}} \end{aligned}$$

for a constant $c > 0$. Summing over n, it follows from the Borel–Cantelli lemma I that with probability one, $M_{t_n} \le \sqrt{3\frac{\gamma}{1+\varepsilon} t_n \log\log t_n}$ for all but finitely many n. Since a.s. $\frac{1}{1+\varepsilon} < \frac{T_{[t]}}{t} < 1+\varepsilon$ for all t sufficiently large, one has that, with probability one,

$$\limsup_{t\to\infty} \frac{|\tilde{S}_{[t]} - B_t|}{\sqrt{t\log\log t}} \le \sqrt{3\frac{\gamma}{1+\varepsilon}}.$$

Letting $\varepsilon \downarrow 0$ one has $\frac{\gamma}{1+\varepsilon} \to 0$, establishing the desired result. ■

EXERCISES

Exercise Set XII

1. (i) If τ_1, τ_2 are stopping times, show that $\tau_1 \vee \tau_2$ and $\tau_1 \wedge \tau_2$ are stopping times. (ii) If $\tau_1 \le \tau_2$ are stopping times, show that $\mathcal{F}_{\tau_1} \subseteq \mathcal{F}_{\tau_2}$.
2. (i) Extend the Markov property for one-dimensional Brownian motion (Proposition 12.2) to k-dimensional Brownian motion with respect to a given filtration. (ii) Prove parts (ii), (iii) of Proposition 12.3.
3. Suppose that X, Y, Z are three random variables with values in arbitrary measurable spaces $(S_i, \mathcal{S}_i)$, $i = 1, 2, 3$. Assume that regular conditional distributions exist; see Chapter II for general conditions. Show that $\sigma(Z)$ is conditionally independent of $\sigma(X)$ given $\sigma(Y)$ if and only if the conditional distribution of Z given $\sigma(Y)$ a.s. coincides with the conditional distribution of Z given $\sigma(X, Y)$.
4. Prove that the event $A(c, d)$ introduced in the proof of Proposition 12.5 is measurable, i.e., the event $[\tau < d, B_t > 0\ \forall \tau < t \le d]$ is measurable.
5. Check the conditions for the application of the optional sampling theorem (Theorem 3.6(b)) for deriving (12.42)–(12.46). [*Hint*: For Brownian motion $\{Y_t^x : t \ge 0\}$ with a drift μ and diffusion coefficient $\sigma^2 > 0$, let $Z_1 = Y_1^x - x$, $Z_k = Y_k^x - Y_{k-1}^x (k \ge 1)$. Then $Z_1, Z_2, \ldots$ are i.i.d. and Corollary 3.8 applies with $a = c, b = d$. This proves $P(\tau < \infty) = 1$. The uniform integrability of $\{Y_{t\wedge\tau}^x : t \ge 0\}$ is immediate, since $c \le Y_{t\wedge\tau}^x \le d$ for all $t \ge 0$.]

6. Let $u' < 0 < v'$. Show that if $F = F_{u',v'}$ is the mean-zero two-point distribution concentrated at $\{u', v'\}$, then $P((U,V) = (u',v')) = 1$ in the Skorokhod embedding of F defined by $\gamma(du\,dv)$.

7. Given any distribution F on $\mathbb{R}$, let $\tau := \inf\{t \ge 0 : B_t = Z\}$, where Z is independent of $B = \{B_t : t \ge 0\}$ and has distribution F. Then $B_\tau = Z$. One can thus embed a random walk with (a nondegenerate) step distribution F (say, with mean zero) in different ways. However, show that $\mathbb{E}\tau = \infty$. [*Hint*: The stable distribution of $\tau_a := \inf\{t \ge 0 : B_t = a\}$ has infinite mean for every $a \ne 0$. To see this, use Corollary 11.3 to obtain $P(\tau_a > t) \ge 1 - 2P(B_t > a) = P(|B_t| \le a) = P(|B_1| \le \frac{a}{\sqrt{t}})$, whose integral over $[0,\infty)$ is divergent.]

8. Prove that $\varphi(\lambda) := \mathbb{E}\exp\{\lambda\tau_{u,v}\} \le \mathbb{E}\exp\{\lambda\tau_{-a,a}\} < \infty$ for $\lambda < \lambda_0(a)$ for some $\lambda_0(a) > 0$, where $a = \max\{-u, v\}$. Here $\tau_{u,v}$ is the first passage time of standard Brownian motion to $\{u, v\}$, $u < 0 < v$. [*Hint*: Use Corollary 3.8 with $X_n := B_n - B_{n-1}$ $(n \ge 1)$.]

9. (i) Show that for every $\lambda \ge 0$, $X_t := \exp\{\sqrt{2\lambda}B_t - \lambda t\}$, $t \ge 0$, is a martingale.

   (ii) Use the optional sampling theorem to prove $\varphi(-\lambda) = 2\left(e^{\sqrt{2\lambda}\,a} + e^{-\sqrt{2\lambda}\,a}\right)^{-1}$, where $\varphi(-\lambda) = \mathbb{E}\exp(-\lambda\tau_{-a,a})$, in the notation of the previous exercise.

10. Refer to the notation of Theorem 12.8.

    (i) Prove that $T_i - T_{i-1}$ $(i \ge 1)$ has a finite moment-generating function in a neighborhood of the origin if F has compact support.

    (ii) Prove that $\mathbb{E}T_1^2 < \infty$ if $\int |z|^5 F(dz) < \infty$. [*Hint*: $\tau_{u,v} \le \tau_{-a,a}$ with $a := \max\{-u, v\} \le v - u$ and $\mathbb{E}\tau^2_{U,V} \le c\theta \int (v-u)^5 F_+(dv)F_-(du)$ for some $c > 0$.]

11. In Theorem 12.7 suppose F is a symmetric distribution. Let X_i $(i \ge 1)$ be i.i.d. with common distribution F and independent of $\{B_t : t \ge 0\}$. Let $\widetilde{T}_1 := \inf\{t \ge 0 : B_t \in \{-X_1, X_1\}\}$, $\widetilde{T}_i := \widetilde{T}_{i-1} + \inf\{t \ge 0 : B_{\widetilde{T}_{i-1}+t} \in \{-X_i, X_i\}\}$ $(i \ge 1)$, $\widetilde{T}_0 = 0$.

    (i) Show that $B_{\widetilde{T}_i} - B_{\widetilde{T}_{i-1}}$ $(i \ge 1)$ are i.i.d. with common distribution F, and $\widetilde{T}_i - \widetilde{T}_{i-1}$ $(i \ge 1)$ are i.i.d.

    (ii) Prove that $\mathbb{E}\widetilde{T}_1 = \mathbb{E}X_1^2$, and $\mathbb{E}\widetilde{T}_1^2 = c\mathbb{E}X_1^4$, where c is a constant to be computed.

    (iii) Compute $\mathbb{E}e^{-\lambda\widetilde{T}_1}$ for $\lambda \ge 0$.

12. Prove that T_i $(i \ge 0)$ defined by (12.52) are $\{\mathcal{F}_t\}$–stopping times, where $\mathcal{F}_t$ is as defined there.

13. (i) Let Z_k, $k \ge 1$, be i.i.d. with finite variance. Prove that $n^{-\frac{1}{2}}\max\{|Z_k| : 1 \le k \le n\} \to 0$ in probability as $n \to \infty$. [*Hint*: $nP(Z_1 > \sqrt{n}\,\varepsilon) \le \frac{1}{\varepsilon^2}\,\mathbb{E}Z_1^2\mathbf{1}[|Z_1| \ge \sqrt{n}\,\varepsilon]$, $\forall\,\varepsilon > 0$].

    (ii) Derive (12.59) [*Hint*: $\varepsilon_n = \max_{1\le k\le n}|\frac{T_k}{k} - 1| \cdot \frac{k}{n} \le \left\{\max_{1\le k\le k_0}|\frac{T_k}{k} - 1|\right\} \cdot \frac{k_0}{n} + \max_{k\ge k_0}\left|\frac{T_k}{k} - 1\right| \;\forall\, k_0 = 1, 2, \ldots$.]

CHAPTER XIII

A Historical Note on Brownian Motion

Historically, the mathematical roots of Brownian motion lie in the central limit theorem (CLT). The first CLT seems to have been obtained in 1733 by DeMoivre[1] for the normal approximation to the binomial distribution (i.e., sum of i.i.d. Bernoulli 0 or 1-valued random variables). In his 1812 treatise Laplace[2] obtained the far reaching generalization to sums of arbitrary independent and identically distributed random variables having finite moments of all orders. Although by the standards of rigor of present day mathematics Laplace's derivation would not be considered complete, the essential ideas behind this remarkable result may be found in his work. The first rigorous proof[3] of the CLT was given by Lyapounov almost 100 years later using characteristic functions under the Lyapounov condition for sums of independent, but not necessarily identically distributed, random variables having finite $(2+\delta)$th moments for some $\delta > 0$. This moment condition was relaxed in 1922 by Lindeberg[4] to prove the more general CLT, and in 1935, Feller[5] showed that the conditions are necessary (as well as sufficient), under uniform asymptotic negligibility of summands. The most

[1] DeMoivre's normal approximation to the binomial first appeared in a pamphlet "Approximatio ad summam terminorum binomii" in 1733. It appeared in book form in the 1756 edition of the Doctrine of Chance, London.

[2] Laplace, P.-S. (1812), "Théorie Analytique des Probabilités", Paris.

[3] Lyapunov, A.M. (1901). Nouvelle forme du théorème sur la limite de probabilités. *Mem. Acad. Imp. Sci. St.-Petersberg* **12** (5), 1–24.

[4] Lindeberg, J.W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Zeitschr.* **15**, 211–225.

[5] Feller, W. (1935). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. *Math. Zeitschr.* **40**, 521–559. Also, ibid (1937), **42**, 301–312.

popular form of the CLT is that for i.i.d. summands with finite second moments due to Paul Lévy.[6]

There are not many results in mathematics that have had such a profound impact as the CLT, not only on probability and statistics but also on many other branches of mathematics, as well as the natural and physical sciences and engineering as a whole. The idea of a stochastic process $\{B_t : t \geq 0\}$ that has independent Gaussian increments also derives from it. One may consider an infinite i.i.d. sequence $\{X_m : m \geq 1\}$ with finite second moments as in the CLT, and consider sums $S_n, S_{2n} - S_n, S_{3n} - S_{2n}, \ldots$, over consecutive disjoint blocks of n of these random variables X_m having mean μ and variance σ^2. The block sums are independent, each approximately Gaussian with mean $n\mu$ and variance $n\sigma^2$. If one scales the sums as $\frac{S_n - n\mu}{\sigma\sqrt{n}}, \frac{S_{2n} - S_n - n\mu}{\sigma\sqrt{n}} \ldots$, then in the limit one should get a process with independent Gaussian increments. If time is scaled so that one unit of time in the new *macroscopic* scale is equal to n units of time in the old scale, the $B_1, B_2 - B_1, B_3 - B_2, \ldots$ are independent Gaussian $\Phi_{0,1}$. Brownian motion is precisely such a process, but constructed for all times $t \geq 0$ and having continuous sample paths. The conception of such a process was previously introduced in a1900 PhD thesis by Bachelier[7] as a model for the movements of stock prices.

Brownian motion is named after the nineteenth-century botanist Robert Brown, who observed under the microscope perpetual irregular motions exhibited by small grains or particles of the size of colloidal molecules immersed in a fluid. Brown[8] himself credited earlier scientists for having made similar observations. After some initial speculation that the movements are those of living organisms was discounted, the movements were attributed to inherent molecular motions. Independently of this debate and unaware of the massive experimental observations that had been made concerning this matter, Einstein[9] published a paper in 1905 in which he derived the *diffusion equation*

$$\frac{\partial C(t,x)}{\partial t} = D\left(\frac{\partial^2 C(t,x)}{\partial x_1^2} + \frac{\partial^2 C(t,x)}{\partial x_2^2} + \frac{\partial^2 C(t,x)}{\partial x_3^2}\right), \quad x = (x_1, x_2, x_3), \qquad (13.1)$$

for the concentration $C(t,x)$ of large solute molecules of uniform size and spherical shape in a stationary liquid at a point x at time t. The argument (at least implicit in the above article) is that a solute molecule is randomly displaced frequently by collisions with the molecules of the surrounding liquid. Regarding the successive

[6] Lévy, P. (1925).

[7] Bachelier, L. (1900). Théorie de las spéculation. *Ann. Sci. École Norm. Sup.* **17**, 21–86. (In: *The Random Character of Stock Market Prices*, Paul H. Cootner, ed. MIT Press, 1964).

[8] Brown, R. (1828). A brief account of microscopical observations made in the months of June, July, and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philos. Magazine N.S.* **14**, 161–173.

[9] Einstein, A. (1905). On the movement of small particles suspended in a stationary liquid demanded by the molecular–kinetic theory of heat. *Ann. der Physik* **17**, 549.

displacements as independent (and identically distributed) with mean vector zero and dispersion matrix $\mathrm{Diag}(d, d, d)$, one deduces a Gaussian distribution of the position of the solute molecule at time t with mean vector zero and a dispersion matrix $2t$ $\mathrm{Diag}(D, D, D)$, where $2D = fd$ with f as the average number of collisions, or displacements, per unit time. The law of large numbers (assuming that the different solute molecules move independently) then provides a Gaussian concentration law that is easily seen to satisfy the equation (13.1), away from the boundary. It is not clear that Laplace was aware of the profound fact that the operator $\Delta = \sum_1^3 \partial^2/\partial x_i^2$ in (13.1) bearing his name is intimately related to the central limit theorem he had derived.

Apprised of the experimental evidence concerning the so-called Brownian movement, Einstein titled his next article[10] on the subject, "On the theory of the Brownian movement." In addition to deriving the form of the equation (13.1), Einstein used classical thermodynamics, namely the Maxwell–Boltzmann steady-state (Gaussian) velocity distribution and Stokes' law of hydrodynamics (for the frictional force on a spherical particle immersed in a liquid) to express the *diffusion coefficient* D as $D = kT/3\pi\eta a$, where a is the radius of the spherical solute molecule, η is the coefficient of viscosity, T is the temperature, and k is the Boltzmann constant. In particular, the *physical parameters* are embodied in a *statistical parameter*. Based on this derivation, Jean Baptiste Perrin[11] estimated k or, equivalently, Avogadro's number, for which he was awarded the Nobel Prize in 1926. Meanwhile, in 1923, Wiener[12] proved that one may take Brownian paths to be continuous almost surely. That is, he constructed the probability measure Q, the so-called *Wiener measure* on $C[0, \infty)$, extending the normal distribution to infinitely many dimensions in the sense that the coordinate process $X_t(\omega) := \omega(t)$, $\omega \in C[0, \infty)$, $t \geq 0$, has independent Gaussian increments, namely, $X_{t+s} - X_t$ has the normal distribution $\Phi_{0,s} \equiv N(0, s)$, $\forall\ 0 \leq t < \infty$, $s > 0$, and $\{X_{t_{i+1}} - X_{t_i} : i = 1, 2, \ldots, m-1\}$ are independent $\forall\ 0 \leq t_1 < t_2 < \cdots < t_m$ ($\forall\ m > 1$). This was a delicate result, especially since the Brownian paths turned out to have very little smoothness beyond continuity. Indeed, in 1933 it was shown by Paley, Wiener, and Zygmund[13] that with probability one, a Brownian path is continuous but nowhere differentiable. This says that a Brownian particle has no velocity, confirming some remarkable empirical observations in the early physics of Brownian motion. In his monograph "Atoms", Perrin exclaims: "The trajectories are confused and complicated so often and so rapidly that it is impossible to follow them; the trajectory actually measured is very much simpler and shorter than the real one. Similarly, the apparent mean speed of a grain during a given time varies *in the wildest way* in magnitude and direction, and does not tend to a limit as the time taken for an observation decreases, as may be easily shown by noting, in the camera lucida, the

[10] Einstein, A. (1906). On the theory of the Brownian movement. *Ann. der Physik* **19**, 371–381. English translation in *Investigations on the Theory of the Brownian Movement* (R. Fürth, ed.), Dover, 1954.

[11] Jean Perrin, *Atoms*, Ox Bow Press, 1990 (French original, 1913).

[12] Wiener, N. (1923). Differential space. *J. Math. Phys.* **2**, 131–174.

[13] Paley, R.E.A.C., Wiener, N. and Zygmund, A. (1933). Notes on random functions. *Math. Zietschr.* **37**, 647–668.

positions occupied by a grain from minute to minute, and then every five seconds, or, better still, by photographing them every twentieth of a second, as has been done by Victor Henri Comandon, and de Broglie when kinematographing the movement. It is impossible to fix a tangent, even approximately, at any point on a trajectory, and we are thus reminded of the continuous underived functions of the mathematicians."

A more dynamical theory of Brownian (particle) motion was given by Ornstein and Uhlenbeck,[14] following the turn-of-the-century work of Langevin.[15]

The so-called Langevin equation used by Ornstein and Uhlenbeck is a *stochastic differential equation* given (in one dimension) by

$$dv(t) = -\beta v(t)dt + \sigma dB(t), \tag{13.2}$$

where $v(t)$ is the velocity of a Brownian molecule of mass m, $-m\beta v$ is the frictional force on it, and $\sigma^2 = 2\beta^2 D$ (D as above). By integrating $v(t)$ one gets a differentiable model of the Brownian molecule. If β and $\sigma^2 \to \infty$ such that $s^2/2\beta^2 = D$ remains a constant, then the position process converges to Einstein's model of Brownian motion (with variance parameter $2D$), providing a scale range for which the models approximately agree.[16] Within the framework of stochastic differential equations one sees that the *steady state velocity distribution* for the Langevin equation is a Gaussian distribution. On physical grounds this can be equated with the Maxwell–Boltzmann velocity distribution known from statistical mechanics and thermodynamics. In this way one may obtain Einstein's fundamental relationship between the physical parameters and statistical parameters mentioned above.

