Springer Finance

Springer Finance

Springer Finance is a programme of books aimed at students, academics, and practitioners working on increasingly technical approaches to the analysis of financial markets. It aims to cover a variety of topics, not only mathematical finance but foreign exchanges, term structure, risk management, portfolio theory, equity derivatives, and financial economics.

M. Ammann, Credit Risk Valuation: Methods, Models, and Applications (2001)
E. Barucci, Financial Markets Theory: Equilibrium, Efficiency and Information (2003)
N.H. Bingham and R. Kiesel, Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives, 2nd Edition (2004)
T.R. Bielecki and M. Rutkowski, Credit Risk: Modeling, Valuation and Hedging (2001)
D. Brigo amd F. Mercurio, Interest Rate Models: Theory and Practice (2001)
R. Buff, Uncertain Volatility Models – Theory and Application (2002)
R.-A. Dana and M. Jeanblanc, Financial Markets in Continuous Time (2003)
G. Deboeck and T. Kohonen (Editors), Visual Explorations in Finance with Self-Organizing Maps (1998)
R.J. Elliott and P.E. Kopp, Mathematics of Financial Markets (1999)
H. Geman, D. Madan, S.R. Pliska and T. Vorst (Editors), Mathematical Finance – Bachelier Congress 2000 (2001)
M. Gundlach and F. Lehrbass (Editors), CreditRisk+ in the Banking Industry (2004)
Y.-K. Kwok, Mathematical Models of Financial Derivatives (1998)
M. Külpmann, Irrational Exuberance Reconsidered: The Cross Section of Stock Returns, 2nd Edition (2004)
A. Pelsser, Efficient Methods for Valuing Interest Rate Derivatives (2000)
J.-L. Prigent, Weak Convergence of Financial Markets (2003)
B. Schmid, Credit Risk Pricing Models: Theory and Practice, 2nd Edition (2004)
S.E. Shreve, Stochastic Calculus for Finance I: The Binomial Asset Pricing Model (2004)
S.E. Shreve, Stochastic Calculus for Finance II: Continuous-Time Models (2004)
M. Yor, Exponential Functionals of Brownian Motion and Related Processes (2001)
R. Zagst, Interest-Rate Management (2002)
Y.-I. Zhu and I.-L Chern, Derivative Securities and Difference Methods (2004)
A. Ziegler, Incomplete Information and Heterogeneous Beliefs in Continuous-Time Finance (2003)
A. Ziegler, A Game Theory Analysis of Options: Corporate Finance and Financial Intermediation in Continuous Time, 2nd Edition (2004)

Steven E. Shreve

Stochastic Calculus for Finance II

Continuous-Time Models

With 28 Figures

Springer

Steven E. Shreve
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15213
USA
shreve@cmu.edu

Mathematics Subject Classification (2000): 60-01, 60H10, 60J65, 91B28

Library of Congress Cataloging-in-Publication Data
Shreve, Steven E.
Stochastic calculus for finance / Steven E. Shreve.
p. cm. — (Springer finance series)
Includes bibliographical references and index.
Contents v. 2. Continuous-time models.
ISBN 0-387-40101-6 (alk. paper)
1. Finance—Mathematical models—Textbooks. 2. Stochastic analysis—Textbooks. I. Title. II. Springer finance.
HG106.S57 2003
332'.01'51922—dc22 2003063342

ISBN 0-387-40101-6 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2

springeronline com

To my students

This page intentionally left blank

Preface

Origin of This Text

This text has evolved from mathematics courses in the Master of Science in Computational Finance (MSCF) program at Carnegie Mellon University. The content of this book has been used successfully with students whose mathematics background consists of calculus and calculus-based probability. The text gives precise statements of results, plausibility arguments, and even some proofs, but more importantly, intuitive explanations developed and refined through classroom experience with this material are provided. Exercises conclude every chapter. Some of these extend the theory and others are drawn from practical problems in quantitative finance.

The first three chapters of Volume I have been used in a half-semester course in the MSCF program. The full Volume I has been used in a full-semester course in the Carnegie Mellon Bachelor's program in Computational Finance. Volume II was developed to support three half-semester courses in the MSCF program.

Dedication

Since its inception in 1994, the Carnegie Mellon Master's program in Computational Finance has graduated hundreds of students. These people, who have come from a variety of educational and professional backgrounds, have been a joy to teach. They have been eager to learn, asking questions that stimulated thinking, working hard to understand the material both theoretically and practically, and often requesting the inclusion of additional topics. Many came from the finance industry, and were gracious in sharing their knowledge in ways that enhanced the classroom experience for all.

This text and my own store of knowledge have benefited greatly from interactions with the MSCF students, and I continue to learn from the MSCF

alumni. I take this opportunity to express gratitude to these students and former students by dedicating this work to them.

Acknowledgments

Conversations with several people, including my colleagues David Heath and Dmitry Kramkov, have influenced this text. Łukasz Kruk read much of the manuscript and provided numerous comments and corrections. Other students and faculty have pointed out errors in and suggested improvements of earlier drafts of this work. Some of these are Jonathan Anderson, Nathaniel Carter, Bogdan Doytchinov, David German, Steven Gillispie, Karel Janeček, Sean Jones, Anatoli Karolik, David Korpi, Andrzej Krause, Rael Limbitco, Petr Luksan, Sergey Myagchilov, Nicki Rasmussen, Isaac Sonin, Massimo Tassan-Solet, David Whitaker and Uwe Wystup. In some cases, users of these earlier drafts have suggested exercises or examples, and their contributions are acknowledged at appropriate points in the text. To all those who aided in the development of this text, I am most grateful.

During the creation of this text, the author was partially supported by the National Science Foundation under grants DMS-9802464, DMS-0103814, and DMS-0139911. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Pittsburgh, Pennsylvania, USA — Steven E. Shreve
April 2004

Contents

Introduction

Background

By awarding Harry Markowitz, William Sharpe, and Merton Miller the 1990 Nobel Prize in Economics, the Nobel Prize Committee brought to worldwide attention the fact that the previous forty years had seen the emergence of a new scientific discipline, the "theory of finance." This theory attempts to understand how financial markets work, how to make them more efficient, and how they should be regulated. It explains and enhances the important role these markets play in capital allocation and risk reduction to facilitate economic activity. Without losing its application to practical aspects of trading and regulation, the theory of finance has become increasingly mathematical, to the point that problems in finance are now driving research in mathematics.

Harry Markowitz's 1952 Ph.D. thesis *Portfolio Selection* laid the groundwork for the mathematical theory of finance. Markowitz developed a notion of mean return and covariances for common stocks that allowed him to quantify the concept of "diversification" in a market. He showed how to compute the mean return and variance for a given portfolio and argued that investors should hold only those portfolios whose variance is minimal among all portfolios with a given mean return. Although the language of finance now involves stochastic (Itô) calculus, management of risk in a quantifiable manner is the underlying theme of the modern theory and practice of quantitative finance.

In 1969, Robert Merton introduced stochastic calculus into the study of finance. Merton was motivated by the desire to understand how prices are set in financial markets, which is the classical economics question of "equilibrium," and in later papers he used the machinery of stochastic calculus to begin investigation of this issue.

At the same time as Merton's work and with Merton's assistance, Fischer Black and Myron Scholes were developing their celebrated option pricing formula. This work won the 1997 Nobel Prize in Economics. It provided a satisfying solution to an important practical problem, that of finding a fair price for a European call option (i.e., the right to buy one share of a given

stock at a specified price and time). In the period 1979–1983, Harrison, Kreps, and Pliska used the general theory of continuous-time stochastic processes to put the Black-Scholes option-pricing formula on a solid theoretical basis, and, as a result, showed how to price numerous other "derivative" securities.

Many of the theoretical developments in finance have found immediate application in financial markets. To understand how they are applied, we digress for a moment on the role of financial institutions. A principal function of a nation's financial institutions is to act as a risk-reducing intermediary among customers engaged in production. For example, the insurance industry pools premiums of many customers and must pay off only the few who actually incur losses. But risk arises in situations for which pooled-premium insurance is unavailable. For instance, as a hedge against higher fuel costs, an airline may want to buy a security whose value will rise if oil prices rise. But who wants to sell such a security? The role of a financial institution is to design such a security, determine a "fair" price for it, and sell it to airlines. The security thus sold is usually "derivative" (i.e., its value is based on the value of other, identified securities). "Fair" in this context means that the financial institution earns just enough from selling the security to enable it to trade in other securities whose relation with oil prices is such that, if oil prices do indeed rise, the firm can pay off its increased obligation to the airlines. An "efficient" market is one in which risk-hedging securities are widely available at "fair" prices.

The Black-Scholes option pricing formula provided, for the first time, a theoretical method of fairly pricing a risk-hedging security. If an investment bank offers a derivative security at a price that is higher than "fair," it may be underbid. If it offers the security at less than the "fair" price, it runs the risk of substantial loss. This makes the bank reluctant to offer many of the derivative securities that would contribute to market efficiency. In particular, the bank only wants to offer derivative securities whose "fair" price can be determined in advance. Furthermore, if the bank sells such a security, it must then address the hedging problem: how should it manage the risk associated with its new position? The mathematical theory growing out of the Black-Scholes option pricing formula provides solutions for both the pricing and hedging problems. It thus has enabled the creation of a host of specialized derivative securities. This theory is the subject of this text.

Relationship between Volumes I and II

Volume II treats the continuous-time theory of stochastic calculus within the context of finance applications. The presentation of this theory is the raison d'être of this work. Volume II includes a self-contained treatment of the probability theory needed for stochastic calculus, including Brownian motion and its properties.

Volume I presents many of the same finance applications, but within the simpler context of the discrete-time binomial model. It prepares the reader for Volume II by treating several fundamental concepts, including martingales, Markov processes, change of measure and risk-neutral pricing in this less technical setting. However, Volume II has a self-contained treatment of these topics, and strictly speaking, it is not necessary to read Volume I before reading Volume II. It is helpful in that the difficult concepts of Volume II are first seen in a simpler context in Volume I.

In the Carnegie Mellon Master's program in Computational Finance, the course based on Volume I is a prerequisite for the courses based on Volume II. However, graduate students in computer science, finance, mathematics, physics and statistics frequently take the courses based on Volume II without first taking the course based on Volume I.

The reader who begins with Volume II may use Volume I as a reference. As several concepts are presented in Volume II, reference is made to the analogous concepts in Volume I. The reader can at that point choose to read only Volume II or to refer to Volume I for a discussion of the concept at hand in a more transparent setting.

Summary of Volume I

Volume I presents the binomial asset pricing model. Although this model is interesting in its own right, and is often the paradigm of practice, here it is used primarily as a vehicle for introducing in a simple setting the concepts needed for the continuous-time theory of Volume II.

Chapter 1, *The Binomial No-Arbitrage Pricing Model*, presents the no-arbitrage method of option pricing in a binomial model. The mathematics is simple, but the profound concept of risk-neutral pricing introduced here is not. Chapter 2, *Probability Theory on Coin Toss Space*, formalizes the results of Chapter 1, using the notions of martingales and Markov processes. This chapter culminates with the risk-neutral pricing formula for European derivative securities. The tools used to derive this formula are not really required for the derivation in the binomial model, but we need these concepts in Volume II and therefore develop them in the simpler discrete-time setting of Volume I. Chapter 3, *State Prices*, discusses the change of measure associated with risk-neutral pricing of European derivative securities, again as a warm-up exercise for change of measure in continuous-time models. An interesting application developed here is to solve the problem of optimal (in the sense of expected utility maximization) investment in a binomial model. The ideas of Chapters 1 to 3 are essential to understanding the methodology of modern quantitative finance. They are developed again in Chapters 4 and 5 of Volume II.

The remaining three chapters of Volume I treat more specialized concepts. Chapter 4, *American Derivative Securities*, considers derivative securities whose owner can choose the exercise time. This topic is revisited in

a continuous-time context in Chapter 8 of Volume II. Chapter 5, *Random Walk*, explains the reflection principle for random walk. The analogous reflection principle for Brownian motion plays a prominent role in the derivation of pricing formulas for exotic options in Chapter 7 of Volume II. Finally, Chapter 6, *Interest-Rate-Dependent Assets*, considers models with random interest rates, examining the difference between forward and futures prices and introducing the concept of a forward measure. Forward and futures prices reappear at the end of Chapter 5 of Volume II. Forward measures for continuous-time models are developed in Chapter 9 of Volume II and used to create forward LIBOR models for interest rate movements in Chapter 10 of Volume II.

Summary of Volume II

Chapter 1, *General Probability Theory*, and Chapter 2, *Information and Conditioning*, of Volume II lay the measure-theoretic foundation for probability theory required for a treatment of continuous-time models. Chapter 1 presents probability spaces, Lebesgue integrals, and change of measure. Independence, conditional expectations, and properties of conditional expectations are introduced in Chapter 2. These chapters are used extensively throughout the text, but some readers, especially those with exposure to probability theory, may choose to skip this material at the outset, referring to it as needed.

Chapter 3, *Brownian Motion*, introduces Brownian motion and its properties. The most important of these for stochastic calculus is quadratic variation, presented in Section 3.4. All of this material is needed in order to proceed, except Sections 3.6 and 3.7, which are used only in Chapter 7, *Exotic Options* and Chapter 8, *Early Exercise*.

The core of Volume II is Chapter 4, *Stochastic Calculus*. Here the Itô integral is constructed and Itô's formula (called the Itô-Doeblin formula in this text) is developed. Several consequences of the Itô-Doeblin formula are worked out. One of these is the characterization of Brownian motion in terms of its quadratic variation (Lévy's theorem) and another is the Black-Scholes equation for a European call price (called the Black-Scholes-Merton equation in this text). The only material which the reader may omit is Section 4.7, *Brownian Bridge*. This topic is included because of its importance in Monte Carlo simulation, but it is not used elsewhere in the text.

Chapter 5, *Risk-Neutral Pricing*, states and proves Girsanov's Theorem, which underlies change of measure. This permits a systematic treatment of risk-neutral pricing and the Fundamental Theorems of Asset Pricing (Section 5.4). Section 5.5, *Dividend-Paying Stocks*, is not used elsewhere in the text. Section 5.6, *Forwards and Futures*, appears later in Section 9.4 and in some exercises.

Chapter 6, *Connections with Partial Differential Equations*, develops the connection between stochastic calculus and partial differential equations. This is used frequently in later chapters.

With the exceptions noted above, the material in Chapters 1–6 is fundamental for quantitative finance is essential for reading the later chapters. After Chapter 6, the reader has choices.

Chapter 7, *Exotic Options*, is not used in subsequent chapters, nor is Chapter 8, *Early Exercise*. Chapter 9, *Change of Numéraire*, plays an important role in Section 10.4, *Forward LIBOR model*, but is not otherwise used. Chapter 10, *Term Structure Models*, and Chapter 11, *Introduction to Jump Processes*, are not used elsewhere in the text.

This page intentionally left blank

1

General Probability Theory

1.1 Infinite Probability Spaces

An infinite probability space is used to model a situation in which a random experiment with infinitely many possible outcomes is conducted. For purposes of the following discussion, there are two such experiments to keep in mind:

(i) choose a number from the unit interval [0,1], and
(ii) toss a coin infinitely many times.

In each case, we need a sample space of possible outcomes. For (i), our sample space will be simply the unit interval $[0, 1]$. A generic element of $[0, 1]$ will be denoted by ω, rather than the more natural choice x, because these elements are the possible outcomes of a random experiment.

For case (ii), we define

$$\Omega_\infty = \text{ the set of infinite sequences of } H\text{s and } T\text{s.} \tag{1.1.1}$$

A generic element of Ω_∞ will be denoted $\omega = \omega_1\omega_2\ldots$, where ω_n indicates the result of the nth coin toss.

The samples spaces listed above are not only infinite but are *uncountably infinite* (i.e., it is not possible to list their elements in a sequence). The first problem we face with an uncountably infinite sample space is that, for most interesting experiments, the probability of any particular outcome is zero. Consequently, we cannot determine the probability of a subset A of the sample space, a so-called *event*, by summing up the probabilities of the elements in A, as we did in equation (2.1.5) of Chapter 2 of Volume I. We must instead define the probabilities of events directly. But in infinite sample spaces there are infinitely many events. Even though we may understand well what random experiment we want to model, some of the events may have such complicated descriptions that it is not obvious what their probabilities should be. It would be hopeless to try to give a formula that determines the probability for every subset of an uncountably infinite sample space. We instead give a formula for

the probability of certain simple events and then appeal to the properties of probability measures to determine the probability of more complicated events. This prompts the following definitions, after which we describe the process of setting up the uniform probability measure on $[0,1]$.

Definition 1.1.1. *Let Ω be a nonempty set, and let $\mathcal{F}$ be a collection of subsets of Ω. We say that $\mathcal{F}$ is a* σ-algebra *(called a* σ-field *by some authors) provided that:*

(i) the empty set $\emptyset$ belongs to $\mathcal{F}$,
(ii) whenever a set A belongs to $\mathcal{F}$, its complement A^c also belongs to $\mathcal{F}$, and
(iii) whenever a sequence of sets $A_1, A_2, \ldots$ belongs to $\mathcal{F}$, their union $\cup_{n=1}^{\infty} A_n$ also belongs to $\mathcal{F}$.

If we have a σ-algebra of sets, then all the operations we might want to do to the sets will give us other sets in the σ-algebra. If we have two sets A and B in a σ-algebra, then by considering the sequence $A, B, \emptyset, \emptyset, \emptyset, \ldots$, we can conclude from (i) and (iii) that $A \cup B$ must also be in the σ-algebra. The same argument shows that if $A_1, A_2, \ldots, A_N$ are finitely many sets in a σ-algebra, then their union must also be in the σ-algebra. Finally, if $A_1, A_2, \ldots$ is a sequence of sets in a σ-algebra, then because

$$\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c \right)^c ,$$

properties (ii) and (iii) applied to the right-hand side show that $\cap_{n=1}^{\infty} A_n$ is also in the σ-algebra. Similarly, the intersection of a finite number of sets in a σ-algebra results in a set in the σ-algebra. Of course, if $\mathcal{F}$ is a σ-algebra, then the whole space Ω must be one of the sets in $\mathcal{F}$ because $\Omega = \emptyset^c$.

Definition 1.1.2. *Let Ω be a nonempty set, and let $\mathcal{F}$ be a σ-algebra of subsets of Ω. A* probability measure $\mathbb{P}$ *is a function that, to every set $A \in \mathcal{F}$, assigns a number in $[0,1]$, called the* probability of A *and written $\mathbb{P}(A)$. We require:*

(i) $\mathbb{P}(\Omega) = 1$, and
(ii) (countable additivity) whenever $A_1, A_2, \ldots$ is a sequence of disjoint sets in $\mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n). \tag{1.1.2}$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space.

If Ω is a finite set and $\mathcal{F}$ is the collection of all subsets of Ω, then $\mathcal{F}$ is a σ-algebra and Definition 1.1.2 boils down to Definition 2.1.1 of Chapter 2 of Volume I. In the context of infinite probability spaces, we must take care that the definition of probability measure just given is consistent with our intuition. The countable additivity condition (ii) in Definition 1.1.2 is designed to take

care of this. For example, we should be sure that $\mathbb{P}(\emptyset) = 0$. That follows from taking

$$A_1 = A_2 = A_3 = \cdots = \emptyset$$

in (1.1.2), for then this equation becomes $\mathbb{P}(\emptyset) = \sum_{n=1}^{\infty} \mathbb{P}(\emptyset)$. The only number in $[0, 1]$ that $\mathbb{P}(\emptyset)$ could be is

$$\mathbb{P}(\emptyset) = 0. \tag{1.1.3}$$

We also still want (2.1.7) of Chapter 2 of Volume I to hold: if A and B are disjoint sets in $\mathcal{F}$, we want to have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \tag{1.1.4}$$

Not only does Definition 1.1.2(ii) guarantee this, it guarantees the *finite additivity* condition that if $A_1, A_2, \ldots, A_N$ are finitely many disjoint sets in $\mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} \mathbb{P}(A_n). \tag{1.1.5}$$

To see this, apply (1.1.2) with

$$A_{N+1} = A_{N+2} = A_{N+3} = \cdots = \emptyset.$$

In the special case that $N = 2$ and $A_1 = A$, $A_2 = B$, we get (1.1.4). From part (i) of Definition 1.1.2 and (1.1.4) with $B = A^c$, we get

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A). \tag{1.1.6}$$

In summary, from Definition 1.1.2, we conclude that a probability measure must satisfy (1.1.3)–(1.1.6).

We now describe by example the process of construction of probability measures on uncountable sample spaces. We do this here for the spaces $[0, 1]$ and Ω_∞ with which we began this section.

Example 1.1.3 (Uniform (Lebesgue) measure on $[0, 1]$*).* We construct a mathematical model for choosing a number at random from the unit interval $[0, 1]$ so that the probability is distributed uniformly over the interval. We define the probability of closed intervals $[a, b]$ by the formula

$$\mathbb{P}[a, b] = b - a, \ 0 \le a \le b \le 1, \tag{1.1.7}$$

(i.e., the probability that the number chosen is between a and b is $b - a$). (This particular probability measure on $[0, 1]$ is called *Lebesgue measure* and in this text is sometimes denoted $\mathcal{L}$. The Lebesgue measure of a subset of $\mathbb{R}$ is its "length.") If $b = a$, then $[a, b]$ is the set containing only the number a, and (1.1.7) says that the probability of this set is zero (i.e., the probability is zero that the number we choose is exactly equal to a). Because single points have zero probability, the probability of an open interval (a, b) is the same as the probability of the closed interval $[a, b]$; we have

$$\mathbb{P}(a,b) = b - a,\ 0 \le a \le b \le 1. \tag{1.1.8}$$

There are many other subsets of $[0,1]$ whose probability is determined by the formula (1.1.7) and the properties of probability measures. For example, the set $\left[0,\frac{1}{3}\right] \cup \left[\frac{2}{3},1\right]$ is not an interval, but we know from (1.1.7) and (1.1.4) that its probability is $\frac{2}{3}$.

It is natural to ask if there is some way to describe the collection of all sets whose probability is determined by formula (1.1.7) and the properties of probability measures. It turns out that this collection of sets is the σ-algebra we get starting with the closed intervals and putting in everything else required in order to have a σ-algebra. Since an open interval can be written as a union of a sequence of closed intervals,

$$(a,b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right],$$

this σ-algebra contains all open intervals. It must also contain the set $\left[0,\frac{1}{3}\right] \cup \left[\frac{2}{3},1\right]$, mentioned at the end of the preceding paragraph, and many other sets.

The σ-algebra obtained by beginning with closed intervals and adding everything else necessary in order to have a σ-algebra is called the *Borel σ-algebra* of subsets of $[0,1]$ and is denoted $\mathcal{B}[0,1]$. The sets in this σ-algebra are called *Borel sets*. These are the subsets of $[0,1]$, the so-called events, whose probability is determined once we specify the probability of the closed intervals. Every subset of $[0,1]$ we encounter in this text is a Borel set, and this can be verified if desired by writing the set in terms of unions, intersections, and complements of sequences of closed intervals.[1] □

Example 1.1.4 (Infinite, independent coin-toss space). We toss a coin infinitely many times and let Ω_∞ of (1.1.1) denote the set of possible outcomes. We assume the probability of head on each toss is $p > 0$, the probability of tail is $q = 1 - p > 0$, and the different tosses are independent, a concept we define precisely in the next chapter. We want to construct a probability measure corresponding to this random experiment.

We first define $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$. These $2^{(2^0)} = 2$ sets form a σ-algebra, which we call $\mathcal{F}_0$:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}. \tag{1.1.9}$$

We next define $\mathbb{P}$ for the two sets

$$\begin{aligned} A_H &= \text{the set of all sequences beginning with } H = \{\omega; \omega_1 = H\}, \\ A_T &= \text{the set of all sequences beginning with } T = \{\omega; \omega_1 = T\}, \end{aligned}$$

[1] See Appendix A, Section A.1 for the construction of the *Cantor set*, which gives some indication of how complicated sets in $\mathcal{B}[0,1]$ can be.

by setting $\mathbb{P}(A_H) = p$, $\mathbb{P}(A_T) = q$. We have now defined $\mathbb{P}$ for $2^{(2^1)} = 4$ sets, and these four sets form a σ-algebra; since $A_H^c = A_T$ we do not need to add anything else in order to have a σ-algebra. We call this σ-algebra $\mathcal{F}_1$:

$$\mathcal{F}_1 = \{\emptyset, \Omega, A_H, A_T\}. \tag{1.1.10}$$

We next define $\mathbb{P}$ for the four sets

$$\begin{aligned}
A_{HH} &= \text{ The set of all sequences beginning with } HH \\
&= \{\omega; \omega_1 = H, \omega_2 = H\}, \\
A_{HT} &= \text{ The set of all sequences beginning with } HT \\
&= \{\omega; \omega_1 = H, \omega_2 = T\}, \\
A_{TH} &= \text{ The set of all sequences beginning with } TH \\
&= \{\omega; \omega_1 = T, \omega_2 = H\}, \\
A_{TT} &= \text{ The set of all sequences beginning with } TT \\
&= \{\omega; \omega_1 = T, \omega_2 = T\}
\end{aligned}$$

by setting

$$\mathbb{P}(A_{HH}) = p^2,\ \mathbb{P}(A_{HT}) = pq,\ \mathbb{P}(A_{TH}) = pq,\ \mathbb{P}(A_{TT}) = q^2. \tag{1.1.11}$$

Because of (1.1.6), this determines the probability of the complements A_{HH}^c, A_{HT}^c, A_{TH}^c, A_{TT}^c. Using (1.1.5), we see that the probabilities of the unions $A_{HH} \cup A_{TH}$, $A_{HH} \cup A_{TT}$, $A_{HT} \cup A_{TH}$, and $A_{HT} \cup A_{TT}$ are also determined. We have already defined the probabilities of the two other pairwise unions $A_{HH} \cup A_{HT} = A_H$ and $A_{TH} \cup A_{TT} = A_T$. We have already noted that the probability of the triple unions is determined since these are complements of the sets in (1.1.11), e.g.,

$$A_{HH} \cup A_{HT} \cup A_{TH} = A_{TT}^c.$$

At this point, we have determined the probability of $2^{(2^2)} = 16$ sets, and these sets form a σ-algebra, which we call $\mathcal{F}_2$:

$$\mathcal{F}_2 = \left\{ \begin{array}{l} \emptyset, \Omega, A_H, A_T, A_{HH}, A_{HT}, A_{TH}, A_{TT}, A_{HH}^c, A_{HT}^c, A_{TH}^c, A_{TT}^c, \\ A_{HH} \cup A_{TH},\ A_{HH} \cup A_{TT},\ A_{HT} \cup A_{TH},\ A_{HT} \cup A_{TT} \end{array} \right\}. \tag{1.1.12}$$

We next define the probability of every set that can be described in terms of the outcome of the first three coin tosses. Counting the sets we already have, this will give us $2^{(2^3)} = 256$ sets, and these will form a σ-algebra, which we call $\mathcal{F}_3$.

By continuing this process, we can define the probability of every set that can be described in terms of finitely many tosses. Once the probabilities of all these sets are specified, there are other sets, not describable in terms of finitely many coin tosses, whose probabilities are determined. For example,

the set containing only the single sequence $HHHH\ldots$ cannot be described in terms of finitely many coin tosses, but it is a subset of A_H, A_{HH}, A_{HHH}, etc. Furthermore,

$$\mathbb{P}(A_H) = p,\ \mathbb{P}(A_{HH}) = p^2,\ \mathbb{P}(A_{HHH}) = p^3, \ldots,$$

and since these probabilities converge to zero, we must have

$$\mathbb{P}(\text{Every toss results in head}) = 0.$$

Similarly, the single sequence $HTHTHT\ldots$, being the intersection of the sets A_H, A_{HT}, A_{HTH}, etc. must have probability less than or equal to each of

$$\mathbb{P}(A_H) = p,\ \mathbb{P}(A_{HT}) = pq,\ \mathbb{P}(A_{HTH}) = p^2q, \ldots,$$

and hence must have probability zero. The same argument shows that every individual sequence in Ω_∞ has probability zero.

We create a σ-algebra, called $\mathcal{F}_\infty$, by putting in every set that can be described in terms of finitely many coin tosses and then adding all other sets required in order to have a σ-algebra. It turns out that once we specify the probability of every set that can be described in terms of finitely many coin tosses, the probability of every set in $\mathcal{F}_\infty$ is determined. There are sets in $\mathcal{F}_\infty$ whose probability, although determined, is not easily computed. For example, consider the set A of sequences $\omega = \omega_1\omega_2\ldots$ for which

$$\lim_{n\to\infty} \frac{H_n(\omega_1\ldots\omega_n)}{n} = \frac{1}{2}, \tag{1.1.13}$$

where $H_n(\omega_1\ldots\omega_n)$ denotes the number of Hs in the first n tosses. In other words, A is the set of sequences of heads and tails for which the long-run average number of heads is $\frac{1}{2}$. Because its description involves all the coin tosses, it was not defined directly at any stage of the process outlined above. On the other hand, it is in $\mathcal{F}_\infty$, and that means its probability is somehow determined by this process and the properties of probability measures. To see that A is in $\mathcal{F}_\infty$, we fix positive integers m and n and define the set

$$A_{n,m} = \left\{\omega; \left|\frac{H_n(\omega_1\ldots\omega_n)}{n} - \frac{1}{2}\right| \le \frac{1}{m}\right\}.$$

This set is in $\mathcal{F}_n$, and once n and m are known, its probability is defined by the process outlined above. By the definition of limit, a coin-toss sequence $\omega = \omega_1\omega_2\ldots$ satisfies (1.1.13) if and only if for every positive integer m there exists a positive integer N such that for all $n \ge N$ we have $\omega \in A_{n,m}$. In other words, the set of ω for which (1.1.13) holds is

$$A = \bigcap_{m=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_{n,m}.$$

The set A is in $\mathcal{F}_\infty$ because it is described in terms of unions and intersections of sequences of sets that are in $\mathcal{F}_\infty$. This does not immediately tell us how to compute $\mathbb{P}(A)$, but it tells us that $\mathbb{P}(A)$ is somehow determined. As it turns out, the Strong Law of Large Numbers asserts that $\mathbb{P}(A) = 1$ if $p = \frac{1}{2}$ and $\mathbb{P}(A) = 0$ if $p \neq \frac{1}{2}$.

Every subset of Ω_∞ we shall encounter will be in $\mathcal{F}_\infty$. Indeed, it is extremely difficult to produce a set not in $\mathcal{F}_\infty$, although such sets exist. □

The observation in Example 1.1.4 that every individual sequence has probability zero highlights a paradox in uncountable probability spaces. We would like to say that something that has probability zero cannot happen. In particular, we would like to say that if we toss a coin infinitely many times, it cannot happen that we get a head on every toss (we are assuming here that the probability for head on each toss is $p > 0$ and $q = 1 - p > 0$). It would be satisfying if events that have probability zero are sure not to happen and events that have probability one are sure to happen. In particular, we would like to say that we are sure to get at least one tail. However, because the sequence that is all heads is in our sample space, and is no less likely to happen than any other particular sequence (every single sequence has probability zero), mathematicians have created a terminology that equivocates. We say that we will get at least one tail *almost surely.* Whenever an event is said to be almost sure, we mean it has probability one, even though it may not include every possible outcome. The outcome or set of outcomes not included, taken all together, has probability zero.

Definition 1.1.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If a set $A \in \mathcal{F}$ satisfies $\mathbb{P}(A) = 1$, we say that the event A occurs* almost surely.

1.2 Random Variables and Distributions

Definition 1.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A* random variable *is a real-valued function X defined on Ω with the property that for every Borel subset B of $\mathbb{R}$, the subset of Ω given by*

$$\{X \in B\} = \{\omega \in \Omega; X(\omega) \in B\} \tag{1.2.1}$$

is in the σ-algebra $\mathcal{F}$. (We sometimes also permit a random variable to take the values $+\infty$ and $-\infty$.)

To get the Borel subsets of $\mathbb{R}$, one begins with the closed intervals $[a, b] \subset \mathbb{R}$ and adds all other sets that are necessary in order to have a σ-algebra. This means that unions of sequences of closed intervals are Borel sets. In particular, every open interval is a Borel set, because an open interval can be written as the union of a sequence of closed intervals. Furthermore, every open *set* (whether or not an interval) is a Borel set because every open set is the union

of a sequence of open intervals. Every closed set is a Borel set because it is the complement of an open set. We denote the collection of Borel subsets of $\mathbb{R}$ by $\mathcal{B}(\mathbb{R})$ and call it the *Borel σ-algebra of* $\mathbb{R}$. Every subset of $\mathbb{R}$ we encounter in this text is in this σ-algebra.

A random variable X is a numerical quantity whose value is determined by the random experiment of choosing $\omega \in \Omega$. We shall be interested in the probability that X takes various values. It is often the case that the probability that X takes a particular value is zero, and hence we shall mostly talk about the probability that X takes a value in some set rather than the probability that X takes a particular value. In other words, we will want to speak of $\mathbb{P}\{X \in B\}$. Definition 1.2.1 requires that $\{X \in B\}$ be in $\mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$, so that we are sure the probability of this set is defined.

Example 1.2.2 (Stock prices). Recall the independent, infinite coin-toss space $(\Omega_\infty, \mathcal{F}_\infty, \mathbb{P})$ of Example 1.1.4. Let us define stock prices by the formulas

$$S_0(\omega) = 4 \text{ for all } \omega \in \Omega_\infty,$$

$$S_1(\omega) = \begin{cases} 8 \text{ if } \omega_1 = H, \\ 2 \text{ if } \omega_1 = T, \end{cases}$$

$$S_2(\omega) = \begin{cases} 16 \text{ if } \omega_1 = \omega_2 = H, \\ 4 \quad \text{if } \omega_1 \neq \omega_2, \\ 1 \quad \text{if } \omega_1 = \omega_2 = T, \end{cases}$$

and, in general,

$$S_{n+1}(\omega) = \begin{cases} 2S_n(\omega) & \text{if } \omega_{n+1} = H, \\ \frac{1}{2}S_n(\omega) & \text{if } \omega_{n+1} = T. \end{cases}$$

All of these are random variables. They assign a numerical value to each possible sequence of coin tosses. Furthermore, we can compute the probabilities that these random variables take various values. For example, in the notation of Example 1.1.4,

$$\mathbb{P}\{S_2 = 4\} = \mathbb{P}\big(A_{HT} \cup A_{TH}\big) = 2pq. \qquad \square$$

In the previous example, the random variables S_0, S_1, S_2, etc., have distributions. Indeed, $S_0 = 4$ with probability one, so we can regard this random variable as putting a unit of mass on the number 4. On the other hand, $\mathbb{P}\{S_2 = 16\} = p^2$, $\mathbb{P}\{S_2 = 4\} = 2pq$, and $\mathbb{P}\{S_2 = 1\} = q^2$. We can think of the distribution of this random variable as three lumps of mass, one of size p^2 located at the number 16, another of size $2pq$ located at the number 4, and a third of size q^2 located at the number 1. We need to allow for the possibility that the random variables we consider don't assign any lumps of mass but rather spread a unit of mass "continuously" over the real line. To do this, we should think of the distribution of a random variable as telling us how much mass is in a set rather than how much mass is at a point. In other words, the distribution of a random variable is itself a probability measure, but it is a measure on subsets of $\mathbb{R}$ rather than subsets of Ω.

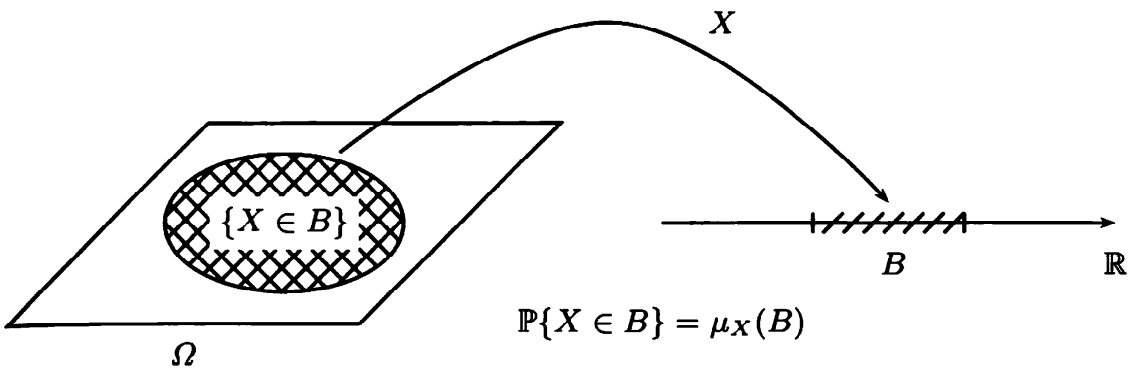

Fig. 1.2.1. Distribution measure of X.

Definition 1.2.3. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The* distribution measure of X *is the probability measure μ_X that assigns to each Borel subset B of $\mathbb{R}$ the mass $\mu_X(B) = \mathbb{P}\{X \in B\}$ (see Figure 1.2.1).*

In this definition, the set B could contain a single number. For example, if $B = \{4\}$, then in Example 1.2.2 we would have $\mu_{S_2}(B) = 2pq$. If $B = [2, 5]$, we still have $\mu_{S_2}(B) = 2pq$, because the only mass that S_2 puts in the interval $[2, 5]$ is the lump of mass placed at the number 4. Definition 1.2.3 for the distribution measure of a random variable makes sense for discrete random variables as well as for random variables that spread a unit of mass "continuously" over the real line.

Random variables have distributions, but distributions and random variables are different concepts. Two different random variables can have the same distribution. A single random variable can have two different distributions. Consider the following example.

Example 1.2.4. Let $\mathbb{P}$ be the uniform measure on $[0, 1]$ described in Example 1.1.3. Define $X(\omega) = \omega$ and $Y(\omega) = 1 - \omega$ for all $\omega \in [0, 1]$. Then the distribution measure of X is uniform, i.e.,

$$\mu_X[a, b] = \mathbb{P}\{\omega; a \le X(\omega) \le b\} = \mathbb{P}[a, b] = b - a, \quad 0 \le a \le b \le 1,$$

by the definition of $\mathbb{P}$. Although the random variable Y is different from the random variable X (if X takes the value $\frac{1}{3}$, Y takes the value $\frac{2}{3}$), Y has the same distribution as X:

$$\begin{aligned}\mu_Y[a, b] &= \mathbb{P}\{\omega; a \le Y(\omega) \le b\} = \mathbb{P}\{\omega; a \le 1 - \omega \le b\} = \mathbb{P}[1 - b, 1 - a] \\ &= (1 - a) - (1 - b) = b - a = \mu_X[a, b], \quad 0 \le a \le b \le 1.\end{aligned}$$

Now suppose we define another probability measure $\widetilde{\mathbb{P}}$ on $[0, 1]$ by specifying

$$\widetilde{\mathbb{P}}[a, b] = \int_a^b 2\omega \, d\omega = b^2 - a^2, \quad 0 \le a \le b \le 1. \tag{1.2.2}$$

Equation (1.2.2) and the properties of probability measures determine $\widetilde{\mathbb{P}}(B)$ for every Borel subset B of $\mathbb{R}$. Note that $\widetilde{\mathbb{P}}[0,1] = 1$, so $\widetilde{\mathbb{P}}$ is in fact a probability measure. Under $\widetilde{\mathbb{P}}$, the random variable X no longer has the uniform distribution. Denoting the distribution measure of X under $\widetilde{\mathbb{P}}$ by $\tilde{\mu}_X$, we have

$$\tilde{\mu}_X[a,b] = \widetilde{\mathbb{P}}\{\omega; a \le X(\omega) \le b\} = \widetilde{\mathbb{P}}[a,b] = b^2 - a^2, \quad 0 \le a \le b \le 1.$$

Under $\widetilde{\mathbb{P}}$, the distribution of Y no longer agrees with the distribution of X. We have

$$\begin{aligned}\tilde{\mu}_Y[a,b] &= \widetilde{\mathbb{P}}\{\omega; a \le Y(\omega) \le b\} = \widetilde{\mathbb{P}}\{\omega; a \le 1-\omega \le b\} = \widetilde{\mathbb{P}}[1-b, 1-a] \\ &= (1-a)^2 - (1-b)^2, \quad 0 \le a \le b \le 1. \end{aligned}$$ □

There are other ways to record the distribution of a random variable rather than specifying the distribution measure μ_X. We can describe the distribution of a random variable in terms of its *cumulative distribution function (cdf)*

$$F(x) = \mathbb{P}\{X \le x\}, \quad x \in \mathbb{R}. \tag{1.2.3}$$

If we know the distribution measure μ_X, then we know the cdf F because $F(x) = \mu_X(-\infty, x]$. On the other hand, if we know the cdf F, then we can compute $\mu_X(x,y] = F(y) - F(x)$ for $x < y$. For $a \le b$, we have

$$[a,b] = \bigcap_{n=1}^{\infty} \left(a - \tfrac{1}{n}, b\right],$$

and so we can compute[2]

$$\mu_X[a,b] = \lim_{n\to\infty} \mu_X\left(a - \tfrac{1}{n}, b\right] = F(b) - \lim_{n\to\infty} F\left(a - \tfrac{1}{n}\right). \tag{1.2.4}$$

Once the distribution measure $\mu_X[a,b]$ is known for every interval $[a,b] \subset \mathbb{R}$, it is determined for every Borel subset of $\mathbb{R}$. Therefore, in principle, knowing the cdf F for a random variable is the same as knowing its distribution measure μ_X.

In two special cases, the distribution of a random variable can be recorded in more detail. The first of these is when there is a *density function* $f(x)$, a nonnegative function defined for $x \in \mathbb{R}$ such that

$$\mu_X[a,b] = \mathbb{P}\{a \le X \le b\} = \int_a^b f(x)\,dx, \quad -\infty < a \le b < \infty. \tag{1.2.5}$$

In particular, because the closed intervals $[-n,n]$ have union $\mathbb{R}$, we must have[3]

[2] See Appendix A, Theorem A.1.1(ii) for more detail.
[3] See Appendix A, Theorem A.1.1(i) for more detail.

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{n\to\infty} \int_{-n}^{n} f(x)\,dx = \lim_{n\to\infty} \mathbb{P}\{-n \le X \le n\}$$
$$= \mathbb{P}\{X \in \mathbb{R}\} = \mathbb{P}(\Omega) = 1. \tag{1.2.6}$$

(For purposes of this discussion, we are not considering random variables that can take the value $\pm\infty$.)

The second special case is that of a *probability mass function*, in which case there is either a finite sequence of numbers $x_1, x_2, \ldots, x_N$ or an infinite sequence $x_1, x_2, \ldots$ such that with probability one the random variable X takes one of the values in the sequence. We then define $p_i = \mathbb{P}\{X = x_i\}$. Each p_i is nonnegative, and $\sum_i p_i = 1$. The mass assigned to a Borel set $B \subset \mathbb{R}$ by the distribution measure of X is

$$\mu_X(B) = \sum_{\{i; x_i \in B\}} p_i, \quad B \in \mathcal{B}(\mathbb{R}). \tag{1.2.7}$$

The distribution of some random variables can be described via a density, as in (1.2.5). For other random variables, the distribution must be described in terms of a probability mass function, as in (1.2.7). There are random variables whose distribution is given by a mixture of a density and a probability mass function, and there are random variables whose distribution has no lumps of mass but neither does it have a density.[4] Random variables of this last type have applications in finance but only at a level more advanced than this part of the text.

Example 1.2.5. (Another random variable uniformly distributed on [0,1].) We construct a uniformly distributed random variable taking values in $[0, 1]$ and defined on infinite coin-toss space Ω_∞. Suppose in the independent coin-toss space of Example 1.1.4 that the probability for head on each toss is $p = \frac{1}{2}$. For $n = 1, 2, \ldots$, we define

$$Y_n(\omega) = \begin{cases} 1 \text{ if } \omega_n = H, \\ 0 \text{ if } \omega_n = T. \end{cases} \tag{1.2.8}$$

We set

$$X = \sum_{n=1}^{\infty} \frac{Y_n}{2^n}.$$

If $Y_1 = 0$, which happens with probability $\frac{1}{2}$, then $0 \le X \le \frac{1}{2}$. If $Y_1 = 1$, which also happens with probability $\frac{1}{2}$, then $\frac{1}{2} \le X \le 1$. If $Y_1 = 0$ and $Y_2 = 0$, which happens with probability $\frac{1}{4}$, then $0 \le X \le \frac{1}{4}$. If $Y_1 = 0$ and $Y_2 = 1$, which also happens with probability $\frac{1}{4}$, then $\frac{1}{4} \le X \le \frac{1}{2}$. This pattern continues; indeed for any interval $[\frac{k}{2^n}, \frac{k+1}{2^n}] \subset [0, 1]$, the probability that the interval contains X is $\frac{1}{2^n}$. In terms of the distribution measure μ_X of X, we write this fact as

$$\mu_X\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right] = \frac{1}{2^n} \text{ whenever } k \text{ and } n \text{ are integers and } 0 \le k \le 2^n - 1.$$

[4] See Appendix A, Section A.3.

Taking unions of intervals of this form and using the finite additivity of probability measures, we see that whenever k, m, and n are integers and $0 \le k \le m \le 2^n$, we have

$$\mu_X\left[\frac{k}{2^n}, \frac{m}{2^n}\right] = \frac{m}{2^n} - \frac{k}{2^n}. \tag{1.2.9}$$

From (1.2.9), one can show that

$$\mu_X[a,b] = b - a, \quad 0 \le a \le b \le 1;$$

in other words, the distribution measure of X is uniform on $[0,1]$.

Example 1.2.6 (Standard normal random variable). Let

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

be the *standard normal density*, and define the *cumulative normal distribution function*

$$N(x) = \int_{-\infty}^{x} \varphi(\xi)\, d\xi.$$

The function $N(x)$ is strictly increasing, mapping $\mathbb{R}$ onto $(0,1)$, and so has a strictly increasing inverse function $N^{-1}(y)$. In other words, $N(N^{-1}(y)) = y$ for all $y \in (0,1)$. Now let Y be a uniformly distributed random variable, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (two possibilities for $(\Omega, \mathcal{F}, \mathbb{P})$ and Y are presented in Examples 1.2.4 and 1.2.5), and set $X = N^{-1}(Y)$. Whenever $-\infty < a \le b < \infty$, we have

$$\begin{aligned}
\mu_X[a,b] &= \mathbb{P}\{\omega \in \Omega; a \le X(\omega) \le b\} \\
&= \mathbb{P}\{\omega \in \Omega; a \le N^{-1}(Y(\omega)) \le b\} \\
&= \mathbb{P}\{\omega \in \Omega; N(a) \le N(N^{-1}(Y(\omega))) \le N(b)\} \\
&= \mathbb{P}\{\omega \in \Omega; N(a) \le Y(\omega) \le N(b)\} \\
&= N(b) - N(a) \\
&= \int_a^b \varphi(x)\, dx.
\end{aligned}$$

The measure μ_X on $\mathbb{R}$ given by this formula is called the *standard normal distribution*. Any random variable that has this distribution, regardless of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which it is defined, is called a *standard normal random variable*. The method used here for generating a standard normal random variable from a uniformly distributed random variable is called the *probability integral transform* and is widely used in Monte Carlo simulation.

Another way to construct a standard normal random variable is to take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, take $\mathbb{P}$ to be the probability measure on $\mathbb{R}$ that satisfies

$$\mathbb{P}[a,b] = \int_a^b \varphi(x)\,dx \text{ whenever } -\infty < a \le b < \infty,$$

and take $X(\omega) = \omega$ for all $\omega \in \mathbb{R}$. □

The second construction of a standard normal random variable in Example 1.2.6 is economical, and this method can be used to construct a random variable with any desired distribution. However, it is not useful when we want to have multiple random variables, each with a specified distribution and with certain dependencies among the random variables. For such cases, we construct (or at least assume there exists) a single probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all the random variables of interest are defined. This point of view may seem overly abstract at the outset, but in the end it pays off handsomely in conceptual simplicity.

1.3 Expectations

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We would like to compute an "average value" of X, where we take the probabilities into account when doing the averaging. If Ω is finite, we simply define this average value by

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega).$$

If Ω is countably infinite, its elements can be listed in a sequence $\omega_1, \omega_2, \omega_3, \ldots$, and we can define $\mathbb{E}X$ as an infinite sum:

$$\mathbb{E}X = \sum_{k=1}^{\infty} X(\omega_k)\mathbb{P}(\omega_k).$$

Difficulty arises, however, if Ω is uncountably infinite. Uncountable sums cannot be defined. Instead, we must think in terms of integrals.

To see how to go about this, we first review the Riemann integral. If $f(x)$ is a continuous function defined for all x in the closed interval $[a,b]$, we define the Riemann integral $\int_a^b f(x)dx$ as follows. First partition $[a,b]$ into subintervals $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$, where $a = x_0 < x_1 < \cdots < x_n = b$. We denote by $\Pi = \{x_0, x_1, \ldots, x_n\}$ the set of partition points and by

$$\|\Pi\| = \max_{1 \le k \le n} (x_k - x_{k-1})$$

the length of the longest subinterval in the partition. For each subinterval $[x_{k-1}, x_k]$, we set $M_k = \max_{x_{k-1} \le x \le x_k} f(x)$ and $m_k = \min_{x_{k-1} \le x \le x_k} f(x)$. The upper Riemann sum is

$$\mathrm{RS}^+_\Pi(f) = \sum_{k=1}^{n} M_k(x_k - x_{k-1}),$$

and the lower Riemann sum (see Figure 1.3.1) is

$$\mathrm{RS}^-_{\Pi}(f) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}).$$

As $\|\Pi\|$ converges to zero (i.e., as we put in more and more partition points, and the subintervals in the partition become shorter and shorter), the upper Riemann sum $\mathrm{RS}^+_{\Pi}(f)$ and the lower Riemann sum $\mathrm{RS}^-_{\Pi}(f)$ converge to the same limit, which we call $\int_a^b f(x)dx$. This is the *Riemann integral.*

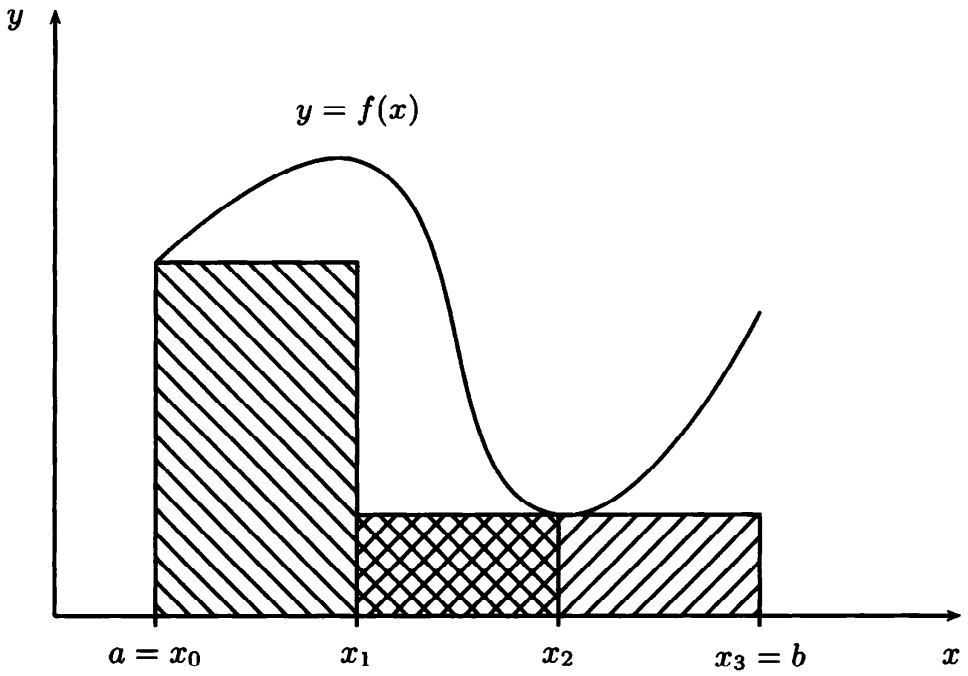

Fig. 1.3.1. Lower Riemann sum.

The problem we have with imitating this procedure to define expectation is that the random variable X, unlike the function f in the previous paragraph, is a function of $\omega \in \Omega$, and Ω is often not a subset of $\mathbb{R}$. In Figure 1.3.2 the "x-axis" is not the real numbers but some abstract space Ω. There is no natural way to partition the set Ω as we partitioned $[a, b]$ above. Therefore, we partition instead the y-axis in Figure 1.3.2. To see how this goes, assume for the moment that $0 \le X(\omega) < \infty$ for every $\omega \in \Omega$, and let $\Pi = \{y_0, y_1, y_2, \dots\}$, where $0 = y_0 < y_1 < y_2 < \dots$. For each subinterval $[y_k, y_{k+1}]$, we set

$$A_k = \{\omega \in \Omega; y_k \le X(\omega) < y_{k+1}\}.$$

We define the lower Lebesgue sum to be (see Figure 1.3.2)

$$\mathrm{LS}^-_{\Pi}(X) = \sum_{k=1}^{\infty} y_k \, \mathbb{P}(A_k).$$

This lower sum converges as $\|\Pi\|$, the maximal distance between the y_k partition points, approaches zero, and we define this limit to be the *Lebesgue integral* $\int_\Omega X(\omega)\,d\mathbb{P}(\omega)$, or simply $\int_\Omega X\,d\mathbb{P}$. The Lebesgue integral might be ∞, because we have not made any assumptions about how large the values of X can be.

We assumed a moment ago that $0 \le X(\omega) < \infty$ for every $\omega \in \Omega$. If the set of ω that violates this condition has zero probability, there is no effect on the integral we just defined. If $\mathbb{P}\{\omega; X(\omega) \ge 0\} = 1$ but $\mathbb{P}\{\omega; X(\omega) = \infty\} > 0$, then we define $\int_\Omega X(\omega)d\mathbb{P}(\omega) = \infty$.

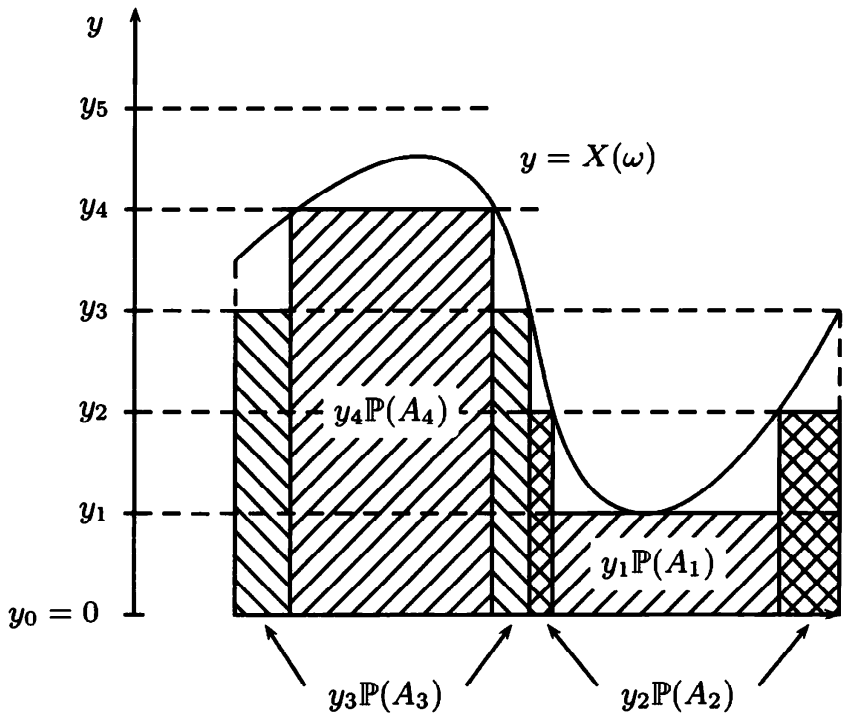

Fig. 1.3.2. Lower Lebesgue sum.

Finally, we need to consider random variables X that can take both positive and negative values. For such a random variable, we define the *positive* and *negative parts* of X by

$$X^+(\omega) = \max\{X(\omega), 0\}, \quad X^-(\omega) = \max\{-X(\omega), 0\}. \tag{1.3.1}$$

Both X^+ and X^- are nonnegative random variables, $X = X^+ - X^-$, and $|X| = X^+ + X^-$. Both $\int_\Omega X^+(\omega)\,d\mathbb{P}(\omega)$ and $\int_\Omega X^-(\omega)\,d\mathbb{P}(\omega)$ are defined by the procedure described above, and provided they are not both ∞, we can define

$$\int_\Omega X(\omega)\,d\mathbb{P}(\omega) = \int_\Omega X^+(\omega)\,d\mathbb{P}(\omega) - \int_\Omega X^-(\omega)\,d\mathbb{P}(\omega). \tag{1.3.2}$$

If $\int_\Omega X^+(\omega)\,d\mathbb{P}(\omega)$ and $\int_\Omega X^-(\omega)\,d\mathbb{P}(\omega)$ are both finite, we say that X is *integrable*, and $\int_\Omega X(\omega)\,d\mathbb{P}(\omega)$ is also finite. If $\int_\Omega X^+(\omega)\,d\mathbb{P}(\omega) = \infty$ and

$\int_\Omega X^-(\omega)\,d\mathbb{P}(\omega)$ is finite, then $\int_\Omega X(\omega)\,d\mathbb{P}(\omega) = \infty$. If $\int_\Omega X^+(\omega)\,d\mathbb{P}(\omega)$ is finite and $\int_\Omega X^-(\omega)\,d\mathbb{P}(\omega) = \infty$, then $\int_\Omega X(\omega)\,d\mathbb{P}(\omega) = -\infty$. If both $\int_\Omega X^+(\omega)\,d\mathbb{P}(\omega) = \infty$ and $\int_\Omega X^-(\omega)\,d\mathbb{P}(\omega) = \infty$, then an "$\infty - \infty$" situation arises in (1.3.2), and $\int_\Omega X(\omega)\,d\mathbb{P}(\omega)$ is not defined.

The Lebesgue integral has the following basic properties.

Theorem 1.3.1. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

(i) If X takes only finitely many values $y_0, y_1, y_2, \ldots, y_n$, then

$$\int_\Omega X(\omega)\,d\mathbb{P}(\omega) = \sum_{k=0}^{n} y_k \mathbb{P}\{X = y_k\}.$$

(ii) **(Integrability)** *The random variable X is integrable if and only if*

$$\int_\Omega |X(\omega)|\,d\mathbb{P}(\omega) < \infty.$$

Now let Y be another random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.

(iii) **(Comparison)** *If $X \le Y$ almost surely (i.e., $\mathbb{P}\{X \le Y\} = 1$), and if $\int_\Omega X(\omega)\,d\mathbb{P}(\omega)$ and $\int_\Omega Y(\omega)\,d\mathbb{P}(\omega)$ are defined, then*

$$\int_\Omega X(\omega)\,d\mathbb{P}(\omega) \le \int_\Omega Y(\omega)\,d\mathbb{P}(\omega).$$

In particular, if $X = Y$ almost surely and one of the integrals is defined, then they are both defined and

$$\int_\Omega X(\omega)\,d\mathbb{P}(\omega) = \int_\Omega Y(\omega)\,d\mathbb{P}(\omega).$$

(iv) **(Linearity)** *If α and β are real constants and X and Y are integrable, or if α and β are nonnegative constants and X and Y are nonnegative, then*

$$\int_\Omega (\alpha X(\omega) + \beta Y(\omega))\,d\mathbb{P}(\omega) = \alpha \int_\Omega X(\omega)\,d\mathbb{P}(\omega) + \beta \int_\Omega Y(\omega)\,d\mathbb{P}(\omega).$$

PARTIAL PROOF: For (i), we consider only the case when X is almost surely nonnegative. If zero is not among the y_ks, we may add $y_0 = 0$ to the list and then relabel the y_ks if necessary so that $0 = y_0 < y_1 < y_2 < \cdots < y_n$. Using these as our partition points, we have $A_k = \{y_k \le X < y_{k+1}\} = \{X = y_k\}$ and the lower Lebesgue sum is

$$\mathrm{LS}_\Pi^-(X) = \sum_{k=0}^{n} y_k \mathbb{P}\{X = y_k\}.$$

If we put in more partition points, the lower Lebesgue sum does not change, and hence this is also the Lebesgue integral.

We next consider part (iii). If $X \leq Y$ almost surely, then $X^+ \leq Y^+$ and $X^- \geq Y^-$ almost surely. Because $X^+ \leq Y^+$ almost surely, for every partition Π, the lower Lebesgue sums satisfy $\mathrm{LS}_\Pi^-(X^+) \leq \mathrm{LS}_\Pi^-(Y^+)$, so

$$\int_\Omega X^+(\omega)\, d\mathbb{P}(\omega) \leq \int_\Omega Y^+(\omega)\, d\mathbb{P}(\omega). \tag{1.3.3}$$

Because $X^- \geq Y^-$ almost surely, we also have

$$\int_\Omega X^-(\omega)\, d\mathbb{P}(\omega) \geq \int_\Omega Y^-(\omega)\, d\mathbb{P}(\omega). \tag{1.3.4}$$

Subtracting (1.3.4) from (1.3.3) and recalling the definition (1.3.2), we obtain the comparison property (iii).

The linearity property (iv) requires a more detailed analysis of the construction of Lebesgue integrals. We do not provide that here.

We can use the comparison property (iii) and the linearity property (iv) to prove (ii) as follows. Because $|X| = X^+ + X^-$, we have $X^+ \leq |X|$ and $X^- \leq |X|$. If $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$, then the comparison property implies $\int_\Omega X^+(\omega)\, d\mathbb{P}(\omega) < \infty$ and $\int_\Omega X^-(\omega)\, d\mathbb{P}(\omega) < \infty$, and X is integrable by definition. On the other hand, if X is integrable, then $\int_\Omega X^+(\omega)\, d\mathbb{P}(\omega) < \infty$ and $\int_\Omega X^-(\omega)\, d\mathbb{P}(\omega) < \infty$. Adding these two quantities and using (iv), we see that $\int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$. □

Remark 1.3.2. We often want to integrate a random variable X over a subset A of Ω rather than over all of Ω. For this reason, we define

$$\int_A X(\omega)\, d\mathbb{P}(\omega) = \int_\Omega \mathbb{I}_A(\omega) X(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{F},$$

where $\mathbb{I}_A$ is the *indicator function (random variable)* given by

$$\mathbb{I}_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A, \\ 0 \text{ if } \omega \notin A. \end{cases}$$

If A and B are disjoint sets in $\mathcal{F}$, then $\mathbb{I}_A + \mathbb{I}_B = \mathbb{I}_{A \cup B}$ and the linearity property (iv) of Theorem 1.3.1 implies that

$$\int_{A \cup B} X(\omega)\, d\mathbb{P}(\omega) = \int_A X(\omega)\, d\mathbb{P}(\omega) + \int_B X(\omega)\, d\mathbb{P}(\omega).$$

Definition 1.3.3. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The* expectation *(or* expected value*) of X is defined to be*

$$\mathbb{E}X = \int_\Omega X(\omega)\, d\mathbb{P}(\omega).$$

This definition makes sense if X is integrable*, i.e.; if*

$$\mathbb{E}|X| = \int_\Omega |X(\omega)|\, d\mathbb{P}(\omega) < \infty$$

or if $X \geq 0$ almost surely. In the latter case, $\mathbb{E}X$ might be ∞.

We have thus managed to define $\mathbb{E}X$ when X is a random variable on an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We restate in terms of expected values the basic properties of Theorem 1.3.1 and add an additional one.

Theorem 1.3.4. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

(i) If X takes only finitely many values $x_0, x_1, \ldots, x_n$, then

$$\mathbb{E}X = \sum_{k=0}^{n} x_k \mathbb{P}\{X = x_k\}.$$

In particular, if Ω is finite, then

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega).$$

(ii) **(Integrability)** *The random variable X is integrable if and only if*

$$\mathbb{E}|X| < \infty.$$

Now let Y be another random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.

(iii) **(Comparison)** *If $X \leq Y$ almost surely and X and Y are integrable or almost surely nonnegative, then*

$$\mathbb{E}X \leq \mathbb{E}Y.$$

In particular, if $X = Y$ almost surely and one of the random variables is integrable or almost surely nonnegative, then they are both integrable or almost surely nonnegative, respectively, and

$$\mathbb{E}X = \mathbb{E}Y.$$

(iv) **(Linearity)** *If α and β are real constants and X and Y are integrable or if α and β are nonnegative constants and X and Y are nonnegative, then*

$$E(\alpha X + \beta Y) = \alpha \mathbb{E}X + \beta \mathbb{E}Y.$$

(v) **(Jensen's inequality)** *If φ is a convex, real-valued function defined on $\mathbb{R}$, and if $\mathbb{E}|X| < \infty$, then*

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X).$$

PROOF: The only new claim is Jensen's inequality, and the proof of that is the same as the proof given for Theorem 2.2.5 of Chapter 2 of Volume I. □

Example 1.3.5. Consider the infinite independent coin-toss space Ω_∞ of Example 1.1.4 with the probability measure $\mathbb{P}$ that corresponds to probability $\frac{1}{2}$ for head on each toss. Let

$$Y_n(\omega) = \begin{cases} 1 \text{ if } \omega_n = H, \\ 0 \text{ if } \omega_n = T. \end{cases}$$

Even though the probability space Ω_∞ is uncountable, this random variable takes only two values, and we can compute its expectation using Theorem 1.3.4(i):

$$\mathbb{E}Y_n = 1 \cdot \mathbb{P}\{Y_n = 1\} + 0 \cdot \mathbb{P}\{Y_n = 0\} = \frac{1}{2}.$$

Example 1.3.6. Let $\Omega = [0,1]$, and let $\mathbb{P}$ be the Lebesgue measure on $[0,1]$ (see Example 1.1.3). Consider the random variable

$$X(\omega) = \begin{cases} 1 \text{ if } \omega \text{ is irrational}, \\ 0 \text{ if } \omega \text{ is rational}. \end{cases}$$

Again the random variable takes only two values, and we can compute its expectation using Theorem 1.3.4(i):

$$\mathbb{E}X = 1 \cdot \mathbb{P}\{\omega \in [0,1]; \omega \text{ is irrational}\} + 0 \cdot \mathbb{P}\{\omega \in [0,1]; \omega \text{ is rational}\}.$$

There are only countably many rational numbers in $[0,1]$ (i.e., they can all be listed in a sequence $x_1, x_2, x_3, \ldots$). Each number in the sequence has probability zero, and because of the countable additivity property (ii) of Definition 1.1.2, the whole sequence must have probability zero. Therefore, $\mathbb{P}\{\omega \in [0,1]; \omega \text{ is rational}\} = 0$. Since $\mathbb{P}[0,1] = 1$, the probability of the set of irrational numbers in $[0,1]$ must be 1. We conclude that $\mathbb{E}X = 1$.

The idea behind this example is that if we choose a number from $[0,1]$ according to the uniform distribution, then with probability one the number chosen will be irrational. Therefore, the random variable X is almost surely equal to 1, and hence its expected value equals 1. As a practical matter, of course, almost any algorithm we devise for generating a random number in $[0,1]$ will generate a rational number. The uniform distribution is often a reasonable idealization of the output of algorithms that generate random numbers in $[0,1]$, but if we push the model too far it can depart from reality.

If we had been working with Riemann rather than Lebesgue integrals, we would have gotten a different result. To make the notation more familiar, we write x rather than ω and $f(x)$ rather than $X(\omega)$, thus defining

$$f(x) = \begin{cases} 1 \text{ if } x \text{ is irrational}, \\ 0 \text{ if } x \text{ is rational}. \end{cases} \tag{1.3.5}$$

We have just seen that the Lebesgue integral of this function over the interval $[0,1]$ is 1.

To construct the Riemann integral, we choose partition points $0 = x_0 < x_1 < x_2 < \cdots < x_n = 1$. We define

$$M_k = \max_{x_{k-1} \le x \le x_k} f(x), \quad m_k = \min_{x_{k-1} \le x \le x_k} f(x).$$

But each interval $[x_{k-1}, x_k]$ contains both rational and irrational numbers, so $M_k = 1$ and $m_k = 0$. Therefore, for this partition $\Pi = \{x_0, x_1, \ldots, x_n\}$, the upper Riemann sum is 1,

$$\mathrm{RS}^+_\Pi(f) = \sum_{k=1}^{n} M_k(x_k - x_{k-1}) = \sum_{k=1}^{n}(x_k - x_{k-1}) = 1,$$

whereas the lower Riemann sum is zero,

$$\mathrm{RS}^-_\Pi(f) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}) = 0.$$

This happens no matter how small we take the subintervals in the partition. Since the upper Riemann sum is always 1 and the lower Riemann sum is always 0, the upper and lower Riemann sums do not converge to the same limit and the Riemann integral is not defined. For the Riemann integral, which discretizes the x-axis rather than the y-axis, this function is too discontinuous to handle. The Lebesgue integral, however, which discretizes the y-axis, sees this as a simple function taking only two values. □

We constructed the Lebesgue integral because we wanted to integrate over abstract probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$, but as Example 1.3.6 shows, after this construction we can take Ω to be a subset of the real numbers and then compare Lebesgue and Riemann integrals. This example further shows that these two integrals can give different results. Fortunately, the behavior in Example 1.3.6 is the worst that can happen. To make this statement precise, we first extend the construction of the Lebesgue integral to all of $\mathbb{R}$, rather than just $[0, 1]$.

Definition 1.3.7. *Let $\mathcal{B}(\mathbb{R})$ be the σ-algebra of Borel subsets of $\mathbb{R}$ (i.e., the smallest σ-algebra containing all the closed intervals $[a, b]$).*[5] *The* Lebesgue measure on $\mathbb{R}$, *which we denote by $\mathcal{L}$, assigns to each set $B \in \mathcal{B}(\mathbb{R})$ a number in $[0, \infty)$ or the value ∞ so that*

(i) $\mathcal{L}[a, b] = b - a$ whenever $a \le b$, and
(ii) if $B_1, B_2, B_3, \ldots$ is a sequence of disjoint sets in $\mathcal{B}(\mathbb{R})$, then we have the countable additivity property

$$\mathcal{L}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathcal{L}(B_n).$$

[5] This concept is discussed in more detail in Appendix A, Section A.2.

Definition 1.3.7 is similar to Definition 1.1.2, except that now some sets have measure greater than 1. The Lebesgue measure of every interval is its length, so that $\mathbb{R}$ and half-lines $[a,\infty)$ and $(-\infty,b]$ have infinite Lebesgue measure, single points have Lebesgue measure zero, and the Lebesgue measure of the empty set is zero. Lebesgue measure has the finite additivity property (see (1.1.5))

$$\mathcal{L}\left(\bigcup_{n=1}^{N} B_n\right) = \sum_{n=1}^{N} \mathcal{L}(B_n)$$

whenever $B_1, B_2, \ldots, B_N$ are disjoint Borel subsets of $\mathbb{R}$.

Now let $f(x)$ be a real-valued function defined on $\mathbb{R}$. For the following construction, we need to assume that for every Borel subset B of $\mathbb{R}$, the set $\{x; f(x) \in B\}$ is also a Borel subset of $\mathbb{R}$. A function f with this property is said to be *Borel measurable*. Every continuous and piecewise continuous function is Borel measurable. Indeed, it is extremely difficult to find a function that is not Borel measurable. We wish to define the Lebesgue integral $\int_{\mathbb{R}} f(x)\, d\mathcal{L}(x)$ of f over $\mathbb{R}$. To do this, we assume for the moment that $0 \le f(x) < \infty$ for every $x \in \mathbb{R}$. We choose a partition $\Pi = \{y_0, y_1, y_2, \ldots\}$, where $0 = y_0 < y_1 < y_2 < \ldots$. For each subinterval $[y_k, y_{k+1})$, we define

$$B_k = \{x \in \mathbb{R}; y_k \le f(x) < y_{k+1}\}.$$

Because of the assumption that f is Borel measurable, even though these sets B_k can be quite complicated, they are Borel subsets of $\mathbb{R}$ and so their Lebesgue measures are defined. We define the lower Lebesgue sum

$$\mathrm{LS}^{-}_{\Pi}(f) = \sum_{k=1}^{\infty} y_k \mathcal{L}(B_k).$$

As $\|\Pi\|$ converges to zero, these lower Lebesgue sums will converge to a limit, which we define to be $\int_{\mathbb{R}} f(x)\, d\mathcal{L}(x)$. It is possible that this integral gives the value ∞.

We assumed a moment ago that $0 \le f(x) < \infty$ for every $x \in \mathbb{R}$. If the set of x where the condition is violated has zero Lebesgue measure, the integral of f is not affected. If $\mathcal{L}\{x \in \mathbb{R}; f(x) < 0\} = 0$ and $\mathcal{L}\{x \in \mathbb{R}; f(x) = \infty\} > 0$, we define $\int_{\mathbb{R}} f(x) d\mathcal{L}(x) = \infty$.

We next consider the possibility that $f(x)$ takes both positive and negative values. In this case, we define

$$f^+(x) = \max\{f(x), 0\}, \quad f^-(x) = \max\{-f(x), 0\}.$$

Because f^+ and f^- are nonnegative, $\int_{\mathbb{R}} f^+(x)\, d\mathcal{L}(x)$ and $\int_{\mathbb{R}} f^-(x)\, d\mathcal{L}(x)$ are defined by the procedure described above. We then define

$$\int_{\mathbb{R}} f(x)\, d\mathcal{L}(x) = \int_{\mathbb{R}} f^+(x)\, d\mathcal{L}(x) - \int_{\mathbb{R}} f^-(x)\, d\mathcal{L}(x),$$

provided this is not $\infty - \infty$. In the case where both $\int_{\mathbb{R}} f^+(x)\,d\mathcal{L}(x)$ and $\int_{\mathbb{R}} f^-(x)\,d\mathcal{L}(x)$ are infinite, $\int_{\mathbb{R}} f(x)\,d\mathcal{L}(x)$ is not defined. If $\int_{\mathbb{R}} f^+(x)\,d\mathcal{L}(x)$ and $\int_{\mathbb{R}} f^-(x)\,d\mathcal{L}(x)$ are finite, we say that f is *integrable*. This is equivalent to the condition $\int_{\mathbb{R}} |f(x)|\,d\mathcal{L}(x) < \infty$. The Lebesgue integral just constructed has the comparison and linearity properties described in Theorem 1.3.1. Moreover, if f takes only finitely many values $y_0, y_1, y_2, \ldots, y_n$, then

$$\int_{\mathbb{R}} f(x)\,d\mathcal{L}(x) = \sum_{k=0}^{n} y_k\,\mathcal{L}\{x \in \mathbb{R}; f(x) = y_k\},$$

provided the computation of the right-hand side does not require that $\infty - \infty$ be assigned a value.