Brownian motion is a central notion throughout the theoretical development of stochastic processes and its applications. This rich history and its remarkable consequences are brought to life under several different guises in major portions of the theory of stochastic processes.

[14] Uhlenbeck, G.E. and Ornstein, L.S. (1930). On the theory of Brownian motion. *Phys. Rev.* **36**, 823–841. Reprinted in *Selected Papers on Noise and Stochastic Processes* (1954). (N. Wax, ed.), Dover. Also see Chandrasekhar, S. (1943). Stochastic problems in physics and astronomy. *Rev. Modern Physics* **15**, 2–91. Reprinted in *Selected Papers on Noise and Stochastic Processes* (1954) (N. Wax, ed.), Dover.

[15] Langevin, P. (1908). Sur La théorie du movement brownien. *C.R. Acad. Sci. Paris* **146**, 530–533.

[16] For a complete dynamical description see Nelson, E. (1967). *Dynamical Theories of Brownian Motion.* Princeton Univ. Press, Princeton, N.J.

APPENDIX A

Measure and Integration

1 MEASURES AND THE CARATHÉODORY EXTENSION.

Let S be a nonempty set. A class $\mathcal{A}$ of subsets of S is a **field**, or an **algebra** if (i) $\emptyset \in \mathcal{A}$, $S \in \mathcal{A}$, (ii) $A \in \mathcal{A} \Longrightarrow A^c \in \mathcal{A}$, (iii) $A, B \in \mathcal{A} \Longrightarrow A \cup B \in \mathcal{A}$. Note that (ii) and (iii) imply that $\mathcal{A}$ is closed under finite unions and finite intersections. If (iii) is replaced by (iii)$'$: $A_n \in \mathcal{A}$ $(n = 1, 2, \dots) \Longrightarrow \cup_{n=1}^{\infty} A_n \in \mathcal{A}$, then $\mathcal{A}$ is said to be a **σ-field**, or a **σ-algebra**. Note that (iii)$'$ implies (iii), and that a σ-field is closed under countable intersections.

A function $\mu : \mathcal{A} \to [0, \infty]$ is said to be a **measure on a field** $\mathcal{A}$ if $\mu(\emptyset) = 0$ and $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ for every sequence of pairwise disjoint sets $A_n \in \mathcal{A}$ $(n = 1, 2, \dots)$ such that $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$. Note that this property, known as **countable additivity**, implies **finite additivity** (by letting $A_n = \emptyset$ for $n \geq m$ for some m, say). A measure μ on a field $\mathcal{A}$ is **σ-finite** if there exists a sequence $A_n \in \mathcal{A}$ $(n = 1, 2, \dots)$ such that $\cup_{n=1}^{\infty} A_n = S$ and $\mu(A_n) < \infty$ for every n.

If μ is a measure on a field $\mathcal{A}$, and $A_n \in \mathcal{A}$ $(n \geq 1)$, $A \subseteq \cup_n A_n, A \in \mathcal{A}$, then $\mu(A) \leq \sum_{n=1}^{\infty} \mu(A_n)$ **(subadditivity)**. To see this write $B_1 = A_1, B_n = A_1^c \cap \dots \cap A_{n-1}^c \cap A_n (n \geq 2)$. Then $B_n (n \geq 1)$ are disjoint, $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} B_n$, so that $\mu(A) = \mu(A \cap (\cup_{n=1}^{\infty} B_n)) = \sum_{n=1}^{\infty} \mu(A \cap B_n) \leq \sum_{n=1}^{\infty} \mu(A_n)$ (since $B_n \subseteq A_n$ for all n).

Let μ be a measure on a σ-field $\mathcal{F}$ on S. Then $\mathcal{F}$ is said to be **μ-complete** if all subsets of μ-null sets in $\mathcal{F}$ belong to $\mathcal{F}$: $N \in \mathcal{F}$, $\mu(N) = 0$, $B \subseteq N \Longrightarrow B \in \mathcal{F}$. In this case the measure μ is also said to be **complete**. Given any measure μ on a σ-field $\mathcal{F}$, it is simple to check that the class of sets

$$\overline{\mathcal{F}} = \{C = A \cup B : A \in \mathcal{F},\ B \subseteq N \text{ for some } N \in \mathcal{F} \text{ such that } \mu(N) = 0\} \qquad (1.1)$$

is a μ-complete σ-field, $\widetilde{\mu}(A \cup B) := \mu(A)$ $(A \in \mathcal{F},\ B \subseteq N,\ \mu(N) = 0)$ is well defined, and $\widetilde{\mu}$ is a measure on $\overline{\mathcal{F}}$ extending μ. This extension of μ is called the **completion of μ.**

We now derive one of the most basic results in measure theory, due to Carathéodory, which provides an extension of a measure μ on a field $\mathcal{A}$ to a measure on the σ-field $\mathcal{F} = \sigma(\mathcal{A})$, the smallest σ-field containing $\mathcal{A}$. First, on the set 2^S of all subsets of S, call a set function $\mu^* : 2^S \to [0, \infty]$ an **outer measure** on S if (1) $\mu^*(\emptyset) = 0$, (2) *(monotonicity)* $A \subseteq B \Longrightarrow \mu^*(A) \leq \mu^*(B)$, and (3) *(subadditivity)* $\mu^*(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$ for every sequence A_n $(n = 1, 2, \dots)$.

Proposition 1.1. Let $\mathcal{A}$ be a class of sets such that $\emptyset \in \mathcal{A}$, $S \in \mathcal{A}$, and let $\mu : \mathcal{A} \to [0, \infty]$ be a function such that $\mu(\emptyset) = 0$. For every set $A \subseteq S$, define

$$\mu^*(A) = \inf \left\{ \sum_n \mu(A_n) : A_n \in \mathcal{A}\ \forall n, A \subseteq \cup_n A_n \right\}. \tag{1.2}$$

Then μ^* is an outer measure on S.

Proof. (1) Since $\emptyset \subseteq \emptyset \in \mathcal{A}$, $\mu^*(\emptyset) = 0$. (2) Let $A \subseteq B$. Then every countable collection $\{A_n : n = 1, 2, \dots\} \subset \mathcal{A}$ that covers B (i.e., $B \subseteq \cup_n A_n$) also covers A. Hence $\mu^*(A) \leq \mu^*(B)$. (3) Let $A_n \subset S$ $(n = 1, 2, \dots)$, and $A = \cup_n A_n$. If $\mu^*(A_n) = \infty$ for some n, then by (2), $\mu^*(A) = \infty$. Assume now that $\mu^*(A_n) < \infty$ $\forall n$. Fix $\varepsilon > 0$ arbitrarily. For each n there exists a sequence $\{A_{n,k} : k = 1, 2, \cdots\} \subset \mathcal{A}$ such that $A_n \subseteq \cup_k A_{n,k}$ and $\sum_k \mu(A_{n,k}) < \mu^*(A_n) + \varepsilon/2^n$ $(n = 1, 2, \dots)$. Then $A \subseteq \cup_n \cup_k A_{n,k}$, and therefore $\mu^*(A) \leq \sum_{n,k} \mu(A_{n,k}) \leq \sum_n \mu^*(A_n) + \varepsilon$. ■

The technically simplest, but rather unintuitive, proof of Carathéodory's theorem given below is based on the following notion. Let μ^* be an outer measure on S. A set $A \subseteq S$ is said to be μ^*-**measurable** if the following "balance conditions" are met:

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c) \qquad \forall\ E \subseteq S. \tag{1.3}$$

Theorem 1.2 (Carathéodory Extension Theorem). (a) Let μ^* be an outer measure on S. The class $\mathcal{M}$ of all μ^*-measurable sets is a σ-field, and the restriction of μ^* to $\mathcal{M}$ is a complete measure. (b) Let μ^* be defined by (1.2), where $\mathcal{A}$ is a field and μ is a measure on $\mathcal{A}$. Then $\sigma(\mathcal{A}) \subseteq \mathcal{M}$ and $\mu^* = \mu$ on $\mathcal{A}$. (c) If a measure μ on a field $\mathcal{A}$ is σ-finite, then it has a unique extension to a measure on $\sigma(\mathcal{A})$, this extension being given by μ^* in (1.2) restricted to $\sigma(\mathcal{A})$.

Proof. (a) To show that $\mathcal{M}$ is a *field,* first note that $A = \emptyset$ trivially satisfies (1.3) and that if A satisfies (1.3), so does A^c. Now, in view of the subadditivity property of μ^*, (1.3) is equivalent to the inequality

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c) \qquad \forall\ E \subseteq S. \tag{1.4}$$

To prove that $\mathcal{M}$ is closed under finite intersections, let $A, B \in \mathcal{M}$. Then $\forall\ E \subseteq S$,

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap B) + \mu^*(E \cap B^c) \qquad \text{(since } B \in \mathcal{M}\text{)}\\ &= \mu^*(E \cap B \cap A) + \mu^*(E \cap B \cap A^c) + \mu^*(E \cap B^c \cap A)\\ &\quad + \mu^*(E \cap B^c \cap A^c) \qquad\qquad \text{(since } A \in \mathcal{M}\text{)}\\ &\geq \mu^*(E \cap (B \cap A)) + \mu^*(E \cap (B \cap A)^c).\end{aligned}$$

For the last inequality, use $(B \cap A)^c = B^c \cup A^c = (B^c \cap A) \cup (B^c \cap A^c) \cup (B \cap A^c)$, and subadditivity of μ^*. By the criterion (1.4), $B \cap A \in \mathcal{M}$. Thus $\mathcal{M}$ is a field.

Next, we show that $\mathcal{M}$ *is a σ-field and μ^* is countably additive on $\mathcal{M}$.* Let $B_n \in \mathcal{M}$ $(n = 1, 2, \dots)$ be a pairwise disjoint sequence in $\mathcal{M}$, and write $C_m = \cup_{n=1}^m B_n$ $(m \geq 1)$. We will first show, *by induction on m,* that

$$\mu^*(E \cap C_m) = \sum_{n=1}^m \mu^*(E \cap B_n) \qquad \forall\ E \subseteq S. \tag{1.5}$$

This is true for $m = 1$, since $C_1 = B_1$. Suppose (1.5) holds for some m. Since $B_{m+1} \in \mathcal{M}$, one has for all $E \subseteq S$,

$$\begin{aligned}\mu^*(E \cap C_{m+1}) &= \mu^*((E \cap C_{m+1}) \cap B_{m+1}) + \mu^*((E \cap C_{m+1}) \cap B_{m+1}^c)\\ &= \mu^*(E \cap B_{m+1}) + \mu^*(E \cap C_m)\\ &= \mu^*(E \cap B_{m+1}) + \sum_{n=1}^m \mu^*(E \cap B_m),\end{aligned}$$

using the induction hypothesis for the last equality. Thus (1.5) holds for $m + 1$ in place of m, and the induction is complete. Next, writing $A = \cup_{n=1}^\infty B_n$ one has, for all $E \subseteq S$,

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap C_m) + \mu^*(E \cap C_m^c) \qquad \text{(since } C_m \in \mathcal{M}\text{)}\\ &= \sum_{n=1}^m \mu^*(E \cap B_n) + \mu^*(E \cap C_m^c) \geq \sum_{n=1}^m \mu^*(E \cap B_n) + \mu^*(E \cap A^c),\end{aligned}$$

since $C_m^c \supset A^c$. Letting $m \to \infty$, one gets

$$\mu^*(E) \geq \sum_{n=1}^\infty \mu^*(E \cap B_n) + \mu^*(E \cap A^c) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c), \tag{1.6}$$

using the subadditivity property for the last inequality. This shows that $A \equiv \cup_{n=1}^\infty B_n \in \mathcal{M}$, i.e., $\mathcal{M}$ is closed under countable disjoint unions. If $\{A_n : n = 1, \dots\}$ is an arbitrary sequence in $\mathcal{M}$, one may express $A \equiv \cup_{n=1}^\infty A_n$ as $A = \cup_{n=1}^\infty B_n$, where

$B_1 = A_1$, $B_2 = A_1^c \cap A_2$, $B_n = A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n$ $(n > 2)$, are pairwise disjoint sets in $\mathcal{M}$. Hence $A \in \mathcal{M}$, proving that $\mathcal{M}$ *is a σ-field.* To prove countable additivity of μ^* on $\mathcal{M}$, let B_n $(n \geq 1)$ be a pairwise disjoint sequence in $\mathcal{M}$ as before, and take $E = A \equiv \cup_{n=1}^{\infty} B_n$ in the first inequality in (1.6) to get $\mu^*(\cup_{n=1}^{\infty} B_n) \geq \sum_{n=1}^{\infty} \mu^*(B_n)$. By the subadditive property of μ^*, it follows that $\mu^*(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mu^*(B_n)$.

We have proved that μ^* is a measure on the σ-field $\mathcal{M}$. Finally, if $A \subseteq N \in \mathcal{M}$, $\mu^*(N) = 0$, then $\mu^*(E \cap A) \leq \mu^*(A) \leq \mu^*(N) = 0$, and $\mu^*(E \cap A^c) \leq \mu^*(E)$, so that (1.4) holds, proving $A \in \mathcal{M}$. Hence $\mathcal{M}$ is μ^*-complete.

(b) Consider now the case in which $\mathcal{A}$ is a field, μ is a measure on $\mathcal{A}$, and μ^* is the outer measure (1.2). To prove $\mathcal{A} \subseteq \mathcal{M}$, let $A \in \mathcal{A}$. Fix $E \subseteq S$ and $\varepsilon > 0$ arbitrarily. There exists $A_n \in \mathcal{A}$ $(n = 1, 2, \dots)$ such that $E \subseteq \cup_{n=1}^{\infty} A_n$ and $\mu^*(E) \geq \sum_{n=1}^{\infty} \mu(A_n) - \varepsilon$. Also,

$$\mu^*(E \cap A) \leq \mu^*\left(A \cap \bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A \cap A_n),$$

$$\mu^*(E \cap A^c) \leq \mu^*\left(A^c \cap \bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A^c \cap A_n),$$

$$\begin{aligned}\mu^*(E \cap A) + \mu^*(E \cap A^c) &\leq \sum_{n=1}^{\infty} \{\mu(A \cap A_n) + \mu(A^c \cap A_n)\} \\ &= \sum_{n=1}^{\infty} \mu(A_n) \leq \mu^*(E) + \varepsilon.\end{aligned}$$

Hence (1.4) holds, proving that $A \in \mathcal{M}$. To prove $\mu = \mu^*$ on $\mathcal{A}$, let $A \in \mathcal{A}$. By definition (1.2), $\mu^*(A) \leq \mu(A)$ (letting $A_1 = A$ and $A_n = \emptyset$ for $n \geq 2$, be a cover of A). On the other hand, $\mu(A) \leq \sum_{n=1}^{\infty} \mu(A_n)$ for every sequence $A_n \in \mathcal{A}$ $(n \geq 1)$ such that $A \subseteq \cup_{n=1}^{\infty} A_n$, so that $\mu^*(A) \geq \mu(A)$ (by subadditivity of μ on $\mathcal{A}$). Hence $\mu^*(A) = \mu(A)$.