Finally, sometimes we have a function $f(x)$ defined for every $x \in \mathbb{R}$ but want to compute its Lebesgue integral over only part of $\mathbb{R}$, say over some set $B \in \mathcal{B}(\mathbb{R})$. We do this by multiplying $f(x)$ by the indicator function of B:

$$\mathbb{I}_B(x) = \begin{cases} 1 \text{ if } x \in B, \\ 0 \text{ if } x \notin B. \end{cases}$$

The product $f(x)\mathbb{I}_B(x)$ agrees with $f(x)$ when $x \in B$ and is zero when $x \notin B$. We define

$$\int_B f(x)\,d\mathcal{L}(x) = \int_{\mathbb{R}} \mathbb{I}_B(x) f(x)\,d\mathcal{L}(x).$$

The following theorem, whose proof is beyond the scope of this book, relates Riemann and Lebesgue integrals on $\mathbb{R}$.

Theorem 1.3.8. (Comparison of Riemann and Lebesgue integrals). *Let f be a bounded function defined on $\mathbb{R}$, and let $a < b$ be numbers.*

(i) The Riemann integral $\int_a^b f(x)dx$ is defined (i.e., the lower and upper Riemann sums converge to the same limit) if and only if the set of points x in $[a,b]$ where $f(x)$ is not continuous has Lebesgue measure zero.

(ii) If the Riemann integral $\int_a^b f(x)dx$ is defined, then f is Borel measurable (so the Lebesgue integral $\int_{[a,b]} f(x)\,d\mathcal{L}(x)$ is also defined), and the Riemann and Lebesgue integrals agree.

A single point in $\mathbb{R}$ has Lebesgue measure zero, and so any finite set of points has Lebesgue measure zero. Theorem 1.3.8 guarantees that if we have a real-valued function f on $\mathbb{R}$ that is continuous except at finitely many points, then there will be no difference between Riemann and Lebesgue integrals of this function.

Definition 1.3.9. *If the set of numbers in $\mathbb{R}$ that fail to have some property is a set with Lebesgue measure zero, we say that the property holds* almost everywhere.

Theorem 1.3.8(i) may be restated as:

The Riemann integral $\int_a^b f(x)dx$ *exists if and only if* $f(x)$ *is almost everywhere continuous on* $[a,b]$.

Because the Riemann and Lebesgue integrals agree whenever the Riemann integral is defined, we shall use the more familiar notation $\int_a^b f(x)\,dx$ to denote the Lebesgue integral rather than $\int_{[a,b]} f(x)\,d\mathcal{L}(x)$. If the set B over which we wish to integrate is not an interval, we shall write $\int_B f(x)\,dx$. When we are developing theory, we shall understand $\int_B f(x)\,dx$ to be a Lebesgue integral; when we need to compute, we will use techniques learned in calculus for computing Riemann integrals.

1.4 Convergence of Integrals

There are several ways a sequence of random variables can converge. In this section, we consider the case of convergence almost surely, defined as follows.

Definition 1.4.1. *Let* $X_1, X_2, X_3, \ldots$ *be a sequence of random variables, all defined on the same probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* X *be another random variable defined on this space. We say that* $X_1, X_2, X_3, \ldots$ converges to X almost surely *and write*

$$\lim_{n\to\infty} X_n = X \text{ almost surely}$$

if the set of $\omega \in \Omega$ *for which the sequence of numbers* $X_1(\omega), X_2(\omega), X_3(\omega), \ldots$ *has limit* $X(\omega)$ *is a set with probability one. Equivalently, the set of* $\omega \in \Omega$ *for which the sequence of numbers* $X_1(\omega), X_2(\omega), X_3(\omega), \ldots$ *does not converge to* $X(\omega)$ *is a set with probability zero.*

Example 1.4.2 (Strong Law of Large Numbers). An intuitively appealing case of almost sure convergence is the *Strong Law of Large Numbers*. On the infinite independent coin-toss space Ω_∞, with the probability measure chosen to correspond to probability $p = \frac{1}{2}$ of head on each toss, we define

$$Y_k(\omega) = \begin{cases} 1 \text{ if } \omega_k = H, \\ 0 \text{ if } \omega_k = T, \end{cases}$$

and

$$H_n = \sum_{k=1}^{n} Y_k,$$

so that H_n is the number of heads obtained in the first n tosses. The Strong Law of Large Numbers is a theorem that asserts that

$$\lim_{n\to\infty} \frac{H_n}{n} = \frac{1}{2} \text{ almost surely.}$$

In other words, the ratio of the number of heads to the number of tosses approaches $\frac{1}{2}$ almost surely. The "almost surely" in this assertion acknowledges the fact that there are sequences of tosses, such as the sequence of all heads, for which the ratio does not converge to $\frac{1}{2}$. We shall ultimately see that there are in fact uncountably many such sequences. However, under our assumptions that the probability of head on each toss is $\frac{1}{2}$ and the tosses are independent, the probability of all these sequences taken together is zero. □

Definition 1.4.3. *Let $f_1, f_2, f_3, \ldots$ be a sequence of real-valued, Borel-measurable functions defined on $\mathbb{R}$. Let f be another real-valued, Borel-measurable function defined on $\mathbb{R}$. We say that $f_1, f_2, f_3, \ldots$* converges to f almost everywhere *and write*

$$\lim_{n\to\infty} f_n = f \text{ almost everywhere}$$

if the set of $x \in \mathbb{R}$ for which the sequence of numbers $f_1(x), f_2(x), f_3(x), \ldots$ does not have limit $f(x)$ is a set with Lebesgue measure zero.

It is clear from these definitions that convergence almost surely and convergence almost everywhere are really the same concept in different notation.

Example 1.4.4. Consider a sequence of normal densities, each with mean zero and the nth having variance $\frac{1}{n}$ (see Figure 1.4.1):

$$f_n(x) = \sqrt{\tfrac{n}{2\pi}}\, e^{-\frac{nx^2}{2}}.$$

If $x \neq 0$, then $\lim_{n\to\infty} f_n(x) = 0$, but

$$\lim_{n\to\infty} f_n(0) = \lim_{n\to\infty} \sqrt{\frac{n}{2\pi}} = \infty.$$

Therefore, the sequence $f_1, f_2, f_3, \ldots$ converges everywhere to the function

$$f^*(x) = \begin{cases} 0 & \text{if } x \neq 0, \\ \infty & \text{if } x = 0, \end{cases}$$

and converges almost everywhere to the identically zero function $f(x) = 0$ for all $x \in \mathbb{R}$. The set of x where the convergence to $f(x)$ does not take place contains only the number 0, and this set has zero Lebesgue measure. □

Often when random variables converge almost surely, their expected values converge to the expected value of the limiting random variable. Likewise, when functions converge almost everywhere, it is often the case that their Lebesgue integrals converge to the Lebesgue integral of the limiting function. This is not always the case, however. In Example 1.4.4, we have a sequence of normal densities for which $\int_{-\infty}^{\infty} f_n(x)dx = 1$ for every n but the almost everywhere limit function f is identically zero. It would not help matters to use the everywhere limit function $f^*(x)$ because any two functions that differ only on a set of zero Lebesgue measure must have the same Lebesgue integral.

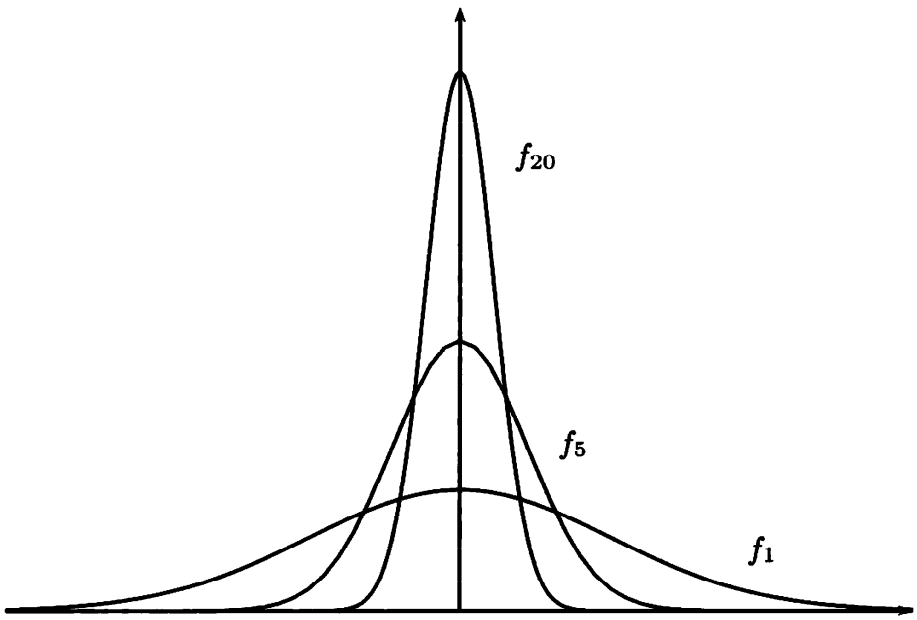

Fig. 1.4.1. Almost everywhere convergence.

Therefore, $\int_{-\infty}^{\infty} f^*(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx = 0$. It cannot be otherwise because $2f^*(x) = f^*(x)$ for every $x \in \mathbb{R}$, and so

$$2\int_{-\infty}^{\infty} f^*(x)\,dx = \int_{-\infty}^{\infty} 2f^*(x) = \int_{-\infty}^{\infty} f^*(x)\,dx.$$

This equation implies that $\int_{-\infty}^{\infty} f^*(x)\,dx = 0$. It would also not help matters to replace the functions f_n by the functions

$$g_n(x) = \begin{cases} f_n(x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The sequence $g_1, g_2, \ldots$ converges to 0 *everywhere*, whereas the integrals $\int_{-\infty}^{\infty} g_n(x)\,dx$ agree with the integrals $\int_{-\infty}^{\infty} f_n(x)\,dx$, and these converge to 1, not 0. The inescapable conclusion is that in this example

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f_n(x)\,dx \neq \int_{-\infty}^{\infty} \lim_{n\to\infty} f_n(x)\,dx;$$

the left-hand side is 1 and the right-hand side is 0.

Incidentally, matters are even worse with the Riemann integral, which is not defined for f^*; upper Riemann sums for f^* are infinite, and lower Riemann sums are zero.

To get the integrals of a sequence of functions to converge to the integral of the limiting function, we need to impose some condition. One condition that guarantees this is that all the functions are nonnegative and they converge to their limit from below. If we think of an integral as the area under a curve, the assumption is that as we go farther out in the sequence of functions, we keep adding area and never taking it away. If we do this, then the area under the

limiting function is the limit of the areas under the functions in the sequence. The precise statement of this result is given in the following theorem.

Theorem 1.4.5 (Monotone convergence). *Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables converging almost surely to another random variable X. If*

$$0 \le X_1 \le X_2 \le X_3 \le \ldots \text{ almost surely,}$$

then

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X.$$

Let $f_1, f_2, f_3, \ldots$ be a sequence of Borel-measurable functions on $\mathbb{R}$ converging almost everywhere to a function f. If

$$0 \le f_1 \le f_2 \le f_3 \le \ldots \text{ almost everywhere,}$$

then

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f_n(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx.$$

The following corollary to the Monotone Convergence Theorem extends Theorem 1.3.4(i).

Corollary 1.4.6. *Suppose the nonnegative random variable X takes countably many values $x_0, x_1, x_2, \ldots$. Then*

$$\mathbb{E}X = \sum_{k=0}^{\infty} x_k \mathbb{P}\{X = x_k\}. \tag{1.4.1}$$

PROOF: Let $A_k = \{X = x_k\}$, so that X can be written as

$$X = \sum_{k=0}^{\infty} x_k \mathbb{I}_{A_k}.$$

Define $X_n = \sum_{k=0}^{n} x_k \mathbb{I}_{A_k}$. Then $0 \le X_1 \le X_2 \le X_3 \le \ldots$ and $\lim_{n\to\infty} X_n = X$ almost surely ("surely," actually). Theorem 1.3.4(i) implies that

$$\mathbb{E}X_n = \sum_{k=0}^{n} x_k \mathbb{P}\{X = x_k\}.$$

Taking the limit on both sides as $n \to \infty$ and using the Monotone Convergence Theorem to justify the first equality below, we obtain

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n = \lim_{n\to\infty} \sum_{k=0}^{n} x_k \mathbb{P}\{X = x_k\} = \sum_{k=0}^{\infty} x_k \mathbb{P}\{X = x_k\}. \qquad \square$$

Remark 1.4.7. If X can take negative as well as positive values, we can apply Corollary 1.4.6 to X^+ and X^- separately and then subtract the resulting equations to again get formula (1.4.1), provided the subtraction does not create an "$\infty - \infty$" situation.

Example 1.4.8 (St. Petersburg paradox). On the infinite independent coin-toss space Ω_∞ with the probability of a head on each toss equal to $\frac{1}{2}$, define a random variable X by

$$X(\omega) = \begin{cases} 2 & \text{if } \omega_1 = H, \\ 4 & \text{if } \omega_1 = T, \omega_2 = H, \\ 8 & \text{if } \omega_1 = \omega_2 = T, \omega_3 = H, \\ \quad \vdots & \\ 2^k & \text{if } \omega_1 = \omega_2 = \cdots = \omega_{k-1} = T, \omega_k = H. \\ \quad \vdots & \end{cases}$$

This defines $X(\omega)$ for every sequence of coin tosses except the sequence that is all tails. For this sequence, we define $X(TTT\ldots) = \infty$. The probability that $X = \infty$ is then the probability of this sequence, which is zero. Therefore, X is finite almost surely. According to Corollary 1.4.6,

$$\begin{aligned} \mathbb{E}X &= 2 \cdot \mathbb{P}\{X = 2\} + 4 \cdot \mathbb{P}\{X = 4\} + 8 \cdot \mathbb{P}\{X = 8\} + \ldots \\ &= 2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 8 \cdot \frac{1}{8} + \ldots \\ &= 1 + 1 + 1 + \cdots = \infty. \end{aligned}$$

The point is that $\mathbb{E}X$ can be infinite, even though X is finite almost surely. □

The following theorem provides another common situation in which we are assured that the limit of the integrals of a sequence of functions is the integral of the limiting function.

Theorem 1.4.9 (Dominated convergence). *Let $X_1, X_2, \ldots$ be a sequence of random variables converging almost surely to a random variable X. If there is another random variable Y such that $\mathbb{E}Y < \infty$ and $|X_n| \leq Y$ almost surely for every n, then*

$$\lim_{n\to\infty} EX_n = EX.$$

Let $f_1, f_2, \ldots$ be a sequence of Borel-measurable functions on $\mathbb{R}$ converging almost everywhere to a function f. If there is another function g such that $\int_{-\infty}^{\infty} g(x)\,dx < \infty$ and $|f_n| \leq g$ almost everywhere for every n, then

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f_n(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx.$$

1.5 Computation of Expectations

Let X be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We have defined the expectation of X to be the Lebesgue integral

$$\mathbb{E}X = \int_\Omega X(\omega)\, d\mathbb{P}(\omega),$$

the idea being to average the values of $X(\omega)$ over Ω, taking the probabilities into account. This level of abstraction is sometimes helpful. For example, the equality

$$\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$$

follows directly from the linearity of integrals. By contrast, if we were to derive this fact using a joint density for X and Y, it would be a tedious, unenlightening computation.

On the other hand, the abstract space Ω is not a pleasant environment in which to actually compute integrals. For computations, we often need to rely on densities of the random variables under consideration, and we integrate these over the real numbers rather than over Ω. In this section, we develop the relationship between integrals over Ω and integrals over $\mathbb{R}$.

Recall that the distribution measure of X is the probability measure μ_X defined on $\mathbb{R}$ by

$$\mu_X(B) = \mathbb{P}\{X \in B\} \text{ for every Borel subset } B \text{ of } \mathbb{R}. \tag{1.5.1}$$

Because μ_X is a probability measure on $\mathbb{R}$, we can use it to integrate functions over $\mathbb{R}$. We have the following fundamental theorem relating integrals over $\mathbb{R}$ to integrals over Ω.

Theorem 1.5.1. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let g be a Borel-measurable function on R. Then*

$$\mathbb{E}|g(X)| = \int_{\mathbb{R}} |g(x)|\, d\mu_X(x), \tag{1.5.2}$$

and if this quantity is finite, then

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)\, d\mu_X(x). \tag{1.5.3}$$

PROOF: The proof proceeds by several steps, which collectively are called the *standard machine.*

Step 1. Indicator functions. Suppose the function $g(x) = \mathbb{I}_B(x)$ is the indicator of a Borel subset of $\mathbb{R}$. Since this function is nonnegative, (1.5.2) and (1.5.3) reduce to the same equation, namely

$$\mathbb{E}\mathbb{I}_B(X) = \int_{\mathbb{R}} \mathbb{I}_B(x)\, d\mu_X(x). \tag{1.5.4}$$

Since the random variable $\mathbb{I}_B(X)$ takes only the two values one and zero, its expectation is

$$\mathbb{E}\mathbb{I}_B(X) = 1 \cdot \mathbb{P}\{X \in B\} + 0 \cdot \mathbb{P}\{X \notin B\} = \mathbb{P}\{X \in B\}.$$

Similarly, the function $\mathbb{I}_B(x)$ of the dummy (not random!) variable x takes only the two values one and zero, so according to Theorem 1.3.1(i) with $\Omega = \mathbb{R}$, $X = \mathbb{I}_B$, and $\mathbb{P} = \mu_X$, its integral is

$$\int_{\mathbb{R}} \mathbb{I}_B(x)\, d\mu_X(x) = 1 \cdot \mu_X\{x; \mathbb{I}_B(x) = 1\} + 0 \cdot \mu_X\{x; \mathbb{I}_B(x) = 0\} = \mu_X(B).$$

In light of (1.5.1), we have gotten the same result in both cases, and (1.5.4) is proved.

Step 2. Nonnegative simple functions. A simple function is a finite sum of indicator functions times constants. In this step, we assume that

$$g(x) = \sum_{k=1}^{n} \alpha_k \mathbb{I}_{B_k}(x),$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are nonnegative constants and $B_1, B_2, \ldots, B_n$ are Borel subsets of $\mathbb{R}$. Because of linearity of integrals,

$$\mathbb{E}g(X) = \mathbb{E}\sum_{k=1}^{n} \alpha_k \mathbb{I}_{B_k}(X) = \sum_{k=1}^{n} \alpha_k \mathbb{E}\mathbb{I}_{B_k}(X) = \sum_{k=1}^{n} \alpha_k \int_R \mathbb{I}_{B_k}(x)\, d\mu_X(x),$$

where we have used (1.5.4) in the last step. But the linearity of integrals also implies that

$$\sum_{k=1}^{n} \alpha_k \int_{\mathbb{R}} \mathbb{I}_{B_k}(x)\, d\mu_X(x) = \int_{\mathbb{R}} \left(\sum_{k=1}^{n} \alpha_k \mathbb{I}_{B_k}(x) \right) d\mu_X(x) = \int_{\mathbb{R}} g(x)\, d\mu_X(x),$$

and we conclude that

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)\, d\mu_X(x)$$

when g is a nonnegative simple function.

Step 3. Nonnegative Borel-measurable functions. Let $g(x)$ be an arbitrary nonnegative Borel-measurable function defined on $\mathbb{R}$. For each positive integer n, define the sets

$$B_{k,n} = \left\{ x; \frac{k}{2^n} \le g(x) < \frac{k+1}{2^n} \right\}, \quad k = 0, 1, 2, \ldots 4^n - 1.$$

For each fixed n, the sets $B_{0,n}, B_{1,n}, \ldots, B_{4^n-1,n}$ correspond to the partition

$$0 < \frac{1}{2^n} < \frac{2}{2^n} < \cdots < \frac{4^n}{2^n} = 2^n.$$

At the next stage $n+1$, the partition points include all those at stage n and new partition points at the midpoints between the old ones. Because of this fact, the simple functions

$$g_n(x) = \sum_{k=0}^{4^n-1} \frac{k}{2^n} \mathbb{I}_{B_{k,n}}(x)$$

satisfy $0 \le g_1 \le g_2 \le \cdots \le g$. Furthermore, these functions become more and more accurate approximations of g as n becomes larger; indeed, $\lim_{n\to\infty} g_n(x) = g(x)$ for every $x \in \mathbb{R}$. From Step 2, we know that

$$\mathbb{E}g_n(X) = \int_{\mathbb{R}} g_n(x)\, d\mu_X(x)$$

for every n. Letting $n \to \infty$ and using the Monotone Convergence Theorem, Theorem 1.4.5, on both sides of the equation, we obtain

$$\mathbb{E}g(X) = \lim_{n\to\infty} \mathbb{E}g_n(X) = \lim_{n\to\infty} \int_{\mathbb{R}} g_n(x)\, d\mu_X(x) = \int_{\mathbb{R}} g(x)\, d\mu_X(x).$$

This proves (1.5.3) when g is a nonnegative Borel-measurable function.

Step 4. General Borel-measurable function. Let $g(x)$ be a general Borel-measurable function, which can take both positive and negative values. The functions

$$g^+(x) = \max\{g(x), 0\} \text{ and } g^-(x) = \max\{-g(x), 0\}$$

are both nonnegative, and from Step 3 we have

$$\mathbb{E}g^+(X) = \int_{\mathbb{R}} g^+(x)\, d\mu_X(x), \quad \mathbb{E}g^-(X) = \int_{\mathbb{R}} g^-(x)\, d\mu_X(x).$$

Adding these two equations, we obtain (1.5.2). If the quantity in (1.5.2) is finite, then

$$\begin{aligned} \mathbb{E}g^+(X) &= \textstyle\int_{\mathbb{R}} g^+(x)\, d\mu_X(x) < \infty, \\ \mathbb{E}g^-(X) &= \textstyle\int_{\mathbb{R}} g^-(x)\, d\mu_X(x) < \infty, \end{aligned}$$

and we can subtract these two equations because this is not an $\infty - \infty$ situation. The result of this subtraction is (1.5.3). □

Theorem 1.5.1 tells us that in order to compute the Lebesgue integral $\mathbb{E}X = \int_\Omega X(\omega)\, d\mathbb{P}(\omega)$ over the abstract space Ω, it suffices to compute the integral $\int_{\mathbb{R}} g(x)\, d\mu_X(x)$ over the set of real numbers. This is still a Lebesgue integral, and the integrator is the distribution measure μ_X rather than the Lebesgue measure. To actually perform a computation, we need to reduce this to something more familiar. Depending on the nature of the random variable X, the distribution measure μ_X on the right-hand side of (1.5.3) can

have different forms. In the simplest case, X takes only finitely many values $x_0, x_1, x_2, \ldots, x_n$, and then μ_X places a mass of size $p_k = \mathbb{P}\{X = x_k\}$ at each number x_k. In this case, formula (1.5.3) becomes

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)\mu_X(dx) = \sum_{k=0}^{n} g(x_k)p_k.$$

The most common case for continuous-time models in finance is when X has a density. This means that there is a nonnegative, Borel-measurable function f defined on $\mathbb{R}$ such that

$$\mu_X(B) = \int_B f(x)\,dx \text{ for every Borel subset } B \text{ of } \mathbb{R}. \tag{1.5.5}$$

This density allows us to compute the measure μ_X of a set B by computing an integral over B. In most cases, the density function f is bounded and continuous or almost everywhere continuous, so that the integral on the right-hand side of (1.5.5) can be computed as a Riemann integral.

If X has a density, we can use this density to compute expectations, as shown by the following theorem.

Theorem 1.5.2. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let g be a Borel-measurable function on $\mathbb{R}$. Suppose that X has a density f (i.e., f is a function satisfying (1.5.5)). Then*

$$\mathbb{E}|g(X)| = \int_{-\infty}^{\infty} |g(x)|f(x)\,dx. \tag{1.5.6}$$

If this quantity is finite, then

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)\,dx. \tag{1.5.7}$$

PROOF: The proof proceeds again by the standard machine.

Step 1. Indicator functions. If $g(x) = \mathbb{I}_B(x)$, then because g is nonnegative, equations (1.5.6) and (1.5.7) are the same and reduce to

$$\mathbb{E}\mathbb{I}_B(X) = \int_B f(x)\,dx.$$

The left-hand side is $\mathbb{P}\{X \in B\} = \mu_X(B)$, and (1.5.5) shows that the two sides are equal.

Step 2. Simple functions. If $g(x) = \sum_{k=1}^{n} \alpha_k \mathbb{I}_{B_k}(x)$, then

$$\begin{aligned}\mathbb{E}g(X) &= \mathbb{E}\left(\sum_{k=1}^{n}\alpha_k\mathbb{I}_{B_k}(X)\right) = \sum_{k=1}^{n}\alpha_k\mathbb{E}\mathbb{I}_{B_k}(X)\\ &= \sum_{k=1}^{n}\alpha_k\int_{-\infty}^{\infty}\mathbb{I}_{B_k}(x)f(x)\,dx = \int_{-\infty}^{\infty}\sum_{k=1}^{n}\alpha_k\mathbb{I}_{B_k}(x)f(x)\,dx\\ &= \int_{-\infty}^{\infty}g(x)f(x)\,dx.\end{aligned}$$

Step 3. Nonnegative Borel-measurable functions. Just as in the proof of Theorem 1.5.1 we construct a sequence of nonnegative simple functions $0 \le g_1 \le g_2 \le \cdots \le g$ such that $\lim_{n\to\infty} g_n(x) = g(x)$ for every $x \in R$. We have already shown that

$$\mathbb{E}g_n(X) = \int_{-\infty}^{\infty} g_n(x)f(x)\,dx$$

for every n. We let $n \to \infty$, using the Monotone Convergence Theorem, Theorem 1.4.5, on both sides of the equation, to obtain (1.5.7).

Step 4. General Borel-measurable functions. Let g be a general Borel-measurable function, which can take positive and negative values. We have just proved that

$$\mathbb{E}g^+(X) = \int_{-\infty}^{\infty} g^+(x)f(x)\,dx, \quad \mathbb{E}g^-(X) = \int_{-\infty}^{\infty} g^-(x)f(x)\,dx.$$

Adding these equations, we obtain (1.5.6). If the expression in (1.5.6) is finite, we can also subtract these equations to obtain (1.5.7). □

1.6 Change of Measure

We pick up the thread of Section 3.1 of Volume I, in which we used a positive random variable Z to change probability measures on a space Ω. We need to do this when we change from the actual probability measure $\mathbb{P}$ to the risk-neutral probability measure $\widetilde{\mathbb{P}}$ in models of financial markets. When Ω is uncountably infinite and $\mathbb{P}(\omega) = \widetilde{\mathbb{P}}(\omega) = 0$ for every $\omega \in \Omega$, it no longer makes sense to write (3.1.1) of Chapter 3 of Volume I,

$$Z(\omega) = \frac{\widetilde{\mathbb{P}}(\omega)}{\mathbb{P}(\omega)}, \tag{1.6.1}$$

because division by zero is undefined. We could rewrite this equation as

$$Z(\omega)\mathbb{P}(\omega) = \widetilde{\mathbb{P}}(\omega), \tag{1.6.2}$$

and now we have a meaningful equation, with both sides equal to zero, but the equation tells us nothing about the relationship among $\mathbb{P}$, $\widetilde{\mathbb{P}}$, and Z. Because

$\mathbb{P}(\omega) = \widetilde{\mathbb{P}}(\omega) = 0$, the value of $Z(\omega)$ could be anything and (1.6.2) would still hold.

However, (1.6.2) does capture the spirit of what we would like to accomplish. To change from $\mathbb{P}$ to $\widetilde{\mathbb{P}}$, we need to reassign probabilities in Ω using Z to tell us where in Ω we should revise the probability upward (where $Z > 1$) and where we should revise the probability downward (where $Z < 1$). However, we should do this set-by-set, rather than ω-by-ω. The process is described by the following theorem.

Theorem 1.6.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let Z be an almost surely nonnegative random variable with $\mathbb{E}Z = 1$. For $A \in \mathcal{F}$, define*

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega). \tag{1.6.3}$$

Then $\widetilde{\mathbb{P}}$ is a probability measure. Furthermore, if X is a nonnegative random variable, then

$$\widetilde{\mathbb{E}}X = \mathbb{E}[XZ]. \tag{1.6.4}$$

If Z is almost surely strictly positive, we also have

$$\mathbb{E}Y = \widetilde{\mathbb{E}}\left[\frac{Y}{Z}\right] \tag{1.6.5}$$

for every nonnegative random variable Y.

The $\widetilde{\mathbb{E}}$ appearing in (1.6.4) is expectation under the probability measure $\widetilde{\mathbb{P}}$ (i.e., $\widetilde{\mathbb{E}}X = \int_\Omega X(\omega)\, d\widetilde{\mathbb{P}}(\omega)$).

Remark 1.6.2. Suppose X is a random variable that can take both positive and negative values. We may apply (1.6.4) to its positive and negative parts $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$, and then subtract the resulting equations to see that (1.6.4) holds for this X as well, provided the subtraction does not result in an $\infty - \infty$ situation. The same remark applies to (1.6.5).

PROOF OF THEOREM 1.6.1: According to Definition 1.1.2, to check that $\widetilde{\mathbb{P}}$ is a probability measure, we must verify that $\widetilde{\mathbb{P}}(\Omega) = 1$ and that $\widetilde{\mathbb{P}}$ is countably additive. We have by assumption

$$\widetilde{\mathbb{P}}(\Omega) = \int_\Omega Z(\omega)\, d\mathbb{P}(\omega) = \mathbb{E}Z = 1.$$

For countable additivity, let $A_1, A_2, \ldots$ be a sequence of disjoint sets in $\mathcal{F}$, and define $B_n = \cup_{k=1}^n A_k$, $B_\infty = \cup_{k=1}^\infty A_k$. Because

$$\mathbb{I}_{B_1} \le \mathbb{I}_{B_2} \le \mathbb{I}_{B_3} \le \cdots$$

and $\lim_{n\to\infty} \mathbb{I}_{B_n} = \mathbb{I}_{B_\infty}$, we may use the Monotone Convergence Theorem, Theorem 1.4.5, to write

$$\widetilde{\mathbb{P}}(B_\infty) = \int_\Omega \mathbb{I}_{B_\infty}(\omega) Z(\omega)\, d\mathbb{P}(\omega) = \lim_{n\to\infty} \int_\Omega \mathbb{I}_{B_n}(\omega) Z(\omega)\, d\mathbb{P}(\omega).$$

But $\mathbb{I}_{B_n}(\omega) = \sum_{k=1}^n \mathbb{I}_{A_k}(\omega)$, and so

$$\int_\Omega \mathbb{I}_{B_n}(\omega) Z(\omega)\, d\mathbb{P}(\omega) = \sum_{k=1}^n \int_\Omega \mathbb{I}_{A_k}(\omega) Z(\omega)\, d\mathbb{P}(\omega) = \sum_{k=1}^n \widetilde{\mathbb{P}}(A_k).$$

Putting these two equations together, we obtain the countable additivity property

$$\widetilde{\mathbb{P}}\left(\bigcup_{k=1}^\infty A_k\right) = \lim_{n\to\infty} \sum_{k=1}^n \widetilde{\mathbb{P}}(A_k) = \sum_{k=1}^\infty \widetilde{\mathbb{P}}(A_k).$$

Now suppose X is a nonnegative random variable. If X is an indicator function $X = \mathbb{I}_A$, then

$$\widetilde{\mathbb{E}} X = \widetilde{\mathbb{P}}(A) = \int_\Omega \mathbb{I}_A(\omega) Z(\omega)\, d\mathbb{P}(\omega) = \mathbb{E}\left[\mathbb{I}_A Z\right] = \mathbb{E}\left[XZ\right],$$

which is (1.6.4). We finish the proof of (1.6.4) using the standard machine developed in Theorem 1.5.1. When $Z > 0$ almost surely, $\frac{Y}{Z}$ is defined and we may replace X in (1.6.4) by $\frac{Y}{Z}$ to obtain (1.6.5). □

Definition 1.6.3. *Let Ω be a nonempty set and $\mathcal{F}$ a σ-algebra of subsets of Ω. Two probability measures $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$ are said to be* equivalent *if they agree which sets in $\mathcal{F}$ have probability zero.*

Under the assumptions of Theorem 1.6.1, including the assumption that $Z > 0$ almost surely, $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are equivalent. Suppose $A \in \mathcal{F}$ is given and $\mathbb{P}(A) = 0$. Then the random variable $\mathbb{I}_A Z$ is $\mathbb{P}$ almost surely zero, which implies

$$\widetilde{\mathbb{P}}(A) = \int_\Omega \mathbb{I}_A(\omega) Z(\omega)\, d\mathbb{P}(\omega) = 0.$$

On the other hand, suppose $B \in \mathcal{F}$ satisfies $\widetilde{\mathbb{P}}(B) = 0$. Then $\frac{1}{Z}\mathbb{I}_B = 0$ almost surely under $\widetilde{\mathbb{P}}$, so

$$\widetilde{\mathbb{E}}\left[\frac{1}{Z}\mathbb{I}_B\right] = 0.$$

Equation (1.6.5) implies $\mathbb{P}(B) = \mathbb{E}\mathbb{I}_B = 0$. This shows that $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ agree which sets have probability zero. Because the sets with probability one are complements of the sets with probability zero, $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ agree which sets have probability one as well. Because $\widetilde{\mathbb{P}}$ and $\mathbb{P}$ are equivalent, we do not need to specify which measure we have in mind when we say an event occurs *almost surely.*

In financial models, we will first set up a sample space Ω, which one can regard as the set of possible scenarios for the future. We imagine this

set of possible scenarios has an actual probability measure $\mathbb{P}$. However, for purposes of pricing derivative securities, we will use a risk-neutral measure $\widetilde{\mathbb{P}}$. We will insist that these two measures are equivalent. They must agree on what is possible and what is impossible; they may disagree on how probable the possibilities are. This is the same situation we had in the binomial model; $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ assigned different probabilities to the stock price paths, but they agreed which stock price paths were possible. In the continuous-time model, after we have $\mathbb{P}$ and $\widetilde{\mathbb{P}}$, we shall determine prices of derivative securities that allow us to set up hedges that work with $\widetilde{\mathbb{P}}$-probability one. These hedges then also work with $\mathbb{P}$-probability one. Although we have used the risk-neutral probability to compute prices, we will have obtained hedges that work with probability one under the actual (and the risk-neutral) probability measure.

It is common to refer to computations done under the actual measure as computations in the *real world* and computations done under the risk-neutral measure as computations in the *risk-neutral world.* This unfortunate terminology raises the question whether prices computed in the "risk-neutral world" are appropriate for the "real world" in which we live and have our profits and losses. Our answer to this question is that *there is only one world* in the models. There is a single sample space Ω representing all possible future states of the financial markets, and there is a single set of asset prices, modeled by random variables (i.e., functions of these future states of the market). We sometimes work in this world assuming that probabilities are given by an empirically estimated actual probability measure and sometimes assuming that they are given by risk-neutral probabilities, but we do not change our view of the world of possibilities. A hedge that works almost surely under one assumption of probabilities works almost surely under the other assumption as well, since the probability measures agree which events have probability one.

The change of measure discussed in Section 3.1 of Volume I is the special case of Theorem 1.6.1 for finite probability spaces, and Example 3.1.2 of Chapter 3 of Volume I provides a case with explicit formulas for $\mathbb{P}$, $\widetilde{\mathbb{P}}$, and Z when the expectations are sums. We give here two examples on uncountable probability spaces.

Example 1.6.4. Recall Example 1.2.4 in which $\Omega = [0,1]$, $\mathbb{P}$ is the uniform (i.e., Lebesgue) measure, and

$$\widetilde{\mathbb{P}}[a,b] = \int_a^b 2\omega\, d\omega = b^2 - a^2,\ 0 \le a \le b \le 1. \tag{1.2.2}$$

We may use the fact that $\mathbb{P}(d\omega) = d\omega$ to rewrite (1.2.2) as

$$\widetilde{\mathbb{P}}[a,b] = \int_{[a,b]} 2\omega\, d\mathbb{P}(\omega). \tag{1.2.2}'$$

Because $\mathcal{B}[0,1]$ is the σ-algebra generated by the closed intervals (i.e., begin with the closed intervals and put in all other sets necessary in order to have a

σ-algebra), the validity of (1.2.2)′ for all closed intervals $[a,b] \subset [0,1]$ implies its validity for all Borel subsets of $[0,1]$:

$$\widetilde{\mathbb{P}}(B) = \int_B 2\omega \, d\mathbb{P}(\omega) \text{ for every Borel set } B \subset \mathbb{R}.$$

This is (1.6.3) with $Z(\omega) = 2\omega$.

Note that $Z(\omega) = 2\omega$ is strictly positive almost surely ($\mathbb{P}\{0\} = 0$), and

$$\widetilde{\mathbb{E}}Z = \int_0^1 2\omega \, d\omega = 1.$$

According to (1.6.4), for every nonnegative random variable $X(\omega)$, we have the equation

$$\int_0^1 X(\omega) \, d\widetilde{\mathbb{P}}(\omega) = \int_0^1 X(\omega) \cdot 2\omega \, d\omega.$$

This suggests the notation

$$d\widetilde{\mathbb{P}}(\omega) = 2\omega \, d\omega = 2\omega \, d\mathbb{P}(\omega). \tag{1.6.6}$$

□

In general, when $\mathbb{P}$, $\widetilde{\mathbb{P}}$, and Z are related as in Theorem 1.6.1, we may rewrite the two equations (1.6.4) and (1.6.5) as

$$\int_\Omega X(\omega) \, d\widetilde{\mathbb{P}}(\omega) = \int_\Omega X(\omega) Z(\omega) \, d\mathbb{P}(\omega),$$
$$\int_\Omega Y(\omega) \, d\mathbb{P}(\omega) = \int_\Omega \frac{Y(\omega)}{Z(\omega)} \, d\widetilde{\mathbb{P}}(\omega).$$

A good way to remember these equations is to formally write $Z(\omega) = \frac{d\widetilde{\mathbb{P}}(\omega)}{d\mathbb{P}(\omega)}$. Equation (1.6.6) is a special case of this notation that captures the idea behind the nonsensical equation (1.6.1) that Z is somehow a "ratio of probabilities." In Example 1.6.4, $Z(\omega) = 2\omega$ is in fact a ratio of densities, with the denominator being the uniform density 1 for all $\omega \in [0,1]$.

Definition 1.6.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\widetilde{\mathbb{P}}$ be another probability measure on $(\Omega, \mathcal{F})$ that is equivalent to $\mathbb{P}$, and let Z be an almost surely positive random variable that relates $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ via (1.6.3). Then Z is called the* Radon-Nikodým derivative *of $\widetilde{\mathbb{P}}$ with respect to $\mathbb{P}$, and we write*

$$Z = \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}.$$

Example 1.6.6 (Change of measure for a normal random variable). Let X be a standard normal random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Two ways of constructing X and $(\Omega, \mathcal{F}, \mathbb{P})$ were described in Example 1.2.6. For purposes of this example, we do not need to know the details about the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, except we note that the set Ω is necessarily uncountably infinite and $\mathbb{P}(\omega) = 0$ for every $\omega \in \Omega$.

When we say X is a standard normal random variable, we mean that

$$\mu_X(B) = \mathbb{P}\{X \in B\} = \int_B \varphi(x)\,dx \text{ for every Borel subset } B \text{ of } \mathbb{R}, \quad (1.6.7)$$

where

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is the standard normal density. If we take $B = (-\infty, b]$, this reduces to the more familiar condition

$$\mathbb{P}\{X \le b\} = \int_{-\infty}^{b} \varphi(x)dx \text{ for every } b \in \mathbb{R}. \quad (1.6.8)$$

In fact, (1.6.8) is equivalent to the apparently stronger statement (1.6.7). Note that $\mathbb{E}X = 0$ and variance $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = 1$.

Let θ be a constant and define $Y = X + \theta$, so that under $\mathbb{P}$, the random variable Y is normal with $\mathbb{E}Y = \theta$ and variance $\text{Var}(Y) = \mathbb{E}(Y - \mathbb{E}Y)^2 = 1$. Although it is not required by the formulas, we will assume θ is positive for the discussion below. We want to change to a new probability measure $\widetilde{\mathbb{P}}$ on Ω under which Y is a standard normal random variable. In other words, we want $\widetilde{\mathbb{E}}Y = 0$ and $\widetilde{\text{Var}}(Y) = \widetilde{\mathbb{E}}(Y - \widetilde{\mathbb{E}}Y)^2 = 1$. We want to do this not by subtracting θ away from Y, but rather by assigning less probability to those ω for which $Y(\omega)$ is sufficiently positive and more probability to those ω for which $Y(\omega)$ is negative. *We want to change the distribution of Y without changing the random variable Y.* In finance, the change from the actual to the risk-neutral probability measure changes the distribution of asset prices without changing the asset prices themselves, and this example is a step in understanding that procedure.

We first define the random variable

$$Z(\omega) = \exp\left\{-\theta X(\omega) - \frac{1}{2}\theta^2\right\} \text{ for all } \omega \in \Omega.$$

This random variable has two important properties that allow it to serve as a Radon-Nikodým derivative for obtaining a probability measure $\widetilde{\mathbb{P}}$ equivalent to $\mathbb{P}$:

(i) $Z(\omega) > 0$ for all $\omega \in \Omega$ ($Z > 0$ almost surely would be good enough), and
(ii) $\mathbb{E}Z = 1$.

Property (i) is obvious because Z is defined as an exponential. Property (ii) follows from the integration

$$\begin{aligned}
\mathbb{E}Z &= \int_{-\infty}^{\infty} \exp\left\{-\theta x - \frac{1}{2}\theta^2\right\} \varphi(x)dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(x^2 + 2\theta x + \theta^2)\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(x+\theta)^2\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}y^2\right\} dy,
\end{aligned}$$

where we have made the change of dummy variable $y = x + \theta$ in the last step. But $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-\frac{1}{2}y^2\}dy$, being the integral of the standard normal density, is equal to one.

We use the random variable Z to create a new probability measure $\widetilde{\mathbb{P}}$ by adjusting the probabilities of the events in Ω. We do this by defining

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{F}. \tag{1.6.9}$$

The random variable Z has the property that if $X(\omega)$ is positive, then $Z(\omega) < 1$ (we are still thinking of θ as a positive constant). This shows that $\widetilde{\mathbb{P}}$ assigns less probability than $\mathbb{P}$ to sets on which X is positive, a step in the right direction of statistically recentering Y. We claim not only that $\widetilde{\mathbb{E}}Y = 0$ but also that, under $\widetilde{\mathbb{P}}$, Y is a standard normal random variable. To see this, we compute

$$\begin{aligned}
\widetilde{\mathbb{P}}\{Y \le b\} &= \int_{\{\omega; Y(\omega) \le b\}} Z(\omega)\, d\mathbb{P}(\omega) \\
&= \int_\Omega \mathbb{I}_{\{Y(\omega) \le b\}} Z(\omega)\, d\mathbb{P}(\omega) \\
&= \int_\Omega \mathbb{I}_{\{X(\omega) \le b-\theta\}} \exp\left\{-\theta X(\omega) - \frac{1}{2}\theta^2\right\} d\mathbb{P}(\omega).
\end{aligned}$$

At this point, we have managed to write $\widetilde{\mathbb{P}}\{Y \le b\}$ in terms of a function of the random variable X, integrated with respect to the probability measure $\mathbb{P}$ under which X is standard normal. According to Theorem 1.5.2,

$$\begin{aligned}
&\int_\Omega \mathbb{I}_{\{X(\omega) \le b-\theta\}} \exp\left\{-\theta X(\omega) - \frac{1}{2}\theta^2\right\} d\mathbb{P}(\omega) \\
&\qquad = \int_{-\infty}^{\infty} \mathbb{I}_{\{x \le b-\theta\}} e^{-\theta x - \frac{1}{2}\theta^2} \varphi(x)dx \\
&\qquad = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b-\theta} e^{-\theta x - \frac{1}{2}\theta^2} e^{-\frac{x^2}{2}} dx
\end{aligned}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b-\theta} e^{-\frac{1}{2}(x+\theta)^2} dx$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-\frac{1}{2}y^2} dy,$$

where we have made the change of dummy variable $y = x + \theta$ in the last step. We conclude that

$$\widetilde{\mathbb{P}}\{Y \leq b\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-\frac{1}{2}y^2} dy,$$

which shows that Y is a standard normal random variable under the probability measure $\widetilde{\mathbb{P}}$. □

Following Corollary 2.4.6 of Chapter 2 of Volume I, we discussed how the existence of a risk-neutral measure guarantees that a financial model is free of arbitrage, the so-called *First Fundamental Theorem of Asset Pricing*. The same argument applies in continuous-time models and in fact underlies the Heath-Jarrow-Morton no-arbitrage condition for term-structure models. Consequently, we are interested in the existence of risk-neutral measures. As discussed earlier in this section, these must be equivalent to the actual probability measure. How can such probability measures $\widetilde{\mathbb{P}}$ arise? In Theorem 1.6.1, we began with the probability measure $\mathbb{P}$ and an almost surely positive random variable Z and constructed the equivalent probability measure $\widetilde{\mathbb{P}}$. It turns out that this is the only way to obtain a probability measure $\widetilde{\mathbb{P}}$ equivalent to $\mathbb{P}$. The proof of the following profound theorem is beyond the scope of this text.

Theorem 1.6.7 (Radon-Nikodým). *Let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ be equivalent probability measures defined on $(\Omega, \mathcal{F})$. Then there exists an almost surely positive random variable Z such that $\mathbb{E}Z = 1$ and*

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega) \text{ for every } A \in \mathcal{F}.$$

1.7 Summary

Probability theory begins with a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ (Definition 1.1.2). Here Ω is the set of all possible outcomes of a random experiment, $\mathcal{F}$ is the collection of subsets of Ω whose probability is defined, and $\mathbb{P}$ is a function mapping $\mathcal{F}$ to $[0, 1]$. The two axioms of probability spaces are $\mathbb{P}(\Omega) = 1$ and *countable additivity*: the probability of a union of disjoint sets is the sum of the probabilities of the individual sets.

The collection of sets $\mathcal{F}$ in the preceding paragraph is a *σ-algebra*, which means that $\emptyset$ belongs to $\mathcal{F}$, the complement of every set in $\mathcal{F}$ is also in $\mathcal{F}$, and the union of any sequence of sets in $\mathcal{F}$ is also in $\mathcal{F}$. The Borel σ-algebra in $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra that contains all the closed interval

$[a,b]$ in $\mathbb{R}$. Every set encountered in practice is a Borel set (i.e., belongs to $\mathcal{B}(\mathbb{R})$).

A *random variable* X is a mapping from Ω to $\mathbb{R}$ (Definition 1.2.1). By definition, it has the property that, for every $B \in \mathcal{B}(\mathbb{R})$, the set $\{\omega \in \Omega; X(\omega) \in B\}$ is in the σ-algebra $\mathcal{F}$. A random variable X together with the probability measure $\mathbb{P}$ on Ω determines a *distribution* on $\mathbb{R}$. This distribution is not the random variable. Different random variables can have the same distribution, and the same random variable can have different distributions. We describe the distribution as a measure μ_X that assigns to each Borel subset B of $\mathbb{R}$ the mass $\mu_X(B) = \mathbb{P}\{X \in B\}$ (Definition 1.2.3). If X has a density $f(x)$, then $\mu_X(B) = \int_B f(x)\,dx$. If X is a discrete random variable, which means that it takes one of countably many values $x_1, x_2, \ldots$, then we define $p_i = \mathbb{P}\{X = x_i\}$ and have $\mu_X(B) = \sum_{\{i;x_i \in B\}} p_i$.

The *expectation* of a random variable X is $\mathbb{E}X = \int_\Omega X(\omega)\,d\mathbb{P}(\omega)$, where the right-hand side is a Lebesgue integral over Ω. Lebesgue integrals are discussed in Section 1.3. They differ from Riemann integrals, which form approximating sums to the integral by partitioning the "x" (horizontal)-axis, because Lebesgue integrals form approximating sums to the integral by partitioning the "y" (vertical)-axis. Lebesgue integrals have the properties one would expect (Theorem 1.3.4):
Comparison. If $X \le Y$ almost surely, then $\mathbb{E}X \le \mathbb{E}Y$;
Linearity. $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}X + \beta \mathbb{E}Y$.
In addition, if φ is a convex function, we have *Jensen's inequality*: $\varphi(\mathbb{E}X) \le \mathbb{E}\varphi(X)$.

If the random variable X has a density $f(x)$, then $\mathbb{E}X = \int_{-\infty}^{\infty} x f(x)\,dx$ and, more generally, $\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f(x)\,dx$ (Theorem 1.5.2). If the random variable is discrete with $p_i = \mathbb{P}\{X = x_i\}$, then $\mathbb{E}g(X) = \sum_i g(x_i) p_i$.

Suppose we have a sequence of random variables $X_1, X_2, X_3, \ldots$ converging almost surely to a random variable X. It is not always true that

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n. \tag{1.7.1}$$

However, if $0 \le X_1 \le X_2 \le X_3 \le \ldots$ almost surely, then (1.7.1) holds (Monotone Convergence Theorem, Theorem 1.4.5). Alternatively, if there exists a random variable Y such that $\mathbb{E}Y < \infty$ and $|X_n| \le Y$ almost surely for every n, then again (1.7.1) holds (Dominated Convergence Theorem, Theorem 1.4.9).

We may start with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and change to a different measure $\widetilde{\mathbb{P}}$. Our motivation for considering two measures is that in finance there is both an actual probability measure and a risk-neutral probability measure. If $\mathbb{P}$ is a probability measure and Z is a nonnegative random variable satisfying $\mathbb{E}Z = 1$, then $\widetilde{\mathbb{P}}$ defined by

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\,d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{F}$$

is also a probability measure (Theorem 1.6.1). If Z is strictly positive almost surely, the two measures are *equivalent*: they agree about which sets have probability zero. For a random variable X, we have the change-of-expectation formula $\widetilde{\mathbb{E}}[X] = \mathbb{E}[XZ]$. If Z is strictly positive almost surely, there is a change-of-expectation formula in the other direction. Namely, if Y is a random variable, then $\mathbb{E}Y = \widetilde{\mathbb{E}}\left[\frac{Y}{Z}\right]$.

1.8 Notes

Probability theory is usually learned in two stages. In the first stage, one learns that a discrete random variable has a probability mass function and a continuous random variable has a density. These can be used to compute expectations and variances, and even conditional expectations, which are discussed in Chapter 2. Furthermore, one learns how transformations of continuous random variables change densities. A well-written book that contains all these things is DeGroot [48].

The second stage of probability theory, which is treated in this chapter, is measure-theoretic. In this stage, one views a random variable as a function from a sample space Ω to the set of real numbers $\mathbb{R}$. Certain subsets of Ω are called events, and the collection of all events forms a σ-algebra $\mathcal{F}$. Each set A in $\mathcal{F}$ has a probability $\mathbb{P}(A)$. This point of view handles both discrete and continuous random variables within the same unifying framework. It is necessary to adopt this point of view in order to understand the change from the actual to the risk-neutral measure in finance.

The measure-theoretic view of probability theory was begun by Kolmogorov [104]. A comprehensive book on measure-theoretic probability is Billingsley [10]. A succinct book on measure-theoretic probability and martingales is Williams [161]. A more detailed book is Chung [35]. All of these are at the level of a Ph.D. course in mathematics.

1.9 Exercises

Exercise 1.1. Using the properties of Definition 1.1.2 for a probability measure $\mathbb{P}$, show the following.

(i) If $A \in \mathcal{F}$, $B \in \mathcal{F}$, and $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
(ii) If $A \in \mathcal{F}$ and $\{A_n\}_{n=1}^{\infty}$ is a sequence of sets in $\mathcal{F}$ with $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$ and $A \subset A_n$ for every n, then $\mathbb{P}(A) = 0$. (This property was used implicitly in Example 1.1.4 when we argued that the sequence of all heads, and indeed any particular sequence, must have probability zero.)

Exercise 1.2. The infinite coin-toss space Ω_∞ of Example 1.1.4 is *uncountably infinite*. In other words, we cannot list all its elements in a sequence.

To see that this is impossible, suppose there were such a sequential list of all elements of Ω_∞:

$$\begin{aligned}\omega^{(1)} &= \omega_1^{(1)}\omega_2^{(1)}\omega_3^{(1)}\omega_4^{(1)}\ldots,\\ \omega^{(2)} &= \omega_1^{(2)}\omega_2^{(2)}\omega_3^{(2)}\omega_4^{(2)}\ldots,\\ \omega^{(3)} &= \omega_1^{(3)}\omega_2^{(3)}\omega_3^{(3)}\omega_4^{(3)}\ldots,\\ &\vdots\end{aligned}$$

An element that does not appear in this list is the sequence whose first component is H if $\omega_1^{(1)}$ is T and is T if $\omega_1^{(1)}$ is H, whose second component is H if $\omega_2^{(2)}$ is T and is T if $\omega_2^{(2)}$ is H, whose third component is H if $\omega_3^{(3)}$ is T and is T if $\omega_3^{(3)}$ is H, etc. Thus, the list does not include every element of Ω_∞.

Now consider the set of sequences of coin tosses in which the outcome on each even-numbered toss matches the outcome of the toss preceding it, i.e.,

$$A = \{\omega = \omega_1\omega_2\omega_3\omega_4\omega_5\ldots; \omega_1 = \omega_2, \omega_3 = \omega_4, \ldots\}.$$

(i) Show that A is uncountably infinite.
(ii) Show that, when $0 < p < 1$, we have $\mathbb{P}(A) = 0$.

Uncountably infinite sets can have any probability between zero and one, including zero and one. The uncountability of the set does not help determine its probability.

Exercise 1.3. Consider the set function $\mathbb{P}$ defined for every subset of $[0,1]$ by the formula that $\mathbb{P}(A) = 0$ if A is a finite set and $\mathbb{P}(A) = \infty$ if A is an infinite set. Show that $\mathbb{P}$ satisfies (1.1.3)–(1.1.5), but $\mathbb{P}$ does not have the countable additivity property (1.1.2). We see then that the finite additivity property (1.1.5) does not imply the countable additivity property (1.1.2).

Exercise 1.4. (i) Construct a standard normal random variable Z on the probability space $(\Omega_\infty, \mathcal{F}_\infty, \mathbb{P})$ of Example 1.1.4 under the assumption that the probability for head is $p = \frac{1}{2}$. (Hint: Consider Examples 1.2.5 and 1.2.6.)

(ii) Define a sequence of random variables $\{Z_n\}_{n=1}^\infty$ on Ω_∞ such that

$$\lim_{n\to\infty} Z_n(\omega) = Z(\omega) \text{ for every } \omega \in \Omega_\infty$$

and, for each n, Z_n depends only on the first n coin tosses. (This gives us a procedure for approximating a standard normal random variable by random variables generated by a finite number of coin tosses, a useful algorithm for Monte Carlo simulation.)

Exercise 1.5. When dealing with double Lebesgue integrals, just as with double Riemann integrals, the order of integration can be reversed. The only

assumption required is that the function being integrated be either nonnegative or integrable. Here is an application of this fact.

Let X be a nonnegative random variable with cumulative distribution function $F(x) = \mathbb{P}\{X \leq x\}$. Show that

$$\mathbb{E}X = \int_0^\infty (1 - F(x))\, dx$$

by showing that

$$\int_\Omega \int_0^\infty \mathbb{I}_{[0,X(\omega))}(x)\, dx\, d\mathbb{P}(\omega)$$

is equal to both $\mathbb{E}X$ and $\int_0^\infty (1 - F(x))\, dx$.

Exercise 1.6. Let u be a fixed number in $\mathbb{R}$, and define the convex function $\varphi(x) = e^{ux}$ for all $x \in \mathbb{R}$. Let X be a normal random variable with mean $\mu = \mathbb{E}X$ and standard deviation $\sigma = \left[\mathbb{E}(X - \mu)^2\right]^{\frac{1}{2}}$, i.e., with density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(i) Verify that

$$\mathbb{E}e^{uX} = e^{u\mu + \frac{1}{2}u^2\sigma^2}.$$

(ii) Verify that Jensen's inequality holds (as it must):

$$\mathbb{E}\varphi(X) \geq \varphi(\mathbb{E}X).$$

Exercise 1.7. For each positive integer n, define f_n to be the normal density with mean zero and variance n, i.e.,

$$f_n(x) = \frac{1}{\sqrt{2n\pi}}\, e^{-\frac{x^2}{2n}}.$$

(i) What is the function $f(x) = \lim_{n\to\infty} f_n(x)$?
(ii) What is $\lim_{n\to\infty} \int_{-\infty}^\infty f_n(x)\, dx$?
(iii) Note that

$$\lim_{n\to\infty} \int_{-\infty}^\infty f_n(x)\, dx \neq \int_{-\infty}^\infty f(x)\, dx.$$

Explain why this does not violate the Monotone Convergence Theorem, Theorem 1.4.5.

Exercise 1.8 (Moment-generating function). Let X be a nonnegative random variable, and assume that

$$\varphi(t) = \mathbb{E}e^{tX}$$

is finite for every $t \in \mathbb{R}$. Assume further that $\mathbb{E}\left[Xe^{tX}\right] < \infty$ for every $t \in \mathbb{R}$. The purpose of this exercise is to show that $\varphi'(t) = \mathbb{E}\left[Xe^{tX}\right]$ and, in particular, $\varphi'(0) = \mathbb{E}X$.

We recall the definition of derivative:

$$\varphi'(t) = \lim_{s \to t} \frac{\varphi(t) - \varphi(s)}{t - s} = \lim_{s \to t} \frac{\mathbb{E}e^{tX} - \mathbb{E}e^{sX}}{t - s} = \lim_{s \to t} \mathbb{E}\left[\frac{e^{tX} - e^{sX}}{t - s}\right].$$

The limit above is taken over a *continuous* variable s, but we can choose a sequence of numbers $\{s_n\}_{n=1}^{\infty}$ converging to t and compute

$$\lim_{s_n \to t} \mathbb{E}\left[\frac{e^{tX} - e^{s_n X}}{t - s_n}\right],$$

where now we are taking a limit of the expectations of the *sequence* of random variables

$$Y_n = \frac{e^{tX} - e^{s_n X}}{t - s_n}.$$

If this limit turns out to be the same, regardless of how we choose the sequence $\{s_n\}_{n=1}^{\infty}$ that converges to t, then this limit is also the same as $\lim_{s \to t} \mathbb{E}\left[\frac{e^{tX} - e^{sX}}{t - s}\right]$ and is $\varphi'(t)$.

The Mean Value Theorem from calculus states that if $f(t)$ is a differentiable function, then for any two numbers s and t, there is a number θ between s and t such that

$$f(t) - f(s) = f'(\theta)(t - s).$$

If we fix $\omega \in \Omega$ and define $f(t) = e^{tX(\omega)}$, then this becomes

$$e^{tX(\omega)} - e^{sX(\omega)} = (t - s)X(\omega)e^{\theta(\omega)X(\omega)}, \tag{1.9.1}$$

where $\theta(\omega)$ is a number depending on ω (i.e., a random variable lying between t and s).

(i) Use the Dominated Convergence Theorem (Theorem 1.4.9) and equation (1.9.1) to show that

$$\lim_{n \to \infty} \mathbb{E}Y_n = \mathbb{E}\left[\lim_{n \to \infty} Y_n\right] = \mathbb{E}\left[Xe^{tX}\right]. \tag{1.9.2}$$

This establishes the desired formula $\varphi'(t) = \mathbb{E}\left[Xe^{tX}\right]$.

(ii) Suppose the random variable X can take both positive and negative values and $\mathbb{E}e^{tX} < \infty$ and $E\left[|X|e^{tX}\right] < \infty$ for every $t \in \mathbb{R}$. Show that once again $\varphi'(t) = E\left[Xe^{tX}\right]$. (Hint: Use the notation (1.3.1) to write $X = X^+ - X^-$.)

Exercise 1.9. Suppose X is a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, A is a set in $\mathcal{F}$, and for every Borel subset B of $\mathbb{R}$, we have

$$\int_A \mathbb{I}_B(X(\omega))\, d\mathbb{P}(\omega) = \mathbb{P}(A) \cdot \mathbb{P}\{X \in B\}. \tag{1.9.3}$$

Then we say that X is *independent* of the event A.

Show that if X is independent of an event A, then

$$\int_A g(X(\omega))\, d\mathbb{P}(\omega) = \mathbb{P}(A) \cdot \mathbb{E}g(X)$$

for every nonnegative, Borel-measurable function g.

Exercise 1.10. Let $\mathbb{P}$ be the uniform (Lebesgue) measure on $\Omega = [0,1]$. Define

$$Z(\omega) = \begin{cases} 0 \text{ if } 0 \le \omega < \frac{1}{2}, \\ 2 \text{ if } \frac{1}{2} \le \omega \le 1. \end{cases}$$

For $A \in \mathcal{B}[0,1]$, define

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega).$$

(i) Show that $\widetilde{\mathbb{P}}$ is a probability measure.

(ii) Show that if $\mathbb{P}(A) = 0$, then $\widetilde{\mathbb{P}}(A) = 0$. We say that $\widetilde{\mathbb{P}}$ is *absolutely continuous* with respect to $\mathbb{P}$.

(iii) Show that there is a set A for which $\widetilde{\mathbb{P}}(A) = 0$ but $\mathbb{P}(A) > 0$. In other words, $\widetilde{\mathbb{P}}$ and $\mathbb{P}$ are not equivalent.

Exercise 1.11. In Example 1.6.6, we began with a standard normal random variable X under a measure $\mathbb{P}$. According to Exercise 1.6, this random variable has the moment-generating function

$$\mathbb{E}e^{uX} = e^{\frac{1}{2}u^2} \text{ for all } u \in \mathbb{R}.$$

The moment-generating function of a random variable determines its distribution. In particular, any random variable that has moment-generating function $e^{\frac{1}{2}u^2}$ must be standard normal.

In Example 1.6.6, we also defined $Y = X + \theta$, where θ is a constant, we set $Z = e^{-\theta X - \frac{1}{2}\theta^2}$, and we defined $\widetilde{\mathbb{P}}$ by the formula (1.6.9):

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{F}.$$

We showed by considering its cumulative distribution function that Y is a standard normal random variable under $\widetilde{\mathbb{P}}$. Give another proof that Y is standard normal under $\widetilde{\mathbb{P}}$ by verifying the moment-generating function formula

$$\widetilde{\mathbb{E}}e^{uY} = e^{\frac{1}{2}u^2} \text{ for all } u \in \mathbb{R}.$$

Exercise 1.12. In Example 1.6.6, we began with a standard normal random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and defined the random variable $Y = X + \theta$, where θ is a constant. We also defined $Z = e^{-\theta X - \frac{1}{2}\theta^2}$ and used Z as the Radon-Nikodým derivative to construct the probability measure $\widetilde{\mathbb{P}}$ by the formula (1.6.9):

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{F}.$$

Under $\widetilde{\mathbb{P}}$, the random variable Y was shown to be standard normal.

We now have a standard normal random variable Y on the probability space $(\Omega, \mathcal{F}, \widetilde{\mathbb{P}})$, and X is related to Y by $X = Y - \theta$. By what we have just stated, with X replaced by Y and θ replaced by $-\theta$, we could define $\widehat{Z} = e^{\theta Y - \frac{1}{2}\theta^2}$ and then use $\widehat{Z}$ as a Radon-Nikodým derivative to construct a probability measure $\widehat{\mathbb{P}}$ by the formula

$$\widehat{\mathbb{P}}(A) = \int_A \widehat{Z}(\omega)\, d\widetilde{\mathbb{P}}(\omega) \text{ for all } A \in \mathcal{F},$$

so that, under $\widehat{\mathbb{P}}$, the random variable X is standard normal. Show that $\widehat{Z} = \frac{1}{Z}$ and $\widehat{\mathbb{P}} = \mathbb{P}$.

Exercise 1.13 (Change of measure for a normal random variable). A nonrigorous but informative derivation of the formula for the Radon-Nikodým derivative $Z(\omega)$ in Example 1.6.6 is provided by this exercise. As in that example, let X be a standard normal random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $Y = X + \theta$. Our goal is to define a strictly positive random variable $Z(\omega)$ so that when we set

$$\widetilde{\mathbb{P}}(A) = \int_A Z(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{F}, \tag{1.9.4}$$

the random variable Y under $\widetilde{\mathbb{P}}$ is standard normal. If we fix $\overline{\omega} \in \Omega$ and choose a set A that contains $\overline{\omega}$ and is "small," then (1.9.4) gives

$$\widetilde{\mathbb{P}}(A) \approx Z(\overline{\omega})\mathbb{P}(A),$$

where the symbol $\approx$ means "is approximately equal to." Dividing by $\mathbb{P}(A)$, we see that

$$\frac{\widetilde{\mathbb{P}}(A)}{\mathbb{P}(A)} \approx Z(\overline{\omega})$$

for "small" sets A containing $\overline{\omega}$. We use this observation to identify $Z(\overline{\omega})$.

With $\overline{\omega}$ fixed, let $x = X(\overline{\omega})$. For $\epsilon > 0$, we define $B(x, \epsilon) = \left[x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}\right]$ to be the closed interval centered at x and having length ϵ. Let $y = x + \theta$ and $B(y, \epsilon) = \left[y - \frac{\epsilon}{2}, y + \frac{\epsilon}{2}\right]$.

(i) Show that

$$\frac{1}{\epsilon}\mathbb{P}\{X \in B(x, \epsilon)\} \approx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{X^2(\overline{\omega})}{2}\right\}.$$

(ii) In order for Y to be a standard normal random variable under $\widetilde{\mathbb{P}}$, show that we must have

$$\frac{1}{\epsilon}\widetilde{\mathbb{P}}\{Y \in B(y, \epsilon)\} \approx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{Y^2(\overline{\omega})}{2}\right\}.$$

(iii) Show that $\{X \in B(x,\epsilon)\}$ and $\{Y \in B(y,\epsilon)\}$ are the same set, which we call $A(\overline{\omega},\epsilon)$. This set contains $\overline{\omega}$ and is "small" when $\epsilon > 0$ is small.

(iv) Show that

$$\frac{\widetilde{\mathbb{P}}(A)}{\mathbb{P}(A)} \approx \exp\left\{-\theta X(\overline{\omega}) - \frac{1}{2}\theta^2\right\}.$$

The right-hand side is the value we obtained for $Z(\overline{\omega})$ in Example 1.6.6.

Exercise 1.14 (Change of measure for an exponential random variable). Let X be a nonnegative random variable defined on a probability space $(\Omega, \mathcal{F}, P)$ with the *exponential distribution*, which is

$$\mathbb{P}\{X \le a\} = 1 - e^{-\lambda a}, \ a \ge 0,$$

where λ is a positive constant. Let $\tilde{\lambda}$ be another positive constant, and define

$$Z = \frac{\tilde{\lambda}}{\lambda} e^{-(\tilde{\lambda}-\lambda)X}.$$

Define $\widetilde{\mathbb{P}}$ by

$$\widetilde{\mathbb{P}}(A) = \int_A Z \, d\mathbb{P} \quad \text{for all } A \in \mathcal{F}.$$

(i) Show that $\widetilde{\mathbb{P}}(\Omega) = 1$.

(ii) Compute the cumulative distribution function

$$\widetilde{\mathbb{P}}\{X \le a\} \text{ for } a \ge 0$$

for the random variable X under the probability measure $\widetilde{\mathbb{P}}$.

Exercise 1.15 (Provided by Alexander Ng). Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and assume X has a density function $f(x)$ that is positive for every $x \in \mathbb{R}$. Let g be a strictly increasing, differentiable function satisfying

$$\lim_{y \to -\infty} g(y) = -\infty, \quad \lim_{y \to \infty} g(y) = \infty,$$

and define the random variable $Y = g(X)$.

Let $h(y)$ be an arbitrary nonnegative function satisfying $\int_{-\infty}^{\infty} h(y)\, dy = 1$. We want to change the probability measure so that $h(y)$ is the density function for the random variable Y. To do this, we define

$$Z = \frac{h\big(g(X)\big)g'(X)}{f(X)}.$$

(i) Show that Z is nonnegative and $\mathbb{E}Z = 1$.

Now define $\widetilde{\mathbb{P}}$ by

$$\widetilde{\mathbb{P}}(A) = \int_A Z \, d\mathbb{P} \quad \text{for all } A \in \mathcal{F}.$$

(ii) Show that Y has density h under $\widetilde{\mathbb{P}}$.

This page intentionally left blank

2

Information and Conditioning

2.1 Information and σ-algebras

The no-arbitrage theory of derivative security pricing is based on contingency plans. In order to price a derivative security, we determine the initial wealth we would need to set up a hedge of a short position in the derivative security. The hedge must specify what position we will take in the underlying security at each future time contingent on how the uncertainty between the present time and that future time is resolved. In order to make these contingency plans, we need a way to mathematically model the information on which our future decisions can be based. In the binomial model, that information was knowledge of the coin tosses between the initial time and the future time. For the continuous-time model, we need to develop somewhat more sophisticated machinery to capture this concept of information.

We imagine as always that some random experiment is performed, and the outcome is a particular ω in the set of all possible outcomes Ω. We might then be given some information—not enough to know the precise value of ω but enough to narrow down the possibilities. For example, the true ω might be the result of three coin tosses, and we are told only the first one. Or perhaps we are told the stock price at time two without being told any of the coin tosses. In such a situation, although we do not know the true ω precisely, we can make a list of sets that are sure to contain it and other sets that are sure not to contain it. These are the sets that are *resolved by the information.*

Indeed, suppose Ω is the set of eight possible outcomes of three coin tosses. If we are told the outcome of the first coin toss only, the sets

$$A_H = \{HHH, HHT, HTH, HTT\}, \quad A_T = \{THH, THT, TTH, TTT\} \tag{2.1.1}$$

are resolved. For each of these sets, once we are told the first coin toss, we know if the true ω is a member. The empty set $\emptyset$ and the whole space Ω are always resolved, even without any information; the true ω does not belong to $\emptyset$ and does belong to Ω. The four sets that are resolved by the first coin toss

form the σ-algebra

$$\mathcal{F}_1 = \{\emptyset, \Omega, A_H, A_T\}.$$

We shall think of this σ-algebra as containing the information learned by observing the first coin toss. More precisely, if instead of being told the first coin toss, we are told, for each set in $\mathcal{F}_1$, whether or not the true ω belongs to the set, we know the outcome of the first coin toss and nothing more.

If we are told the first two coin tosses, we obtain a finer resolution. In particular, the four sets

$$\begin{array}{ll} A_{HH} = \{HHH, HHT\}, & A_{HT} = \{HTH, HTT\}, \\ A_{TH} = \{THH, THT\}, & A_{TT} = \{TTH, TTT\}, \end{array} \tag{2.1.2}$$

are resolved. Of course, the sets in $\mathcal{F}_1$ are still resolved. Whenever a set is resolved, so is its complement, which means that A_{HH}^c, A_{HT}^c, A_{TH}^c, and A_{TT}^c are resolved. Whenever two sets are resolved, so is their union, which means that $A_{HH} \cup A_{TH}$, $A_{HH} \cup A_{TT}$, $A_{HT} \cup A_{TH}$, and $A_{HT} \cup A_{TT}$ are resolved. We have already noted that the two other pairwise unions, $A_H = A_{HH} \cup A_{HT}$ and $A_T = A_{TH} \cup A_{TT}$, are resolved. The triple unions are also resolved, and these are the complements already mentioned, e.g.,

$$A_{HH} \cup A_{HT} \cup A_{TH} = A_{TT}^c.$$

In all, we have 16 resolved sets that together form a σ-algebra we call $\mathcal{F}_2$; i.e.,

$$\mathcal{F}_2 = \left\{ \begin{array}{l} \emptyset, \Omega, A_H, A_T, A_{HH}, A_{HT}, A_{TH}, A_{TT}, A_{HH}^c, A_{HT}^c, A_{TH}^c, A_{TT}^c, \\ A_{HH} \cup A_{TH},\ A_{HH} \cup A_{TT},\ A_{HT} \cup A_{TH},\ A_{HT} \cup A_{TT} \end{array} \right\}. \tag{2.1.3}$$

We shall think of this σ-algebra as containing the information learned by observing the first two coin tosses.

If we are told all three coin tosses, we know the true ω and every subset of Ω is resolved. There are 256 subsets of Ω and, taken all together, they constitute the σ-algebra $\mathcal{F}_3$:

$$\mathcal{F}_3 = \text{ The set of all subsets of } \Omega.$$

If we are told nothing about the coin tosses, the only resolved sets are $\emptyset$ and Ω. We form the so-called *trivial σ-field* $\mathcal{F}_0$ with these two sets:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}.$$

We have then four σ-algebras, $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$, indexed by time. As time moves forward, we obtain finer resolution. In other words, if $n < m$, then $\mathcal{F}_m$ contains every set in $\mathcal{F}_n$ and even more. This means that $\mathcal{F}_m$ contains more information than $\mathcal{F}_n$. The collection of σ-algebras $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ is an example of a *filtration*. We give the continuous-time formulation of this situation in the following definition.

Definition 2.1.1. *Let Ω be a nonempty set. Let T be a fixed positive number, and assume that for each $t \in [0,T]$ there is a σ-algebra $\mathcal{F}(t)$. Assume further that if $s \le t$, then every set in $\mathcal{F}(s)$ is also in $\mathcal{F}(t)$. Then we call the collection of σ-algebras $\mathcal{F}(t)$, $0 \le t \le T$, a* filtration.

A filtration tells us the information we will have at future times. More precisely, when we get to time t, we will know for each set in $\mathcal{F}(t)$ whether the true ω lies in that set.

Example 2.1.2. Suppose our sample space is $\Omega = C_0[0,T]$, the set of continuous functions defined on $[0,T]$ taking the value zero at time zero. Suppose one of these functions $\overline{\omega}$ is chosen at random and we get to observe it up to time t, where $0 \le t \le T$. That is to say, we know the value of $\overline{\omega}(s)$ for $0 \le s \le t$, but we do not know the value of $\overline{\omega}(s)$ for $t < s \le T$. Certain subsets of Ω are resolved. For example, the set $\{\omega \in \Omega; \max_{0 \le s \le t} \omega(s) \le 1\}$ is resolved. We would put this in the σ-algebra $\mathcal{F}(t)$. Other subsets of Ω are not resolved by time t. For example, if $t < T$, the set $\{\omega \in \Omega; \omega(T) > 0\}$ is not resolved by time t. Indeed, the sets that are resolved by time t are just those sets that can be described in terms of the path of ω up to time t.[1] Every reasonable[2] subset of $\Omega = C_0[0,T]$ is resolved by time T. By contrast, at time zero we see only the value of $\overline{\omega}(0)$, which is equal to zero by the definition of Ω. We learn nothing about the outcome of the random experiment of choosing $\overline{\omega}$ by observing this. The only sets resolved at time zero are $\emptyset$ and Ω, and consequently $\mathcal{F}(0) = \{\emptyset, \Omega\}$. □

Example 2.1.2 provides the simplest setting in which we may construct a Brownian motion. It remains only to assign probability to the sets in $\mathcal{F} = \mathcal{F}(T)$, and then the paths $\omega \in C_0[0,T]$ will be the paths of the Brownian motion.