(c) Suppose μ is a σ-finite measure on the field $\mathcal{A}$, and μ^* its extension to the σ-field $\sigma(\mathcal{A})$ $(\subseteq \mathcal{M})$ as derived in (b). Let ν be another extension of μ to $\sigma(\mathcal{A})$. Since one may express $S = \cup_{n=1}^{\infty} A_n$ with $A_n \in \mathcal{A}$ pairwise disjoint and $\mu(A_n) < \infty$ $\forall n$, it is enough to consider the restrictions of μ and ν to $A_n \cap \sigma(\mathcal{A}) \equiv \{A_n \cap A : A \in \sigma(\mathcal{A})\}$ for each n separately. In other words, it is enough to prove that $\mu = \nu$ on $\sigma(\mathcal{A})$ in the case $\mu(S) < \infty$. But for this case, the class $\mathcal{C} = \{A \in \mathcal{F} : \mu(A) = \nu(A)\}$ is a λ-system, and it contains the π-system $\mathcal{A}$. Hence, by the π-λ theorem, $\mathcal{C} = \sigma(\mathcal{A})$. ■

***Example 1** (Lebesgue–Stieltjes Measures).* Let $S = \mathbb{R}$ and $\mathcal{B}(\mathbb{R})$ the Borel σ-field. A measure μ on $\mathcal{B}(\mathbb{R})$ is said to be a **Lebesgue–Stieltjes** (or **L–S**) **measure** if $\mu((a, b]) < \infty$ $\forall -\infty < a < b < \infty$. Given such a measure one may define its **distribution function** $F_\mu : \mathbb{R} \to \mathbb{R}$ by

$$F_\mu(x) = \begin{cases} -\mu((x,0]) + c & \text{if } x < 0 \\ \mu((0,x]) + c & \text{if } x > 0, \end{cases} \tag{1.7}$$

where c is an arbitrary constant. Note that

$$\mu((a,b]) = F_\mu(b) - F_\mu(a) \qquad (-\infty < a < b < \infty). \tag{1.8}$$

Moreover, F_μ is nondecreasing and right-continuous. Conversely, *given a function F which is nondecreasing and right-continuous on $\mathbb{R}$, there exists a unique Lebesgue–Stieltjes measure μ whose distribution function is F.* To prove this, first, fix an interval $S = (c,d]$, $-\infty < c < d < \infty$. The class of all finite unions $\cup_{j=1}^m (a_j, b_j]$ of pairwise disjoint intervals $(a_j, b_j]$ $(c \le a_j < b_j \le d)$ is a field $\mathcal{A}$ on S. Define the set function μ on $\mathcal{A}$ first by (1.8) on intervals $(a,b]$, and then on disjoint unions above as $\sum_{j=1}^m \mu((a_j,b_j])$. It is simple to check that this is *well-defined,* i.e., if $(c_i, d_i]$, $1 \le i \le n$, is another representation of $\cup_{j=1}^m (a_j, b_j]$ as a union of disjoint intervals, then $\sum_{i=1}^n [F(d_i) - F(c_i)] = \sum_{j=1}^m [F(b_j) - F(a_j)]$ (Show this by splitting each $(a_j, b_j]$ by $(c_i, d_i]$, $1 \le i \le n$). *Finite additivity of* μ on $\mathcal{A}$ is then a consequence of the definition of μ. In view of this, to prove *countable additivity* of μ on $\mathcal{A}$, it is enough to show that if $I_j = (a_j, b_j]$ $(j = 1, 2, \dots)$ is a sequence of pairwise disjoint intervals whose union is $(a,b]$, then $\mu((a,b]) \equiv F(b) - F(a) = \sum_{j=1}^\infty \mu(I_j)$. Clearly, $\sum_{j=1}^n \mu(I_j) = \mu(\cup_{j=1}^n I_j) \le \mu((a,b])$ for all n, so that $\sum_{j=1}^\infty \mu(I_j) \le \mu((a,b])$. For the opposite inequality, fix $\varepsilon > 0$ and find $\delta > 0$ such that $F(a+\delta) - F(a) < \varepsilon$ (by right-continuity of F). Also, find $\delta_j > 0$ such that $F(b_j + \delta_j) - F(b_j) < \varepsilon/2^j$ $(j = 1, 2, \dots)$. Then $\{(a_j, b_j + \delta_j) : j \ge 1\}$ is an open cover of the compact interval $[a+\delta, b]$, so that there exists a finite subcover: $[a+\delta, b] \subseteq \cup_{j=1}^m (a_j, b_j + \delta_j)$, say. Then $\mu((a,b]) = F(b) - F(a) \le F(b) - F(a+\delta) + \varepsilon \le \sum_{j=1}^m [F(b_j + \delta_j) - F(a_j)] + \varepsilon \le \sum_{j=1}^m [F(b_j) - F(a_j)] + 2\varepsilon \le \sum_{j=1}^\infty [F(b_j) - F(a_j)] + 2\varepsilon \le \sum_{j=1}^\infty \mu(I_j) + 2\varepsilon$. This proves that μ is a measure on $\mathcal{A}$. Now use Carathéodory's extension theorem to extend uniquely μ to $\sigma(\mathcal{A}) = \mathcal{B}((c,d])$. Since $\mathbb{R} = \cup_{n=-\infty}^\infty (n, n+1]$, one may construct μ on each of $(n, n+1]$ and then piece (or add) them together to construct the unique L-S measure on $\mathcal{B}(\mathbb{R})$ with the given distribution function F.

(a) As a very special L-S measure, one constructs **Lebesgue measure** m on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ specified by

$$m((a,b]) = b - a,$$

with distribution function $F(x) = x$.

(b) For an example of a L-S measure with a continuous distribution, but that does not have a density with respect to Lebesgue measure, consider the representation of a real x in $(0,1]$ by its *ternary* representation $x = \sum_{n=1}^\infty a_n 3^{-n}$, where $a_n \in \{0,1,2\}$. By requiring that there be an infinite number of 2's among $\{a_n\}$ one gets a one-to-one correspondence $x \mapsto \{a_n\}$. The **Cantor set** C is defined to be the set of all x in whose ternary expansion the digit 1 does not occur. That is, C is obtained by first omitting the *middle third* $(1/3, 2/3]$ of $(0,1]$, then omitting the middle thirds of the two remaining intervals $(0, 1/3]$, $(2/3, 1]$, then omitting the middle thirds of the four remaining intervals, and so on. The Lebesgue measure of the omitted set is $\sum_{n=1}^{\infty} 2^{n-1}3^{-n} = 1$. Hence the remaining set C has Lebesgue measure zero. Define the **Cantor function**

$$F(x) = \sum_{n=1}^{\infty} \frac{a_n}{2} 2^{-n} \qquad \text{for } x = \textstyle\sum_{n=1}^{\infty} a_n 3^{-n} \in C,$$

and extend F to $[0, 1]$ by letting $F(0) = 0$, and F constant between the endpoints of every omitted interval. Then F is continuous and nondecreasing, and the corresponding L–S probability measure μ is the **Cantor measure** on $(0, 1]$, which is **nonatomic** (i.e., $\mu(\{x\}) = 0\ \forall\, x$) and **singular** in the sense that $\mu(C) = 1$ and $m(C) = 0$, where m is Lebesgue measure.

2 INTEGRATION AND BASIC CONVERGENCE THEOREMS

Let $\mathcal{S}$ be a σ-field on S. We say that $(S, \mathcal{S})$ is a **measurable space**. Denote by L the *class of* all (extended) real-valued *measurable functions* $f : S \to \overline{\mathbb{R}} = [-\infty, \infty]$, i.e., $f^{-1}(B) \in \mathcal{S}\ \forall\, B \in \mathcal{B}(\mathbb{R})$ and $f^{-1}(\{-\infty\}) \in \mathcal{S}$, $f^{-1}(\{+\infty\}) \in \mathcal{S}$. The subclass of nonnegative *measurable functions* is denoted by L^+. A **simple function** is of the form $f = \sum_1^m a_j \mathbf{1}_{A_j}$, where $m \geq 1$, $a_j \in \mathbb{R}\ \forall\, j$, A_j's are pairwise disjoint sets in $\mathcal{S}$. The class of all simple functions is denoted by L_s, and the subclass of nonnegative simple functions by L_s^+.

In general, if $(S_i, \mathcal{S}_i)$ $(i = 1, 2)$ are measurable spaces, a *map,* or function, $f : S_1 \to S_2$ *is* said to be **measurable** if $f^{-1}(B) \in \mathcal{S}_1\ \forall\, B \in \mathcal{S}_2$. In particular, if S_i is a metric space with Borel σ-field $\mathcal{S}_i$ $(i = 1, 2)$, then a *continuous map* $f : S_1 \to S_2$ is measurable, since $f^{-1}(B)$ is an open subset of S_1 if B is an open subset of S_2, and since $\mathcal{F} \equiv \{B \in \mathcal{S}_2 : f^{-1}(B) \in \mathcal{S}_1\}$ is a σ-field (containing the class of all open subsets of S_2). It is simple to check that *compositions* of measurable maps are measurable. As an example, let $(S, \mathcal{S})$ be a measurable space, and let f, g be measurable maps on S into $\mathbb{R}^k$. Then $\alpha f + \beta g$ is measurable for all constants $\alpha, \beta \in \mathbb{R}$. To see this, consider the map $h(x, y) \mapsto \alpha x + \beta y$ on $\mathbb{R}^k \times \mathbb{R}^k$ into $\mathbb{R}^k$. Since h is continuous, it is measurable. Also, $\varphi(s) := (f(s), g(s))$ is a measurable map on S into $S \times S$, with the **product σ-field** $\mathcal{S} \otimes \mathcal{S}$ (i.e., the smallest σ-field on $S \times S$ containing the class of all **measurable rectangles** $A \times B$, $A \in \mathcal{S}$, $B \in \mathcal{S}$). Now $\alpha f + \beta g$ equals the composition $h \circ \varphi$ and is therefore measurable.

Example 1. (a) Let $f_1 f_2, \ldots, f_k$ be (extended) real-valued measurable functions on a measurable space $(S, \mathcal{S})$. Then $M \equiv \max\{f_1, \ldots, f_k\}$ is measurable, since $[M \leq x] = \cap_{j=1}^k [f_j \leq x]\ \forall\, x \in \mathbb{R}^1$. Similarly, $\min\{f_1, \ldots, f_k\}$ is measurable. (b) Let f_n $(n \geq 1)$ be a sequence of (extended) real-valued measurable functions on a measurable space $(S, \mathcal{S})$. Then $h \equiv \liminf f_n$ is measurable. For $g_n := \inf\{f_j : j \geq n\}$ is measurable $(n \geq 1)$, since $[g_n \geq x] = \cap_{j=n}^{\infty} [f_j \geq x]$. Also, $g_n \uparrow h$, so that $[h \leq x] = \cap_{n=1}^{\infty} [g_n \leq x]$ $\forall\, x \in \mathbb{R}$. Similarly, $\limsup f_n$ is measurable.

Let μ be a σ-finite measure on $(S, \mathcal{S})$ (i.e., on $\mathcal{S}$). For $f \in L_s^+$ define the **integral** $\int f d\mu$, or simply $\int f$ when there is no ambiguity about the underlying measure μ, by

$$\int f \equiv \int f \, d\mu := \sum_{j=1}^{m} a_j \mu(A_j), \tag{2.1}$$

with the convention $0 \cdot \infty = 0$. If $f = \sum_1^n b_i \mathbf{1}_{B_i}$ is another representation of f, then $a_j = b_i$ on $A_j \cap B_i$, and using the finite additivity of μ, one has $\sum_1^m a_j \mu(A_j) = \sum_j \sum_i a_j \mu(A_j \cap B_i) = \sum_j \sum_i b_i \mu(A_j \cap B_i) = \sum_i b_i \mu(B_i)$. Thus $\int f$ is well defined for $f \in L_s^+$. Using a similar splitting where necessary, one can prove the following properties of the integral on L_s^+:

$$\begin{aligned}
&\text{(i)} \quad \int cf = c \int f && \forall\, c \geq 0, \\
&\text{(ii)} \quad \int f \leq \int g && \text{if } f \leq g, \\
&\text{(iii)} \quad \int (f + g) = \int f + \int g.
\end{aligned} \tag{2.2}$$

For an arbitrary $f \in L^+$ (set of all extended nonnegative measurable functions on S) define

$$\int f := \lim \int f_n, \tag{2.3}$$

where $f_n \in L_s^+$ and $f_n \uparrow f$. To show that $\int f$ is *well defined* let us first observe that there does exist $f_n \in L_s^+$, $f_n \uparrow f$. For example, let f_n be the so-called **standard approximation**,

$$f_n = \sum_{k=1}^{n2^n} (k-1) 2^{-n} \mathbf{1}_{\left[\frac{k-1}{2^n} \leq f < \frac{k}{2^n}\right]} + n \mathbf{1}_{[f \geq n]}. \tag{2.4}$$

Secondly, suppose $f_n, g_n \in L_s^+$, $f_n \uparrow f$, $g_n \uparrow f$. We will show that

$$\lim_n \int g_n = \lim_n \int f_n. \tag{2.5}$$

For this fix $c \in (0,1)$ and $m \geq 1$. One may write $g_m = \sum_1^k a_j \mathbf{1}_{A_j}$, $\int c g_m = c \sum_1^k a_j \mu(A_j) = c \int g_m$. Let $B_n = \{x \in S : f_n(x) \geq c g_m(x)\}$. Then $f_n = f_n \mathbf{1}_{B_n} + f_n \mathbf{1}_{B_n^c}$, so that (by (2.2)), and using $B_n \uparrow S$,

$$\int f_n = \int f_n \mathbf{1}_{B_n} + \int f_n \mathbf{1}_{B_n^c} \geq \int f_n \mathbf{1}_{B_n} \geq \int c g_m \mathbf{1}_{B_n}$$

$$= c\sum_{j=1}^{k} a_j\mu(A_j \cap B_n) \uparrow c\sum_{j=1}^{k} a_j\mu(A_j) = c\int g_m \quad \text{as } n \uparrow \infty.$$

Hence $\lim_n \int f_n \geq c\int g_m\ \forall\, c \in (0,1)$, which implies $\lim_n \int f_n \geq \int g_m$. Letting $m \uparrow \infty$, we obtain $\lim_n \int f_n \geq \lim_m \int g_m$. Reversing the roles of f_n and g_n, we then get (2.5), and $\int f$ is well defined.

As simple consequences of the definition (2.3) and the order property (2.2)(ii), one obtains the following results.

Proposition 2.1. (a) Let $f \in L^+$. Then

$$\int f = \sup\left\{\int g : g \in L_s^+,\, g \leq f\right\}, \qquad \forall\, f \in L^+. \tag{2.6}$$

(b) Let $f, g \in L^+$, $c \geq 0$. Then (2.2)(i)—(iii) hold.

Proof. (a). Clearly, $\int f$ is dominated by the right hand side of (2.6), by the definition (2.3). For the reverse inequality, let $g \in L_s^+$, $g \leq f$. Then $g_n := \max\{g, f_n\} \uparrow f$ with f_n as in (2.3). Since $g_n \in L_s^+$, it follows that $\int g \leq \int g_n \to \int f$. Hence $\int g \leq \int f$.

(b) (i) and (iii) in (2.2) follow from the definition (2.3), while (ii) follows from (2.6). ■

A useful convergence result for functions in L^+ is the following.

Proposition 2.2. Let f_k, $f \in L^+$, $f_n \uparrow f$. Then $\int f_k \uparrow \int f$.

Proof. By Proposition 2.1(b), $\lim_k \int f_k \leq \int f$ (order property of integrals). Next, let $g_{k,n} \in L_s^+$, $g_{k,n} \uparrow f_k$ as $n \uparrow \infty$. Define $g_n = \max\{g_{k,n} : k = 1, \ldots, n\} \in L_s^+$, $g_n \uparrow g$, say. But $g_n \geq g_{k,n}\ \forall\, k \leq n$, so that $g \geq f_k\ \forall\, k$, implying $g \geq f$. On the other hand, $g_n \leq f\ \forall\, n$. Thus $g = f$ and, therefore, $\lim_n \int g_n = \int f$. But $g_n \leq f_n\ \forall\, n$ which implies $\lim_n \int f_n \geq \lim_n \int g_n = \int f$. ■

Let $f \in L$, and set $f^+ = \max\{f, 0\}$, $f^- = -\min\{f, 0\}$. Then f^+, $f^- \in L^+$ and $f = f^+ - f^-$. If at least one of $\int f^+$, $\int f^-$ is finite, we say that the **integral of** f **exists** and define

$$\int f = \int f^+ - \int f^-. \tag{2.7}$$

If $\int f^+$ and $\int f^-$ are both finite, then $0 \leq \int |f| = \int f^+ + \int f^- < \infty$ (since $|f| = f^+ + f^-$, Proposition 2.1(b) applies), and f is said to be **integrable** (with respect to μ). The following result is now simple to prove.

Proposition 2.3. Let $f, g \in L$ be integrable and $\alpha, \beta \in \mathbb{R}^1$. Then (i) αf, βg, $\alpha f + \beta g$ are integrable and $\int(\alpha f + \beta g) = \alpha \int f + \beta \int g$ *(linearity),* and (ii) $f \leq g$ implies $\int f \leq \int g$ *(order).*

Proof. (i) First, let $\alpha \geq 0$. Then $(\alpha f)^+ = \alpha f^+$, $(\alpha f)^- = \alpha f^-$, so that $\int \alpha f = \int \alpha f^+ - \int \alpha f^- = \alpha \int f^+ - \alpha \int f^- = \alpha \int f$, by Proposition 2.1(b). Now let $\alpha < 0$. Then $(\alpha f)^+ = -\alpha f^-$, $(\alpha f)^- = -\alpha f^+$. Hence $\int \alpha f = \int -\alpha f^- - \int -\alpha f^+ = -\alpha \int f^- - (-\alpha) \int f^+ = \alpha(\int f^+ - \int f^-) = \alpha \int f$. Next if f, g are integrable, then writing $h = f + g$, we have $|h| \leq |f| + |g|$, so that $\int |h| \leq \int |f| + \int |g| < \infty$. Since $h = f + g = f^+ + g^+ - f^- - g^- = h^+ - h^-$, one has $h^+ + f^- + g^- = h^- + f^+ + g^+$ and, by Proposition 2.1(b), $\int h^+ + \int f^- + \int g^- = \int h^- + \int f^+ + \int g^+$. Therefore, $\int h \equiv \int h^+ - \int h^- = \int f^+ - \int f^- + \int g^+ - \int g^- = \int f + \int g$. This proves (i). To prove (ii) note that $f \leq g$ implies $f^+ \leq g^+$, $f^- \geq g^-$. Hence $\int f \equiv \int f^+ - \int f^- \leq \int g^+ - \int g^- \equiv \int g$. ■

Our next task is to show that the integral of a function f remains unaffected if it is modified arbitrarily (but measurably) on a μ-null set. First, note that if $f \in L^+$, then

$$\int f = 0 \quad \text{iff} \quad f = 0 \text{ a.e. } (\mu) \quad (f \in L^+), \tag{2.8}$$

where **a.e.** (μ) is short-hand for *almost everywhere with respect to* μ, or *outside a* μ*-null set.* To prove (2.8), let $N = \{x : f(x) > 0\}$. Then one has $f = f\mathbf{1}_N + f\mathbf{1}_{N^c} = f \cdot \mathbf{1}_N$ ($f = 0$ on N^c). If $\mu(N) = 0$, then for all $g \in L_s^+$, $g \leq f$, one has $g = 0$ on N^c, so that $\int g = \int g_{\mathbf{1}_N} + \int g_{\mathbf{1}_{N_c}} = \int g_{\mathbf{1}_N} = 0$, implying $\int f = 0$. Conversely, if $\int f = 0$, then $\mu(N) = 0$. For otherwise there exists $\varepsilon > 0$ such that writing $N_\varepsilon := \{x : f(x) > \varepsilon\}$, one has $\mu(N_\varepsilon) > 0$. In that case, $g := \varepsilon \mathbf{1}_{N_\varepsilon} \leq f$ and $\int f \geq \int g = \varepsilon \mu(N_\varepsilon) > 0$, a contradiction.

As a consequence of (2.8), one has the result that *if $f = g$ a.e., and f, g are integrable, then* $\int f = \int g$. To see this note that $|\int f - \int g| = |\int (f - g)| \leq \int |f - g| = 0$, since $|f - g| = 0$ a.e.