The discussion preceding Definition 2.1.1 suggests that the σ-algebras in a filtration can be built by taking unions and complements of certain fundamental sets in the way $\mathcal{F}_2$ was constructed from the four sets A_{HH}, A_{HT}, A_{TH}, and A_{TT}. If this were the case, it would be enough to work with these so-called *atoms* (indivisible sets in the σ-algebra) and not consider all the other sets. In uncountable sample spaces, however, there are sets that cannot be constructed as countable unions of atoms (and uncountable unions are forbidden because we cannot add up probabilities of such unions). For example, let us fix $t \in (0,T)$ in Example 2.1.2. Now choose a continuous function $f(u)$, defined only for $0 \le u \le t$ and satisfying $f(0) = 0$. The set of continuous functions $\omega \in C_0[0,T]$ that agree with f on $[0,t]$ and that are free to take any values on $(t,T]$ form an atom in $\mathcal{F}_t$. In symbols, this atom is

[1] For technical reasons, we would not include in $\mathcal{F}(t)$ sets such as $\{\omega \in \Omega; \max_{0 \le s \le t} \omega(s) \in B\}$ if B is a subset of $\mathbb{R}$ that is not Borel measurable. This technical issue can safely be ignored.

[2] Once again, there are pathological sets such as $\{\omega \in \Omega; \omega(T) \in B\}$, where B is a subset of $\mathbb{R}$ that is not Borel measurable. These are not included in $\mathcal{F}(T)$, but that shall not concern us.

$$\{\omega \in C_0[0,T]; \omega(u) = f(u) \text{ for all } u \in [0,t]\}.$$

Each time we choose a new function $f(u)$, defined for $0 \leq u \leq t$, we get a new atom. However, there is no way to obtain the important set $\{\omega \in \Omega; \omega(t) > 0\}$ by taking countable unions of these atoms. Moreover, it is usually the case that the atoms have zero probability. Consequently, in what follows we work with all the sets of $\mathcal{F}(t)$, especially those with positive probability, not with just the atoms.

Besides observing the evolution of an economy over time, which is the idea behind Example 2.1.2, there is a second way we might acquire information about the value of ω. Let X be a random variable. We assume throughout that there is a "formula" for X, and we know this formula even before the random experiment is performed. Because we already know this formula, we are waiting only to learn the value of ω to substitute into the formula so we can evaluate $X(\omega)$. But suppose that rather than being told the value of ω we are told only the value of $X(\omega)$. This resolves certain sets. For example, if we know the value of $X(\omega)$, then we know if ω is in the set $\{X \leq 1\}$ (yes if $X(\omega) \leq 1$ and no if $X(\omega) > 1$). Indeed, every set of the form $\{X \in B\}$, where B is a subset of $\mathbb{R}$, is resolved. Again, for technical reasons, we restrict attention to subsets B that are Borel measurable.

Definition 2.1.3. *Let X be a random variable defined on a nonempty sample space Ω. The* σ-algebra generated by X, *denoted $\sigma(X)$, is the collection of all subsets of Ω of the form $\{X \in B\}$,*[3] *where B ranges over the Borel subsets of* $\mathbb{R}$.

Example 2.1.4. We return to the three-period model of Example 1.2.1 of Chapter 1. In that model, Ω is the set of eight possible outcomes of three coin tosses, and

$$\begin{aligned} &S_2(HHH) = S_2(HHT) = 16, \\ &S_2(HTH) = S_2(HTT) = S_2(THH) = S_2(THT) = 4, \\ &S_2(TTH) = S_2(TTT) = 1. \end{aligned}$$

In Example 1.2.2 of Chapter 1, we wrote S_2 as a function of the first two coin tosses alone, but now we include the irrelevant third toss in the argument to get the full picture. If we take B to be the set containing the single number 16, then $\{S_2 \in B\} = \{HHH, HHT\} = A_{HH}$, where we are using the notation of (2.1.2). It follows that A_{HH} belongs to the σ-algebra $\sigma(S_2)$. Similarly, we can take B to contain the single number 4 and conclude that $A_{HT} \cup A_{TH}$ belongs to $\sigma(S_2)$, and we can take B to contain the single number 1 to see that A_{TT} belongs to $\sigma(S_2)$. Taking $B = \emptyset$, we obtain $\emptyset$. Taking $B = \mathbb{R}$, we obtain Ω. Taking $B = [4, 16]$, we obtain the set $A_{HH} \cup A_{HT} \cup A_{TH}$. In short, as B ranges over the Borel subsets of $\mathbb{R}$, we will obtain the list of sets

[3] We recall that $\{X \in B\}$ is shorthand notation for the subset $\{\omega \in \Omega; X(\omega) \in B\}$ of Ω.

$$\emptyset, \Omega, A_{HH}, A_{HT} \cup A_{TH}, A_{TT}$$

and all unions and complements of these. This is the σ-algebra $\sigma(S_2)$.

Every set in $\sigma(S_2)$ is in the σ-algebra $\mathcal{F}_2$ of (2.1.3), the information contained in the first two coin tosses. On the other hand, A_{HT} and A_{TH} appear separately in $\mathcal{F}_2$ and only their union appears in $\sigma(S_2)$. This is because seeing the first two coin tosses allows us to distinguish an initial head followed by a tail from an initial tail followed by a head, but knowing only the value of S_2 does not permit this. There is enough information in $\mathcal{F}_2$ to determine the value of S_2 and even more. We say that S_2 is *$\mathcal{F}_2$-measurable.* □

Definition 2.1.5. *Let X be a random variable defined on a nonempty sample space Ω. Let $\mathcal{G}$ be a σ-algebra of subsets of Ω. If every set in $\sigma(X)$ is also in $\mathcal{G}$, we say that X is* $\mathcal{G}$-measurable.

A random variable X is $\mathcal{G}$-measurable if and only if the information in $\mathcal{G}$ is sufficient to determine the value of X. If X is $\mathcal{G}$-measurable, then $f(X)$ is also $\mathcal{G}$-measurable for any Borel-measurable function f; if the information in $\mathcal{G}$ is sufficient to determine the value of X, it will also determine the value of $f(X)$. If X and Y are $\mathcal{G}$-measurable, then $f(X, Y)$ is $\mathcal{G}$-measurable for any Borel-measurable function $f(x, y)$ of two variables. In particular, $X + Y$ and XY are $\mathcal{G}$-measurable.

A portfolio position $\Delta(t)$ taken at time t must be $\mathcal{F}(t)$-measurable (i.e., must depend only on information available to the investor at time t). We revisit a concept first encountered in Definition 2.4.1 of Chapter 2 of Volume I.

Definition 2.1.6. *Let Ω be a nonempty sample space equipped with a filtration $\mathcal{F}(t)$, $0 \le t \le T$. Let $X(t)$ be a collection of random variables indexed by $t \in [0, T]$. We say this collection of random variables is an* adapted stochastic process *if, for each t, the random variable $X(t)$ is $\mathcal{F}(t)$-measurable.*

In the continuous-time models of this text, asset prices, portfolio processes (i.e., positions), and wealth processes (i.e., values of portfolio processes) will all be adapted to a filtration that we regard as a model of the flow of public information.

2.2 Independence

When a random variable is measurable with respect to a σ-algebra $\mathcal{G}$, the information contained in $\mathcal{G}$ is sufficient to determine the value of the random variable. The other extreme is when a random variable is independent of a σ-algebra. In this case, the information contained in the σ-algebra gives no clue about the value of the random variable. Independence is the subject of the present section. In the more common case, when we have a σ-algebra $\mathcal{G}$ and a random variable X that is neither measurable with respect to $\mathcal{G}$ nor

independent of $\mathcal{G}$, the information in $\mathcal{G}$ is not sufficient to evaluate X, but we can estimate X based on the information in $\mathcal{G}$. We take up this case in the next section.

In contrast to the concept of measurability, we need a probability measure in order to talk about independence. Consequently, independence can be affected by changes of probability measure; measurability is not.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that two sets A and B in $\mathcal{F}$ are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

For example, in $\Omega = \{HH, HT, TH, TT\}$ with $0 \leq p \leq 1$, $q = 1 - p$, and

$$\mathbb{P}(HH) = p^2,\ \mathbb{P}(HT) = pq,\ \mathbb{P}(TH) = pq,\ \mathbb{P}(TT) = q^2,$$

the sets

$$A = \{\text{head on first toss}\} = \{HH, HT\}$$

and

$$B = \{\text{head on the second toss}\} = \{HH, TH\}$$

are independent. Indeed,

$$\mathbb{P}(A \cap B) = \mathbb{P}(HH) = p^2 \text{ and } \mathbb{P}(A)\mathbb{P}(B) = (p^2 + pq)(p^2 + pq) = p^2.$$

Independence of sets A and B means that knowing that the outcome ω of a random experiment is in A does not change our estimation of the probability that it is in B. If we know the first toss results in head, we still have probability p for a head on the second toss.

In a similar way, we want to define independence of two random variables X and Y to mean that if ω occurs and we know the value of $X(\omega)$ (without actually knowing ω), then our estimation of the distribution of Y is the same as when we did not know the value of $X(\omega)$. The formal definitions are the following.

Definition 2.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{G}$ and $\mathcal{H}$ be sub-σ-algebras of $\mathcal{F}$ (i.e., the sets in $\mathcal{G}$ and the sets in $\mathcal{H}$ are also in $\mathcal{F}$). We say these two* σ-algebras are independent *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \text{ for all } A \in \mathcal{G},\ B \in \mathcal{H}.$$

Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say these two random variables are independent *if the σ-algebras they generate, $\sigma(X)$ and $\sigma(Y)$, are independent. We say that the* random variable X is independent of the σ-algebra $\mathcal{G}$ *if $\sigma(X)$ and $\mathcal{G}$ are independent.*

Recall that $\sigma(X)$ is the collection of all sets of the form $\{X \in C\}$, where C ranges over the Borel subsets of $\mathbb{R}$. Similarly, every set in $\sigma(Y)$ is of the form $\{Y \in D\}$. Definition 2.2.1 says that X and Y are independent if and only if

$$\mathbb{P}\{X \in C \text{ and } Y \in D\} = \mathbb{P}\{X \in C\} \cdot \mathbb{P}\{Y \in D\}$$

for all Borel subsets C and D of $\mathbb{R}$.

Example 2.2.2. Recall the space Ω of three independent coin tosses on which the stock price random variables of Figure 1.2.2 of Chapter 1 are constructed. Let the probability measure $\mathbb{P}$ be given by

$$\mathbb{P}(HHH) = p^3, \ \mathbb{P}(HHT) = p^2q, \mathbb{P}(HTH) = p^2q, \mathbb{P}(HTT) = pq^2,$$
$$\mathbb{P}(THH) = p^2q, \ \mathbb{P}(THT) = pq^2, \ \mathbb{P}(TTH) = pq^2, \ \mathbb{P}(TTT) = q^3.$$

Intuitively, the random variables S_2 and S_3 are not independent because if we know that S_2 takes the value 16, then we know that S_3 is either 8 or 32 and is not 2 or .50. To formalize this, we consider the sets $\{S_3 = 32\} = \{HHH\}$ and $\{S_2 = 16\} = \{HHH, HHT\}$, whose probabilities are $\mathbb{P}\{S_3 = 32\} = p^3$ and $\mathbb{P}\{S_2 = 16\} = p^2$. In order to have independence, we must have

$$\mathbb{P}\{S_2 = 16 \text{ and } S_3 = 32\} = \mathbb{P}\{S_2 = 16\} \cdot \mathbb{P}\{S_3 = 32\} = p^5.$$

But $\mathbb{P}\{S_2 = 16 \text{ and } S_3 = 32\} = \mathbb{P}\{HHH\} = p^3$, so independence requires $p = 1$ or $p = 0$. Indeed, if $p = 1$, then after learning that $S_2 = 16$, we do not revise our estimate of the distribution of S_3; we already knew it would be 32. If $p = 0$, then S_2 cannot be 16, and we do not have to worry about revising our estimate of the distribution of S_3 if this occurs because it will not occur.

As the previous discussion shows, in the interesting cases of $0 < p < 1$, the random variables S_2 and S_3 are not independent. However, the random variables S_2 and $\frac{S_3}{S_2}$ are independent. Intuitively, this is because S_2 depends on the first two tosses, and $\frac{S_3}{S_2}$ depends on the third toss only. The σ-algebra generated by S_2 comprises $\emptyset$, Ω_3, the atoms (fundamental sets)

$$\{S_2 = 16\} = \{HHH, HHT\},$$
$$\{S_2 = 4\} = \{HTH, HTT, THH, THT\},$$
$$\{S_2 = 1\} = \{TTH, TTT\},$$

and their unions. The σ-algebra generated by $\frac{S_3}{S_2}$ comprises $\emptyset$, Ω_3, and the atoms

$$\left\{\frac{S_3}{S_2} = 2\right\} = \{HHH, HTH, THH, TTH\},$$
$$\left\{\frac{S_3}{S_2} = \frac{1}{2}\right\} = \{HHT, HTT, THT, TTT\}.$$

To verify independence, we can conduct a series of checks of the form

$$\mathbb{P}\left\{S_2 = 16 \text{ and } \frac{S_3}{S_2} = 2\right\} = \mathbb{P}\{S_2 = 16\} \cdot \mathbb{P}\left\{\frac{S_3}{S_2} = 2\right\}.$$

The left-hand side of this equality is

$$\mathbb{P}\left\{S_2 = 16 \text{ and } \frac{S_3}{S_2} = 2\right\} = \mathbb{P}\{HHH\} = p^3,$$

and the right-hand side is

$$\begin{aligned}\mathbb{P}\{S_2 = 16\} \cdot \mathbb{P}\left\{\frac{S_3}{S_2} = 2\right\} &= \mathbb{P}\{HHH, HHT\} \cdot \mathbb{P}\{HHH, HTH, THH, TTH\} \\ &= p^2 \cdot p.\end{aligned}$$

Indeed, for every $A \in \sigma(S_2)$ and every $B \in \sigma\left(\frac{S_3}{S_2}\right)$, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \qquad \square$$

We shall often need independence of more than two random variables. We make the following definition.

Definition 2.2.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots$ be a sequence of sub-σ-algebras of $\mathcal{F}$. For a fixed positive integer n, we say that the n* σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n$ are independent *if*

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \cdots \cdot \mathbb{P}(A_n) \quad \text{for all } A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2, \ldots, A_n \in \mathcal{G}_n.$$

Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say the n random variables $X_1, X_2, \ldots, X_n$ are independent *if the σ-algebras $\sigma(X_1)$, $\sigma(X_2), \ldots, \sigma(X_n)$ are independent. We say* the full sequence of σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots$ is independent *if, for every positive integer n, the n σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n$ are independent. We say the* full sequence of random variables $X_1, X_2, X_3, \ldots$ is independent *if, for every positive integer n, the n random variables $X_1, X_2, \ldots, X_n$ are independent.*

Example 2.2.4. The infinite independent coin-toss space $(\Omega_\infty, \mathcal{F}, \mathbb{P})$ of Example 1.1.4 of Chapter 1 exhibits the kind of independence described in Definition 2.2.3. Let $\mathcal{G}_k$ be the σ-algebra of information associated with the kth toss. In other words, $\mathcal{G}_k$ comprises the sets $\emptyset$, Ω_∞, and the atoms

$$\{\omega \in \Omega_\infty; \omega_k = H\} \text{ and } \{\omega \cdots \in \Omega_\infty; \omega_k = T\}.$$

Note that $\mathcal{G}_k$ is different from $\mathcal{F}_k$ in Example 1.1.4 of Chapter 1, the σ-algebra associated with the first k tosses. Under the probability measure constructed in Example 1.1.4 of Chapter 1, the full sequence of σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots$ is independent. Now recall the sequence of the random variables of (1.2.8) of Chapter 1:

$$Y_k(\omega) = \begin{cases} 1 \text{ if } \omega_k = H, \\ 0 \text{ if } \omega_k = T. \end{cases}$$

The full sequence of random variables $Y_1, Y_2, Y_3, \ldots$ is likewise independent. $\square$

The definition of independence of random variables, which was given in terms of independence of σ-algebras that they generate, is a strong condition that is conceptually useful but difficult to check in practice. We illustrate the first point with the following theorem and thereafter give a second theorem that simplifies the verification that two random variables are independent. Although this and the next section treat only the case of a pair of random variables, there are analogues of these results for n random variables.

Theorem 2.2.5. *Let X and Y be independent random variables, and let f and g be Borel-measurable functions on $\mathbb{R}$. Then $f(X)$ and $g(Y)$ are independent random variables.*

PROOF: Let A be in the σ-algebra generated by $f(X)$. This σ-algebra is a sub-σ-algebra of $\sigma(X)$. To see this, recall that, by definition, every set A in $\sigma(f(X))$ is of the form $\{\omega \in \Omega; f(X(\omega)) \in C\}$, where C is a Borel subset of $\mathbb{R}$. We define $D = \{x \in \mathbb{R}; f(x) \in C\}$ and then have

$$A = \{\omega \in \Omega; f(X(\omega)) \in C\} = \{\omega \in \Omega, X(\omega) \in D\}. \tag{2.2.1}$$

The set on the right-hand side of (2.2.1) is in $\sigma(X)$, so $A \in \sigma(X)$.

Let B be in the σ-algebra generated by $g(Y)$. This σ-algebra is a sub-σ-algebra of $\sigma(Y)$, so $B \in \sigma(Y)$. Since X and Y are independent, we have $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. □

Definition 2.2.6. *Let X and Y be random variables. The pair of random variables (X, Y) takes values in the plane $\mathbb{R}^2$, and the* joint distribution measure of (X, Y) *is given by*[4]

$$\mu_{X,Y}(C) = \mathbb{P}\{(X, Y) \in C\} \text{ for all Borel sets } C \subset \mathbb{R}^2. \tag{2.2.2}$$

This is a probability measure (i.e., a way of assigning measure between 0 and 1 to subsets of $\mathbb{R}^2$ so that $\mu_{X,Y}(\mathbb{R}^2) = 1$ and the countable additivity property is satisfied). The joint cumulative distribution function of (X, Y) *is*

$$F_{X,Y}(a, b) = \mu_{X,Y}\big((-\infty, a] \times (-\infty, b]\big) = \mathbb{P}\{X \le a, Y \le b\}, \ a \in \mathbb{R}, b \in \mathbb{R}. \tag{2.2.3}$$

We say that a nonnegative, Borel-measurable function $f_{X,Y}(x, y)$ is a joint density *for the pair of random variables (X, Y) if*

$$\mu_{X,Y}(C) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}_C(x, y) f_{X,Y}(x, y)\, dy\, dx \text{ for all Borel sets } C \subset \mathbb{R}^2. \tag{2.2.4}$$

[4] One way to generate the σ-algebra of Borel subsets of $\mathbb{R}^2$ is to start with the collection of closed rectangles $[a_1, b_1] \times [a_2, b_2]$ and then add all other sets necessary in order to have a σ-algebra. Any set in this resulting σ-algebra is called a *Borel subset of* $\mathbb{R}^2$. All subsets of $\mathbb{R}^2$ normally encountered belong to this σ-algebra.

Condition (2.2.4) holds if and only if

$$F_{X,Y}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx \text{ for all } a \in \mathbb{R}, b \in \mathbb{R}. \tag{2.2.5}$$

The *distribution measures* (generally called the *marginal distribution measures* in this context) of X and Y are

$$\begin{aligned}\mu_X(A) = \mathbb{P}\{X \in A\} &= \mu_{X,Y}(A \times \mathbb{R}) \text{ for all Borel subsets } A \subset \mathbb{R},\\ \mu_Y(B) = \mathbb{P}\{Y \in B\} &= \mu_{X,Y}(\mathbb{R} \times B) \text{ for all Borel subsets } B \subset \mathbb{R}.\end{aligned}$$

The *(marginal) cumulative distribution functions* are

$$\begin{aligned}F_X(a) = \mu_X(-\infty, a] &= \mathbb{P}\{X \le a\} \text{ for all } a \in \mathbb{R},\\ F_Y(b) = \mu_Y(-\infty, b] &= \mathbb{P}\{Y \le b\} \text{ for all } b \in \mathbb{R}.\end{aligned}$$

If the joint density $f_{X,Y}$ exists, then the marginal densities exist and are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

The *marginal densities*, if they exist, are nonnegative, Borel-measurable functions that satisfy

$$\begin{aligned}\mu_X(A) &= \int_A f_X(x)\,dx \text{ for all Borel subsets } A \subset \mathbb{R},\\ \mu_Y(B) &= \int_B f_Y(y)\,dy \text{ for all Borel subsets } B \subset \mathbb{R}.\end{aligned}$$

These last conditions hold if and only if

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\,dx \text{ for all } a \in \mathbb{R}, \tag{2.2.6}$$

$$F_Y(b) = \int_{-\infty}^{b} f_Y(y)\,dy \text{ for all } b \in \mathbb{R}. \tag{2.2.7}$$

Theorem 2.2.7. *Let X and Y be random variables. The following conditions are equivalent.*

(i) X and Y are independent.
(ii) The joint distribution measure factors:

$$\mu_{X,Y}(A \times B) = \mu_X(A) \cdot \mu_Y(B) \text{ for all Borel subsets } A \subset \mathbb{R}, B \subset \mathbb{R}. \tag{2.2.8}$$

(iii) The joint cumulative distribution function factors:

$$F_{X,Y}(a,b) = F_X(a) \cdot F_Y(b) \text{ for all } a \in \mathbb{R}, b \in \mathbb{R}. \tag{2.2.9}$$

(iv) The joint moment-generating function factors:

$$\mathbb{E}e^{uX+vY} = \mathbb{E}e^{uX} \cdot \mathbb{E}e^{vY} \tag{2.2.10}$$

for all $u \in \mathbb{R}$, $v \in \mathbb{R}$ for which the expectations are finite.

If there is a joint density, each of the conditions above is equivalent to the following.

(v) The joint density factors:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \text{ for almost every } x \in \mathbb{R}, y \in \mathbb{R}. \tag{2.2.11}$$

The conditions above imply but are not equivalent to the following.

(vi) The expectation factors:

$$\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y, \tag{2.2.12}$$

provided $\mathbb{E}|XY| < \infty$.

OUTLINE OF PROOF: We sketch the various steps that constitute the proof of this theorem.
(i)⇒(ii) Assume that X and Y are independent. Then

$$\begin{aligned}\mu_{X,Y}(A \times B) &= \mathbb{P}\{X \in A \text{ and } Y \in B\} \\ &= \mathbb{P}\big(\{X \in A\} \cap \{Y \in B\}\big) \\ &= \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{Y \in B\} \\ &= \mu_X(A) \cdot \mu_Y(B).\end{aligned}$$

(ii)⇒(i) A typical set in $\sigma(X)$ is of the form $\{X \in A\}$, and a typical set in $\sigma(Y)$ is of the form $\{Y \in B\}$. Assume (ii). Then

$$\begin{aligned}\mathbb{P}\big(\{X \in A\} \cap \{Y \in B\}\big) &= \mathbb{P}\{X \in A \text{ and } Y \in B\} \\ &= \mu_{X,Y}(A \times B) \\ &= \mu_X(A) \cdot \mu_Y(B) \\ &= \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{Y \in B\}.\end{aligned}$$

This shows that every set in $\sigma(X)$ is independent of every set in $\sigma(Y)$.
(ii)⇒(iii) Assume (2.2.8). Then

$$\begin{aligned}F_{X,Y}(a,b) &= \mu_{X,Y}\big((-\infty,a] \times (-\infty,b]\big) \\ &= \mu_X(-\infty,a] \cdot \mu_Y(-\infty,b] \\ &= F_X(a) \cdot F_Y(b).\end{aligned}$$

(iii)⇒(ii) Equation (2.2.9) implies that (2.2.8) holds whenever A is of the form $A = (-\infty,a]$ and B is of the form $B = (-\infty,b]$. This is enough to

establish (2.2.8) for all Borel sets A and B, but the details of this are beyond the scope of the text.

(iii)⇒(v) If there is a joint density, then (iii) implies

$$\int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx = \int_{-\infty}^{a} f_X(x)\,dx \cdot \int_{\infty}^{b} f_Y(y)\,dy.$$

Differentiating first with respect to a and then with respect to b, we obtain

$$f_{X,Y}(a,b) = f_X(a)\cdot f_Y(b),$$

which is just (2.2.11) with different dummy variables.

(v)⇒(iii) Assume there is a joint density. If we also assume (2.2.11), we can integrate both sides to get

$$\begin{aligned} F_{X,Y}(a,b) &= \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx \\ &= \int_{-\infty}^{a}\int_{-\infty}^{b} f_X(x)\cdot f_Y(y)\,dy\,dx \\ &= \int_{-\infty}^{a} f_X(x)\,dx \cdot \int_{-\infty}^{b} f_Y(y)\,dy \\ &= F_X(a)\cdot F_Y(b). \end{aligned}$$

(i)⇒(iv) We first use the "standard machine" as in the proof of Theorem 1.5.1 of Chapter 1, starting with the case when h is the indicator function of a Borel subset of $\mathbb{R}^2$, to show that, for every real-valued, Borel-measurable function $h(x,y)$ on $\mathbb{R}^2$, we have

$$\mathbb{E}|h(X,Y)| = \int_{\mathbb{R}^2} |h(x,y)|\,d\mu_{X,Y}(x,y),$$

and if this quantity is finite, then

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)\,d\mu_{X,Y}(x,y). \tag{2.2.13}$$

This is true for any pair of random variables X and Y, whether or not they are independent. If X and Y are independent, then the joint distribution $\mu_{X,Y}$ is a product of marginal distributions, and this permits us to rewrite (2.2.13) as

$$\mathbb{E}h(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\,d\mu_Y(y)\,d\mu_X(x). \tag{2.2.14}$$

We now fix numbers u and v and take $h(x,y) = e^{ux+vy}$. Equation (2.2.14) reduces to

$$\begin{aligned}\mathbb{E}e^{uX+vY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{ux+vy}\,d\mu_Y(y)\,d\mu_X(x)\\ &= \int_{-\infty}^{\infty} e^{ux}\,d\mu_X(x)\cdot\int_{-\infty}^{\infty} e^{vy}\,d\mu_Y(y)\\ &= \mathbb{E}e^{uX}\cdot\mathbb{E}e^{vY},\end{aligned}$$

where we have used Theorem 1.5.1 of Chapter 1 for the last step. The proof **(iv)⇒(i)** is beyond the scope of this text.
(i)⇒(vi) In the special case when $h(x,y) = xy$, (2.2.14) reduces to

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} x\,d\mu_X(x)\cdot\int_{-\infty}^{\infty} y\,d\mu_Y(y) = \mathbb{E}X\cdot\mathbb{E}Y,$$

where again we have used Theorem 1.5.1 of Chapter 1 for the last step. □

Example 2.2.8 (Independent normal random variables). Random variables X and Y are independent and standard normal if they have the joint density

$$f_{X,Y}(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)} \text{ for all } x\in\mathbb{R}, y\in\mathbb{R}.$$

This is the product of the marginal densities

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \text{ and } f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}.$$

We use the notation

$$N(a) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{1}{2}x^2}\,dx \tag{2.2.15}$$

for the standard normal cumulative distribution function. The joint cumulative distribution function for (X,Y) factors:

$$\begin{aligned}F_{X,Y}(a,b) &= \int_{-\infty}^{a}\int_{-\infty}^{b} f_X(x)f_Y(y)\,dy\,dx\\ &= \int_{-\infty}^{a} f_X(x)\,dx\cdot\int_{-\infty}^{b} f_Y(y)\,dy\\ &= N(a)\cdot N(b).\end{aligned}$$

The joint distribution μ_X is the probability measure on $\mathbb{R}^2$ that assigns a measure to each Borel set $C\subset\mathbb{R}^2$ equal to the integral of $f_{X,Y}(x,y)$ over C. If $C = A\times B$, where $A\in\mathcal{B}(\mathbb{R})$ and $B\in\mathcal{B}(\mathbb{R})$, then $\mu_{X,Y}$ factors:

$$\begin{aligned}\mu_{X,Y}(A\times B) &= \int_A\int_B f_X(x)f_Y(y)\,dy\,dx\\ &= \int_A f_X(x)\,dx\cdot\int_B f_Y(y)\,dy\\ &= \mu_X(A)\cdot\mu_Y(B).\end{aligned}$$

□

We give an example to show that property (vi) of Theorem 2.2.7 does not imply independence. We precede this with a definition.

Definition 2.2.9. *Let X be a random variable whose expected value is defined. The* variance of X, *denoted* $\mathrm{Var}(X)$, *is*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

Because $(X - \mathbb{E}X)^2$ is nonnegative, $\mathrm{Var}(X)$ *is always defined, although it may be infinite. The* standard deviation of X *is* $\sqrt{\mathrm{Var}(X)}$. *The linearity of expectations shows that*

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2.$$

Let Y be another random variable and assume that $\mathbb{E}X$, $\mathrm{Var}(X)$, $\mathbb{E}Y$ *and* $Var(Y)$ *are all finite. The* covariance of X and Y *is*

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right].$$

The linearity of expectations shows that

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y.$$

In particular, $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$ if and only if $\mathrm{Cov}(X, Y) = 0$. *Assume, in addition to the finiteness of expectations and variances, that* $\mathrm{Var}(X) > 0$ *and* $Var(Y) > 0$. *The* correlation coefficient of X and Y *is*

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var(X)Var}(Y)}}.$$

If $\rho(X, Y) = 0$ (or equivalently, $\mathrm{Cov}(X, Y) = 0$*), we say that X* and Y are uncorrelated.

Property (vi) of Theorem 2.2.7 implies that independent random variables are uncorrelated. The converse is not true, even for normal random variables, although it is true of *jointly normal* random variables (see Definition 2.2.11 below).

Example 2.2.10 (Uncorrelated, dependent normal random variables). Let X be a standard normal random variable and let Z be independent of X and satisfy[5]

[5] To construct such random variables, we can choose $\Omega = \{(\omega_1, \omega_2); 0 \leq \omega_1 \leq 1,\ 0 \leq \omega_2 \leq 1\}$ to be the unit square and choose $\mathbb{P}$ to be the two-dimensional Lebesgue measure according to which $\mathbb{P}(A)$ is equal to the area of A for every Borel subset of Ω. We then set $X(\omega_1, \omega_2) = N^{-1}(\omega_1)$, which is a standard normal random variable under $\mathbb{P}$ (see Example 1.2.6 for a discussion of this probability integral transform). We set $Z(\omega_1, \omega_2)$ to be -1 if $0 \leq \omega_2 \leq \frac{1}{2}$ and to be 1 if $\frac{1}{2} < \omega_2 \leq 1$.

$$\mathbb{P}\{Z=1\}=\frac{1}{2} \text{ and } \mathbb{P}\{Z=-1\}=\frac{1}{2}. \tag{2.2.16}$$

Define $Y = ZX$. We show below that, like X, the random variable Y is standard normal. Furthermore, X and Y are uncorrelated, but they are not independent. The pair (X, Y) does not have a joint density.

Let us first determine the distribution of Y. We compute

$$\begin{aligned} F_Y(b) &= \mathbb{P}\{Y \le b\} \\ &= \mathbb{P}\{Y \le b \text{ and } Z=1\} + \mathbb{P}\{Y \le b \text{ and } Z=-1\} \\ &= \mathbb{P}\{X \le b \text{ and } Z=1\} + \mathbb{P}\{-X \le b \text{ and } Z=-1\}. \end{aligned}$$

Because X and Z are independent, we have

$$\begin{aligned} &\mathbb{P}\{X \le b \text{ and } Z=1\} + \mathbb{P}\{-X \le b \text{ and } Z=-1\} \\ &\quad = \mathbb{P}\{Z=1\} \cdot \mathbb{P}\{X \le b\} + \mathbb{P}\{Z=-1\} \cdot \mathbb{P}\{-X \le b\} \\ &\quad = \frac{1}{2} \cdot \mathbb{P}\{X \le b\} + \frac{1}{2} \cdot \mathbb{P}\{-X \le b\}. \end{aligned}$$

Because X is a standard normal random variable, so is $-X$. Therefore, $\mathbb{P}\{X \le b\} = \mathbb{P}\{-X \le b\} = N(b)$. It follows that $F_Y(b) = N(b)$; in other words, Y is a standard normal random variable.

Since $\mathbb{E}X = \mathbb{E}Y = 0$, the covariance of X and Y is

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] = \mathbb{E}\big[ZX^2\big].$$

Because Z and X are independent, so are Z and X^2, and we may use Theorem 2.2.7(vi) to write

$$\mathbb{E}\big[ZX^2\big] = \mathbb{E}Z \cdot \mathbb{E}\big[X^2\big] = 0 \cdot 1 = 0.$$

Therefore, X and Y are uncorrelated.

The random variables X and Y cannot be independent for if they were, then $|X|$ and $|Y|$ would also be independent (Theorem 2.2.5). But $|X| = |Y|$. In particular,

$$\mathbb{P}\{|X| \le 1, |Y| \le 1\} = \mathbb{P}\{|X| \le 1\} = N(1) - N(-1),$$

and

$$\mathbb{P}\{|X| \le 1\} \cdot \mathbb{P}\{|Y| \le 1\} = \big(N(1) - N(-1)\big)^2.$$

These two expressions are not equal, as they would be for independent random variables.

Finally, we want to examine the joint distribution measure $\mu_{X,Y}$ of (X, Y). Since $|X| = |Y|$, the pair (X, Y) takes values only in the set

$$C = \{(x, y); x = \pm y\}.$$

In other words, $\mu_{X,Y}(C) = 1$ and $\mu_{X,Y}(C^c) = 0$. But C has zero area. It follows that for any nonnegative function f, we must have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{I}_C(x,y) f(x,y)\, dy\, dx = 0.$$

One way of thinking about this is to observe that if we want to integrate a function $\mathbb{I}_C(x,y)f(x,y)$ over the plane $\mathbb{R}^2$, we could first fix x and integrate out the y-variable, but since $f(x,y)\mathbb{I}_C(x,y)$ is zero except when $y = x$ and $y = -x$, we will get zero. When we next integrate out the x-variable, we will be integrating the zero function, and the end result will be zero. There cannot be a joint density for (X,Y) because with this choice of the set C, the left-hand side of (2.2.4) is one but the right-hand side is zero. Of course, X and Y have marginal densities because they are both standard normal. Moreover, the joint cumulative distribution function exists (as it always does). In this case, it is

$$\begin{aligned}
&F_{X,Y}(a,b)\\
&= \mathbb{P}\{X \le a \text{ and } Y \le b\}\\
&= \mathbb{P}\{X \le a, X \le b, \text{ and } Z = 1\} + \mathbb{P}\{X \le a, -X \le b, \text{ and } Z = -1\}\\
&= \mathbb{P}\{Z = 1\} \cdot \mathbb{P}\{X \le \min(a,b)\} + \mathbb{P}\{Z = -1\} \cdot \mathbb{P}\{-b \le X \le a\}\\
&= \frac{1}{2}N(\min(a,b)) + \frac{1}{2}\max\{N(a) - N(-b), 0\}.
\end{aligned}$$

There is no joint density $f_{X,Y}(x,y)$ that permits us to write this function in the form (2.2.5). □

Definition 2.2.11. *Two random variables X and Y are said to be* jointly normal *if they have the joint density*

$$\begin{aligned}
&f_{X,Y}(x,y)\\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right.\right.\\
&\qquad\qquad \left.\left. + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}, \qquad (2.2.17)
\end{aligned}$$

where $\sigma_1 > 0$, $\sigma_2 > 0$, $|\rho| < 1$, and μ_1, μ_2 are real numbers. More generally, a random column vector $\mathbf{X} = (X_1, \ldots, X_n)^{\text{tr}}$, where the superscript tr *denotes transpose, is* jointly normal *if it has joint density*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)C^{-1}(\mathbf{x}-\mu)^{\text{tr}}\right\}. \qquad (2.2.18)$$

In equation (2.2.18), $\mathbf{x} = (x_1, \ldots, x_n)$ is a row vector of dummy variables, $\mu = (\mu_1, \ldots, \mu_n)$ is the row vector of expectations, and C is the positive definite matrix of covariances.

In the case of (2.2.17), X is normal with expectation μ_1 and variance σ_1^2, Y is normal with expectation μ_2 and variance σ_2^2, and the correlation between X and Y is ρ. The density factors (equivalently, X and Y are independent) if and only if $\rho = 0$. In the case (2.2.18), the density factors into the product of n normal densities (equivalently, the components of $\mathbf{X}$ are independent) if and only if C is a diagonal matrix (all the covariances are zero).

Linear combinations of jointly normal random variables (i.e., sums of constants times the random variables) are jointly normal. Since independent normal random variables are jointly normal, a general method for creating jointly normal random variables is to begin with a set of independent normal random variables and take linear combinations. Conversely, any set of jointly normal random variables can be reduced to linear combinations of independent normal random variables. We do this reduction for a pair of correlated normal random variables in Example 2.2.12 below.

Since the distribution of jointly normal random variables is characterized in terms of means and covariances, and joint normality is preserved under linear combinations, it is not necessary to deal directly with the density when making linear changes of variables. The following example illustrates this point.

Example 2.2.12. Let (X, Y) be jointly normal with the density (2.2.17). Define $W = Y - \frac{\rho\sigma_2}{\sigma_1}X$. Then X and W are independent. To verify this, it suffices to show that X and W have covariance zero since they are jointly normal. We compute

$$\begin{aligned}
\mathrm{Cov}(X, W) &= \mathbb{E}\big[(X - \mathbb{E}X)(W - \mathbb{E}W)\big] \\
&= \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] - \mathbb{E}\left[\frac{\rho\sigma_2}{\sigma_1}(X - \mathbb{E}X)^2\right] \\
&= \mathrm{Cov}(X, Y) - \frac{\rho\sigma_2}{\sigma_1}\sigma_1^2 \\
&= 0.
\end{aligned}$$

The expectation of W is $\mu_3 = \mu_2 - \frac{\rho\sigma_2\mu_1}{\sigma_1}$, and the variance is

$$\begin{aligned}
\sigma_3^2 &= \mathbb{E}\big[(W - \mathbb{E}W)^2\big] \\
&= \mathbb{E}\big[(Y - \mathbb{E}Y)^2\big] - \frac{2\rho\sigma_2}{\sigma_1}\mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] + \frac{\rho^2\sigma_2^2}{\sigma_1^2}\mathbb{E}\big[(X - \mathbb{E}X)^2\big] \\
&= (1 - \rho^2)\sigma_2^2.
\end{aligned}$$

The joint density of X and W is

$$f_{X,W}(x, w) = \frac{1}{2\pi\sigma_1\sigma_3}\exp\left\{-\frac{(x - \mu_1)^2}{2\sigma_1^2} - \frac{(w - \mu_3)^2}{2\sigma_3^2}\right\}.$$

Note finally that we have decomposed Y into the linear combination

$$Y = \frac{\rho\sigma_2}{\sigma_1}X + W \tag{2.2.19}$$

of a pair of independent normal random variables X and W. □

2.3 General Conditional Expectations

We consider a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$. If X is $\mathcal{G}$-measurable, then the information in $\mathcal{G}$ is sufficient to determine the value of X. If X is independent of $\mathcal{G}$, then the information in $\mathcal{G}$ provides no help in determining the value of X. In the intermediate case, we can use the information in $\mathcal{G}$ to estimate but not precisely evaluate X. The *conditional expectation of X given $\mathcal{G}$* is such an estimate.

We have already discussed conditional expectations in the binomial model. Let Ω be the set of all possible outcomes of N coin tosses, and assume these coin tosses are independent with probability p for head and probability $q = 1 - p$ for tail on each toss. Let $\mathbb{P}(\omega)$ denote the probability of a sequence of coin tosses under these assumptions. Let n be an integer, $1 \leq n \leq N-1$, and let X be a random variable. Then the conditional expectation of X under $\mathbb{P}$, based on the information at time n, is (see Definition 2.3.1 of Chapter 2)

$$\begin{aligned}&\mathbb{E}_n[X](\omega_1 \dots \omega_n)\\ &= \sum_{\omega_{n+1} \dots \omega_N} p^{\#H(\omega_{n+1} \dots \omega_N)} q^{\#T(\omega_{n+1} \dots \omega_N)} X(\omega_1 \dots \omega_n \omega_{n+1} \dots \omega_N). \quad (2.3.1)\end{aligned}$$

In the special cases $n = 0$ and $n = N$, we define

$$\mathbb{E}_0 X = \sum_{\omega_0 \dots \omega_N} p^{\#H(\omega_0 \dots \omega_N)} q^{\#T(\omega_0 \dots \omega_N)} X(\omega_0 \dots \omega_N) = \mathbb{E}X, \quad (2.3.2)$$

$$E_N[X](\omega_0 \dots \omega_N) = X(\omega_0 \dots \omega_N). \quad (2.3.3)$$

In (2.3.2), we have the estimate of X based on no information, and in (2.3.3) we have the estimate based on full information.

We need to generalize (2.3.1)–(2.3.3) in a way suitable for a continuous-time model. Toward that end, we examine (2.3.1) within the context of a three-period example. Consider the general three-period model of Figure 2.3.1. We assume the probability of head on each toss is p and the probability of tail is $q = 1 - p$, and we compute

$$\mathbb{E}_2[S_3](HH) = pS_3(HHH) + qS_3(HHT), \quad (2.3.4)$$

$$\mathbb{E}_2[S_3](HT) = pS_3(HTH) + qS_3(HTT), \quad (2.3.5)$$

$$\mathbb{E}_2[S_3](TH) = pS_3(THH) + qS_3(THT), \quad (2.3.6)$$

$$\mathbb{E}_2[S_3](TT) = pS_3(TTH) + qS_3(TTT). \quad (2.3.7)$$

Recall the σ-algebra $\mathcal{F}_2$ of (2.1.3), which is built up from the four fundamental sets (we call them *atoms* because they are indivisible within the σ-algebra) A_{HH}, A_{HT}, A_{TH}, and A_{TT} of (2.1.2). We multiply (2.3.4) by $\mathbb{P}(A_{HH}) = p^2$, multiply (2.3.5) by $\mathbb{P}(A_{HT}) = pq$, multiply (2.3.6) by $\mathbb{P}(A_{TH}) = pq$, and multiply (2.3.7) by $\mathbb{P}(A_{TT}) = q^2$. The resulting equations may be written as

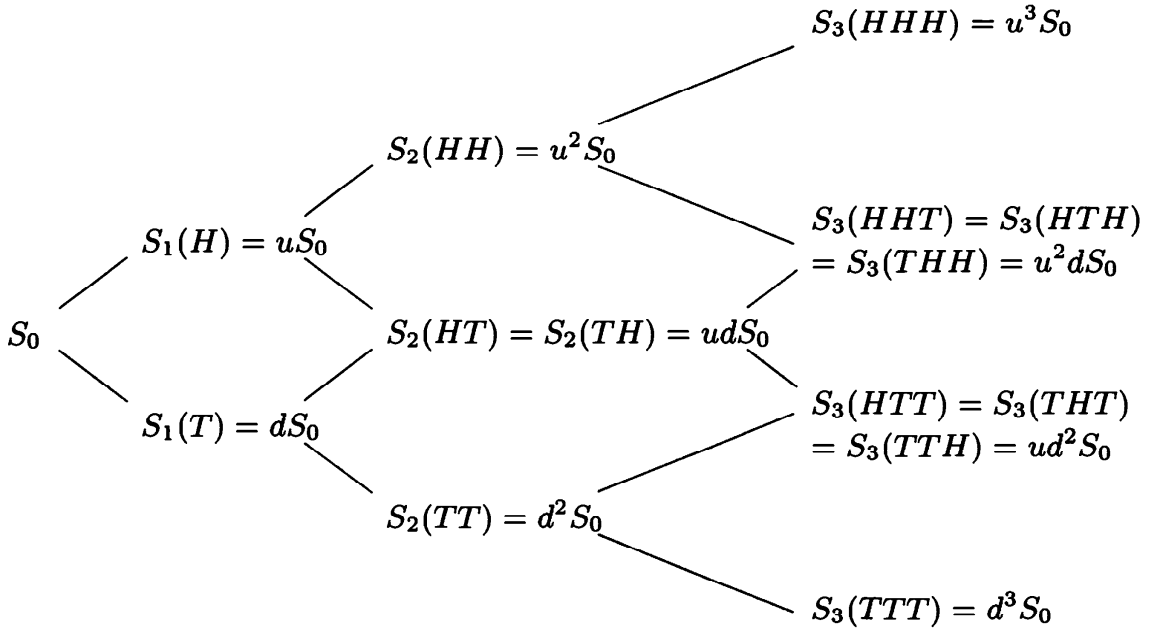

Fig. 2.3.1. General three-period model.

$$E_2[S_3](HH)\mathbb{P}(A_{HH}) = \sum_{\omega \in A_{HH}} S_3(\omega)\mathbb{P}(\omega), \tag{2.3.8}$$

$$E_2[S_3](HT)\mathbb{P}(A_{HT}) = \sum_{\omega \in A_{HT}} S_3(\omega)\mathbb{P}(\omega), \tag{2.3.9}$$

$$E_2[S_3](TH)\mathbb{P}(A_{TH}) = \sum_{\omega \in A_{TH}} S_3(\omega)\mathbb{P}(\omega), \tag{2.3.10}$$

$$E_2[S_3](TT)\mathbb{P}(A_{TT}) = \sum_{\omega \in A_{TT}} S_3(\omega)\mathbb{P}(\omega). \tag{2.3.11}$$

We could divide each of these equations by the probability of the atom appearing as the second factor on the left-hand sides and thereby recover the formulas (2.3.4)–(2.3.7) for the conditional expectations. However, in the continuous-time model, atoms typically have probability zero, and such a step cannot be performed. We therefore take an alternate route here to lay the groundwork for the continuous-time model.

On each of the atoms of $\mathcal{F}_2$, the conditional expectation $\mathbb{E}_2[S_3]$ is constant because the conditional expectation does not depend on the third toss and the atom is created by holding the first two tosses fixed. It follows that the left-hand sides of (2.3.8)–(2.3.11) may be written as integrals of the integrand $\mathbb{E}_2[S_3]$ over the atom. For this purpose, we shall write $\mathbb{E}_2[S_3](\omega) = \mathbb{E}_2[S_3](\omega_1\omega_2\omega_3)$, including the third toss in the argument, even though it is irrelevant. The right-hand sides of these equations are sums, which are Lebesgue integrals on a finite probability space. Using Lebesgue integral notation, we rewrite (2.3.8)–(2.3.11) as

$$\int_{A_{HH}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{HH}} S_3(\omega)\, d\mathbb{P}(\omega), \tag{2.3.12}$$

$$\int_{A_{HT}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{HT}} S_3(\omega)\, d\mathbb{P}(\omega), \tag{2.3.13}$$

$$\int_{A_{TH}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{TH}} S_3(\omega)\, d\mathbb{P}(\omega), \tag{2.3.14}$$

$$\int_{A_{TT}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{TT}} S_3(\omega)\, d\mathbb{P}(\omega). \tag{2.3.15}$$

In other words, on each of the atoms the value of the conditional expectation has been chosen to be that constant that yields the same average over the atom as the random variable S_3 being estimated.

We turn our attention now to the other sets in $\mathcal{F}_2$. The full list appears in (2.1.3), and every set on the list, except for the empty set, is a finite union of atoms. If we add equations (2.3.12) and (2.3.13), we obtain

$$\int_{A_H} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_H} S_3(\omega)\, d\mathbb{P}(\omega).$$

Similarly, but adding various combinations of (2.3.12)–(2.3.15), we see that

$$\int_A \mathbb{E}_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_A S_3(\omega)\, d\mathbb{P}(\omega) \tag{2.3.16}$$

for every set $A \in \mathcal{F}_2$, except possibly for $A = \emptyset$. However, if $A = \emptyset$, equation (2.3.16) still holds, with both sides equal to zero. We call (2.3.16) the *partial-averaging property* of conditional expectations because it says that the conditional expectation and the random variable being estimated give the same value when averaged over "parts" of Ω (those "parts" that are sets in the conditioning σ-algebra $\mathcal{F}_2$).

We take (2.3.16) as the defining property of conditional expectations. The precise definition is the following.

Definition 2.3.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let X be a random variable that is either nonnegative or integrable. The* conditional expectation of X given $\mathcal{G}$, *denoted $\mathbb{E}[X|\mathcal{G}]$, is any random variable that satisfies*

(i) **(Measurability)** *$\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, and*
(ii) **(Partial averaging)**

$$\int_A \mathbb{E}[X|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) = \int_A X(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{G}. \tag{2.3.17}$$

If $\mathcal{G}$ is the σ-algebra generated by some other random variable W (i.e., $\mathcal{G} = \sigma(W)$), we generally write $\mathbb{E}[X|W]$ rather than $\mathbb{E}[X|\sigma(W)]$.

Property (i) in Definition 2.3.1 guarantees that, although the estimate of X based on the information in $\mathcal{G}$ is itself a random variable, the value of the estimate $\mathbb{E}[X|\mathcal{G}]$ can be determined from the information in $\mathcal{G}$. Property (i) captures the fact that the estimate $\mathbb{E}[X|\mathcal{G}]$ of X is *based on the information in* $\mathcal{G}$. Note in (2.3.4)–(2.3.7) that the conditional expectation $\mathbb{E}_2[S_3]$ is constant on the atoms of $\mathcal{F}_2$; this is property (i) for this case.

The second property ensures that $\mathbb{E}[X|\mathcal{G}]$ is indeed an estimate of X. It gives the same averages as X over all the sets in $\mathcal{G}$. If $\mathcal{G}$ has many sets, which provide a fine resolution of the uncertainty inherent in ω, then this partial-averaging property over the "small" sets in $\mathcal{G}$ says that $\mathbb{E}[X|\mathcal{G}]$ is a good estimator of X. If $\mathcal{G}$ has only a few sets, this partial-averaging property guarantees only that $\mathbb{E}[X|\mathcal{G}]$ is a crude estimate of X.

Definition 2.3.1 raises two immediate questions. First, does there always exist a random variable $\mathbb{E}[X|\mathcal{G}]$ satisfying properties (i) and (ii)? Second, if there is a random variable satisfying these properties, is it unique? The answer to the first question is yes, and the proof of the existence of $\mathbb{E}[X|\mathcal{G}]$ is based on the Radon-Nikodým Theorem, Theorem 1.6.7 (see Appendix B). The answer to the second question is a qualified yes, as we now explain. Suppose Y and Z both satisfy conditions (i) and (ii) of Definition 2.3.1. Because both Y and Z are $\mathcal{G}$-measurable, their difference $Y - Z$ is as well, and thus the set $A = \{Y - Z > 0\}$ is in $\mathcal{G}$. From (2.3.17), we have

$$\int_A Y(\omega)\, d\mathbb{P}(\omega) = \int_A X(\omega)\, d\mathbb{P}(\omega) = \int_A Z(\omega)\, d\mathbb{P}(\omega),$$

and thus

$$\int_A \big(Y(\omega) - Z(\omega)\big)\, d\mathbb{P}(\omega) = 0.$$

The integrand is strictly positive on the set A, so the only way this equation can hold is for A to have probability zero (i.e., $Y \leq Z$ almost surely). We can reverse the roles of Y and Z in this argument and conclude that $Z \leq Y$ almost surely. Hence $Y = Z$ almost surely. This means that although different procedures might result in different random variables when determining $\mathbb{E}[X|\mathcal{G}]$, these different random variables will agree almost surely. The set of ω for which the random variables are different has zero probability.

In this more general context, conditional expectations still have the five fundamental properties developed in Theorem 2.3.2 of Chapter 2 of Volume I. We restate them in the present context.

Theorem 2.3.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$.*

(i) **(Linearity of conditional expectations)** *If X and Y are integrable random variables and c_1 and c_2 are constants, then*

$$\mathbb{E}[c_1 X + c_2 Y|\mathcal{G}] = c_1 \mathbb{E}[X|\mathcal{G}] + c_2 \mathbb{E}[Y|\mathcal{G}]. \tag{2.3.18}$$

This equation also holds if we assume that X and Y are nonnegative (rather than integrable) and c_1 and c_2 are positive, although both sides may be $+\infty$.

(ii) **(Taking out what is known)** *If X and Y are integrable random variables, Y and XY are integrable, and X is $\mathcal{G}$-measurable, then*

$$\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}]. \tag{2.3.19}$$

This equation also holds if we assume that X is positive and Y is nonnegative (rather than integrable), although both sides may be $+\infty$.

(iii) **(Iterated conditioning)** *If $\mathcal{H}$ is a sub-σ algebra of $\mathcal{G}$ ($\mathcal{H}$ contains less information than $\mathcal{G}$) and X is an integrable random variable, then*

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big] = \mathbb{E}\big[X\big|\mathcal{H}\big]. \tag{2.3.20}$$

This equation also holds if we assume that X is nonnegative (rather than integrable), although both sides may be $+\infty$.

(iv) **(Independence)** *If X is integrable and independent of $\mathcal{G}$, then*

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}X. \tag{2.3.21}$$

This equation also holds if we assume that X is nonnegative (rather than integrable), although both sides may be $+\infty$.

(v) **(Conditional Jensen's inequality)** *If $\varphi(x)$ is a convex function of a dummy variable x and X is integrable, then*

$$\mathbb{E}\big[\varphi(X)|\mathcal{G}\big] \geq \varphi\big(E[X|\mathcal{G}]\big). \tag{2.3.22}$$

DISCUSSION AND SKETCH OF PROOF: We take each of these properties in turn.

(i) Linearity allows us to separate the estimation of random variables into estimation of separate pieces and then add the estimates of the pieces to estimate the whole. To verify that $\mathbb{E}[c_1X + c_2Y|\mathcal{G}]$ is given by the right-hand side of (2.3.18), we observe that the right-hand side is $\mathcal{G}$-measurable because $\mathbb{E}[X|\mathcal{G}]$ and $\mathbb{E}[Y|\mathcal{G}]$ are $\mathcal{G}$-measurable and then must check the partial-averaging property (ii) of Definition 2.3.1. Using the fact that $E[X|\mathcal{G}]$ and $E[Y|\mathcal{G}]$ themselves satisfy the partial-averaging property, we have for every $A \in \mathcal{G}$ that

$$\begin{aligned}
&\int_A \big(c_1\mathbb{E}[X|\mathcal{G}](\omega) + c_2\mathbb{E}[Y|\mathcal{G}](\omega)\big)\, d\mathbb{P}(\omega) \\
&= c_1 \int_A \mathbb{E}[X|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) + c_2 \int_A \mathbb{E}[Y|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) \\
&= c_1 \int_A X(\omega)\, d\mathbb{P}(\omega) + c_2 \int_A Y(\omega)\, d\mathbb{P}(\omega) \\
&= \int_A \big(c_1X(\omega) + c_2Y(\omega)\big)\, d\mathbb{P}(\omega),
\end{aligned}$$

which shows that $c_1\mathbb{E}[X|\mathcal{G}]+c_2\mathbb{E}[Y|\mathcal{G}]$ satisfies the partial-averaging property that characterizes $\mathbb{E}[c_1X+c_2Y|\mathcal{G}]$ and hence is $\mathbb{E}[c_1X+c_2Y|\mathcal{G}]$.

(ii) Taking out what is known permits us to remove X from the estimation problem if its value can be determined from the information in $\mathcal{G}$. To estimate XY, it suffices to estimate Y alone and then multiply the estimate by X. To prove (2.3.19), we observe first that $X\mathbb{E}[Y|\mathcal{G}]$ is $\mathcal{G}$-measurable because both X and $\mathbb{E}[Y|\mathcal{G}]$ are $\mathcal{G}$-measurable. We must check the partial-averaging property.

Let us first consider the case when X is a $\mathcal{G}$-measurable indicator random variable (i.e., $X=\mathbb{I}_B$, where B is a set in $\mathcal{G}$). Using the fact that $\mathbb{E}[Y|\mathcal{G}]$ itself satisfies the partial-averaging property, we have for every set $A\in\mathcal{G}$ that

$$\begin{aligned}\int_A X(\omega)\mathbb{E}[Y|\mathcal{G}](\omega)\,d\mathbb{P}(\omega) &= \int_{A\cap B}\mathbb{E}[Y|\mathcal{G}](\omega)\,d\mathbb{P}(\omega)\\ &= \int_{A\cap B} Y(\omega)\,d\mathbb{P}(\omega)\\ &= \int_A X(\omega)Y(\omega)\,d\mathbb{P}(\omega). \qquad (2.3.23)\end{aligned}$$

Having proved (2.3.23) for $\mathcal{G}$-measurable indicator random variables X, we may use the standard machine developed in the proof of Theorem 1.5.1 of Chapter 1 to obtain this equation for all $\mathcal{G}$-measurable random variables X for which XY is integrable. This shows that $X\mathbb{E}[Y|\mathcal{G}]$ satisfies the partial-averaging condition that characterizes $\mathbb{E}[XY|\mathcal{G}]$, and hence $X\mathbb{E}[Y|\mathcal{G}]$ is the conditional expectation $\mathbb{E}[XY|\mathcal{G}]$.

(iii) If we estimate X based on the information in $\mathcal{G}$ and then estimate the estimate based on the smaller amount of information in $\mathcal{H}$, we obtain the random variable we would have gotten by estimating X directly based on the smaller amount of information in $\mathcal{H}$. To prove this, we observe first that $\mathbb{E}[X|\mathcal{H}]$ is $\mathcal{H}$-measurable by definition. The partial-averaging property that characterizes $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$ is

$$\int_A \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big](\omega)\,d\mathbb{P}(\omega)=\int_A \mathbb{E}[X|\mathcal{G}](\omega)\,\mathbb{P}(\omega)\text{ for all }A\in\mathcal{H}.$$

In order to prove (2.3.20), we must show that we can replace $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$ on the left-hand side of this equation by $\mathbb{E}[X|\mathcal{H}]$. But when $A\in\mathcal{H}$, it is also in $\mathcal{G}$, and the partial-averaging properties for $\mathbb{E}[X|\mathcal{H}]$ and $\mathbb{E}[X|\mathcal{G}]$ imply

$$\int_A \mathbb{E}[X|\mathcal{H}](\omega)\,d\mathbb{P}(\omega)=\int_A X(\omega)\,d\mathbb{P}(\omega)=\int_A \mathbb{E}[X|\mathcal{G}](\omega)\,d\mathbb{P}(\omega).$$

This shows that $\mathbb{E}[X|\mathcal{H}]$ satisfies the partial-averaging property that characterizes $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$, and hence $\mathbb{E}[X|\mathcal{H}]$ is $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$.

(iv) If X is independent of the information in $\mathcal{G}$, then the best estimate we can give of X is its expected value. This is also the estimate we would give based on no information. To prove this, we observe first that $\mathbb{E}X$ is $\mathcal{G}$-measurable.

Indeed, $\mathbb{E}X$ is not random and so is measurable with respect to every σ-algebra. We need to verify that $\mathbb{E}X$ satisfies the partial-averaging property that characterizes $\mathbb{E}[X|\mathcal{G}]$; i.e.,

$$\int_A \mathbb{E}X \, d\mathbb{P}(\omega) = \int_A X(\omega) \, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{G}. \tag{2.3.24}$$

Let us consider first the case when X is an indicator random variable independent of $\mathcal{G}$ (i.e., $X = \mathbb{I}_B$, where the set B is independent of $\mathcal{G}$). For all $A \in \mathcal{G}$, we have then

$$\int_A X(\omega) \, d\mathbb{P}(\omega) = \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) = \mathbb{P}(A)\mathbb{E}X = \int_A \mathbb{E}X \, d\mathbb{P}(\omega),$$

and (2.3.24) holds. We complete the proof using the standard machine developed in the proof of Theorem 1.5.1 of Chapter 1.
(v) Using the linearity of conditional expectations, we can repeat the proof of Theorem 2.2.5 of Chapter 2 to prove the conditional Jensen's inequality. □

We note that $\mathbb{E}[X|\mathcal{G}]$ is an unbiased estimator of X:

$$\mathbb{E}\big(\mathbb{E}[X|\mathcal{G}]\big) = \mathbb{E}X. \tag{2.3.25}$$

This equality is just the partial-averaging property (2.3.17) with $A = \Omega$.

Example 2.3.3. Let X and Y be a pair of jointly normal random variables with joint density (2.2.17). As in Example 2.2.12, define $W = Y - \frac{\rho\sigma_2}{\sigma_1}X$ so that X and W are independent and (2.2.19) holds:

$$Y = \frac{\rho\sigma_2}{\sigma_1}X + W. \tag{2.2.19}$$

In Example 2.2.12, we saw that W is normal with mean $\mu_3 = \mu_2 - \frac{\rho\sigma_2\mu_1}{\sigma_1}$ and variance $\sigma_3^2 = (1-\rho^2)\sigma_2^2$. Let us take the conditioning σ-algebra to be $\mathcal{G} = \sigma(X)$. (When $\mathcal{G}$ is generated by a random variable X, it is customary to write $\mathbb{E}[\cdots|X]$ rather than $\mathbb{E}[\cdots|\sigma(X)]$.) We estimate Y, based on X, using (2.2.19) above and properties (i) (Linearity) and (iv) (Independence) from Theorem 2.3.2 to get the linear regression equation

$$\mathbb{E}[Y|X] = \frac{\rho\sigma_2}{\sigma_1}X + \mathbb{E}W = \frac{\rho\sigma_2}{\sigma_1}(X - \mu_1) + \mu_2. \tag{2.3.26}$$

Note that the right-hand side of (2.3.26) is random but is $\sigma(X)$-measurable (i.e., if we know the information in $\sigma(X)$, which is the same as knowing the value of X, then we can evaluate $\mathbb{E}[Y|X]$). Subtracting (2.3.26) from (2.2.19), we see that the error made by the estimator is

$$Y - \mathbb{E}[Y|X] = W - \mathbb{E}W.$$

The error is random, with expected value zero (the estimator is unbiased), and is independent of the estimate $\mathbb{E}[Y|X]$ (because $\mathbb{E}[Y|X]$ is $\sigma(X)$-measurable and W is independent of $\sigma(X)$). The independence between the error and the conditioning random variable X is a consequence of the joint normality in the example. In general, the error and the conditioning random variable are uncorrelated, but not necessarily independent; see Exercise 2.8. □

The Independence Lemma, Lemma 2.5.3 of Chapter 2 of Volume I, now takes the following more general form.

Lemma 2.3.4 (Independence). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Suppose the random variables $X_1, \ldots, X_K$ are $\mathcal{G}$-measurable and the random variables $Y_1, \ldots, Y_L$ are independent of $\mathcal{G}$. Let $f(x_1, \ldots, x_K, y_1, \ldots, y_L)$ be a function of the dummy variables $x_1, \ldots, x_K$ and $y_1, \ldots, y_L$, and define*

$$g(x_1, \ldots, x_K) = \mathbb{E}f(x_1, \ldots, x_K, Y_1, \ldots, Y_L). \tag{2.3.27}$$

Then

$$\mathbb{E}\big[f(X_1, \ldots, X_K, Y_1, \ldots, Y_L)\big|\mathcal{G}\big] = g(X_1, \ldots, X_K). \tag{2.3.28}$$

As with Lemma 2.5.3 of Volume I, the idea here is that since the information in $\mathcal{G}$ is sufficient to determine the values of $X_1, \ldots, X_K$, we should hold these random variables constant when estimating $f(X_1, \ldots, X_K, Y_1, \ldots, Y_K)$. The other random variables, $Y_1, \ldots, Y_L$, are independent of $\mathcal{G}$, and so we should integrate them out without regard to the information in $\mathcal{G}$. These two steps, holding $X_1, \ldots, X_K$ constant and integrating out $Y_1, \ldots, Y_L$, are accomplished by (2.3.27). We get an estimate that depends on the values of $X_1, \ldots, X_K$ and, to capture this fact, we replaced the dummy (nonrandom) variables $x_1, \ldots, x_K$ by the random variables $X_1, \ldots, X_K$ at the last step. Although Lemma 2.5.3 of Volume I has a relatively simple proof, the proof of Lemma 2.3.4 requires some measure-theoretic ideas beyond the scope of this text, and will not be given.

Example 2.3.3 continued. Continuing with the notation of Example 2.3.3, suppose we want to estimate some function $f(x, y)$ of the random variables X and Y based on knowledge of X. We cannot use the Independence Lemma directly because X and Y are not independent. However, we can write Y as $Y = \frac{\rho\sigma_2}{\sigma_1}X + W$. Because X is $\sigma(X)$-measurable, W is independent of $\sigma(X)$ and W is normal with mean μ_3 and variance σ_3^2, the Independence Lemma tells us how to compute $\mathbb{E}[f(X, Y)|X]$. We should first replace the random variable X by a dummy variable x and then take the expectation (i.e., integrate with respect to the distribution of W). Thus, we define

$$\begin{aligned} g(x) &= \mathbb{E}f\left(x, \frac{\rho\sigma_1}{\sigma_1}x + W\right) \\ &= \frac{1}{\sigma_3\sqrt{2\pi}} \int_{-\infty}^{\infty} f\left(x, \frac{\rho\sigma_1}{\sigma_2}x + w\right) \exp\left\{-\frac{(w-\mu_3)^2}{2\sigma_3^2}\right\} dw. \end{aligned}$$

Then

$$\mathbb{E}\big[f(X,Y)\big|X\big] = g(X).$$

Our final answer is random but $\sigma(X)$-measurable, as it should be. □

We have all the tools required to introduce martingales and Markov processes in a continuous-time framework. The definitions are provided below. Examples will be given after we construct Brownian motion and Itô integrals in the next chapters.

Definition 2.3.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let T be a fixed positive number, and let $\mathcal{F}(t)$, $0 \leq t \leq T$, be a filtration of sub-σ-algebras of $\mathcal{F}$. Consider an adapted stochastic process $M(t)$, $0 \leq t \leq T$.*

(i) If

$$\mathbb{E}[M(t)|\mathcal{F}(s)] = M(s) \text{ for all } 0 \leq s \leq t \leq T,$$

we say this process is a martingale. *It has no tendency to rise or fall.*

(ii) If

$$\mathbb{E}[M(t)|\mathcal{F}(s)] \geq M(s) \text{ for all } 0 \leq s \leq t \leq T,$$

we say this process is a submartingale. *It has no tendency to fall; it may have a tendency to rise.*

(iii) If

$$\mathbb{E}[M(t)|\mathcal{F}(s)] \leq M(s) \text{ for all } 0 \leq s \leq t \leq T,$$

we say this process is a supermartingale. *It has no tendency to rise; it may have a tendency to fall.*

Definition 2.3.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let T be a fixed positive number, and let $\mathcal{F}(t)$, $0 \leq t \leq T$, be a filtration of sub-σ-algebras of $\mathcal{F}$. Consider an adapted stochastic process $X(t)$, $0 \leq t \leq T$. Assume that for all $0 \leq s \leq t \leq T$ and for every nonnegative, Borel-measurable function f, there is another Borel-measurable function g such that*

$$\mathbb{E}[f(X(t))|\mathcal{F}(s)] = g(X(s)). \tag{2.3.29}$$

Then we say that the X is a Markov process.

Remark 2.3.7. In Definition 2.3.6, the function f is permitted to depend on t, and the function g will depend on s. These dependencies are not indicated in (2.3.29) because we wish there to emphasize how the dependence on the sample point ω works (i.e., the right-hand side depends on ω only through the random variable $X(s)$). If we indicate the dependence on time by writing $f(t,x)$ rather than $f(x)$, we can write $f(s,x)$ rather than $g(x)$ (we do not need different symbols f and g because the time variables t and s indicate we are dealing with different functions of x at the different times) and can rewrite (2.3.29) as

$$\mathbb{E}[f(t, X(t))|\mathcal{F}(s)] = f(s, X(s)),\ 0 \leq s \leq t \leq T. \tag{2.3.30}$$

Ultimately, we shall see that when we regard $f(t, x)$ as a function of two variables this way, (2.3.30) implies that it satisfies a partial differential equation. This partial differential equation gives us a way to determine $f(s, x)$ if we know $f(t, x)$. The Black-Scholes-Merton partial differential equation is a special case of this. □

2.4 Summary

In measure-theoretic probability, information is modeled using σ-algebras. The information associated with a σ-algebra $\mathcal{G}$ can be thought of as follows. A random experiment is performed and an outcome ω is determined, but the value of ω is not revealed. Instead, for each set in the σ-algebra $\mathcal{G}$, we are told whether ω is in the set. The more sets there are on $\mathcal{G}$, the more information this provides. If $\mathcal{G}$ is the trivial σ-algebra containing only $\emptyset$ and Ω, this provides no information.

A random variable X is *$\mathcal{G}$-measurable* if and only if the set $\{X \in B\} = \{\omega \in \Omega; X(\omega) \in B\}$ is in $\mathcal{G}$ for every Borel subset of $\mathbb{R}$. In this case, the information in $\mathcal{G}$ is enough to determine the value of the random variable $X(\omega)$, even though it may not be enough to determine the value ω of the outcome of the random experiment.

At the other extreme, the information in a σ-algebra $\mathcal{G}$ may be irrelevant to the determination of the value of X. In this case, we say that $\mathcal{G}$ and X are *independent.* This idea is captured mathematically by Definition 2.2.3, which says that X and $\mathcal{G}$ are independent if, for every set $A \in \mathcal{G}$ and every Borel subset B of $\mathbb{R}$, we have

$$\mathbb{P}\{\omega \in \Omega; \omega \in A \text{ and } X(\omega) \in B\} = \mathbb{P}(A) \cdot \mathbb{P}\{\omega \in \Omega; X(\omega) \in B\}.$$

Two random variables X and Y are independent if and only if the *σ algebra generated by X*, defined to be the collection of sets of the form $\{X \in B\}$, is independent of the σ-algebra generated by Y. In other words, X and Y are independent if and only if

$$\mathbb{P}\{X \in B \text{ and } Y \in C\} = \mathbb{P}\{X \in B\} \cdot \mathbb{P}\{X \in C\} \text{ for all } B \in \mathcal{B}(\mathbb{R}), C \in \mathcal{B}(\mathbb{R}),$$

where $\mathcal{B}(\mathbb{R})$ denotes the σ-algebra of Borel subsets of $\mathbb{R}$. There are several equivalent ways to describe independence between two random variables having to do with factoring the joint cumulative distribution function, factoring the joint moment-generating function, and factoring the joint density (if there is a joint density). These are set out in Theorem 2.2.7. Independence implies uncorrelatedness, but uncorrelated random variables do not need to be independent. Jointly normally distributed random variables (Definition 2.2.11) are uncorrelated if and only if they are independent, but normally distributed random variables do not need to be *jointly* normal.

Often we find ourselves between the two extremes of random variables X that are $\mathcal{G}$-measurable and random variables X that are independent of $\mathcal{G}$. In such a case, the information in $\mathcal{G}$ is relevant to the determination of the value of X but is not sufficient to completely determine it. We then want to use the information in $\mathcal{G}$ to estimate X. We denote our estimate by $\mathbb{E}[X|\mathcal{G}]$ and call this the *conditional expectation of X given $\mathcal{G}$*. This is itself a random variable, but one that is $\mathcal{G}$-measurable (i.e., one that we can evaluate using only the information in $\mathcal{G}$). To be sure this is a good estimate of X, we require that it satisfy the *partial-averaging property* (see Definition 2.3.1(ii)):

$$\int_A \mathbb{E}[X|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) = \int_A X(\omega)\, d\mathbb{P}(\omega) \text{ for every } A \in \mathcal{G}.$$

Conditional expectations behave in many ways like expectations, except that expectations do not depend on ω and conditional expectations do. The principal properties of conditional expectations are provided in Theorem 2.3.2, and these are reported briefly here.

Linearity: $\mathbb{E}[c_1X + c_2Y|\mathcal{G}] = c_1\mathbb{E}[X|\mathcal{G}] + c_2\mathbb{E}[Y|\mathcal{G}]$.

Taking out what is known: $\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}]$ if X is $\mathcal{G}$-measurable.

Iterated conditioning: $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big] = \mathbb{E}[X|\mathcal{H}]$ if $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$.

Independence: $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}X$ if X is independent of $\mathcal{G}$.

Jensen's inequality: $\mathbb{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbb{E}[X|\mathcal{G}])$ if φ is convex.

In continuous-time finance, we work within the framework of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We normally have a fixed final time T and then have a *filtration*, which is a collection of σ-algebras $\{\mathcal{F}(t); 0 \leq t \leq T\}$ indexed by the time variable t. We interpret $\mathcal{F}(t)$ as the information available at time t. For $0 \leq s \leq t \leq T$, every set in $\mathcal{F}(s)$ is also in $\mathcal{F}(t)$. In other words, information increases over time. Within this context, an *adapted stochastic process* is a collection of random variables $\{X(t); 0 \leq t \leq T\}$ also indexed by time such that, for every t, $X(t)$ is $\mathcal{F}(t)$-measurable; the information at time t is sufficient to evaluate the random variable $X(t)$. We think of $X(t)$ as the price of some asset at time t and $\mathcal{F}(t)$ as the information obtained by watching all the prices in the market up to time t.

Two important classes of adapted stochastic processes are *martingales* and *Markov processes*. These are defined in Definitions 2.3.5 and 2.3.6, respectively. A martingale has the property that

$$\mathbb{E}[M(t)|\mathcal{F}(s)] = M(s) \text{ for all } 0 \leq s \leq t \leq T.$$

If $\mathbb{E}[M(t)|\mathcal{F}(s)] \geq M(s)$ when $0 \leq s \leq t \leq T$, we have a *submartingale*. If the inequality is reversed, we have a *supermartingale*. A Markov process has the property that whenever $0 \leq s \leq t \leq T$ and we are given a function f, there is another function g such that

$$\mathbb{E}[f(X(t))|\mathcal{F}(s)] = g(X(s)).$$

The important feature here is that the estimate of $f(X(t))$ made at time s depends only on the process value $X(s)$ at time s and not on the path of the process before time s.

A useful tool for establishing that a process is Markov is the *Independence Lemma*, Lemma 2.3.4. The simplest version of this says that if X is a $\mathcal{G}$-measurable random variable and Y is independent of $\mathcal{G}$, then

$$\mathbb{E}[f(X,Y)|\mathcal{G}] = g(X),$$

where $g(x) = \mathbb{E}f(x,Y)$.