From here on, all functions f, g, h, f_n, g_n, h_n, etc., are assumed to be measurable, unless specified otherwise.

An important notion in measure theory is that of convergence in measure. Let f_n $(n \geq 1)$, f be measurable functions on a measure space $(S, \mathcal{S}, \mu)$. The sequence $\{f_n\}_{n \geq 1}$ **converges in measure** to f if

$$\mu\left([|f_n - f| > \varepsilon]\right) \longrightarrow 0 \qquad \text{as } n \to \infty, \forall\, \varepsilon > 0. \tag{2.9}$$

Proposition 2.4. (a) If $f_n \to f$ in measure then there exists a subsequence $\{f_{n_k}\}_{k \geq 1}$ that converges a.e. to f. (b) If $\mu(S) < \infty$, then the convergence $f_n \to f$ a.e. implies $f_n \to f$ in measure.

Proof. (a) Assume $f_n \to f$ in measure. For each k one can find n_k such that $\mu([|f_{n_k} - f| > 1/2^k]) < 1/2^k$. Now, for any given $\varepsilon > 0$, $[\limsup_k |f_{n_k} - f| > \varepsilon] \subseteq \cap_{m=1}^{\infty} \cup_{k=m}^{\infty} [|f_{n_k} - f| > 1/2^k] = N$, say. But $\mu(N) \leq \sum_{k=m}^{\infty} 2^{-k} = 2^{-m+1} \to 0$ as $m \to \infty$, i.e., $\mu(N) = 0$, proving $f_{n_k} \to f$ a.e., as $k \to \infty$. (b) Suppose $\mu(S) < \infty$, and $f_n \to f$ a.e. If f_n does not converge in measure to f, there exist $\varepsilon > 0$, $\delta > 0$, and a subsequence $\{f_{n_k}\}_{k \geq 1}$ such that $\mu([|f_{n_k} - f| > \varepsilon]) > \delta \; \forall\, k = 1, 2, \ldots$. But writing $A_k = [|f_{n_k} - f| > \varepsilon]$, one then has $B_m \equiv \cup_{k=m}^{\infty} A_k \downarrow B = [|f_{n_k} - f| > \varepsilon$ for infinitely many $k]$. Since $B_m^c \uparrow B^c$, it follows from countable additivity of μ that $\mu(B_m^c) \uparrow \mu(B^c)$, so that $\mu(B_m) = \mu(S) - \mu(B_m^c) \to \mu(S) - \mu(B^c) = \mu(B)$. Since $\mu(B_m) > \delta \; \forall\, m$, one obtains $\mu(B) \geq \delta$, which contradicts the fact that $f_n \to f$ a.e. ■

Theorem 2.5 *(Basic Convergence Theorems for Integrals).*

(a) *(Monotone Convergence Theorem).* Suppose f_n $(n \geq 1)$, f are nonnegative a.e. and $f_n \uparrow f$ a.e., then $\int f_n \uparrow \int f$.

(b) *(Fatou's Lemma).* If $g_n \geq 0$ a.e., then $\int \liminf g_n \leq \liminf \int g_n$.

(c) *(Lebesgue's Dominated Convergence Theorem).* If $f_n \to f$ in μ-measure and $|f_n| \leq h$ a.e., where h is integrable, then $\lim_n \int |f_n - f| = 0$. In particular, $\int f_n \to \int f$.

Proof. (a) Since a countable union of μ-null sets is μ-null, there exists N such that $\mu(N) = 0$ and $f_n \geq 0$, $f \geq 0$, $f_n \uparrow f$ on N^c. Setting $f_n = 0$ $(n \geq 1)$ and $f = 0$ on N does not change the integrals $\int f_n$, $\int f$. Hence one may apply Proposition 2.2.

(b) As in (a), one may assume $g_n \geq 0$ on S $(\forall n \geq 1)$. Let $f_n = \inf\{g_k : k \geq n\}$. Then $0 \leq f_n \uparrow \liminf g_n = f$, say, and $\int f_n \uparrow \int f$ (by (a)). Also $f_n \leq g_n \; \forall\, n$, so that $\int g_n \geq \int f_n \; \forall\, n$, implying, in particular, $\liminf \int g_n \geq \liminf \int f_n = \lim \int f_n = \int f$.

(c) First assume $f_n \to f$ a.e. Apply Fatou's lemma to $g_n := 2h - |f_n - f|$, $0 \leq g_n \to 2h$ a.e., to get $\int 2h \leq \liminf \int g_n = \int 2h - \limsup \int |f_n - f|$, proving $\lim \int |f_n - f| = 0$. Now assume $f_n \to f$ in μ-measure. If $\int |f_n - f|$ does not converge to zero, there exist $\delta > 0$ and a subsequence $1 < n_1 < n_2 < \cdots$ such that $\int |f_{n_k} - f| > \delta \; \forall\, k$. Then there exists, by Proposition 2.4(a), a further subsequence of $\{n_k : k \geq 1\}$, say $\{n'_k : k \geq 1\}$, such that $f_{n'_k} \to f$ a.e. as $k \to \infty$, to which the above result applies to yield $\int |f_{n'_k} - f| \to 0$ as $k \to \infty$, contradicting $\int |f_{n'_k} - f| > \delta \; \forall\, k$. ■

The next result provides useful approximations to functions in the complex Banach space $L^p = L^p(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), \mu)$, with norm $\|f\| := \left(\int_{\mathbb{R}^k} |f(x)|^p \mu(dx)\right)^{\frac{1}{p}}$, where $1 \leq p < \infty$. The result can of course be specialized to the real Banach space L^p.

Proposition 2.6. Let μ be a measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ that is finite on compact subsets of $\mathbb{R}^k$. Then the set of infinitely differentiable functions with compact support is dense in L^p.

Proof. For simplicity of notation we will take $k = 1$. The general case is similar. It is easy to see by considering real and imaginary parts separately, and then splitting a real-valued function f as $f = f^+ - f^-$, that it is enough to consider real-valued, nonnegative $f \in L^p$. Given $\varepsilon > 0$, find $N > 0$ such that $\int_{\{x:|x|\geq N\}} f^p d\mu < \frac{\varepsilon}{5}$. Set $f_N = f\mathbf{1}_{(-N,N]}$. Since $f_{N,M} := f_N \wedge M \equiv \min\{f_N, M\} \uparrow f$ as $M \uparrow \infty$, and $|f_{N,M} - f_N|^p \leq 2^p|f_N|^p \leq 2^p|f|^p$, there exists M such that $\|f_N - f_{N,M}\| < \frac{\varepsilon}{5}$. Because $f_{N,M}$ is bounded, there exists a simple function $g = \sum_{j=1}^m x_j \mathbf{1}_{B_j}$, where $x_j > 0, B_j$ Borel, $B_j \subseteq (-N,N], \mu(B_j) < \infty, 1 \leq j \leq m$, such that $\sup\{|f_{N,M}(x) - g(x)| : x \in \mathbb{R}\} < \frac{\varepsilon}{5}$ (use the standard approximation (2.4)). Then $\|f_{N,M} - g\| < \frac{\varepsilon}{5}$.

We will now approximate g by a μ-a.e. continuous step function. For this, first note that the set of all finite unions of disjoint intervals of the form $(a,b]$, $-N \leq a < b \leq N$, is a field $\mathcal{F}_0$ on $(-N,N]$ such that $\sigma(\mathcal{F}_0) = \mathcal{B}((-N,N])$. Hence by Carathéodory's extension theorem, one can find a sequence of such disjoint intervals whose union contains B_j and approximates it as closely as desired. Since $\mu(B_j) < \infty$, one may take a finite subset of these intervals, say $(a_{ij}, b_{ij}], 1 \leq i \leq n_j$, such that $A_j = \cup_{i=1}^{n_j}(a_{ij}, b_{ij}]$ satisfies $\mu(B_j \Delta A_j) < m^{-1}(\frac{\varepsilon}{5c})^p$, for $c := \max\{x_1, \ldots, x_m\}$ $(1 \leq j \leq m)$. Since the set $\{x : \mu(\{x\}) > 0\}$ is countable, one may use the approximation of $(a_{ij}, b_{ij}]$ from above and below, if necessary, to ensure that $\mu(\{a_{ij}\}) = 0 = \mu(\{b_{ij}\}), 1 \leq i \leq n_j, j = 1, \ldots m$. Note that, with $h = \sum_{j=1}^m x_j \mathbf{1}_{A_j}$ and $g = \sum_{j=1}^m x_j \mathbf{1}_{B_j}$, as above, one has $\|h - g\|^p \leq mc^p[m^{-1}(\frac{\varepsilon}{5})^p] = (\frac{\varepsilon}{5})^p$, so that $\|h - g\| < \varepsilon/5$. Finally, let ψ be an infinitely differentiable probability density on $\mathbb{R}$ with compact support (e.g. see (5.7) in Chapter V). Define $\psi_n(x) = n\psi(nx)(n = 1, 2, \ldots)$. Then the probabilities $\psi_n(x)dx$ converge weakly to δ_0 as $n \to \infty$. Hence the functions

$$h_n(x) := \int_{\mathbb{R}} h(x-y)\psi_n(y)dy = \int_{\mathbb{R}} h(y)\psi_n(x+y)dy, \quad n \geq 1, \tag{2.10}$$

are infinitely differentiable with compact support, and $h_n(x) \to h(x)$ at all points x of continuity of h. Since the set of possible discontinuities of h, namely $\{a_{ij} : 1 \leq i \leq n_j, 1 \leq j \leq m\} \cup \{b_{ij} : 1 \leq i \leq n_j, 1 \leq j \leq m\}$ has μ-measure zero, $h_n \to h$ μ-almost everywhere. Also h_n, h have compact support and are uniformly bounded by $c = \max\{x_1, \ldots, x_m\}$. Hence $h_n \to h$ in L^p, and there exists n_0 such that $\|h_{n_0} - h\| < \frac{\varepsilon}{5}$. Therefore,

$$\begin{aligned}\|h_{n_0} - f\| &\leq \|h_{n_0} - f\| + \|h - g\| + \|g - f_{N,M}\| + \|f_{N,M} - f_N\| + \|f_N - f\| \\ &< 5(\frac{\varepsilon}{5}) = \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary the proof is complete. ■

Remark 2.1. Note that the proof shows that if μ is finite on compacts sets, then step functions are dense in $L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$. Indeed rational-valued step functions with supporting intervals whose endpoints are dyadic rationals are dense in L^p. In particular, it follows that L^p is **separable**. The argument extends to $\mathbb{R}^k$ for $k > 1$ as well.

3 PRODUCT MEASURES

Let $(S_i, \mathcal{S}_i)$ $(i = 1, 2)$ be measurable spaces. The **product σ-field** $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2$ on the Cartesian product space $S = S_1 \times S_2$ is the smallest σ-field containing all sets of the form $A \times B$, with $A \in \mathcal{S}_1$ and $B \in \mathcal{S}_2$, called **measurable rectangles**. Let μ_i be a σ-finite measure on $(S_i, \mathcal{S}_i)$ $(i = 1, 2)$. Define the set function μ on the class $\mathcal{R}$ of all measurable rectangles

$$\mu(A \times B) := \mu_1(A)\mu_2(B) \qquad (A \in \mathcal{S}_1, B \in \mathcal{S}_2). \tag{3.1}$$

Theorem 3.1. There exists a unique extension of μ from $\mathcal{R}$ to a σ-finite measure on the product σ-field $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2$.

Proof. For the proof we need first the fact that if $C \in \mathcal{S}$, then the **x-section** $C_x := \{y \in S_2 : (x, y) \in C\}$ belongs to $\mathcal{S}_2$, $\forall\, x \in S_1$. The class $\mathcal{C}$ of all sets C for which this is true contains $\mathcal{R}$, since $(A \times B)_x = B$ (if $x \in A$), or $\emptyset$ (if $x \notin A$). Since it is easy to check that $\mathcal{C}$ is a λ-system containing the π-system $\mathcal{R}$, it follows by the $\pi - \lambda$ Theorem that $\mathcal{C} \supset \mathcal{S}$.

Similarly, if f is an extended real-valued measurable function on the product space $(S, \mathcal{S})$ then every **x-section of** f defined by $f_x(y) = f(x, y)$, $y \in S_2$, is a measurable function on $(S_2, \mathcal{S}_2)$, $\forall\, x \in S$. For if $D \in \mathcal{B}(\mathbb{R})$ and $x \in S_1$, then $f_x^{-1}(D) \equiv [y : f(x, y) \in D] = (f^{-1}(D))_x \in \mathcal{S}_2$.

Next, for $C \in \mathcal{S}$ the function $x \mapsto \mu_2(C_x)$ is measurable on $(S_1, \mathcal{S}_1)$. This is clearly true for $C \in \mathcal{R}$, and the general assertion again follows from the π-λ theorem. Now define μ on $\mathcal{C}$ by

$$\mu(C) := \int_{S_1} \mu_2(C_x)\mu_1(dx). \tag{3.2}$$

If $C = \cup_n C_n$, where $C_n \in \mathcal{S}$ $(n = 1, 2, \dots)$ are pairwise disjoint, then $C_x = \cup_n (C_n)_x$ and, by countable additivity of μ_2, $\mu_2(C_x) = \sum_n \mu_2((C_n)_x)$, so that $\mu(C) = \int_{S_1} \sum_n \mu_2((C_n)_x)\mu_1(dx) = \sum_n \int_{S_1} \mu_2((C_n)_x)\mu_1(dx) = \sum_n \mu(C_n)$. Here the interchange of the order of summation and integration is valid, by the monotone convergence theorem. Thus (3.2) defines a measure on $\mathcal{S}$, extending (3.1). The measure μ is clearly σ-finite. If ν is another σ-finite measure on $(S, \mathcal{S})$ such that $\nu(A \times B) = \mu_1(A)\mu_2(B)$ $\forall\, A \times B \in \mathcal{R}$, then the class of sets $C \in \mathcal{S}$ such that $\mu(C) = \nu(C)$ is easily seen to be a σ-field and therefore contains $\sigma(\mathcal{R}) = \mathcal{S}$. ■

The measure μ in Theorem 3.1 is called the **product measure** and denoted by $\mu_1 \times \mu_2$. The measure space $(S, \mathcal{S}, \mu)$ with $S = S_1 \times S_2$, $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2$, $\mu = \mu_2 \times \mu_2$, is called the **product measure space**.

Next note that instead of (3.2), one can define the measure $\widetilde{\mu}$ by

$$\widetilde{\mu}(C) = \int_{S_2} \mu_1(C^y)\mu_2(dy), \qquad C \in \mathcal{S}, \tag{3.3}$$

where $C^y = \{x \in S_1 : (x,y) \in C|$ is the ***y*-section** of C, for $y \in S_2$. But $\widetilde{\mu} = \mu$ on $\mathcal{R}$ and therefore, by the uniqueness of the extension, $\widetilde{\mu} = \mu$ on $\mathcal{S}$. It follows that if $f = \mathbf{1}_C$ for some $C \in \mathcal{S}$ and $f^y(x) := f(x,y)$, then

$$\int_S f d\mu = \int_{S_1} \left\{ \int_{S_2} f_x(y)\mu_2(dy) \right\} \mu_1(dx) = \int_{S_2} \left\{ \int_{S_1} f^y(x)\mu_1(dx) \right\} \mu_2(dy). \quad (3.4)$$

This equality of the iterated integrals with different orders of integration immediately extends to nonnegative simple functions. For arbitrary nonnegative $\mathcal{S}$-measurable f one uses an approximation $f_n \uparrow f$ by simple functions f_n and applies the monotone convergence theorem to arrive at the following important result.

***Theorem 3.2** (Fubini–Tonelli Theorem).* (a) Let f be a nonnegative measurable function on the product measure space $(S, \mathcal{S}, \mu)$, where $S = S_1 \times S_2$, $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2$, $\mu = \mu_1 \times \mu_2$. Then (3.4) holds. (b) If f is μ-integrable, then (3.4) holds.

Proof. We have outlined above a proof of (a). For (b), use $f = f^+ - f^-$, linearity of the integral (with respect to μ, μ_1, μ_2), and (a). ■

Given k (≥ 2) σ-finite measure spaces $(S_i, \mathcal{S}_i, \mu_i)$, $1 \leq i \leq k$, the above definitions and results can be extended to define the product measure space $(S, \mathcal{S}, \mu)$ with (1) $S = S_1 \times \cdots \times S_k$ the **Cartesian product** of $S_1, \ldots, S_k$, and (2) $\mathcal{S} = \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_k$, the **product σ-field**, i.e., the smallest σ-field on S containing the class $\mathcal{R}$ of all measurable rectangles $A_1 \times A_2 \times \cdots \times A_k$ $(A_i \in \mathcal{S}_i,\ 1 \leq i \leq k)$, and (3) $\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_k$, the σ-finite **product measure** on $\mathcal{S}$ satisfying

$$\mu(A_1 \times A_2 \times \cdots \times A_k) = \mu_1(A_1)\mu_2(A_2)\cdots\mu_k(A_k) \qquad (A_i \in \mathcal{S}_i,\ a \leq i \leq k). \quad (3.5)$$

Example 1. The **Lebesgue measure on** $\mathbb{R}^k$ is the product measure $\mathbf{m} = m_1 \times m_2 \times \cdots \times m_k$ defined by taking $S_i = \mathbb{R}$, $\mathcal{S}_i = \mathcal{B}(\mathbb{R})$, $m_i =$ Lebesgue measure on $\mathbb{R}$, $1 \leq i \leq k$.

4 RIESZ REPRESENTATION ON $C(S)$

Suppose that S is a compact metric space with Borel σ-field $\mathcal{B}$. If μ is a finite measure on $(S, \mathcal{B})$, then the linear functional $\ell_\mu(f) = \int_S f\, d\mu$, $f \in C(S)$, is clearly a linear functional on $C(S)$. Moreover ℓ_μ is a **positive linear functional** in the sense that $\ell_\mu(f) \geq 0$ for all $f \in C(S)$ such that $f(x) \geq 0$ for all $x \in S$. Additionally, giving $C(S)$ the uniform norm $\|f\| = \sup\{|f(x)| : x \in S\}$, one has that ℓ_μ is a **bounded linear functional** in the sense that $\sup_{\|f\| \leq 1, f \in C(S)} |\ell_\mu(f)| < \infty$. In view of linearity this boundedness is easily checked to be equivalent to continuity of $\ell_\mu : C(S) \to \mathbb{R}$. The **Riesz representation theorem** for $C(S)$ asserts that these are the only bounded positive linear functionals on $C(S)$.

***Theorem 4.1** (Riesz Representation Theorem on $C(S)$).* Let S be a compact metric space. If ℓ is a bounded positive linear functional on $C(S)$, then there is a unique finite measure μ on $(S, \mathcal{B})$ such that for all $f \in C(S)$,

$$\ell(f) = \int_S f \, d\mu.$$

Moreover, μ is regular in the sense that

$$\mu(A) = \inf\{\mu(G) : G \supseteq A, G \text{ open}\} = \sup\{\mu(F) : F \subseteq A, F \text{ closed}\}, \quad A \in \mathcal{B}.$$

Observe that the uniqueness assertion follows trivially from the fact that $C(S)$ is a measure-determining class of functions for finite measures. The proof will follow from a sequence of lemmas.

For a function $f \in C(S)$, the smallest closed set outside of which f is zero is called the **support** of f and is denoted by $\text{supp}(f)$. Note that if $f \in C(S)$ satisfies $0 \le f \le \mathbf{1}_A$, then $\text{supp}(f) \subseteq \overline{A}$. For open sets $G \subseteq S$ it is convenient to introduce notation $g \prec G$ to denote a function $g \in C(S)$ **subordinate** to G in the sense that $0 \le g \le 1$ and $\text{supp}(g) \subseteq G$. With this notation we will see that the desired measure may be expressed explicitly for open $G \subseteq S$ as

$$\mu(G) = \sup\{\ell(g) : g \prec G\}. \tag{4.1}$$

Note that since S is open (and closed), one has $\mu(S) < \infty$ from the boundedness of ℓ. With μ defined for open sets by (4.1), for arbitrary $A \subseteq S$ let

$$\mu^*(A) = \inf\{\mu(G) : G \supseteq A, \ G \text{ open}\}. \tag{4.2}$$

***Lemma 1** (Urysohn).* Given F closed, G open, $F \subset G$, there exists $g \in C(S)$ such that $g \prec G, g = 1$ on F.

Proof. For each $x \in F$ there exists $\varepsilon_x > 0$ such that $B(x : 2\varepsilon_x) \subset G$, where $B(x : \delta)$ is the open ball with center x and radius $\delta > 0$. Then $\{B(x : \varepsilon_x) : x \in F\}$ is an open cover of the compact set F. let $\{B(x_i : \varepsilon_{x_i}) : i = 1, \dots, N\}$ be a subcover. Write $V = U_{i=1}^N B(x_i : \varepsilon_{x_i})$. By the proof of "(ii) implies (iii)" in Alexandrov's Theorem (Theorem 5.1), there exists $g \in C(S)$ such that $g = 1$ on F, $g = 0$ on V^c, $0 \le g \le 1$. Since $\text{supp}(g) \subset \overline{V} \subset G$; the proof is complete. ■

***Lemma 2** (Partition of Unity).* Suppose that $G_1, \dots, G_N$ are open subsets of a compact metric space S, and assume that $F \subseteq G_1 \cup G_2 \cup \cdots \cup G_N$ is a closed subset of S. Then there are functions $g_n \prec G_n, n = 1, \dots, N$, such that $\sum_{n=1}^N g_n = 1$ on F.

Proof. For each $x \in F$ there is an open set U_x containing x whose closure is contained in G_n for some n (depending on x). Since F is compact there are points $x_1, \dots, x_m$ with $U_{x_1} \cup \cdots \cup U_{x_m} \supseteq F$. For $1 \le n \le N$, let H_n denote the union of sets $\overline{U}_{x_j}$ contained in G_n. By the Urysohn lemma, there exist functions $h_n \prec G_n$, $h_n = 1$ on

H_n. Take $g_1 = h_1, g_2 = (1-h_1)h_2, \dots, g_N = \Pi_{n=1}^{N-1}(1-h_n)h_N$. Then $g_n \prec G_n$ for each n and, by induction,

$$\sum_{n=1}^{N} g_n = 1 - \Pi_{n=1}^{N}(1-h_n).$$

Notice that, since $F \subseteq H_1 \cup \cdots \cup H_N$, one has $h_n(x) = 1$ for some n, at each $x \in F$. ■

Lemma 3. μ^* is an outer measure and each Borel-measurable subset of S is μ^*-measurable.

Proof. For the first part we will in fact show that

$$\mu^*(A) = \inf\left\{\sum_{n=1}^{\infty} \mu(G_n) : \cup_{n=1}^{\infty} G_n \supset A, G_n \text{open}\right\}.$$

from which it follows by Proposition 1.1 that μ^* is an outer measure. For this formula it suffices to check that for any given sequence G_n, $n \geq 1$, of open sets one has $\mu(\cup_{n=1}^{\infty} G_n) \leq \sum_{n=1}^{\infty} \mu(G_n)$. Let $G = \cup_{n=1}^{\infty} G_n$ and $g \prec G, g \in C(S)$. The support $\operatorname{supp}(g) \subseteq S$ is compact, and hence $\operatorname{supp}(g) \subset \cup_{n=1}^{N} G_n$, for some N. By Lemma 2, there are functions $g_n \in C(S)$, $1 \leq n \leq N$, such that $g_n \prec G_n$, and $\sum_{n=1}^{N} g_n = 1$ on $\operatorname{supp}(g)$. Now $g = \sum_{n=1}^{N} g_n g$ and $g_n g \prec G_n$, so that

$$\ell(g) = \sum_{n=1}^{N} l(g_n g) \leq \sum_{n=1}^{N} \mu(G_n) \leq \sum_{n=1}^{\infty} \mu(G_n).$$

Since $g \prec G$ is arbitrary, it follows that $\mu(G) \leq \sum_{n=1}^{\infty} \mu(G_n)$ as desired. For the second part of the lemma it suffices to check that each open set is μ^*-measurable. That is, if G is an open set, then for any $E \subseteq S$ one must check $\mu^*(E) \geq \mu^*(E \cap G) + \mu^*(E \cap G^c)$. If E is also open then given $\varepsilon > 0$ there is a $g \in C(S), g \prec E \cap G$, such that $\ell(g) > \mu(E \cap G) - \varepsilon$. Similarly, $E \cap \operatorname{supp}(g)^c$ is open and there is a $\tilde{g} \in C(S), \tilde{g} \prec E \cap \operatorname{supp}(g)^c$, such that $\ell(\tilde{g}) > \mu(E \cap \operatorname{supp}(g)^c) - \varepsilon$. But now $g + \tilde{g} \prec E$ and $\mu(E) > \ell(g) + \ell(\tilde{g}) > \mu(E \cap G) + \mu(E \cap \operatorname{supp}(g)^c) - 2\varepsilon \geq \mu^*(E \cap G) + \mu^*(E \cap G^c) - 2\varepsilon$. Since ε is arbitrary, the desired Carathéodory balance condition (1.3) holds for open E. For arbitrary $E \subseteq S$ let $\varepsilon > 0$ and select an open set $U \supseteq E$ such that $\mu(U) < \mu^*(E) + \varepsilon$. Then $\mu^*(E) + \varepsilon \geq \mu(U) \geq \mu^*(U \cap G) + \mu^*(U \cap G^c) \geq \mu^*(E \cap G) + \mu^*(E \cap G^c)$. ■

From here one readily obtains a measure space $(S, \mathcal{B}, \mu)$ by restricting μ^* to $\mathcal{B}$. The proof of the theorem is completed with the following lemma.

Lemma 4. For closed $F \subseteq S$,

$$\mu(F) = \inf\{\ell(h) : h \geq \mathbf{1}_F\}.$$

Moreover,

$$\ell(f) = \int_S f \, d\mu \qquad \forall f \in C(S).$$

Proof. For closed $F \subseteq S$ and an arbitrary $h \in C(S)$ with $h \geq \mathbf{1}_F$ consider, for $\varepsilon > 0$, the open set $G_\varepsilon = \{x \in S : h(x) > 1 - \varepsilon\}$. Let $g \in C(S)$, $g \prec G_\varepsilon$. Then $\ell(g) \leq (1-\varepsilon)^{-1}\ell(h)$. It now follows that $\mu(F) \leq \mu(G_\varepsilon) \leq (1-\varepsilon)^{-1}\ell(h)$, and hence, since $\varepsilon > 0$ is arbitrary, $\mu(F) \leq \ell(h)$. To see that $\mu(F)$ is the greatest lower bound, let $\varepsilon > 0$ and let $G \supset F$ be an open set with $\mu(G) - \mu(F) < \varepsilon$. By Urysohn's lemma there is an $h \in C(S)$, $h \prec G$, with $h \geq \mathbf{1}_F$. Thus, using the definition of μ, $\ell(h) \leq \mu(G) \leq \mu(F) + \varepsilon$. To establish that μ furnishes the desired representation of ℓ, let $f \in C(S)$. In view of the linearity of ℓ, it suffices to check that $\ell(f) \leq \int_S f \, d\mu$; since the same inequality would then be true with f replaced by $-f$, and hence the reverse inequality follows. Let $m = \min\{f(x) : x \in S\}, M = \max\{f(x) : x \in S\}$. For $\varepsilon > 0$, partition $[m, M]$ as $y_0 < m < y_1 < \cdots < y_n = M$ such that $y_j - y_{j-1} < \varepsilon$. Let $A_j = f^{-1}(y_{j-1}, y_j] \cap \operatorname{supp}(f), 1 \leq j \leq n$. Then $A_1, \ldots, A_n$ is a partition of $\operatorname{supp}(f)$ into disjoint Borel-measurable subsets. Let $G_j \supset A_j$ be an open set with $\mu(G_j) < \mu(A_j) + \frac{\varepsilon}{n}$, $j = 1, \ldots, n$, with $f(x) < y_j + \varepsilon, x \in G_j$. Apply Lemma 2 to obtain $g_j \prec G_j$ with $\sum_{j=1}^n g_j = 1$ on $\operatorname{supp}(f)$. Then $f = \sum_{j=1}^n g_j f$, and since $g_j f \leq (y_j + \varepsilon) g_j$, and $y_j - \varepsilon < f(x)$ on A_j, one has

$$\begin{aligned}
\ell(f) &= \sum_{j=1}^n \ell(g_j f) \leq \sum_{j=1}^n (y_j + \varepsilon)\ell(g_j) \leq \sum_{j=1}^n (y_j + \varepsilon)\mu(G_j) \\
&\leq \sum_{j=1}^n (y_j + \varepsilon)\mu(A_j) + \sum_{j=1}^n (y_j + \varepsilon)\frac{\varepsilon}{n} \\
&\leq \sum_{j=1}^n (y_j - \varepsilon)\mu(A_j) + 2\varepsilon\mu(\operatorname{supp}(f)) + (M + \varepsilon)\varepsilon \\
&\leq \sum_{j=1}^n \int_{A_j} f d\mu + \{2\mu(\operatorname{supp}(f)) + M + \varepsilon\}\varepsilon \\
&= \int_S f d\mu + \{2\mu(\operatorname{supp}(f)) + M + \varepsilon\}\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the desired inequality is established. ■

Example 1. To associate the Riemann integral of continuous functions f on the k-dimensional unit $S = [-1, 1]^k$ with a measure and the corresponding Lebesgue integral, apply the Riesz representation theorem to the bounded linear functional defined by

$$\ell(f) = \int_{-1}^{1} \cdots \int_{-1}^{1} f(x_1, \ldots, x_k) dx_1 \cdots dx_k. \tag{4.3}$$

APPENDIX B

Topology and Function Spaces

We begin with an important classical result. Let $C[0,1]$ denote the set of all real-valued continuous functions on $[0,1]$, endowed with the **sup norm:** $\|f\| = \max\{|f(x)| : x \in [0,1]\}$. With the distance $d(f,g) = \|f-g\|$, $C[0,1]$ is a real **Banach space** i.e., it is a vector space (with respect to real scalars) and it is a complete (normed) metric space. Recall that a **norm** $\| \quad \| : V \to [0,\infty)$ on a vector space V satisfies: $\|g\| = 0$ iff $g = 0$, $\|\alpha g\| = |\alpha| \cdot \|g\|$ (α scalar, $g \in V$), and $\|f+g\| \le \|f\| + \|g\|$. Also, a subset A of a metric space is **complete** if every Cauchy sequence in A has a convergent subsequence in A.

***Theorem 1.2** (Weirstrass Polynomial Approximation Theorem).* Polynomials are dense in $C[0,1]$.

Proof. Let $g \in C[0,1]$. Define a sequence h_n $(n \ge 1)$ of polynomials on $[0,1]$ as

$$h_n(p) = \sum_{i=0}^{n} g\left(\frac{i}{n}\right)\binom{n}{i} p^i (1-p)^{n-i} \qquad (p \in [0,1]),\ n \ge 1. \tag{1.4}$$

Then, for each p one may write $h_n(p) = \mathbb{E}g(X/n)$, where X is a binomial random variable $B(n,p)$. Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that $|g(p') - g(p'')| \le \varepsilon/2$, if $|p' - p''| \le \delta$ and p', $p'' \in [0,1]$. Hence

$$\begin{aligned}|h_n(p) - g(p)| &= |\mathbb{E}g(X/n) - g(p)| \\ &\le \frac{\varepsilon}{2} P\left(\left|\frac{X}{n} - p\right| \le \delta\right) + 2\|g\| P\left(\left|\frac{X}{n} - p\right| > \delta\right)\end{aligned}$$

$$\leq \frac{\varepsilon}{2} + 2\|g\| \frac{p(1-p)}{n\delta^2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all p if $n \geq \frac{\|g\|}{\varepsilon\delta^2}$. ■

Instead of $[0,1]$, we now consider an arbitrary compact Hausdorff space S. Recall that a topological space S (with a topology $\mathcal{T}$ of open sets) is **Hausdorff** if for every pair $x, y \in S$, $x \neq y$, there exist disjoint open sets U, V such that $x \in U$, $y \in V$. A topological space is **compact** if every open cover of S has a finite subcover. That is, if $\{V_\lambda : \lambda \in \Lambda\}$ is a collection of open sets such that $\cup_{\lambda\in\Lambda} V_\lambda = S$, then there exists a finite set $\{\lambda_1, \dots, \lambda_k\} \subseteq \Lambda$ such that $\cup\{V_{\lambda_i} : 1 \leq i \leq k\} = S$. By taking complements it is immediately seen that S is compact if and only if it has the **finite intersection property:** if $\{C_\lambda : \lambda \in \Lambda\}$ is a collection of closed sets whose intersection is empty, then it has a finite subcollection whose intersection is empty. The following are a few useful related notions. A topological space S is called **locally compact** if every point $x \in S$ has a compact neighborhood. The space S is called **σ-compact** if it is a countable union of compact sets. A *subset* D of a topological space $(S, \mathcal{T})$ is **compact** if it is *compact* as a topological space with the **relative topology** defined by $D \cap \mathcal{T}$.

It is simple to check, using the finite intersection property, that a real-valued continuous function on a compact space S attains its supremum (and infimum). From this it follows that the space $C(S)$ of real-valued continuous functions on S is a Banach space under the norm (called the **supnorm**) $\|f\| := \max\{|f(x)| : x \in S\}$. It is also an **algebra** i.e., it is a vector space that is also closed under (pointwise) multiplication $(f, g) \mapsto fg \; \forall f, g \in C(S)$. A **subalgebra** of $C(S)$ is a vector subspace that is also closed under multiplication. A subset $\mathcal{H}$ of $C(S)$ is said to **separate points** if for every pair of points $x \neq y$ in S there is a function $f \in \mathcal{H}$ such that $f(x) \neq f(y)$.

***Theorem 1.3** (Stone–Weierstrass Theorem).* Let S be a compact Hausdorff space, and $\mathcal{H}$ a subalgebra of $C(S)$. If $\mathcal{H}$ includes constant functions and separates points, then $\mathcal{H}$ is dense in S, i.e., $\overline{\mathcal{H}} = C(S)$.

Proof. Step 1. If $f \in \mathcal{H}$, then $|f| \in \overline{\mathcal{H}}$. To prove this use Theorem 1.2 to find a sequence h_n of polynomials converging uniformly to the function $h(p) = \sqrt{p}$ on $[0,1]$. Now if $f \in \mathcal{H}$ then all polynomial functions of $g \equiv f/\|f\|$ belong to $\mathcal{H}$. In particular, $h_n \circ g^2 \in \mathcal{H}$ ($g^2(x) \equiv (g(x))^2 \in [0,1]$). But $h_n \circ g^2$ converges uniformly on S to $h \circ g^2 = |f|/\|f\|$, so that the functions $(\|f\|)h_n \circ g^2$ in $\mathcal{H}$ converge uniformly to $|f|$.

Step 2. If $f, g \in \overline{\mathcal{H}}$ then $\max\{f, g\}$*,* $\min\{f, g\} \in \overline{\mathcal{H}}$*.* To see this, write $\max\{f, g\} = \frac{1}{2}(f + g + |f - g|)$, $\min\{f, g\} = \frac{1}{2}(f + g - |f - g|)$ and apply Step 1.