2.5 Notes

In the measure-theoretic view of probability theory, a conditional expectation is itself a random variable, measurable with respect to the conditioning σ-algebra. This point of view is indispensable for treating the rather complicated conditional expectations that arise in martingale theory. It was invented by Kolmogorov [104]. The term *martingale* was apparently first used by Ville [158], who assigned the name to a betting strategy. The concept dates back to 1934 work of Lévy. The first systematic treatment of martingales was provided by Doob [53].

2.6 Exercises

Exercise 2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a general probability space, and suppose a random variable X on this space is measurable with respect to the trivial σ-algebra $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Show that X is not random (i.e., there is a constant c such that $X(\omega) = c$ for all $\omega \in \Omega$). Such a random variable is called *degenerate.*

Exercise 2.2. Independence of random variables can be affected by changes of measure. To illustrate this point, consider the space of two coin tosses $\Omega_2 = \{HH, HT, TH, TT\}$, and let stock prices be given by

$$S_0 = 4,\ S_1(H) = 8,\ S_1(T) = 2,$$
$$S_2(HH) = 16,\ S_2(HT) = S_2(TH) = 4,\ S_2(TT) = 1.$$

Consider two probability measures given by

$$\widetilde{\mathbb{P}}(HH) = \tfrac{1}{4},\ \widetilde{\mathbb{P}}(HT) = \tfrac{1}{4},\ \widetilde{\mathbb{P}}(TH) = \tfrac{1}{4},\ \widetilde{\mathbb{P}}(TT) = \tfrac{1}{4},$$
$$\mathbb{P}(HH) = \tfrac{4}{9},\ \mathbb{P}(HT) = \tfrac{2}{9},\ \mathbb{P}(TH) = \tfrac{2}{9},\ \mathbb{P}(TT) = \tfrac{1}{9}.$$

Define the random variable

$$X = \begin{cases} 1 \text{ if } S_2 = 4, \\ 0 \text{ if } S_2 \neq 4. \end{cases}$$

(i) List all the sets in $\sigma(X)$.
(ii) List all the sets in $\sigma(S_1)$.
(iii) Show that $\sigma(X)$ and $\sigma(S_1)$ are independent under the probability measure $\widetilde{\mathbb{P}}$.
(iv) Show that $\sigma(X)$ and $\sigma(S_1)$ are not independent under the probability measure $\mathbb{P}$.
(v) Under $\mathbb{P}$, we have $\mathbb{P}\{S_1 = 8\} = \frac{2}{3}$ and $\mathbb{P}\{S_1 = 2\} = \frac{1}{3}$. Explain intuitively why, if you are told that $X = 1$, you would want to revise your estimate of the distribution of S_1.

Exercise 2.3 (Rotating the axes). Let X and Y be independent standard normal random variables. Let θ be a constant, and define random variables

$$V = X\cos\theta + Y\sin\theta \text{ and } W = -X\sin\theta + Y\cos\theta.$$

Show that V and W are independent standard normal random variables.

Exercise 2.4. In Example 2.2.8, X is a standard normal random variable and Z is an independent random variable satisfying

$$\mathbb{P}\{Z = 1\} = \mathbb{P}\{Z = -1\} = \frac{1}{2}.$$

We defined $Y = XZ$ and showed that Y is standard normal. We established that although X and Y are uncorrelated, they are not independent. In this exercise, we use moment-generating functions to show that Y is standard normal and X and Y are not independent.

(i) Establish the joint moment-generating function formula

$$\mathbb{E}e^{uX+vY} = e^{\frac{1}{2}(u^2+v^2)} \cdot \frac{e^{uv} + e^{-uv}}{2}.$$

(ii) Use the formula above to show that $\mathbb{E}e^{vY} = e^{\frac{1}{2}v^2}$. This is the moment-generating function for a standard normal random variable, and thus Y must be a standard normal random variable.
(iii) Use the formula in (i) and Theorem 2.2.7(iv) to show that X and Y are not independent.

Exercise 2.5. Let (X, Y) be a pair of random variables with joint density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{2|x|+y}{\sqrt{2\pi}} \exp\left\{-\frac{(2|x|+y)^2}{2}\right\} & \text{if } y \geq -|x|, \\ 0 & \text{if } y < -|x|. \end{cases}$$

Show that X and Y are standard normal random variables and that they are uncorrelated but not independent.

Exercise 2.6. Consider a probability space Ω with four elements, which we call a, b, c, and d (i.e., $\Omega = \{a, b, c, d\}$). The σ-algebra $\mathcal{F}$ is the collection of all subsets of Ω; i.e., the sets in $\mathcal{F}$ are

$$\begin{aligned}&\Omega,\ \{a,b,c\},\ \{a,b,d\},\ \{a,c,d\},\ \{b,c,d\},\\&\{a,b\},\ \{a,c\},\ \{a,d\},\ \{b,c\},\ \{b,d\},\ \{c,d\},\\&\{a\},\ \{b\},\ \{c\},\ \{d\},\ \emptyset.\end{aligned}$$

We define a probability measure $\mathbb{P}$ by specifying that

$$\mathbb{P}\{a\} = \frac{1}{6},\ \mathbb{P}\{b\} = \frac{1}{3},\ \mathbb{P}\{c\} = \frac{1}{4},\ \mathbb{P}\{d\} = \frac{1}{4},$$

and, as usual, the probability of every other set in $\mathcal{F}$ is the sum of the probabilities of the elements in the set, e.g., $\mathbb{P}\{a, b, c\} = \mathbb{P}\{a\} + \mathbb{P}\{b\} + \mathbb{P}\{c\} = \frac{3}{4}$.

We next define two random variables, X and Y, by the formulas

$$\begin{aligned}&X(a) = 1, X(b) = 1, X(c) = -1, X(d) = -1,\\&Y(a) = 1, Y(b) = -1, Y(c) = 1, Y(d) = -1.\end{aligned}$$

We then define $Z = X + Y$.

(i) List the sets in $\sigma(X)$.
(ii) Determine $\mathbb{E}[Y|X]$ (i.e., specify the values of this random variable for a, b, c, and d). Verify that the partial-averaging property is satisfied.
(iii) Determine $\mathbb{E}[Z|X]$. Again, verify the partial-averaging property.
(iv) Compute $\mathbb{E}[Z|X] - \mathbb{E}[Y|X]$. Citing the appropriate properties of conditional expectation from Theorem 2.3.2, explain why you get X.

Exercise 2.7. Let Y be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Based on the information in $\mathcal{G}$, we can form the estimate $\mathbb{E}[Y|\mathcal{G}]$ of Y and define the error of the estimation $\text{Err} = Y - \mathbb{E}[Y|\mathcal{G}]$. This is a random variable with expectation zero and some variance $\text{Var}(\text{Err})$. Let X be some other $\mathcal{G}$-measurable random variable, which we can regard as another estimate of Y. Show that

$$\text{Var}(\text{Err}) \leq \text{Var}(Y - X).$$

In other words, the estimate $\mathbb{E}[Y|\mathcal{G}]$ minimizes the variance of the error among all estimates based on the information in $\mathcal{G}$. (Hint: Let $\mu = \mathbb{E}(Y - X)$. Compute the variance of $Y - X$ as

$$\mathbb{E}\left[(Y - X - \mu)^2\right] = \mathbb{E}\left[\left((Y - \mathbb{E}[Y|\mathcal{G}]) + (\mathbb{E}[Y|\mathcal{G}] - X - \mu)\right)^2\right].$$

Multiply out the right-hand side and use iterated conditioning to show the cross-term is zero.)

Exercise 2.8. Let X and Y be integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $Y = Y_1 + Y_2$, where $Y_1 = \mathbb{E}[Y|X]$ is $\sigma(X)$-measurable and $Y_2 = Y - E[Y|X]$. Show that Y_2 and X are uncorrelated. More generally, show that Y_2 is uncorrelated with every $\sigma(X)$-measurable random variable.

Exercise 2.9. Let X be a random variable.

(i) Give an example of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X defined on this probability space, and a function f so that the σ-algebra generated by $f(X)$ is not the trivial σ-algebra $\{\emptyset, \Omega\}$ but is strictly smaller than the σ-algebra generated by X.

(ii) Can the σ-algebra generated by $f(X)$ ever be strictly larger than the σ-algebra generated by X?

Exercise 2.10. Let X and Y be random variables (on some unspecified probability space $(\Omega, \mathcal{F}, \mathbb{P})$), assume they have a joint density $f_{X,Y}(x, y)$, and assume $\mathbb{E}|Y| < \infty$. In particular, for every Borel subset C of $\mathbb{R}^2$, we have

$$\mathbb{P}\{(X, Y) \in C\} = \int_C f_{X,Y}(x, y)\, dx\, dy.$$

In elementary probability, one learns to compute $\mathbb{E}[Y|X = x]$, which is a *nonrandom* function of the *dummy variable* x, by the formula

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy, \tag{2.6.1}$$

where $f_{Y|X}(y|x)$ is the *conditional density* defined by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

The denominator in this expression, $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, \eta) d\eta$, is the *marginal density* of X, and we must assume it is strictly positive for every x. We introduce the symbol $g(x)$ for the function $\mathbb{E}[Y|X = x]$ defined by (2.6.1); i.e.,

$$g(x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = \int_{-\infty}^{\infty} \frac{y f_{X,Y}(x, y)}{f_X(x)}\, dy.$$

In measure-theoretic probability, conditional expectation is a *random variable* $\mathbb{E}[Y|X]$. This exercise is to show that when there is a joint density for (X, Y), this random variable can be obtained by substituting the random variable X in place of the dummy variable x in the function $g(x)$. In other words, this exercise is to show that

$$\mathbb{E}[Y|X] = g(X).$$

(We introduced the symbol $g(x)$ in order to avoid the mathematically confusing expression $E[Y|X = X]$.)

Since $g(X)$ is obviously $\sigma(X)$-measurable, to verify that $\mathbb{E}[Y|X] = g(X)$, we need only check that the partial-averaging property is satisfied. For every Borel-measurable function h mapping $\mathbb{R}$ to $\mathbb{R}$ and satisfying $\mathbb{E}|h(X)| < \infty$, we have

$$\mathbb{E}h(X) = \int_{-\infty}^{\infty} h(x) f_X(x) dx. \tag{2.6.2}$$

This is Theorem 1.5.2 in Chapter 1. Similarly, if h is a function of both x and y, then

$$\mathbb{E}h(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y) dx dy \tag{2.6.3}$$

whenever (X,Y) has a joint density $f_{X,Y}(x,y)$. You may use both (2.6.2) and (2.6.3) in your solution to this problem.

Let A be a set in $\sigma(X)$. By the definition of $\sigma(X)$, there is a Borel subset B of $\mathbb{R}$ such that $A = \{\omega \in \Omega; X(\omega) \in B\}$ or, more simply, $A = \{X \in B\}$. Show the partial-averaging property

$$\int_A g(X) d\mathbb{P} = \int_A Y d\mathbb{P}.$$

Exercise 2.11. (i) Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let W be a nonnegative $\sigma(X)$-measurable random variable. Show there exists a function g such that $W = g(X)$. (Hint: Recall that every set in $\sigma(X)$ is of the form $\{X \in B\}$ for some Borel set $B \subset \mathbb{R}$. Suppose first that W is the indicator of such a set, and then use the standard machine.)

(ii) Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let Y be a nonnegative random variable on this space. We do not assume that X and Y have a joint density. Nonetheless, show there is a function g such that $\mathbb{E}[Y|X] = g(X)$.

This page intentionally left blank

3

Brownian Motion

3.1 Introduction

In this chapter, we define Brownian motion and develop its basic properties. The definition of Brownian motion is provided in Section 3.3. Section 3.2 precedes it to give some intuition. For us, the most important properties of Brownian motion are that it is a martingale (Theorem 3.3.4) and that it accumulates quadratic variation at rate one per unit time (Theorem 3.4.3). The notion of quadratic variation is profound. It makes stochastic calculus different from ordinary calculus. For this reason, we begin already in Section 3.2 to talk about it.

Sections 3.5–3.7 develop properties of Brownian motion we shall need later but not in the development of stochastic calculus in Chapter 4. Therefore, the reader can go to Chapter 4 after completing Section 3.4. The Markov property is the concept used to relate stochastic calculus to partial differential equations. For Brownian motion, this property is presented in Section 3.5. The first passage time of Brownian motion to a level is presented in Section 3.6 and used in Chapter 8 to analyze a perpetual American put on a geometric Brownian motion. This is in the spirit of the perpetual American put analysis for the binomial model, which is given in Section 5.4 of Volume I. The reflection principle for Brownian motion developed in Section 3.7 is used in Chapter 7 to price exotic options.

3.2 Scaled Random Walks

3.2.1 Symmetric Random Walk

To create a Brownian motion, we begin with a symmetric random walk, one path of which is shown in Figure 3.2.1. To construct a symmetric random walk, we repeatedly toss a fair coin (p, the probability of H on each toss, and $q = 1 - p$, the probability of T on each toss, are both equal to $\frac{1}{2}$). We denote

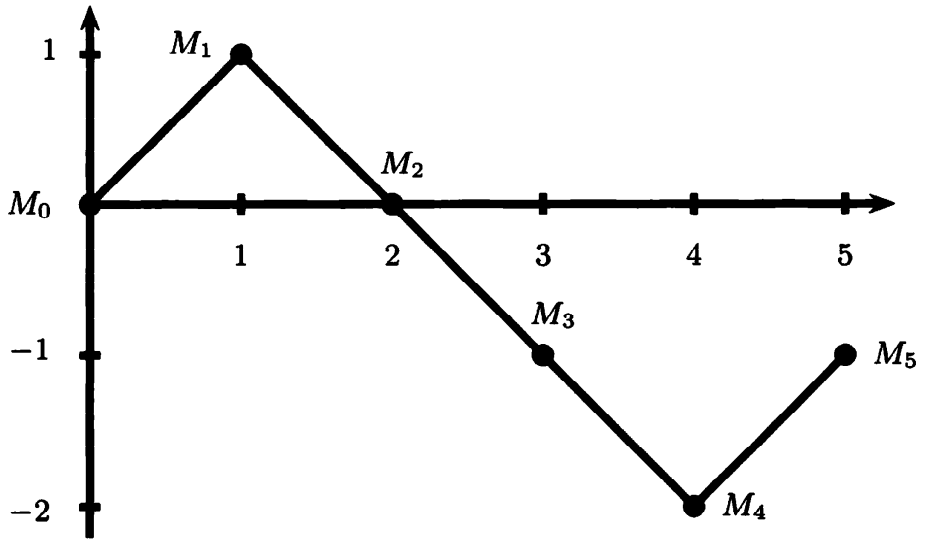

Fig. 3.2.1. Five steps of a random walk.

the successive outcomes of the tosses by $\omega = \omega_1\omega_2\omega_3\ldots$. In other words, ω is the infinite sequence of tosses, and ω_n is the outcome of the nth toss. Let

$$X_j = \begin{cases} 1 & \text{if } \omega_j = H, \\ -1 & \text{if } \omega_j = T, \end{cases} \tag{3.2.1}$$

and define $M_0 = 0$,

$$M_k = \sum_{j=1}^{k} X_j, \quad k = 1, 2, \ldots. \tag{3.2.2}$$

The process M_k, $k = 0, 1, 2, \ldots$ is a *symmetric random walk.* With each toss, it either steps up one unit or down one unit, and each of the two possibilities is equally likely.

3.2.2 Increments of the Symmetric Random Walk

A random walk has *independent increments.* This means that if we choose nonnegative integers $0 = k_0 < k_1 < \cdots < k_m$, the random variables

$$M_{k_1} = (M_{k_1} - M_{k_0}),\ (M_{k_2} - M_{k_1}), \ldots, (M_{k_m} - M_{k_{m-1}})$$

are independent. Each of these random variables,

$$M_{k_{i+1}} - M_{k_i} = \sum_{j=k_i+1}^{k_{i+1}} X_j, \tag{3.2.3}$$

is called an *increment* of the random walk. It is the change in the position of the random walk between times k_i and k_{i+1}. Increments over nonoverlapping time intervals are independent because they depend on different coin tosses.

Moreover, each increment $M_{k_{i+1}} - M_{k_i}$ has expected value 0 and variance $k_{i+1} - k_i$. It is easy to see that the expected value is zero because the expected value of each X_j appearing on the right-hand side of (3.2.3) is zero. We also have $\text{Var}(X_j) = \mathbb{E}X_j^2 = 1$, and since the different X_j are independent, we have from (3.2.3) that

$$\text{Var}\big(M_{k_{i+1}} - M_{k_i}\big) = \sum_{j=k_i+1}^{k_{i+1}} \text{Var}(X_j) = \sum_{j=k_i+1}^{k_{i+1}} 1 = k_{i+1} - k_i. \tag{3.2.4}$$

The variance of the symmetric random walk accumulates at rate one per unit time, so that the variance of the increment over any time interval k to ℓ for nonnegative integers $k < \ell$ is $\ell - k$.

3.2.3 Martingale Property for the Symmetric Random Walk

To see that the symmetric random walk is a martingale, we choose nonnegative integers $k < \ell$ and compute

$$\begin{aligned}
\mathbb{E}\big[M_\ell \big| \mathcal{F}_k\big] &= \mathbb{E}\big[(M_\ell - M_k) + M_k \big| \mathcal{F}_k\big] \\
&= \mathbb{E}\big[M_\ell - M_k \big| \mathcal{F}_k\big] + \mathbb{E}\big[M_k \big| \mathcal{F}_k\big] \\
&= \mathbb{E}\big[M_\ell - M_k \big| \mathcal{F}_k\big] + M_k \\
&= \mathbb{E}\big[M_\ell - M_k\big] + M_k \;=\; M_k.
\end{aligned} \tag{3.2.5}$$

Here we have used the notation $\mathbb{E}[\cdots|\mathcal{F}_k]$ of Chapter 2 to denote the conditional expectation based on the information at time k, which in this case is knowledge of the first k coin tosses. The second equality is a result of the linearity of conditional expectations (Theorem 2.3.2(i)). The third equality is because M_k depends only on the first k coin tosses (it is $\mathcal{F}_k$-measurable, where, in the language of Definition 2.1.5, $\mathcal{F}_k$ is the σ-algebra of information corresponding to the first k coin tosses). The fourth equality follows from independence (Theorem 2.3.2(iv)).

3.2.4 Quadratic Variation of the Symmetric Random Walk

Finally, we consider the *quadratic variation* of the symmetric random walk. The quadratic variation up to time k is defined to be

$$[M, M]_k = \sum_{j=1}^{k} \big(M_j - M_{j-1}\big)^2 = k. \tag{3.2.6}$$

Note that this is computed path-by-path. The quadratic variation up to time k along a path is computed by taking all the one-step increments $M_j - M_{j-1}$ along that path (these are equal to X_j, which is either 1 or -1, depending on

the path), squaring these increments, and then summing them. Since $(M_j - M_{j-1})^2 = 1$, regardless of whether $M_j - M_{j-1}$ is 1 or -1, the sum in (3.2.6) is equal to $\sum_{j=1}^{k} 1 = k$, as reported in that equation.

We note that $[M, M]_k$ is the same as $\text{Var}(M_k)$ (set $k_{i+1} = k$ and $k_i = 0$ in (3.2.4)), but the computations of these two quantities are quite different. $\text{Var}(M_k)$ is computed by taking an average over all paths, taking their probabilities into account. If the random walk were not symmetric (i.e., if p were different from q), this would affect $\text{Var}(M_k)$. By contrast, $[M, M]_k$ is computed along a single path, and the probabilities of up and down steps do not enter the computation. One can compute the variance of a random walk only theoretically because it requires an average over all paths, realized and unrealized. However, from tick-by-tick price data, one can compute the quadratic variation along the realized path rather explicitly. For a random walk, there is the somewhat unusual feature that $[M, M]_k$ does not depend on the particular path chosen, but we shall see later that the quadratic variation for a random process generally does depend on the path along which it is computed.

3.2.5 Scaled Symmetric Random Walk

To approximate a Brownian motion, we speed up time and scale down the step size of a symmetric random walk. More precisely, we fix a positive integer n and define the *scaled symmetric random walk*

$$W^{(n)}(t) = \frac{1}{\sqrt{n}} M_{nt}, \tag{3.2.7}$$

provided nt is itself an integer. If nt is not an integer, we define $W^{(n)}(t)$ by linear interpolation between its values at the nearest points s and u to the left and right of t for which ns and nu are integers. We shall obtain a Brownian motion in the limit as $n \to \infty$. Figure 3.2.2 shows a simulated path of $W^{(100)}$ up to time 4; this was generated by 400 coin tosses with a step up or down of size $\frac{1}{10}$ on each coin toss.

Like the random walk, the scaled random walk has independent increments. If $0 = t_0 < t_1 < \cdots < t_m$ are such that each nt_j is an integer, then

$$\big(W^{(n)}(t_1) - W^{(n)}(t_0)\big),\ \big(W^{(n)}(t_2) - W^{(n)}(t_1)\big), \ldots, \big(W^{(n)}(t_m) - W^{(n)}(t_{m-1})\big)$$

are independent. These random variables depend on different coin tosses. For example, $W^{(100)}(0.20) - W^{(100)}(0)$ depends on the first 20 coin tosses and $W^{(100)}(0.70) - W^{(100)}(0.20)$ depends on the next 50 tosses. Furthermore, if $0 \le s \le t$ are such that ns and nt are integers, then

$$\mathbb{E}\big(W^{(n)}(t) - W^{(n)}(s)\big) = 0, \quad \text{Var}\big(W^{(n)}(t) - W^{(n)}(s)\big) = t - s. \tag{3.2.8}$$

This is because $W^{(n)}(t) - W^{(n)}(s)$ is the sum of $n(t-s)$ independent random variables, each with expected valuc zero and variance $\frac{1}{n}$. For example, $W^{(100)}(0.70) - W^{(100)}(0.20)$ is the sum of 50 independent random

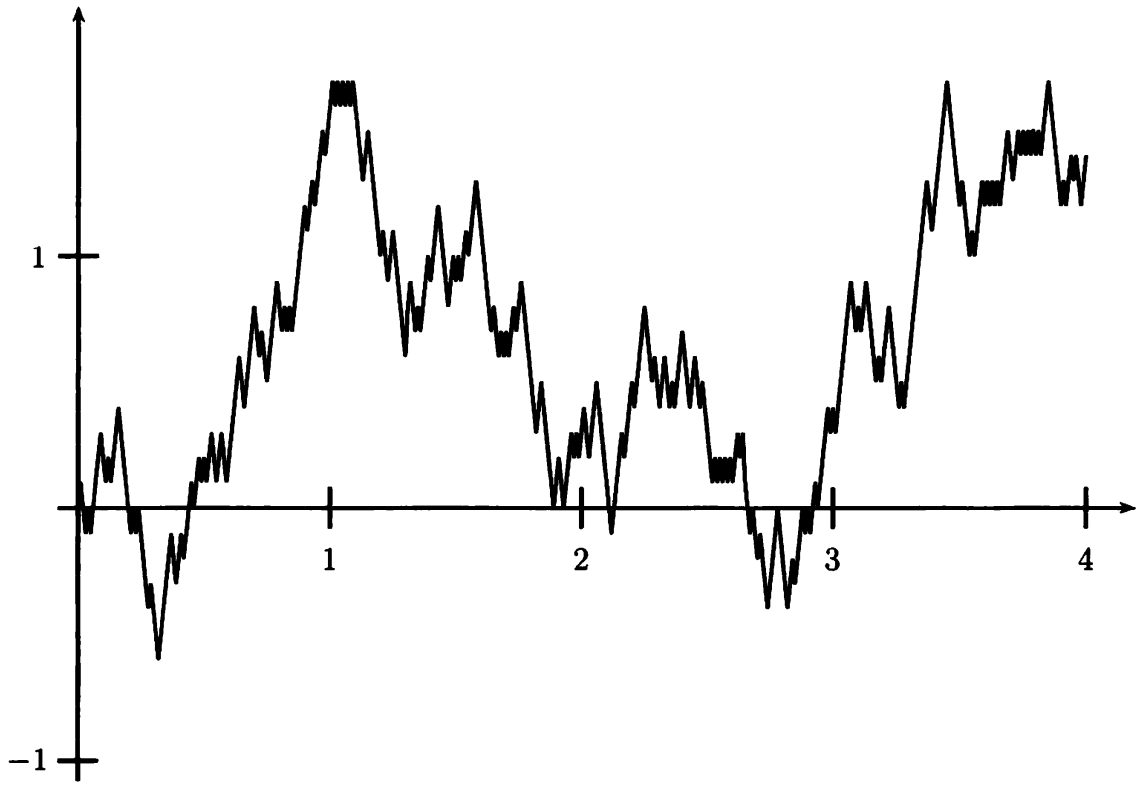

Fig. 3.2.2. A sample path of $W^{(100)}$.

variables, each of which takes the value $\frac{1}{10}$ or $-\frac{1}{10}$. Each of these random variables has expected value zero and variance $\frac{1}{100}$, so the variance of $W^{(100)}(0.70) - W^{(100)}(0.20)$ is $50 \cdot \frac{1}{100} = 0.50$.

Let $0 \leq s \leq t$ be given, and decompose $W^{(n)}(t)$ as

$$W^{(n)}(t) = \big(W^{(n)}(t) - W^{(n)}(s)\big) + W^{(n)}(s).$$

If s and t are chosen so that ns and nt are integers, then the first term on the right-hand side is independent of $\mathcal{F}(s)$, the σ-algebra of information available at time s (which is knowledge of the first ns coin tosses), and $W^{(n)}(s)$ is $\mathcal{F}(s)$-measurable (i.e., it depends only on the first ns coin tosses). We may prove the martingale property for the scaled random walk as we did for the random walk in (3.2.5):

$$\mathbb{E}\big[W^{(n)}(t)\big|\mathcal{F}(s)\big] = W^{(n)}(s) \tag{3.2.9}$$

for $0 \leq s \leq t$ such that ns and nt are integers.

Finally, we consider the *quadratic variation* of the scaled random walk. For $W^{(100)}$, the quadratic variation up to a time, say 1.37, is defined to be

$$\begin{aligned}\big[W^{(100)}, W^{(100)}\big](1.37) &= \sum_{j=1}^{137}\left[W^{(100)}\left(\frac{j}{100}\right) - W^{(100)}\left(\frac{j-1}{100}\right)\right]^2 \\ &= \sum_{j=1}^{137}\left[\frac{1}{10}X_j\right]^2 = \sum_{j=1}^{137}\frac{1}{100} = 1.37.\end{aligned}$$

In general, for $t \geq 0$ such that nt is an integer,

$$[W^{(n)}, W^{(n)}](t) = \sum_{j=1}^{nt} \left[W^{(n)}\left(\frac{j}{n}\right) - W^{(n)}\left(\frac{j-1}{n}\right) \right]^2$$
$$= \sum_{j=1}^{nt} \left[\frac{1}{\sqrt{n}} X_j \right]^2 = \sum_{j=1}^{nt} \frac{1}{n} = t. \tag{3.2.10}$$

If we go from time 0 to time t along the path of the scaled random walk, evaluating the increment over each time step and squaring these increments before summing them, we obtain t, the length of the time interval over which we are doing the computation. This is a path-by-path computation, not an average over all possible paths, and could in principle depend on the particular path along which we do the computation. However, along each path we get the same answer t. Note that $\operatorname{Var} W^{(n)}(t)$ is also t (the second equation in (3.2.8) with $s = 0$), but this latter quantity is an average over all possible paths.

3.2.6 Limiting Distribution of the Scaled Random Walk

In Figure 3.2.2 we see a single sample path of the scaled random walk. In other words, we have fixed a sequence of coin tosses $\omega = \omega_1\omega_2\ldots$ and drawn the path of the resulting process as time t varies. Another way to think about the scaled random walk is to fix the time t and consider the set of all possible paths evaluated at that time t. In other words, we can fix t and think about the scaled random walk corresponding to different values of ω, the sequence of coin tosses. For example, set $t = 0.25$ and consider the set of possible values of $W^{(100)}(0.25) = \frac{1}{10} M_{25}$. This random variable is generated by 25 coin tosses, and since the unscaled random walk M_{25} can take the value of any odd integer between -25 and 25, the scaled random walk $W^{(100)}(0.25)$ can take any of the values

$$-2.5,\ -2.3,\ -2.1,\ \ldots,\ -0.3,\ -0.1,\ 0.1,\ 0.3,\ \ldots,\ 2.1,\ 2.3,\ 2.5.$$

In order for $W^{(100)}(0.25)$ to take the value 0.1, we must get 13 heads and 12 tails in the 25 tosses. The probability of this is

$$\mathbb{P}\{W^{(100)}(0.25) = 0.1\} = \frac{25!}{13!\,12!} \left(\frac{1}{2}\right)^{25} = 0.1555. \tag{3.2.11}$$

We plot this information in Figure 3.2.3 by drawing a histogram bar centered at 0.1 with area 0.1555. Since this bar has width 0.2, its height must be $\frac{0.1555}{0.2} = 0.7775$. Figure 3.2.3 shows similar histogram bars for all possible values of $W^{(100)}(0.25)$ between -1.5 and 1.5.

The random variable $W^{(100)}(0.25)$ has expected value zero and variance 0.25. Superimposed on the histogram in Figure 3.2.3 is the normal density

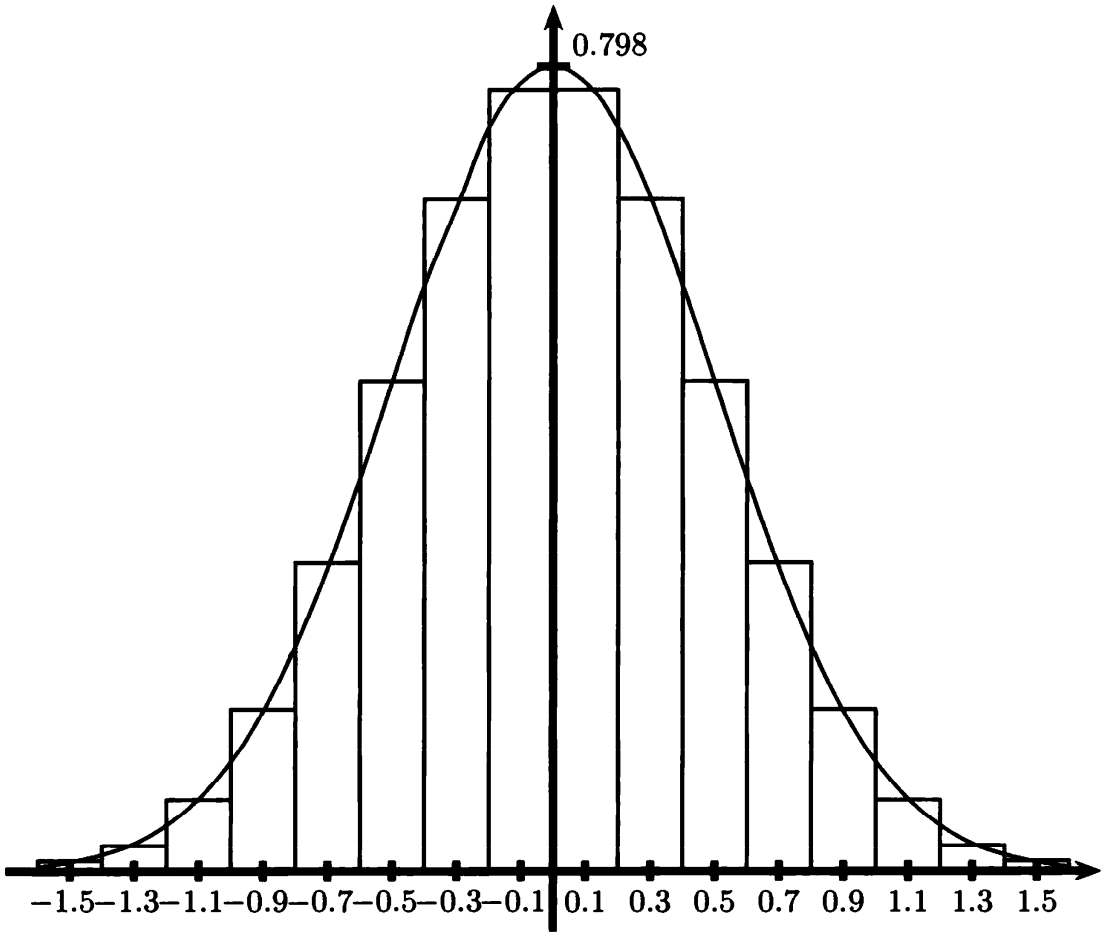

Fig. 3.2.3. Distribution of $W^{(100)}(0.25)$ and normal curve $y = \frac{2}{\sqrt{2\pi}} e^{-2x^2}$.

with this mean and variance. We see that the distribution of $W^{(100)}(0.25)$ is nearly normal. If we were given a continuous bounded function $g(x)$ and asked to compute $\mathbb{E}g\big(W^{(100)}(0.25)\big)$, a good approximation would be obtained by multiplying $g(x)$ by the normal density shown in Figure 3.2.3 and integrating:

$$\mathbb{E}g\big(W^{(100)}(0.25)\big) \approx \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) e^{-2x^2}\, dx. \tag{3.2.12}$$

The Central Limit Theorem asserts that the approximation in (3.2.12) is valid. We provide the version of it that applies to our context.

Theorem 3.2.1 (Central limit). *Fix $t \geq 0$. As $n \to \infty$, the distribution of the scaled random walk $W^{(n)}(t)$ evaluated at time t converges to the normal distribution with mean zero and variance t.*

OUTLINE OF PROOF: One can identify distributions by identifying their moment-generating functions. For the normal density

$$f(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$$

with mean zero and variance t, the moment-generating function is

$$\begin{aligned}\varphi(u) &= \int_{-\infty}^{\infty} e^{ux} f(x)\,dx \\ &= \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} \exp\left\{ux - \frac{x^2}{2t}\right\} dx \\ &= e^{\frac{1}{2}u^2 t} \cdot \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-ut)^2}{2t}\right\} dx \\ &= e^{\frac{1}{2}u^2 t} \end{aligned} \tag{3.2.13}$$

because $\frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-ut)^2}{2t}}$ is a normal density with mean ut and variance t and hence integrates to 1.

If t is such that nt is an integer, then the moment-generating function for $W^{(n)}(t)$ is

$$\begin{aligned}\varphi_n(u) = \mathbb{E}e^{uW^{(n)}(t)} &= \mathbb{E}\exp\left\{\frac{u}{\sqrt{n}} M_{nt}\right\} \\ &= \mathbb{E}\exp\left\{\frac{u}{\sqrt{n}} \sum_{j=1}^{nt} X_j\right\} = \mathbb{E}\prod_{j=1}^{nt} \exp\left\{\frac{u}{\sqrt{n}} X_j\right\}. \end{aligned} \tag{3.2.14}$$

Because the random variables are independent, the right-hand side of (3.2.14) may be written as

$$\prod_{j=1}^{nt} \mathbb{E}\exp\left\{\frac{u}{\sqrt{n}} X_j\right\} = \prod_{j=1}^{nt} \left(\frac{1}{2}e^{\frac{u}{\sqrt{n}}} + \frac{1}{2}e^{-\frac{u}{\sqrt{n}}}\right) = \left(\frac{1}{2}e^{\frac{u}{\sqrt{n}}} + \frac{1}{2}e^{-\frac{u}{\sqrt{n}}}\right)^{nt}.$$

We need to show that, as $n \to \infty$,

$$\varphi_n(u) = \left(\frac{1}{2}e^{\frac{u}{\sqrt{n}}} + \frac{1}{2}e^{-\frac{u}{\sqrt{n}}}\right)^{nt}$$

converges to the moment-generating function $\varphi(u) = e^{\frac{1}{2}u^2 t}$ in (3.2.13). To do this, it suffices to consider the logarithm of $\varphi_n(u)$ and show that

$$\log \varphi_n(u) = nt \log\left(\frac{1}{2}e^{\frac{u}{\sqrt{n}}} + \frac{1}{2}e^{-\frac{u}{\sqrt{n}}}\right)$$

converges to $\log \varphi(u) = \frac{1}{2}u^2 t$.