Step 3. Let $x \neq y \in S$, α and β real numbers. Then there exists $f \in \mathcal{H}$ such that $f(x) = \alpha$, $f(y) = \beta$. For this, find $g \in \mathcal{H}$ such that $a \equiv g(x) \neq b \equiv g(y)$. Let $f = \alpha + \frac{\beta-\alpha}{b-a}(g - a)$.

Step 4. Let $f \in C(S)$. Given any $x \in S$ and $\varepsilon > 0$, there exists $g \in \overline{\mathcal{H}}$ such that $g(x) = f(x)$ and $g(y) < f(y) + \varepsilon \; \forall y \in S$. To prove this, fix f, x, ε as above. By Step 3, for each $y \neq x$ there exists $g_y \in \mathcal{H}$ such that $g_y(x) = f(x)$, $g_y(y) = f(y) + \varepsilon/2$. Then y belongs to the open set $\mathcal{O}_y = \{z : g_y(z) < f(y) + \varepsilon\}$, and $S = \cup\{\mathcal{O}_y : y \in S\backslash\{x\}\}$.

Let $\{\mathcal{O}_{y_1}, \dots, \mathcal{O}_{y_k}\}$ be a subcover of S. Define $g = \min\{g_{y_1}, \dots, g_{y_k}\}$. Then $g \in \overline{\mathcal{H}}$ (by Step 2), $g(x) = f(x)$, and $g(y) < f(y) + \varepsilon \ \forall\, y \in S$.

Step 5. To complete the proof of the theorem, fix $f \in C(S)$, $\varepsilon > 0$. For each $x \in S$, let $f_x = g$ be the function obtained in Step 4. Then $V_x := [z \in S : f_x(z) > f(z) - \varepsilon]$, $x \in S$, form an open cover of S (since $x \in V_x$). Let $\{V_{x_1}, \dots, V_{x_m}\}$ be a finite subcover. Then $\overline{f}_\varepsilon \equiv \max\{f_{x_1}, f_{x_2}, \dots, f_{x_m}\} \in \overline{\mathcal{H}}$ (by Step 2), and $f(z) - \varepsilon < \overline{f}_\varepsilon(z) < f(z) + \varepsilon$ $\forall\, z \in S$. ■

Among many important applications of Theorem 1.3, let us mention two.

Corollary 1.4. Let S be a compact subset of $\mathbb{R}^m$ $(m \geq 1)$. Then the set $\mathcal{P}_m$ of all polynomials in m variables is dense in $C(S)$.

Proof. The set $\mathcal{P}_m$ is clearly an algebra that includes all constant functions. Also, let $\mathbf{x} = (x_1, \dots, x_m) \neq \mathbf{y} = (y_1, \dots, y_m) \in S$. Define $f(\mathbf{z}) = (z_1 - x_1)^2 + \cdots + (z_m - x_m)^2$. Then $f \in \mathcal{P}_m$, $f(\mathbf{x}) = 0$, $f(\mathbf{y}) > 0$. Hence Theorem 1.3 applies. ■

Corollary 1.5 ***(Separability of $C(S)$).*** Let (S, ρ) be a compact metric space. Then $C(S)$ is a separable metric space.

Proof. First observe that S is separable. For there exist finitely many open balls $\{B(x_{j,n} : 1/n) : 1 \leq j \leq k_n\}$ that cover S. Here $B(x : \varepsilon) = \{y \in S : \rho(x, y) < \varepsilon\}$ is a ball with center x and radius ε. Clearly, $\{x_{j,n} : j = 1, \dots, k_n;\ n = 1, 2, \dots\}$ is a countable dense subset of S. To prove separability of $C(S)$, let $\{x_n : n = 1, 2, \dots\}$ be a countable dense subset of S. Denote by $B_{n,k}$ the ball with center x_n and radius $1/k$ $(k = 1, 2, \dots;\ n = 1, 2, \dots)$. Also, let $h_m(u) = 1 - mu$, $0 \leq u < 1/m$, $h_m(u) = 0$ for $u \geq 1/m$, define a sequence of continuous functions on $[0, \infty)$ $(m = 1, 2, \dots)$. Define $f_{n,k,m}(x) := h_m(\rho(x, B_{n,k}))$ $(n \geq 1,\ k \geq 1,\ m \geq 1)$, and let $\mathcal{M}$ be the set of all (finite) linear combinations of *monomials* of the form $f^{j_1}_{n_1 k_1, m_1} \cdots f^{j_r}_{n_r, k_r, m_r}$ $(r \geq 1;\ j_1, \dots, j_r$ nonnegative integers). Then $\mathcal{M}$ is a subalgebra of $C(S)$ that includes constant functions and separates points: if $x \neq y$ then there exists $B_{n,k}$ such that $x \in B_{n,k}$ and $y \notin \overline{B}_{n,k}$, implying $f_{n,k,m}(x) = 1$, $f_{n,k,m}(y) < 1$, if m is sufficiently large. By Theorem 1.3, $\mathcal{M}$ is dense in $C(S)$. The countable subset of $\mathcal{M}$ comprising linear combinations of the monomials with rational scalars is dense in $\mathcal{M}$ and therefore in $\overline{\mathcal{M}} = C(S)$. ■

Remark 1.1. Let $C(S : \mathbb{C})$ denote the set of all complex-valued continuous functions on a compact metric space S. Under the sup norm $\|f\| := \sup\{|f(x)| : x \in S\}$, $C(S : \mathbb{C})$ is a (complex) Banach space. Letting $\{f_n : n = 1, 2, \dots\}$ be a dense sequence in the real Banach space $C(S)$, the countable set $\{f_n + i f_m : n \geq 1, m \geq 1\}$ is clearly dense in $C(S : \mathbb{C})$. Hence $C(S : \mathbb{C})$ is separable.

The next result concerns the **product topology** of the Cartesian product $S = \times_{\lambda \in \Lambda} S_\lambda$ of an arbitrary collection of compact spaces S_λ $(\lambda \in \Lambda)$. This topology comprises all arbitrary unions of sets of the form $V = [\mathbf{x} \equiv (x_\lambda : \lambda \in \Lambda) : x_{\lambda_i} \in V_{\lambda_i}, 1 \le i \le k]$, $\lambda_i \in \Lambda$, V_{λ_i} an open subset of S_{λ_i} $(1 \le i \le k)$, for some $k \ge 1$.

***Theorem 1.6** (Tychonov's Theorem).* Let S_λ be compact for all $\lambda \in \Lambda$. Then $S = \times_{\lambda\in\Lambda} S_\lambda$ is compact under the product topology.

Proof. We will give a proof when Λ is denumerable, say $\Lambda = \{1, 2, \dots\}$, and (S_n, ρ_n) are compact metric spaces, $n \ge 1$. The proof of the general case requires invoking the axiom of choice, and may be found in Folland.[1]

Let $\mathbf{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, \dots)$, $n \ge 1$, be a sequence in S. We will find a convergent subsequence. Let $\{x_1^{(n(1))} : n \ge 1\}$ be a subsequence of $\{x_1^{(n)} : n \ge 1\}$, converging to some $x_1 \in S_1$, $n(1) > n$ $\forall n$. Let $\{x_2^{(n(2))} : n \ge 1\}$ be a subsequence of $\{x_2^{(n(1))} : n \ge 1\}$, converging to some $x_2 \in S_2$, $n(2) > n(1)$ $\forall n$. In general, let $\{x_k^{(n(k))} : n \ge 1\}$ be a subsequence of $\{x_k^{(n(k-1))} : n \ge 1\}$, converging to $x_k \in S_k$ $(k = 1, 2, \dots)$. Then the *diagonal subsequence* $\mathbf{x}^{(1(1))}, \mathbf{x}^{(2(2))}, \dots, \mathbf{x}^{(k(k))}, \dots$ converges to $\mathbf{x} = (x_1, x_2, \dots, x_k, \dots)$. ■

Definition 1.1. A family $\mathcal{C}$ of continuous functions defined on a topological space S is said to be **equicontinuous** at a point $x \in S$ if for every $\varepsilon > 0$ there is a neighborhood U of x such that for every $f \in \mathcal{C}$, $|f(y) - f(x)| < \varepsilon$ for all $y \in U$. If $\mathcal{C}$ is equicontinuous at each $x \in S$ then $\mathcal{C}$ is called **equicontinuous**. Also $\mathcal{C}$ is said to be **uniformly bounded** if there is a number $M > 0$ such that $|f(x)| \le M$ for all $f \in \mathcal{C}$ and all $x \in S$.

The next concept is especially useful in the context of $C(S)$ viewed as a metric space with the uniform metric.

Definition 1.2. A subset A of a metric space is said to be **totally bounded** if, for every $\delta > 0$ there is a covering of A by finitely many balls of radius δ.

Lemma 5. If A is a complete and totally bounded subset of a metric space then A is compact.

Proof. Let $\{x_n : n \ge 1\}$ be a sequence in A. Since A may be covered by finitely many balls of radii $1/2$, one of these, denoted by B_1, must contain x_n for infinitely many n, say $n \in N_1$. Next $A \cap B_1$ may be covered by finitely many balls of radii $1/4$. One of these balls, denoted by B_2, contains $\{x_n : n \in N_1\}$ for infinitely many n, say $n \in N_2 \subseteq N_1$. Continuing in this way, by selecting distinct points $n_1 < n_2 < \cdots$ from $N_1, N_2, \dots$, one may extract a subsequence $\{x_{n_m} : n_m \in N_m\}$ which is Cauchy and, since A is complete, converges in A. Now suppose that $\{U_\lambda : \lambda \in \Lambda\}$ is an open cover of A. In view of the total boundedness of A, if for some $\varepsilon > 0$ one can show that every ball of radius ε which meets A is a subset of some U_λ, then a finite subcover

[1] Folland, G.B. (1984). *Real Analysis,* p. 130. Wiley.

exists. To see that this is indeed the case, suppose not. That is, suppose for every $n \geq 1$ there is a ball B_n of radius at most 2^{-n} which meets A but is not a subset of any U_λ. For each n there is an $x_n \in B_n \cap A$. Since there is a convergent subsequence to $x \in A$, one has $x \in U_\lambda$ for some $\lambda \in \Lambda$. Since U_λ is open and since x is a limit point of the sequence x_n, it follows that $x \in B_n \subseteq U_\lambda$ for n sufficiently large. This is a contradiction to the construction of $B_n, n \geq 1$. ■

***Theorem 1.7** (Arzelà-Ascoli).* A collection $\mathcal{C} \subset C[a,b]$ is relatively compact for the uniform metric on $C[a,b]$ if and only if $\mathcal{C}$ is uniformly bounded and equicontinuous.

Proof. Assume that $\mathcal{C}$ is uniformly bounded and equicontinuous. In view of Lemma 5, it is enough to show the closure of $\mathcal{C}$ is totally bounded and complete to prove relative compactness. The completeness follows from the completeness of $C[a,b]$. For total boundedness it is sufficient to check that $\mathcal{C}$ is totally bounded, since this will be preserved in the closure. Let $\delta > 0$. By equicontinuity, for each $x \in [a,b]$ there is an open set U_x containing x such that $|f(y) - f(x)| < \delta/4$ for all $y \in U_x$, and all $f \in \mathcal{C}$. By compactness of $[a,b]$, there are finitely many points $x_1, \dots, x_n$ in $[a,b]$ such that $\cup_{j=1}^n U_{x_j} = [a,b]$. Now $\{f(x_j) : f \in \mathcal{C},\ j = 1, \dots, n\}$ is a bounded set. Thus there are numbers $y_1, \dots, y_m$ such that for each $f \in \mathcal{C}$, and each j, $|f(x_j) - y_k| < \delta/4$ for some $1 \leq k \leq m$. Let $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$. The set Y^X of functions from X into Y is a finite set and $\mathcal{C} = \cup_{g \in Y^X} \mathcal{C}_g$, where $\mathcal{C}_g := \{f \in \mathcal{C} : |f(x_j) - g(x_j)| < \delta/4, 1 \leq j \leq n\}$. Now, to complete the proof of total boundedness, let us see that this covering of $\mathcal{C}$ is by sets $\mathcal{C}_g$ of diameter at most δ. Let $f, h \in \mathcal{C}_g$. Then $|f - h| < \delta/2$ on X. For $x \in [a,b]$, one has $x \in U_{x_j}$ for some j, and therefore $|f(x) - h(x)| \leq |f(x) - f(x_j)| + |f(x_j) - h(x_j)| + |h(x_j) - h(x)| < \delta$.

To prove necessity, let us first observe that if $\overline{\mathcal{C}}$ is compact, then $\overline{\mathcal{C}}$ is totally bounded. For suppose not. Then there is a $\delta > 0$ such that there is no finite cover by balls of radii δ. Thus, for arbitrary but fixed $g_1 \in \overline{\mathcal{C}}$, there is a $g_2 \in \overline{\mathcal{C}}$ such that $\|g_1 - g_2\| := \max_{a \leq x \leq b} |g_1(x) - g_2(x)| > \delta$. This is because otherwise, the ball centered at g_1 would be a cover of $\overline{\mathcal{C}}$. Proceeding by induction, having found $g_1, \dots, g_n$, there must be a $g_{n+1} \in \overline{\mathcal{C}}$ such that $\|g_k - g_{n+1}\| > \delta$ for $k = 1, \dots n$. Thus, there is an infinite sequence $g_1, g_2, \dots$ in $\overline{\mathcal{C}}$ such that $\|g_j - g_k\| > \delta$ for $j \neq k$. Thus $\overline{\mathcal{C}}$ cannot be compact. Now, since $\overline{\mathcal{C}}$ is totally bounded, given any $\varepsilon > 0$ there exist $g_1, \dots, g_n \in \overline{\mathcal{C}}$ such that for any $f \in \overline{\mathcal{C}}$ one has $\|f - g_k\| < \frac{\varepsilon}{3}$ for some $1 \leq k \leq n$. Since each g_k is a continuous function on the compact interval $[a,b]$, it is bounded. Let $M = \max_{1 \leq k \leq n, a \leq x \leq b} |g_k(x)| + \frac{\varepsilon}{3}$. Then, for $f \in \overline{\mathcal{C}}$, one has $|f(x)| \leq |g_k(x)| + \frac{\varepsilon}{3} \leq M$ for all $a \leq x \leq b$. Thus $\overline{\mathcal{C}}$ is uniformly bounded. Since each g_k is continuous and hence, uniformly continuous on $[a,b]$, there is a $\delta_k > 0$ such that $|g_k(x) - g_k(y)| < \frac{\varepsilon}{3}$ if $|x - y| < \delta_k$. Let $\delta = \min\{\delta_1, \dots, \delta_n\}$. Then for $f \in \overline{\mathcal{C}}$ one has for suitably chosen g_k, $|f(x) - f(y)| \leq \|f - g_k\| + |g_k(x) - g_k(y)| + \|g_k - f\| < \varepsilon$ if $|x - y| < \delta$. Thus $\overline{\mathcal{C}}$ and hence, $\mathcal{C} \subseteq \overline{\mathcal{C}}$ is equicontinuous. ■

Note that the theorem holds for a compact metric space S in place of $[a,b]$, with virtually the same proof.

APPENDIX C

Hilbert Spaces and Applications in Measure Theory

1 HILBERT SPACES

Let H be a real vector space endowed with an **inner-product** $(x, y) \mapsto \langle x, y \rangle$, i.e.,

(i) $\langle x, y \rangle = \langle y, x \rangle$ *(symmetry),*

(ii) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle \; \forall \, \alpha, \beta \in \mathbb{R}$ *(linearity),* and

(iii) $\langle x, x \rangle \geq 0 \; \forall \, x$, with equality iff $x = 0$ *(positive definiteness).* One writes $\|x\|^2 = \langle x, x \rangle$.

Among the basic inequalities on H are the **parallelogram law**

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2, \tag{1.1}$$

which is easy to check, and the **Cauchy–Schwarz inequality**,

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|. \tag{1.2}$$

To prove this, fix x, y. If x or y is 0, this inequality is trivial. Assume then that x, y are nonzero. Since for all $u \in \mathbb{R}$,

$$0 \leq \|x + uy\|^2 = \|x\|^2 + u^2\|y\|^2 + 2u\langle x, y \rangle,$$

minimizing the right side with $u = -\langle x, y \rangle / \|y\|^2$, one gets $0 \leq \|x\|^2 - \langle x, y \rangle^2 / \|y\|^2$, from which (1.2) follows. One can now derive the **triangle inequality**

$$\|x + y\| \leq \|x\| + \|y\|, \tag{1.3}$$

by observing that $\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle \leq \|x\|^2 + \|y\|^2 + 2\|x\| \cdot \|y\| = (\|x\| + \|y\|)^2$, in view of (1.2). Thus $\| \cdot \|$ is a **norm**: (a) $\|x\| \geq 0$, with equality iff $x = 0$, (b) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $x \in H$ and real scalar α, and (c) $\|x+y\| \leq \|x\| + \|y\|$ for all x, $y \in H$. If H is a *complete* metric space under the metric $d(x,y) = \|x-y\|$, then H is said to be a **(real) Hilbert space**.

Lemma 1. Let M be a closed linear subspace of the Hilbert space H. Then for each $x \in H$, the distance $d \equiv d(x, M) \equiv \inf\{d(x,y) : y \in M\}$ is attained at some $z : d = d(x,z)$.

Proof. Let z_n be such that $d(x, z_n) \to d$ as $n \to \infty$, $z_n \in M\ \forall n$. By (1.1), with $x - z_n$, $x - z_m$ for x and y, respectively, one has

$$\begin{aligned}\|z_n - z_m\|^2 &= 2\|x - z_n\|^2 + 2\|x - z_m\|^2 - \|2x - z_n - z_m\|^2 \\ &= 2\|x - z_n\|^2 + 2\|x - z_m\|^2 - 4\|x - \frac{1}{2}(z_n + z_m)\|^2 \\ &\leq 2\|x - z_n\|^2 + 2\|x - z_m\|^2 - 4d^2 \longrightarrow 2d^2 + 2d^2 - 4d^2 = 0,\end{aligned}$$

showing that $\{z_n : n \geq 1\}$ is a Cauchy sequence in M. Letting $z = \lim z_n$, one gets the desired result. ■

Theorem 1.1 *(Projection Theorem).* Let M be a closed linear subspace of a real Hilbert space H. Then each $x \in H$ has a unique representation: $x = y + z$, $y \in M$, $z \in M^{\perp} \equiv \{w \in H : \langle w, v\rangle = 0\ \forall\, v \in M\}$.

Proof. Let $x \in H$. Let $y \in M$ be such that $d \equiv d(x, M) = d(x, y)$. Define $z = x - y$. Then $x = y + z$. For all $u \in \mathbb{R}$ and $w \in M$, $w \neq 0$, one has

$$d^2 \leq \|x - (y + uw)\|^2 = \|x - y\|^2 + u^2\|w\|^2 - 2u\langle x - y, w\rangle. \tag{1.4}$$

If $\langle x - y, w\rangle \neq 0$, one may set $u = \langle x - y, w\rangle / \|w\|^2$ to get $d^2 \leq \|x - y\|^2 - \langle x - y, w\rangle^2 / \|w\|^2 < d^2$, which is impossible, implying $\langle x - y, w\rangle = 0\ \forall w \in M$. Hence $z \in M^{\perp}$. To prove uniqueness of the decomposition, suppose $x = w + v$, $w \in M$, $v \in M^{\perp}$. Then $w + v = y + z$, and $w - y = z - v$. But $w - y \in M$ and $z - v \in M^{\perp}$, implying $w - y = 0$, $z - v = 0$. ■

The function $x \mapsto y$ in Theorem 1.1 is called the **(orthogonal) projection** onto M, and $x \to z$ is the orthogonal projection onto $M^{\perp}$. It is simple to check that these projections are linear maps (on H onto M, and on H onto $M^{\perp}$).

We will denote by H^* the set of all real-valued continuous linear functions (functionals) on H. Note that if ℓ_1, $\ell_2 \in H^*$ and α, $\beta \in \mathbb{R}$, then $\alpha\ell_1 + \beta\ell_2 \in H^*$, i.e., H^* is a real vector space. It turns out that H^* is **isomorphic** to H. To see this, note that for each $y \in H$, the functional ℓ_y, defined by $\ell_y(x) = \langle x, y\rangle$, belongs to H^*. Conversely, one has the following result.

Theorem 1.2 *(Riesz Representation Theorem on Hilbert Spaces).* If $\ell \in H^*$, there exists a unique $y \in H$ such that $\ell(x) = \langle x, y\rangle \; \forall\, x \in H$.

Proof. Since $\ell = 0$ is given by $\ell(x) = \langle x, 0\rangle$, and corresponds to ℓ_0, assume $\ell \neq 0$. Then $M \equiv \{x \in H : \ell(x) = 0\}$ is a closed proper linear subspace of H, $M \neq H$, and therefore $M^\perp \neq \{0\}$. Let $z \in M^\perp$, $\|z\| = 1$. Consider, for any given $x \in H$, the element $w = \ell(x)z - \ell(z)x \in H$, and note that $\ell(w) = 0$. Thus $w \in M$, so that $0 = \langle w, z\rangle = \ell(x) - \ell(z)\langle x, z\rangle$, implying $\ell(x) = \ell(z)\langle x, z\rangle = \langle x, y\rangle$, where $y = \ell(z)z$.

To prove uniqueness of the representation, suppose $\langle x, y_1\rangle = \langle x, y_2\rangle \; \forall\, x \in H$. With $x = y_1 - y_2$ one gets $0 = \langle x, y_1 - y_2\rangle = \|y_1 - y_2\|^2$, so that $y_1 = y_2$. ■

A **complex vector space** H is a vector space with the complex scalar field $\mathbb{C}$. An **inner product** on such a space is a function $\langle\ ,\ \rangle$ on $H \times H$ into $\mathbb{C}$ satisfying (i) $\langle x, y\rangle = \overline{\langle y, x\rangle}$, ($\overline{\alpha}$ is the complex conjugate of $\alpha \in \mathbb{C}$), (ii) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$, (iii) $\|x\|^2 \equiv \langle x, x\rangle > 0 \; \forall\, x \neq 0$. If H, with distance $d(x, y) = \|x - y\|$, is a complete metric space, it is called a **complex Hilbert space**. The parallelogram law (1.1) follows easily in this case. For the Cauchy–Schwarz inequality (1.2), take $u = -\frac{\langle x,y\rangle}{\|y\|^2}$ in the relations (for arbitrary $x, y \in H$, and $u \in \mathbb{C}$)

$$0 \leq \|x + uy\|^2 = \|x\|^2 + |u|^2 \cdot \|y\|^2 + \overline{u}\langle x, y\rangle + u\overline{\langle x, y\rangle}$$

to get $0 \leq \|x\|^2 - |\langle x, y\rangle|^2/\|y\|^2$, from which (1.2) follows. The proof of the lemma remains unchanged for complex H. The triangle inequality (1.3) follows as in the case of real H. In the proof of the projection theorem, (1.4) changes (for $u \in \mathbb{C}$) to

$$d^2 \leq \|x - y\|^2 + |u|^2 \cdot \|w\|^2 - \overline{u}\langle x - y, w\rangle - u\overline{\langle x - y, w\rangle},$$

so that taking $u = \langle x - y, w\rangle/\|w\|^2$, one gets $d^2 \leq \|x - y\|^2 - |\langle x - y, w\rangle|^2/\|w\|^2$, which implies $\langle x - y, w\rangle = 0 \; \forall\, w \in M$. The rest of the proof remains intact. For the proof of the Riesz representation, the relation $0 = \ell(x) - \ell(z)\langle x, z\rangle$ implies $\ell(x) = \ell(z)\langle x, z\rangle = \langle x, \overline{\ell(z)}z\rangle = \langle x, y\rangle$ with $y = \overline{\ell(z)}z$ (instead of $\ell(z)z$ in the case of real H). Thus Theorems 1.1, 1.2 hold for complex Hilbert spaces H also.

A set $\{x_i : i \in I\} \subseteq H$ is **orthonormal** if $\langle x_i, x_j\rangle = 0$ for all $i \neq j$, and $\|x_i\| = 1$. An orthonormal set is **complete** if $\langle x, x_i\rangle = 0$ for all $i \in I$ implies $x = 0$. A complete orthonormal subset of a Hilbert space is called an **orthonormal basis** of H. Suppose H is a separable Hilbert space with a dense set $\{y_n : n = 1, 2, \ldots\}$. By the following **Gram–Schmidt procedure** one can construct a countable orthonormal basis for H. Without loss of generality assume $y_n \neq 0$, $n \geq 1$. Let $x_1 = y_1/\|y_1\|, u_2 = y_2 - \langle y_2, x_1\rangle x_1, x_2 = u_2/\|u_2\|$, assuming $u_2 \neq 0$. If $u_2 = 0$, replace y_2 by the first y in the sequence such that $y - \langle y, x_1\rangle x_1 \neq 0$, and relabel $y = y_2$. Having constructed $u_2, x_2, \ldots, u_n, x_n$ in this manner, define $u_{n+1} = y_{n+1} - \sum_{j=1}^{n}\langle y_{n+1}, x_j\rangle x_j$, $x_{n+1} = y_{n+1}/\|y_{n+1}\|$, assuming $u_{n+1} \neq 0$ (if $u_{n+1} = 0$ then find the first y in the sequence that is not linearly dependent on $\{x_1, \ldots, x_n\}$ and relabel it y_{n+1}). The process terminates after a finite number of steps if H is finite dimensional. Otherwise, one obtains a complete orthonormal sequence $\{x_n : n = 1, 2, \ldots\}$. Completeness follows from the fact that if $\langle x, x_n\rangle = 0$ for all n, then $\langle x, y_n\rangle = 0$ for all n, so that $x \in \{y_n : n \geq 1\}^{\perp} = \{0\}$ (since $\{y_n : n \geq 1\}$ is dense in H, and $A^{\perp}$ is a closed set for all $A \subseteq H$). A complete orthonormal set is called a **complete orthonormal basis**, in view of Theorem 1.3 below.

Lemma 2 (Bessel's Inequality). *Let $\{x_1, x_2, \ldots\}$ be a finite or countable orthonormal subset of a Hilbert space H. Then $\sum_n |\langle x, x_n\rangle|^2 \leq \|x\|^2$ for all $x \in H$.*

Proof. One has

$$\begin{aligned}\|x - \sum_n \langle x, x_n\rangle x_n\|^2 &= \|x\|^2 - 2\,\mathrm{Re}\left\langle x, \sum_n \langle x, x_n\rangle x_n\right\rangle + \left\|\sum_n \langle x, x_n\rangle x_n\right\|^2 \\ &= \|x\|^2 - 2\,\mathrm{Re}\sum_n |\langle x, x_n\rangle|^2 + \sum_n |\langle x, x_n\rangle|^2 = \|x\|^2 - \sum_n |\langle x, x_n\rangle|^2.\end{aligned}$$

The inequality is proven since the expression is nonnegative. ■

Theorem 1.3. *Let $\{x_1, x_2, \ldots\}$ be a complete orthonormal set in a separable Hilbert space. Then for all x one has (a) (Fourier Expansion) $x = \sum_n \langle x, x_n\rangle x_n$, and (b) (Parseval's Equation) $\|x\|^2 = \sum_n |\langle x, x_n\rangle|^2$, and $\langle x, y\rangle = \sum_n \langle x, x_n\rangle \overline{\langle y, y_n\rangle}$, for all $x, y \in H$.*

Proof. (a) In view of Bessel's inequality, the series $\sum_n |\langle x, x_n\rangle|^2$ converges, so that $\left\|\sum_M^N \langle x, x_n\rangle x_n\right\|^2 \to 0$ as $M, N \to \infty$. Hence $\sum_n \langle x, x_n\rangle x_n$ converges to $z \in H$, say. Since $\langle x_m, x_n\rangle = 0$, $n \neq m$, and $\langle x_m, x_m\rangle = 1$, it follows that $\langle z, x_m\rangle = \langle x, x_m\rangle$ for all m. Therefore, $\langle z - x, x_m\rangle = 0$ for all m, and hence by completeness of $\{x_n : n \geq 1\}$, $x = z = \sum_n \langle x, x_n\rangle x_n$. Also, the first calculation in (b) follows, since $\|x\|^2 = \|\sum_n \langle x, x_n\rangle x_n\|^2 = \sum_n |\langle x, x_n\rangle|^2$. More generally, one has $x = \sum_n \langle x, x_n\rangle x_n$, $y = \sum_n \langle y, x_n\rangle x_n$, so that $\langle x, y\rangle = \sum_n \langle x, x_n\rangle \overline{\langle y, x_n\rangle}$, using (i) convergence of $\sum_{n=1}^N \langle x, x_n\rangle x_n$ to x and that of $\sum_{n=1}^N \langle y, x_n\rangle x_n$ to y as $N \to \infty$, and (ii) the continuity of the inner product $(u, v) \to \langle u, v\rangle$ as a function on $H \times H$. ■

For a real Hilbert space, the conjugation sign in (b) is dropped.

2 LEBESGUE DECOMPOSITION AND THE RADON–NIKODYM THEOREM

In the subsection we will give von Neumann's elegant proof of one of the most important results in measure theory.

Let μ, ν be measures on a measurable space $(S, \mathcal{S})$. One says that ν **is absolutely continuous** with respect to μ, $\nu \ll \mu$ in symbols, if $\nu(B) = 0\ \forall\, B \in \mathcal{S}$ for which $\mu(B) = 0$. At the other extreme, ν is **singular** with respect to μ if there exists $A \in \mathcal{S}$ such that $\nu(A) = 0$ and $\mu(A^C) = 0$, that is, if μ and ν concentrate their entire masses on disjoint sets: one then writes $\nu \perp \mu$. Note that $\nu \perp \mu$ implies $\mu \perp \nu$. However, $\nu \ll \mu$ does not imply $\mu \ll \nu$.

Theorem 2.1 *(Lebesgue Decomposition and the Radon–Nikodym Theorem)****.*** Let μ, ν be σ-finite measures on $(S, \mathcal{S})$. Then (a) there exist unique measures $\nu_a \ll \mu$ and $\nu_s \perp \mu$ such that $\nu = \nu_a + \nu_s$ *(Lebesgue decomposition),* and there exists a μ-a.e. unique nonnegative measurable h such that $\nu_a(B) = \int_B h\, d\mu\ \forall\, B \in \mathcal{S}$. (b) In particular, if $\nu \ll \mu$, then there exists a μ-a.e. unique $h \geq 0$ such that $\nu(B) = \int_B h\, d\mu$ $\forall\, B \in \mathcal{S}$ *(Radon–Nikodym theorem).*

Proof. First consider the case of *finite* μ, ν. Write $\lambda = \mu + \nu$. On the real Hilbert space $L^2(\lambda) \equiv L^2(S, \mathcal{S}, \lambda)$, define the linear functional

$$\ell(f) = \int_S f\, d\nu \qquad f \in L^2(\lambda). \tag{2.1}$$

By the Cauchy–Schwarz inequality, writing $\|f\| = (\int |f|^2 d\lambda)^{1/2}$, we have

$$|\ell(f)| \leq \int_S |f| d\lambda \leq \|f\| \cdot (\lambda(S))^{\frac{1}{2}}. \tag{2.2}$$

Thus ℓ is a continuous linear functional on $L^2(\lambda)$. By the Riesz representation theorem (Theorem 1.2), there exists $g \in L^2(\lambda)$ such that

$$\ell(f) \equiv \int_S f\, d\nu = \int_s fg\, d\lambda \qquad (f \in L^2(\lambda)). \tag{2.3}$$

In particular, for $f = \mathbf{1}_B$,

$$\nu(B) = \int_B g\, d\lambda, \qquad \forall\, B \in \mathcal{S}. \tag{2.4}$$

Letting $B = \{x \in S : g(x) > 1\} = E$, say, one gets $\lambda(E) = 0 = \nu(E)$. For if $\lambda(E) > 0$, then (2.4) implies $\nu(E) = \int_E g\, d\lambda > \lambda(E)$, which is impossible. Similarly, letting $F = \{x : g(x) < 0\}$, one shows that $\lambda(F) = 0$. Modifying g on a λ-null set if necessary, we take g to satisfy $0 \leq g \leq 1$ on S. Consider the sets $S_1 = \{x : 0 \leq g(x) < 1\}$ and $S_2 = S_1^c = \{x : g(x) = 1\}$, and define the following measures ν_1, ν_2:

$$\nu_1(B) := \nu(B \cap S_1), \quad \nu_2(B) := \nu(B \cap S_2), \qquad B \in \mathcal{S}. \tag{2.5}$$

Now, using $\lambda = \mu + \nu$, one may rewrite (2.4) as

$$\int_B (1-g)\,d\nu = \int_B g\,d\mu \qquad (B \in \mathcal{S}). \tag{2.6}$$

For $B = S_2$, the left side is zero, while the right side is $\mu(S_2)$, i.e., $\mu(S_2) = 0$. Since $\nu_2(S_2^c) = 0$ by definition, one has $\nu_2 \perp \mu$. On the other hand, on S_1, $1 - g > 0$, so that $\mu(B) = 0 \Longrightarrow \int_{B\cap S_1}(1-g)d\nu = 0 \Longrightarrow \nu(B \cap S_1) = 0$, i.e., $\nu_1(B) = 0$. Hence $\nu_1 \ll \mu$. Thus we have a Lebesgue decomposition $\nu = \nu_a + \nu_s$, with $\nu_a = \nu_1$, $\nu_s = \nu_2$. Its *uniqueness* follows from Corollary 2.3 below. Multiplying both sides of (2.6) by 1, g, g^2, ..., g^n, and adding, we get

$$\int_B (1-g^{n+1})\,d\nu = \int_B (g + g^2 + \cdots + g^{n+1})\,d\mu \qquad (B \in \mathcal{S}). \tag{2.7}$$

Since $1 - g^{n+1} \uparrow 1$ (as $n \uparrow \infty$) on S_1, denoting by h the increasing limit of $g + g^2 + \cdots + g^{n+1}$, one gets

$$\nu_a(B) \equiv \nu_1(B) = \nu(B \cap S_1) = \int_{B\cap S_1} h\,d\mu = \int_B h\,d\mu \qquad (B \in \mathcal{S}),$$

completing the proof of (a). Now (b) is a special case of (a). The *uniqueness* of the function h in this case does not require Proposition 2.2 below. For if $\int_B h\,d\mu = \int_B h'\,d\mu$ $\forall\, B \in \mathcal{S}$, then $\int_B (h-h')\,d\mu = 0\ \forall\, B \in \mathcal{S}$. In particular $\int_{\{h>h'\}}(h-h')\,d\mu = 0$ and $\int_{\{h\leq h'\}}(h'-h)\,d\mu = 0$, so that $\int |h-h'|d\mu = 0$.

For the general case of σ-finite measures μ, ν, let $\{A_n : n \geq 1\}$ be a sequence of pairwise disjoint sets in $\mathcal{S}$ such that $\cup_{n=1}^{\infty} A_n = S$, $\mu(A_n) < \infty$, $\nu(A_n) < \infty\ \forall\, n$. Applying the above result separately to each A_n and adding up one gets the desired result, using the monotone convergence theorem. ■

For the next result, call ν a finite **signed measure** if $\nu : \mathcal{S} \to (-\infty, \infty)$ satisfies $\nu(\emptyset) = 0$, and $\nu(\cup_n B_n) = \sum_n \nu(B_n)$ for every pairwise disjoint sequence B_n $(n \geq 1)$ in $\mathcal{S}$. If ν takes one of the two values $-\infty$, ∞, but not both, ν is said to be **σ-finite signed measure** if there exists a sequence of pairwise disjoint sets $B_n \in \mathcal{S}$ such that ν is a finite signed measure on each B_n $(n \geq 1)$, and $S = \cup_n B_n$.

Proposition 2.2 (Hahn–Jordan Decomposition). Suppose ν is a σ-finite signed measure on $(S, \mathcal{S})$. Then (a) there exists a set $C \in \mathcal{S}$ such that $\nu(C \cap B) \geq 0$ $\forall\, B \in \mathcal{S}$, and $\nu(C^c \cap B) \leq 0\ \forall\, B \in \mathcal{S}$ *(Hahn decomposition),* and (b) defining the measures $\nu^+(B) := \nu(C \cap B)$, $\nu^-(B) := -\nu(C^c \cap B)$, one has $\nu = \nu^+ - \nu^-$ *(Jordan decomposition).*

Proof. First assume that ν is finite, and let $u = \sup\{\nu(B) : B \in \mathcal{S}\}$. Let $B_n \in \mathcal{S}$ $(n \geq 1)$ be such that $\nu(B_n) \to u$. We will construct a set $C \in \mathcal{S}$ such that $\nu(C) = u$. For each m, consider the partition Γ_m of S by 2^m sets of the form $B_1' \cap B_2' \cap \cdots \cap B_m'$ with $B_i' = B_i$

or B_i^c, $1 \le i \le m$. Let A_m be the union of those among these sets whose ν-measures are nonnegative. Clearly, $\nu(A_m) \ge \nu(B_m)$. Expressing $A_m \cup A_{m+1}$ as a (disjoint) union of certain members of the partition Γ_{m+1} and noting that those sets in Γ_{m+1} that make up $A_{m+1} \backslash A_m$ all have nonnegative ν-measures, one has $\nu(A_m \cup A_{m+1}) \ge \nu(A_m)$. By the same argument, $\nu(A_m \cup A_{m+1} \cup A_{m+2}) \ge \nu(A_m \cup A_{m+1}) \ge \nu(A_m)$, and so on, so that $\nu(\cup_{i=m}^n A_i) \ge \nu(A_m)\ \forall\, n \ge m$, implying that $C_m \equiv \cup_{i=m}^\infty A_i$ satisfies $\nu(C_m) \ge \nu(A_m) \ge \nu(B_m)$. Hence $\nu(C) = u$, where $C = \lim_{m\to\infty} C_m$. We will now show that $\nu(B \cap C) \ge 0$ and $\nu(B \cap C^c) \le 0\ \forall\, B \in \mathcal{S}$. First note that $u < \infty$, since ν is finite. Now if $\nu(B \cap C) < 0$ for some B, then $\nu(C \backslash (B \cap C)) > u$, which is impossible. Similarly, if $\nu(B \cap C^c) > 0$ for some B, then $\nu(C \cup (B \cap C^c)) > u$. We have proved the Hahn decomposition (a). The Jordan decomposition (b) follows immediately from this.

If ν is σ-finite, then S is a disjoint union of sets A_n $(n \ge 1)$ such that $\nu_n(B) \equiv \nu(A_n \cap B)$, $B \in \mathcal{S}$, is a finite signed measure for all $n \ge 1$. The Hahn–Jordan decomposition $\nu_n = \nu_n^+ - \nu_n^-$ leads to the corresponding decomposition of $\nu = \nu^+ - \nu^-$, with $\nu^+ = \sum_n \nu_n^+$, $\nu^- = \sum_n \nu_n^-$. ■

The measure $|\nu| := \nu^+ + \nu^-$ is called the **total variation** of a σ-finite signed measure ν.

Corollary 2.3. The Hahn–Jordan decomposition of a σ-finite signed measure ν is the unique decomposition of ν as the difference between two mutually singular σ-finite measures.

Proof. It is enough to assume that ν is finite. Let $\nu = \gamma_1 - \gamma_2$ where $\gamma_1 \perp \gamma_2$ are measures, with $\gamma_1(D^c) = 0$, $\gamma_2(D) = 0$ for some $D \in \mathcal{S}$. Clearly, $\gamma_1(D) = \gamma_1(S) = \sup\{\nu(B) : B \in \mathcal{S}\} = \nu(D) = u$, say. As in the proof of Proposition 2.2, it follows that $\gamma_1(B) = \nu(B \cap D)$, $\gamma_2(B) = -\nu(B \cap D^c)$ for all $B \in \mathcal{S}$. If C is as in Proposition 2.2, then $u = \nu(C)$. Suppose, if possible, $\nu^+(B) \equiv \nu(B \cap C) > \gamma_1(B) = \nu(B \cap D)$, i.e., $\nu(B \cap C \cap D^c) + \nu(B \cap C \cap D) > \nu(B \cap D \cap C) + \nu(B \cap D \cap C^c)$, or, $\nu(B \cap C \cap D^c) > \nu(B \cap D \cap C^c) = \gamma_1(B \cap C^c) \ge 0$. But then $\nu(D \cup (B \cap C \cap D^c)) > \nu(D) = \gamma_1(D) = u$, a contradiction. Hence $\nu^+(B) \le \gamma_1(B)\ \forall\, B \in \mathcal{S}$. Similarly, $\gamma_1(B) \le \nu^+(B)\ \forall\, B \in \mathcal{S}$. ■

One may take ν to be a σ-finite signed measure in Theorem 2.1 (and μ a σ-finite measure). Then ν is *absolutely continuous with respect to* μ, $\nu \ll \mu$, if $\mu(B) = 0 \Longrightarrow \nu(B) = 0$ $(B \in \mathcal{S})$. Use the Hahn–Jordan decomposition $\nu = \nu^+ - \nu^-$, and apply Theorem 2.1 separately to ν^+ and ν^- : $\nu^+ = (\nu^+)_a + (\nu^+)_s$, $\nu^- = (\nu^-)_a + (\nu^-)_s$. Then let $\nu_a = (\nu^+)_a - (\nu^-)_a$, $\nu_s = (\nu^+)_s - (\nu^-)_s$.

Corollary 2.4. Theorem 2.1 extends to σ-finite signed measures ν, with $\nu = \nu_a + \nu_s$, where ν_a and ν_s are σ–finite signed measures, $\nu_a \ll \mu$, $\nu_s \perp \mu$. Also, there exists a measurable function h, unique up to a μ–null set, such that $\nu_a(B) = \int_B h\, d\mu\ \forall\, B \in \mathcal{S}$. If $\nu \ll \mu$, then $\nu(B) = \int_B h\, d\mu\ \forall\, B \in \mathcal{S}$.

References

The following is a somewhat extensive listing of supplementary and/or follow-up textbook references covering some of the same topics and/or further applications of material introduced in this basic course.

Aldous, D. (1989): *Probability Approximations via the Poisson Clumping Heuristic*, Springer-Verlag, NY.

Asmussen, S., Hering, H. (1983): *Branching Processes*, Birkhäuser, Boston.

Athreya, K.B., Lahiri, S.N. (2006): *Measure Theory and Probability Theory*, Springer Texts in Statistics, Springer-Verlag, NY.

Athreya, K.B., Ney, P.E. (1972): *Branching Processes*, Springer-Verlag, NY.

Bass, R. (1995): *Probabilistic Techniques in Analysis*, Springer-Verlag, NY

Bauer, H. (1972): *Probability Theory and Elements of Measure Theory*, English transl., Holt-Rinehart-Winston, NY.

Bhattacharya, R. Waymire, E. (1990): *Stochastic Processes with Applications*, Wiley, NY.

Bhattacharya, R., Waymire, E. (2007): *Stochastic Processes: Theory and Applications*, Springer-Verlag, NY (in preparation).

Bhattacharya, R., Majumdar, M. (2007): *Random Dynamical Systems: Theory and applications*, Cambridge University Press, Cambridge.

Billingsley, P. (1968): *Convergence of Probability Measures*, Wiley, NY.

Billingsley, P. (1995): *Probability and Measure*, 3rd ed, Wiley, NY.

Bingham, N.H., C.M. Goldie, J.L. Teugels (1987): *Regular Variation*, Encyclopedia of Mathematics and Its Applications, Cambridge University Press, Cambridge.

Breiman, L. (1968): *Probability*, Addison Wesley, Reading, MA. Reprint SIAM, Philadelphia, PA.

Chung, K.L. (1974): *A Course in Probability Theory*, 2nd ed, Academic Press, NY

Chung, K.L., Williams, R.J. (1990): *Introduction to Stochastic Integration*, 2nd ed, Birkhauser, Boston.

Davis, M., Etheridge, A. (2006): *Louis Bachelier's Theory of Specultation: The origins of modern finance*, Princeton University Press, Princeton, NJ.

Dieudonné, J. (1960): *Foundations of Modern Analysis*, Academic Press, NY.

Durrett, R. (1984): *Brownian Motion and Martingales in Analysis*, Wadsworth, Belmont, CA.

Durrett, R. (1995): *Probability Theory and Examples*, 2nd ed. Wadsworth, Brooks & Cole, Pacific, Grove, CA.

Dembo, A., Zeitouni, O. (1998): *Large Deviations Techniques and Applications*, 2nd ed. Springer, NY.

DeMoivre, A. (1967): *The Doctrine of Chances*, reprinted from original 1756 imprint, Chelsea, NY.

Deuschel, J.D., Stroock, D.W. (1989): *Large Deviations*, Academic Press, Boston.

Doob, J.L. (1953): *Stochastic Processes*, Wiley, NY.

Doyle, P., Snell, L. (1984): *Random Walks and Electrical Networks*, Carus Mathematical Monographs, The Math. Assoc. America, Washington, DC.

Dudley, R.M. (1966): *Real Analysis and Probability*, Wadsworth, Brooks & Cole, Pacific Grove, CA.

Dym, H., McKean, H.P. (1972): *Fourier Series and Integrals*, Academic Press, NY.

Ellis, R.S. (1985): *Entropy, Large Deviations, and Statistical Mechanics*, Springer-Verlag, NY.

Ethier, S.N., Kurtz, T.G. (1985): *Markov Processes: Characterization and Convergence*, Wiley, NY.

Feller, W. (1968, 1971): *An Introduction to Probability Theory and Its Applications*, vol. 1 3rd ed, vol. 2, 2nd ed, Wiley, NY

Folland, G. (1984): *Real Analysis*, Wiley, NY

Freedman, D. (1971): *Brownian Motion and Diffusion*, Holden-Day, SF., Reprint Springer, NY.

Gnedenko, B.V., Kolmogorov, A.N. (1968): *Limit Distributions for Sums of Independent Random Variables*, English transl., 2nd ed., Addison-Wesley, Reading, MA (1st ed. 1949, Russian).

Grimmett, G. (1999): *Percolation*, 2nd edition, Springer-Verlag, Berlin.

Hall, P., Heyde, C.C. (1980): Martingale Limit Theory and Its Application, Academic Press, NY.

Halmos, P. (1974): *Measure Theory*, Springer Graduate Texts in Mathematics, Springer-Verlag, NY.

Harris, T.E. (1963): *The Theory of Branching Processes*, Springer-Verlag, Berlin.

Itô, K., McKean, H.P. (1965): *Diffusion Processes and Their Sample Paths*, Reprint Springer, Berlin.

Jacod, J., Protter, P. (2003) *Probability Essentials*, 2nd edition, Springer Universitext Series, Springer-Verlag, NY.

Jagers, P. (1975): *Branching Processes with Applications to Biology*, Wiley, NY.

Kac, M. *Probability, Number Theory and Statistical Physics: Selected Papers*, ed. K. Baclawski and M.D. Donsker, MIT Press, Cambridge.

Kallenberg, O. (2001): *Foundations of Modern Probability*, 2nd ed, Springer-Verlag, NY.

Karlin, S. Taylor, H.M. (1975), *A First Course in Probability*, Academic Press, NY.

Karlin, S. Taylor, H.M. (1981), *A First Course in Probability*, Academic Press, NY.

Karatzas, I., Shreve, S.E. (1991): *Brownian motion and stochastic calculus*, 2nd edition, Springer-Verlag, NY.

Laplace, P.-S. (1812): *Théorie Analytique des Probabilités*, Reprinted in Oeuvres Complète de Laplace **7**, Gauthier-Villars, Paris.

Lawler, G. (2005): *Conformally Invariant Processes in the Plane*, American Mathematical Society, Providence.

Lévy, P. (1925): *Calcul des Probabilités.* Gauthier–Villars, Paris.

Lévy, P. (1954): *Théorie de l'addition des variables aléatories*, 2nd ed, Gauthier–Villars, Paris (1st ed. 1937).

Liggett, T.M. (1985): *Interacting Particle Systems*, Springer-Verlag, NY.

Loève, M. (1977-78): *Probability Theory*, vols 1,2 4th ed. Springer, NY (1st ed. 1955).

Lukacs, E. (1970): *Characteristic Functions*, 2nd ed. Griffin, London.

McKean, H.P. (1969): *Stochastic Integrals*, Academic Press, NY.

Neveu, J. (1971): *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco.

Neveu, J. (1975): *Discrete Parameter Martingales*, North-Holland, Amsterdam.

Nualart, D. (1995): *The Malliavin Calculus and Related Topics*, Springer, NY.

Oksendal, B. (1998): *Stochastic Differential Equations*, 5th ed. Springer, Berlin

Parthasarathy, K.R. (1967): *Probability Measures on Metric Spaces*, Academic Press, NY.

Pitman, J. (1993): *Probability*, Springer-Verlag, NY.

Pollard, D. (2002): *A User's Guide to Measure Theoretic Probability*, Cambridge University Press, Cambridge.

Protter, P. (1990): *Stochastic Integration and Differential Equations*, Springer-Verlag, Berlin.

Resnick, S. (1987): *Extreme Values, Regular Variation, and Point Processes*, Springer-Verlag, NY.

Revuz, D., Yor, M. (1999): *Continuous Martingales and Brownian Motion*, 3rd ed. Springer, Berlin.

Rogers, LC.G., Williams, D. (2000): *Diffusions, Markov Processes and Martingales*, vol. 1 (2nd ed.), vol. 2 Cambridge.

Royden, H.L. (1988): *Real Analysis*, 3rd ed. MacMillan, NY.

Rudin, W. (1974): *Real and Complex Analysis, 2nd ed.* McGraw-Hill, NY.

Shreve, S. (2004): *Stochastic Calculus for Finance*, Springer-Verlag, NY.

Spitzer, F. (1976): *Principles of Random Walk*, 2nd ed. Springer, NY

Samorodnitsky, G., Taqqu, M. (1994): *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, N Y.

Stroock, D.W. (1993): *Probability Theory: An Analytic View*, Cambridge Univ. Press.

Stroock, D.W., Varadhan, S.R.S. (1979): *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin.

Varadhan, S.R.S. (1984): *Large Deviations and Applications*, SIAM, Philadelphia.

Williams, D. (1991): *Probability with Martingales*, Cambridge Univ. Press, Cambridge.

Index

Symbol Index

Universitext *(continued from p. ii)*

Jennings: Modern Geometry with Applications
Jones/Morris/Pearson: Abstract Algebra and Famous Impossibilities
Kac/Cheung: Quantum Calculus
Kannan/Krueger: Advanced Analysis
Kelly/Matthews: The Non-Euclidean Hyperbolic Plane
Kostrikin: Introduction to Algebra
Kuo: Introduction to Stochastic Integration
Kurzweil/Stellmacher: The Theory of Finite Groups: An Introduction
Lorenz: Algebra: Volume I: Fields and Galois Theory
Luecking/Rubel: Complex Analysis: A Functional Analysis Approach
MacLane/Moerdijk: Sheaves in Geometry and Logic
Marcus: Number Fields
Martinez: An Introduction to Semiclassical and Microlocal Analysis
Matsuki: Introduction to the Mori Program
McCarthy: Introduction to Arithmetical Functions
McCrimmon: A Taste of Jordan Algebras
Meyer: Essential Mathematics for Applied Fields
Mines/Richman/Ruitenburg: A Course in Constructive Algebra
Moise: Introductory Problems Course in Analysis and Topology
Morris: Introduction to Game Theory
Nicolaescu: An Invitation to Morse Theory
Poizat: A Course In Model Theory: An Introduction to Contemporary Mathematical Logic
Polster: A Geometrical Picture Book
Porter/Woods: Extensions and Absolutes of Hausdorff Spaces
Procesi: Lie Groups
Radjavi/Rosenthal: Simultaneous Triangularization
Ramsay/Richtmyer: Introduction to Hyperbolic Geometry
Rautenberg: A Concise Introduction to Mathematical Logic
Reisel: Elementary Theory of Metric Spaces
Ribenboim: Classical Theory of Algebraic Numbers
Rickart: Natural Function Algebras
Rotman: Galois Theory
Rubel/Colliander: Entire and Meromorphic Functions
Runde: A Taste of Topology
Sagan: Space-Filling Curves
Samelson: Notes on Lie Algebras
Schiff: Normal Families
Shapiro: Composition Operators and Classical Function Theory
Simonnet: Measures and Probability
Smith: Power Series From a Computational Point of View
Smith/Kahanpää/ Kekäläinen/Traves: An Invitation to Algebraic Geometry
Smorynski: Self-Reference and Modal Logic
Stillwell: Geometry of Surfaces
Stroock: An Introduction to the Theory of Large Deviations
Sunder: An Invitation to von Neumann Algebras
Tondeur: Foliations on Riemannian Manifolds
Toth: Finite Möbius Groups, Minimal Immersions of Spheres, and Moduli
Van Brunt: The Calculus of Variations
Weintraub: Galois Theory
Wong: Weyl Transforms
Zhang: Matrix Theory: Basic Results and Techniques
Zong: Sphere Packings
Zong: Strange Phenomena in Convex and Discrete Geometry