For this final computation, we make the change of variable $x = \frac{1}{\sqrt{n}}$ so that

$$\lim_{n\to\infty} \log \varphi_n(u) = t \lim_{x \downarrow 0} \frac{\log\left(\frac{1}{2}e^{ux} + \frac{1}{2}e^{-ux}\right)}{x^2}.$$

If we were to substitute $x = 0$ into the expression on the right-hand side, we would obtain $\frac{0}{0}$, and in this situation, we may use L'Hôpital's rule. The derivative of the numerator with respect to x is

$$\frac{\partial}{\partial x}\log\left(\frac{1}{2}e^{ux}+\frac{1}{2}e^{-ux}\right)=\frac{\frac{u}{2}e^{ux}-\frac{u}{2}e^{-ux}}{\frac{1}{2}e^{ux}+\frac{1}{2}e^{-ux}},$$

and the derivative of the denominator is

$$\frac{\partial}{\partial x}x^2=2x.$$

Therefore,

$$\lim_{n\to\infty}\log\varphi_n(u)=t\lim_{x\downarrow 0}\frac{\frac{u}{2}e^{ux}-\frac{u}{2}e^{-ux}}{2x\left(\frac{1}{2}e^{ux}+\frac{1}{2}e^{-ux}\right)}=\frac{t}{2}\lim_{x\downarrow 0}\frac{\frac{u}{2}e^{ux}-\frac{u}{2}e^{-ux}}{x},$$

where we have used the fact that $\lim_{x\downarrow 0}\left(\frac{1}{2}e^{ux}+\frac{1}{2}e^{-ux}\right)=1$. If we were to substitute $x=0$ into the expression on the right-hand side, we would again obtain $\frac{0}{0}$. In this situation, we apply L'Hôpital's rule again. The derivative of the numerator is

$$\frac{\partial}{\partial x}\left(\frac{u}{2}e^{ux}-\frac{u}{2}e^{-ux}\right)=\frac{u^2}{2}e^{ux}+\frac{u^2}{2}e^{-ux},$$

and the derivative of the denominator is $\frac{\partial}{\partial x}x=1$. Hence,

$$\lim_{n\to\infty}\log\varphi_n(u)=\frac{t}{2}\lim_{x\downarrow 0}\left(\frac{u^2}{2}e^{ux}+\frac{u^2}{2}e^{-ux}\right)=\frac{1}{2}u^2t,$$

as desired. □

3.2.7 Log-Normal Distribution as the Limit of the Binomial Model

The Central Limit Theorem, Theorem 3.2.1, can be used to show that the limit of a properly scaled binomial asset-pricing model leads to a stock price with a log-normal distribution. We present this limiting argument here under the assumption that the interest rate r is zero. The case of a nonzero interest rate is outlined in Exercise 3.8. These results show that the binomial model is a discrete-time version of the geometric Brownian motion model, which is the basis for the Black-Scholes-Merton option-pricing formula.

Let us build a model for a stock price on the time interval from 0 to t by choosing an integer n and constructing a binomial model for the stock price that takes n steps per unit time. We assume that n and t are chosen so that nt is an integer. We take the up factor to be $u_n=1+\frac{\sigma}{\sqrt{n}}$ and the down factor to be $d_n=1-\frac{\sigma}{\sqrt{n}}$. Here σ is a positive constant that will turn out to be the volatility of the limiting stock price process. The risk-neutral probabilities are then (see (1.1.8) of Chapter 1 of Volume I)

$$\tilde{p}=\frac{1+r-d_n}{u_n-d_n}=\frac{\sigma/\sqrt{n}}{2\sigma/\sqrt{n}}=\frac{1}{2},\quad \tilde{q}=\frac{u_n-1-r}{u_n-d_n}=\frac{\sigma/\sqrt{n}}{2\sigma/\sqrt{n}}=\frac{1}{2}.$$

The stock price at time t is determined by the initial stock price $S(0)$ and the result of the first nt coin tosses. The sum of the number of heads H_{nt} and number of tails T_{nt} in the first nt coin tosses is nt, a fact that we write as

$$nt = H_{nt} + T_{nt}.$$

The random walk M_{nt} is the number of heads minus the number of tails in these nt coin tosses:

$$M_{nt} = H_{nt} - T_{nt}.$$

Adding these two equations and dividing by 2, we see that

$$H_{nt} = \frac{1}{2}(nt + M_{nt}).$$

Subtracting them and dividing by 2, we see further that

$$T_{nt} = \frac{1}{2}(nt - M_{nt}).$$

In the model with up factor u_n and down factor d_n, the stock price at time t is

$$S_n(t) = S(0)u_n^{H_{nt}}d_n^{T_{nt}} = S(0)\left(1 + \frac{\sigma}{\sqrt{n}}\right)^{\frac{1}{2}(nt+M_{nt})}\left(1 - \frac{\sigma}{\sqrt{n}}\right)^{\frac{1}{2}(nt-M_{nt})}. \tag{3.2.15}$$

We wish to identify the distribution of this random variable as $n \to \infty$.

Theorem 3.2.2. *As $n \to \infty$, the distribution of $S_n(t)$ in (3.2.15) converges to the distribution of*

$$S(t) = S(0)\exp\left\{\sigma W(t) - \frac{1}{2}\sigma^2 t\right\}, \tag{3.2.16}$$

where $W(t)$ is a normal random variable with mean zero and variance t.

The distribution of $S(t)$ in (3.2.16) is called *log-normal*. More generally, any random variable of the form ce^X, where c is a constant and X is normally distributed, is said to have a log-normal distribution. In the case at hand, $X = \sigma W(t) - \frac{1}{2}\sigma^2 t$ is normal with mean $-\frac{1}{2}\sigma^2 t$ and variance $\sigma^2 t$.

PROOF OF THEOREM 3.2.2: It suffices to show that the distribution of

$$\log S_n(t) = \log S(0) + \frac{1}{2}(nt + M_{nt})\log\left(1 + \frac{\sigma}{\sqrt{n}}\right) + \frac{1}{2}(nt - M_{nt})\log\left(1 - \frac{\sigma}{\sqrt{n}}\right) \tag{3.2.17}$$

converges to the distribution of

$$\log S(t) = \log S(0) + \sigma W(t) - \frac{1}{2}\sigma^2 t,$$

where $W(t)$ is a normal random variable with mean zero and variance t. To do this, we need the Taylor series expansion of $f(x) = \log(1+x)$. We compute $f'(x) = (1+x)^{-1}$ and $f''(x) = -(1+x)^{-2}$ and evaluate them to obtain $f'(0) = 1$ and $f''(0) = -1$. According to Taylor's Theorem,

$$\log(1+x) = f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + O(x^3) = x - \frac{1}{2}x^2 + O(x^3),$$

where $O(x^3)$ indicates a term of order x^3. We apply this to (3.2.17) first with $x = \frac{\sigma}{\sqrt{n}}$ and then with $x = -\frac{\sigma}{\sqrt{n}}$. Our intention is to then let $n \to \infty$, and so we need to keep track of which terms have powers of n in the denominator and which terms do not. The former ones will have limit zero and the latter ones will not. We use the $O(\cdot)$ notation to do this. Not every term of the form $O(n^{-\frac{3}{2}})$ in the following equation is the same; their only common feature is that they have $n^{\frac{3}{2}}$ in their denominators. In particular, from (3.2.17) we have

$$\begin{aligned}
\log S(t) &= \log S(0) + \frac{1}{2}(nt + M_{nt})\left(\frac{\sigma}{\sqrt{n}} - \frac{\sigma^2}{2n} + O(n^{-\frac{3}{2}})\right) \\
&\quad + \frac{1}{2}(nt - M_{nt})\left(-\frac{\sigma}{\sqrt{n}} - \frac{\sigma^2}{2n} + O(n^{-\frac{3}{2}})\right) \\
&= \log S(0) + nt\left(-\frac{\sigma^2}{2n} + O(n^{-\frac{3}{2}})\right) + M_{nt}\left(\frac{\sigma}{\sqrt{n}} + O(n^{-\frac{3}{2}})\right) \\
&= \log S(0) - \frac{1}{2}\sigma^2 t + O(n^{-\frac{1}{2}}) + \sigma W^{(n)}(t) + O(n^{-1})W^{(n)}(t).
\end{aligned}$$

The term $W^{(n)}(t) = \frac{1}{\sqrt{n}}M_{nt}$ appears in two places in the last line. By the Central Limit Theorem, Theorem 3.2.1, its distribution converges to the distribution of a normal random variable with mean zero and variance t, a random variable we call $W(t)$. However, in one of its appearances, $W^{(n)}(t)$ is multiplied by a term that has n in the denominator, and this will have limit zero. The term $O(n^{-\frac{1}{2}})$ also has limit zero as $n \to \infty$. We conclude that as $n \to \infty$ the distribution of $\log S(t)$ approaches the distribution of $\log S(0) - \frac{1}{2}\sigma^2 t + \sigma W(t)$, which is what we set out to prove. □

3.3 Brownian Motion

3.3.1 Definition of Brownian Motion

We obtain Brownian motion as the limit of the scaled random walks $W^{(n)}(t)$ of (3.2.7) as $n \to \infty$. The Brownian motion inherits properties from these random walks. This leads to the following definition.

Definition 3.3.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For each $\omega \in \Omega$, suppose there is a continuous function $W(t)$ of $t \geq 0$ that satisfies $W(0) = 0$ and that depends on ω. Then $W(t)$, $t \geq 0$, is a* Brownian motion *if for all $0 = t_0 < t_1 < \cdots < t_m$ the increments*

$$W(t_1) = W(t_1) - W(t_0),\ W(t_2) - W(t_1),\ \ldots,\ W(t_m) - W(t_{m-1}) \quad (3.3.1)$$

are independent and each of these increments is normally distributed with

$$\mathbb{E}\big[W(t_{i+1}) - W(t_i)\big] = 0, \quad (3.3.2)$$

$$\mathrm{Var}\big[W(t_{i+1}) - W(t_i)\big] = t_{i+1} - t_i. \quad (3.3.3)$$

One difference between Brownian motion $W(t)$ and a scaled random walk, say $W^{(100)}(t)$, is that the scaled random walk has a natural time step $\frac{1}{100}$ and is linear between these time steps, whereas the Brownian motion has no linear pieces. The other difference is that, while the scaled random walk $W^{(100)}(t)$ is only approximately normal for each t (see Figure 3.2.3), the Brownian motion is exactly normal. This is a consequence of the Central Limit Theorem, Theorem 3.2.1. Not only is $W(t) = W(t) - W(0)$ normally distributed for each t, but the increments $W(t) - W(s)$ are normally distributed for all $0 \leq s < t$.

There are two ways to think of ω in Definition 3.3.1. One is to think of ω as the Brownian motion path. A random experiment is performed, and its outcome is the path of the Brownian motion. Then $W(t)$ is the value of this path at time t, and this value of course depends on which path resulted from the random experiment. Alternatively, one can think of ω as something more primitive than the path itself, akin to the outcome of a sequence of coin tosses, although now the coin is being tossed "infinitely fast." Once the sequence of coin tosses has been performed and the result ω obtained, then the path of the Brownian motion can be drawn. If the tossing is done again and a different ω is obtained, then a different path will be drawn.

In either case, the sample space Ω is the set of all possible outcomes of a random experiment, $\mathcal{F}$ is the σ-algebra of subsets of Ω whose probabilities are defined, and $\mathbb{P}$ is a probability measure. For each $A \in \mathcal{F}$, the probability of A is a number $\mathbb{P}(A)$ between zero and one. The distributional statements about Brownian motion pertain to $\mathbb{P}$.

For example, we might wish to determine the probability of the set A containing all $\omega \in \Omega$ that result in a Brownian motion path satisfying $0 \leq W(0.25) \leq 0.2$. Let us first consider this matter for the scaled random walk $W^{(100)}$. If we were asked to determine the set $\{\omega : 0 \leq W^{(100)}(0.25) \leq 0.2\}$, we would note that in order for the scaled random walk $W^{(100)}$ to fall between 0 and 0.2 at time 0.25, the unscaled random walk $M_{25} = 10W^{(100)}(0.25)$ must fall between 0 and 2 after 25 tosses. Since M_{25} can only be an odd number, it falls between 0 and 2 if and only if it is equal to 1 or, equivalently, if and only if $W^{(100)}(0.25) = 0.1$. To achieve this, the coin tossing must result in 13 heads and 12 tails in the first 25 tosses. Therefore, A is the set of all infinite sequences of coin tosses with the property that in the first 25 tosses there

are 13 heads and 12 tails. The probability that one of these sequences occurs, given by (3.2.11), is $\mathbb{P}(A) = 0.1555$.

For the Brownian motion W, there is also a set of outcomes ω to the random experiment that results in a Brownian motion path satisfying $0 \leq W(0.25) \leq 0.2$. We choose not to describe this set as concretely as we just did for the scaled random walk $W^{(100)}$. Nonetheless, there is such a set of $\omega \in \Omega$, and the probability of this set is

$$\mathbb{P}\{0 \leq W(0.25) \leq 0.2\} = \frac{2}{\sqrt{2\pi}} \int_0^{0.2} e^{-2x^2} dx.$$

In place of the area in the histogram bar centered at 0.1 in Figure 3.2.3, which is 0.1555, we now have the area under the normal curve between 0 and 0.2 in that figure. These two areas are nearly the same.

3.3.2 Distribution of Brownian Motion

Because the increments

$$W(t_1) = W(t_1) - W(t_0), W(t_2) - W(t_1), \ldots, W(t_m) - W(t_{m-1})$$

of (3.3.1) are independent and normally distributed, the random variables $W(t_1), W(t_2), \ldots, W(t_m)$ are jointly normally distributed. The joint distribution of jointly normal random variables is determined by their means and covariances. Each of the random variables $W(t_i)$ has mean zero. For any two times, $0 \leq s < t$, the covariance of $W(s)$ and $W(t)$ is

$$\begin{aligned}
\mathbb{E}\big[W(s)W(t)\big] &= \mathbb{E}\big[W(s)\big(W(t) - W(s)\big) + W^2(s)\big] \\
&= \mathbb{E}\big[W(s)\big] \cdot \mathbb{E}\big[W(t) - W(s)\big] + \mathbb{E}\big[W^2(s)\big] \\
&= 0 + \mathrm{Var}\big[W(s)\big] = s,
\end{aligned}$$

where we have used the independence of $W(s)$ and $W(t) - W(s)$ in the second equality. Hence, the *covariance matrix for Brownian motion* (i.e., for the m-dimensional random vector $\big(W(t_1), W(t_2), \ldots, W(t_m)\big)$) is

$$\begin{bmatrix}
\mathbb{E}\big[W^2(t_1)\big] & \mathbb{E}\big[W(t_1)W(t_2)\big] & \cdots & \mathbb{E}\big[W(t_1)W(t_m)\big] \\
\mathbb{E}\big[W(t_2)W(t_1)\big] & \mathbb{E}\big[W^2(t_2)\big] & \cdots & \mathbb{E}\big[W(t_2)W(t_m)\big] \\
\vdots & \vdots & & \vdots \\
\mathbb{E}\big[W(t_m)W(t_1)\big] & \mathbb{E}\big[W(t_m)W(t_2)\big] & \cdots & \mathbb{E}\big[W^2(t_m)\big]
\end{bmatrix} = \begin{bmatrix}
t_1 & t_1 & \cdots & t_1 \\
t_1 & t_2 & \cdots & t_2 \\
\vdots & \vdots & & \vdots \\
t_1 & t_2 & \cdots & t_m
\end{bmatrix}. \tag{3.3.4}$$

The moment-generating function of this random vector can be computed using the moment-generating function (3.2.13) for a zero-mean normal random variable with variance t and the independence of the increments in (3.3.1). To assist in this computation, we note first that

$$
\begin{aligned}
&u_3W(t_3)+u_2W(t_2)+u_1W(t_1)\\
&=u_3\big(W(t_3)-W(t_2)\big)+(u_2+u_3)\big(W(t_2)-W(t_1)\big)\\
&\qquad+(u_1+u_2+u_3)W(t_1)
\end{aligned}
$$

and more generally

$$
\begin{aligned}
&u_mW(t_m)+u_{m-1}W(t_{m-1})+u_{m-2}W(t_{m-2})+\cdots+u_1W(t_1)\\
&=u_m\big(W(t_m)-W(t_{m-1})\big)+(u_{m-1}+u_m)\big(W(t_{m-1})-W(t_{m-2})\big)\\
&\qquad+(u_{m-2}+u_{m-1}+u_m)\big(W(t_{m-2})-W(t_{m-3})\big)+\dots\\
&\qquad\cdots+(u_1+u_2+\cdots+u_m)W(t_1).
\end{aligned}
$$

We use these facts to compute the moment-generating function of the random vector $\big(W(t_1),W(t_2),\dots,W(t_m)\big)$:

$$
\begin{aligned}
&\varphi(u_1,u_2,\dots,u_m)\\
&=\mathbb{E}\exp\{u_mW(t_m)+u_{m-1}W(t_{m-1})+\cdots+u_1W(t_1)\}\\
&=\mathbb{E}\exp\big\{u_m\big(W(t_m)-W(t_{m-1})\big)+(u_{m-1}+u_m)\big(W(t_{m-1})-W(t_{m-2})\big)+\\
&\qquad\cdots+(u_1+u_2+\cdots+u_m)W(t_1)\big\}\\
&=\mathbb{E}\exp\big\{u_m\big(W(t_m)-W(t_{m-1})\big)\big\}\\
&\qquad\cdot\mathbb{E}\exp\big\{(u_{m-1}+u_m)\big(W(t_{m-1})-W(t_{m-2})\big)\big\}\\
&\qquad\cdots\mathbb{E}\exp\big\{(u_1+u_2+\cdots+u_m)W(t_1)\big\}\\
&=\exp\left\{\frac{1}{2}u_m^2(t_m-t_{m-1})\right\}\cdot\exp\left\{\frac{1}{2}(u_{m-1}+u_m)^2(t_{m-1}-t_{m-2})\right\}\\
&\qquad\cdots\exp\left\{\frac{1}{2}(u_1+u_2+\cdots+u_m)^2t_1\right\}.
\end{aligned}
$$

In conclusion, the *moment-generating function for Brownian motion* (i.e., for the m-dimensional random vector $\big(W(t_1),W(t_2),\dots,W(t_m)\big)$) is

$$
\begin{aligned}
&\varphi(u_1,u_2,\dots,u_m)\\
&=\mathbb{E}\exp\{u_mW(t_m)+u_{m-1}W(t_{m-1})+\cdots+u_1W(t_1)\}\\
&=\exp\left\{\frac{1}{2}(u_1+u_2+\cdots+u_m)^2t_1+\frac{1}{2}(u_2+u_3+\cdots+u_m)^2(t_2-t_1)+\right.\\
&\qquad\left.\cdots+\frac{1}{2}(u_{m-1}+u_m)^2(t_{m-1}-t_{m-2})+\frac{1}{2}u_m^2(t_m-t_{m-1})\right\}. \qquad (3.3.5)
\end{aligned}
$$

The distribution of the Brownian increments in (3.3.1) can be specified by specifying the joint density or the joint moment-generating function of the random variables $W(t_1),W(t_2),\dots,W(t_m)$. This leads to the following theorem.

Theorem 3.3.2 (Alternative characterizations of Brownian motion). *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. For each $\omega\in\Omega$, suppose there is a*

continuous function $W(t)$ of $t \geq 0$ that satisfies $W(0) = 0$ and that depends on ω. The following three properties are equivalent.

(i) For all $0 = t_0 < t_1 < \cdots < t_m$, the increments

$$W(t_1) = W(t_1) - W(t_0),\ W(t_2) - W(t_1),\ \ldots,\ W(t_m) - W(t_{m-1})$$

are independent and each of these increments is normally distributed with mean and variance given by (3.3.2) and (3.3.3).

(ii) For all $0 = t_0 < t_1 < \cdots < t_m$, the random variables $W(t_1), W(t_2), \ldots, W(t_m)$ are jointly normally distributed with means equal to zero and covariance matrix (3.3.4).

(iii) For all $0 = t_0 < t_1 < \cdots < t_m$, the random variables $W(t_1), W(t_2), \ldots, W(t_m)$ have the joint moment-generating function (3.3.5).

If any of (i), (ii), or (iii) holds (and hence they all hold), then $W(t)$, $t \geq 0$, is a Brownian motion.

3.3.3 Filtration for Brownian Motion

In addition to the Brownian motion itself, we will need some notation for the amount of information available at each time. We do that with a filtration.

Definition 3.3.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which is defined a Brownian motion $W(t)$, $t \geq 0$. A* filtration for the Brownian motion *is a collection of σ-algebras $\mathcal{F}(t)$, $t \geq 0$, satisfying:*

(i) **(Information accumulates)** *For $0 \leq s < t$, every set in $\mathcal{F}(s)$ is also in $\mathcal{F}(t)$. In other words, there is at least as much information available at the later time $\mathcal{F}(t)$ as there is at the earlier time $\mathcal{F}(s)$.*

(ii) **(Adaptivity)** *For each $t \geq 0$, the Brownian motion $W(t)$ at time t is $\mathcal{F}(t)$-measurable. In other words, the information available at time t is sufficient to evaluate the Brownian motion $W(t)$ at that time.*

(iii) **(Independence of future increments)** *For $0 \leq t < u$, the increment $W(u) - W(t)$ is independent of $\mathcal{F}(t)$. In other words, any increment of the Brownian motion after time t is independent of the information available at time t.*

Let $\Delta(t)$, $t \geq 0$, be a stochastic process. We say that $\Delta(t)$ is adapted *to the filtration $\mathcal{F}(t)$ if for each $t \geq 0$ the random variable $\Delta(t)$ is $\mathcal{F}(t)$-measurable.*[1]

Properties (i) and (ii) in the definition above guarantee that the information available at each time t is at least as much as one would learn from observing the Brownian motion up to time t. Property (iii) says that this

[1] The adapted processes we encounter will serve as integrands, and for this one needs them to be jointly measurable in t and ω so that their integrals are defined and are themselves adapted processes. This is a technical requirement that we shall ignore in this text.

information is of no use in predicting future movements of the Brownian motion. In the asset-pricing models we build, property (iii) leads to the efficient market hypothesis.

There are two possibilities for the filtration $\mathcal{F}(t)$ for a Brownian motion. One is to let $\mathcal{F}(t)$ contain only the information obtained by observing the Brownian motion itself up to time t. The other is to include in $\mathcal{F}(t)$ information obtained by observing the Brownian motion and one or more other processes. However, if the information in $\mathcal{F}(t)$ includes observations of processes other than the Brownian motion W, this additional information is not allowed to give clues about the future increments of W because of property (iii).

3.3.4 Martingale Property for Brownian Motion

Theorem 3.3.4. *Brownian motion is a martingale.*

PROOF: Let $0 \leq s \leq t$ be given. Then

$$\begin{aligned}
\mathbb{E}\big[W(t)\big|\mathcal{F}(s)\big] &= \mathbb{E}\big[\big(W(t) - W(s)\big) + W(s)\big|\mathcal{F}(s)\big] \\
&= \mathbb{E}\big[W(t) - W(s)\big|\mathcal{F}(s)\big] + \mathbb{E}\big[W(s)\big|\mathcal{F}(s)\big] \\
&= \mathbb{E}\big[W(t) - W(s)\big] + W(s) \\
&= W(s).
\end{aligned}$$

The justifications for the steps in this equality are the same as the justifications for (3.2.5). □

3.4 Quadratic Variation

We computed the quadratic variation of the scaled random walk $W^{(n)}$ up to time T in (3.2.10), and this quadratic variation turned out to be T. This was computed by taking each of the steps of the scaled random walk between times 0 and T, squaring them, and summing them.

For Brownian motion, there is no natural step size. If we are given $T > 0$, we could simply choose a step size, say $\frac{T}{n}$ for some large n, and compute the quadratic variation up to time T with this step size. In other words, we could compute

$$\sum_{j=0}^{n-1} \left[W\left(\frac{(j+1)T}{n}\right) - W\left(\frac{jT}{n}\right) \right]^2. \tag{3.4.1}$$

We are interested in this quantity for small step sizes, and so as a last step we could evaluate the limit as $n \to \infty$. If we do this, we will get T, the same final answer as for the scaled random walk in (3.2.10). This is proved in Theorem 3.4.3 below.

The paths of Brownian motion are unusual in that their quadratic variation is not zero. This makes stochastic calculus different from ordinary calculus and is the source of the volatility term in the Black-Scholes-Merton partial differential equation. These matters will be discussed in the next chapter.

3.4.1 First-Order Variation

Before proving that Brownian motion accumulates T units of quadratic variation between times 0 and T, we digress slightly to discuss *first-order variation* (as opposed to *quadratic variation*, which is *second-order variation*). Consider the function $f(t)$ in Figure 3.4.1. We wish to compute the amount of up and down oscillation undergone by this function between times 0 and T, with the down moves adding to rather than subtracting from the up moves. We call this the *first-order variation* $\mathrm{FV}_T(f)$. For the function f shown, it is

$$\begin{aligned}\mathrm{FV}_T(f) &= [f(t_1) - f(0)] - [f(t_2) - f(t_1)] + [f(T) - f(t_2)] \\ &= \int_0^{t_1} f'(t)\,dt + \int_{t_1}^{t_2} (-f'(t))\,dt + \int_{t_2}^{T} f'(t)\,dt \\ &= \int_0^T |f'(t)|\,dt. \end{aligned} \tag{3.4.2}$$

The middle term

$$-[f(t_2) - f(t_1)] = f(t_1) - f(t_2)$$

is included in a way that guarantees that the magnitude of the down move of the function $f(t)$ between times t_1 and t_2 is added to rather than subtracted from the total.

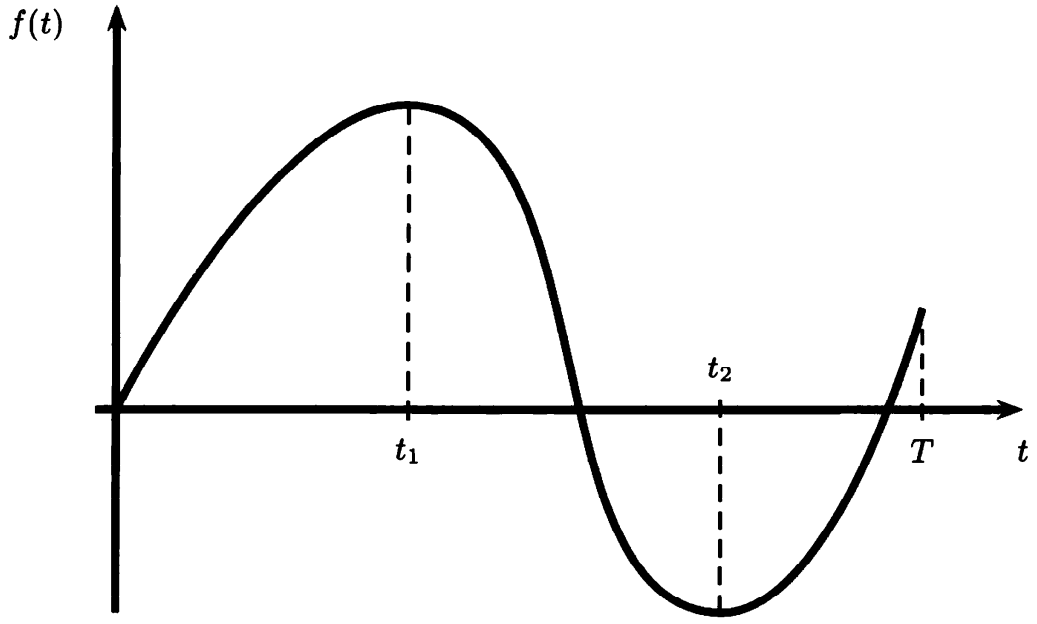

Fig. 3.4.1. Computing the first-order variation.

In general, to compute the first-order variation of a function up to time T, we first choose a *partition* $\Pi = \{t_0, t_1, \ldots, t_n\}$ of $[0, T]$, which is a set of times

$$0 = t_0 < t_1 < \cdots < t_n = T.$$

These will serve to determine the step size. We do not require the partition points $t_0 = 0, t_1, t_2, \ldots, t_n = T$ to be equally spaced, although they are allowed to be. The maximum step size of the partition will be denoted $\|\Pi\| = \max_{j=0,\ldots,n-1}(t_{j+1} - t_j)$. We then define

$$\mathrm{FV}_T(f) = \lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} |f(t_{j+1}) - f(t_j)|. \tag{3.4.3}$$

The limit in (3.4.3) is taken as the number n of partition points goes to infinity and the length of the longest subinterval $t_{j+1} - t_j$ goes to zero.

Our first task is to verify that the definition (3.4.3) is consistent with the formula (3.4.2) for the function shown in Figure 3.4.1. To do this, we use the Mean Value Theorem, which applies to any function $f(t)$ whose derivative $f'(t)$ is defined everywhere. The Mean Value Theorem says that in each subinterval $[t_j, t_{j+1}]$ there is a point t_j^* such that

$$\frac{f(t_{j+1}) - f(t_j)}{t_{j+1} - t_j} = f'(t_j^*). \tag{3.4.4}$$

In other words, somewhere between t_j and t_{j+1}, the tangent line is parallel to the chord connecting the points $\big(t_j, f(t_j)\big)$ and $\big(t_{j+1}, f(t_{j+1})\big)$ (see Figure 3.4.2).

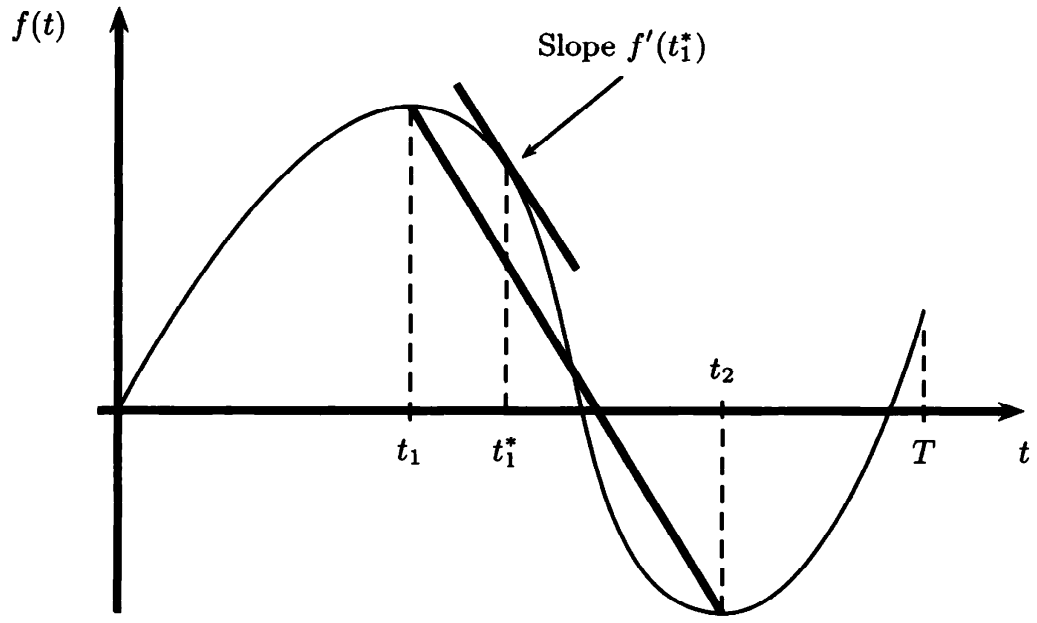

Fig. 3.4.2. Mean Value Theorem.

Multiplying (3.4.4) by $t_{j+1} - t_j$, we obtain

$$f(t_{j+1}) - f(t_j) = f'(t_j^*)(t_{j+1} - t_j).$$

The sum on the right-hand side of (3.4.3) may thus be written as

$$\sum_{j=0}^{n-1} |f'(t_j^*)|(t_{j+1} - t_j),$$

which is a Riemann sum for the integral of the function $|f'(t)|$. Therefore,

$$\mathrm{FV}_T(f) = \lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} |f'(t_j^*)|(t_{j+1} - t_j) = \int_0^T |f'(t)|\, dt,$$

and we have rederived (3.4.2).

3.4.2 Quadratic Variation

Definition 3.4.1. *Let $f(t)$ be a function defined for $0 \le t \le T$. The* quadratic variation *of f up to time T is*

$$[f, f](T) = \lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} [f(t_{j+1}) - f(t_j)]^2, \tag{3.4.5}$$

where $\Pi = \{t_0, t_1, \ldots, t_n\}$ and $0 = t_0 < t_1 < \cdots < t_n = T$.

Remark 3.4.2. Suppose the function f has a continuous derivative. Then

$$\sum_{j=0}^{n-1} [f(t_{j+1}) - f(t_j)]^2 = \sum_{j=0}^{n-1} |f'(t_j^*)|^2 (t_{j+1} - t_j)^2 \le \|\Pi\| \cdot \sum_{j=0}^{n-1} |f'(t_j^*)|^2 (t_{j+1} - t_j),$$

and thus

$$\begin{aligned}
[f, f](T) &\le \lim_{\|\Pi\|\to 0} \left[\|\Pi\| \cdot \sum_{j=0}^{n-1} |f'(t_j^*)|^2 (t_{j+1} - t_j) \right] \\
&= \lim_{\|\Pi\|\to 0} \|\Pi\| \cdot \lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} |f'(t_j^*)|^2 (t_{j+1} - t_j) \\
&= \lim_{\|\Pi\|\to 0} \|\Pi\| \cdot \int_0^T |f'(t)|^2\, dt \;=\; 0.
\end{aligned}$$

In the last step of this argument, we use the fact that $f'(t)$ is continuous to ensure that $\int_0^T |f'(t)|^2 dt$ is finite. If $\int_0^T |f'(t)|^2 dt$ is infinite, then

$$\lim_{\|\Pi\|\to 0} \left[\|\Pi\| \cdot \sum_{j=0}^{n-1} |f'(t_j^*)|^2 (t_{j+1} - t_j) \right]$$

leads to a $0 \cdot \infty$ situation, which can be anything between 0 and ∞. □

Most functions have continuous derivatives, and hence their quadratic variations are zero. For this reason, one never considers quadratic variation in ordinary calculus. The paths of Brownian motion, on the other hand, cannot be differentiated with respect to the time variable. For functions that do not have derivatives, the Mean Value Theorem can fail and Remark 3.4.2 no longer applies. Consider, for example, the absolute value function $f(t) = |t|$ in Figure 3.4.3. The chord connecting $\big(t_1, f(t_1)\big)$ and $\big(t_2, f(t_2)\big)$ has slope $\frac{1}{5}$, but nowhere between t_1 and t_2 does the derivative of $f(t) = |t|$ equal $\frac{1}{5}$. Indeed, this derivative is always -1 for $t < 0$, is always 1 for $t > 0$, and is undefined at $t = 0$, where the the graph of the function $f(t) = |t|$ has a "point." Figure 3.2.2 suggests correctly that the paths of Brownian motion are very "pointy." Indeed, for a Brownian motion path $W(t)$, there is no value of t for which $\frac{d}{dt}W(t)$ is defined.

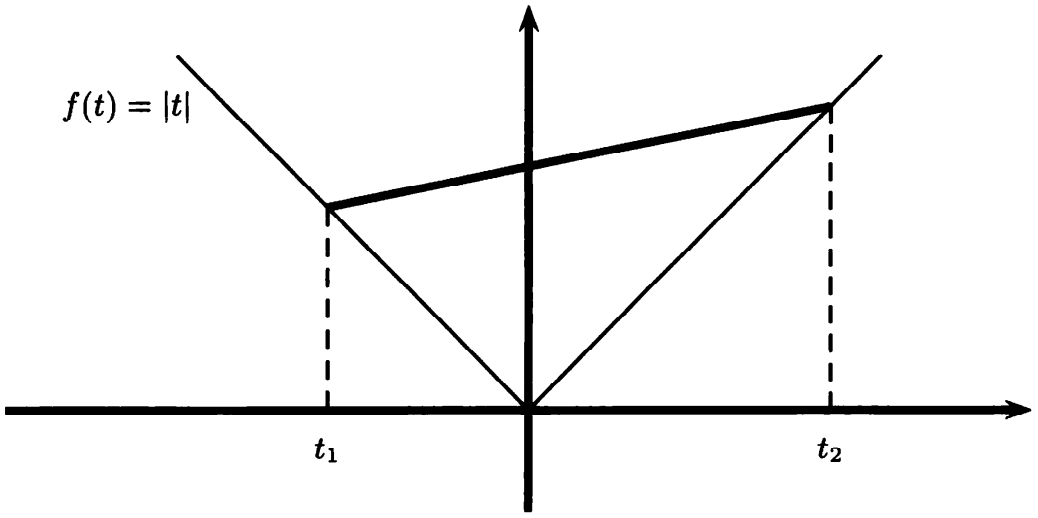

Fig. 3.4.3. Absolute value function.

Theorem 3.4.3. *Let W be a Brownian motion. Then $[W,W](T) = T$ for all $T \geq 0$ almost surely.*

We recall that the terminology *almost surely* means that there can be some paths of the Brownian motion for which the assertion $[W,W](T) = T$ is not true. However, the set of all such paths has zero probability. The set of paths for which the assertion of the theorem is true has probability one.

PROOF OF THEOREM 3.4.3: Let $\Pi = \{t_0, t_1, \ldots, t_n\}$ be a partition of $[0,T]$. Define the *sampled quadratic variation* corresponding to this partition to be

$$Q_\Pi = \sum_{j=0}^{n-1} \big(W(t_{j+1}) - W(t_j)\big)^2.$$

We must show that this sampled quadratic variation, which is a random variable (i.e., it depends on the path of the Brownian motion along which it is

computed) converges to T as $\|\Pi\| \to 0$. We shall show that it has expected value T, and its variance converges to zero. Hence, it converges to its expected value T, regardless of the path along which we are doing the computation.[2]

The sampled quadratic variation is the sum of independent random variables. Therefore, its mean and variance are the sums of the means and variances of these random variables. We have

$$\mathbb{E}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] = \mathrm{Var}\left[W(t_{j+1}) - W(t_j)\right] = t_{j+1} - t_j, \tag{3.4.6}$$

which implies

$$\mathbb{E}Q_\Pi = \sum_{j=0}^{n-1} \mathbb{E}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] = \sum_{j=0}^{n-1} (t_{j+1} - t_j) = T,$$

as desired. Moreover,

$$\begin{aligned}
&\mathrm{Var}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] \\
&\quad = \mathbb{E}\left[\left(\left(W(t_{j+1}) - W(t_j)\right)^2 - (t_{j+1} - t_j)\right)^2\right] \\
&\quad = \mathbb{E}\left[\left(W(t_{j+1}) - W(t_j)\right)^4\right] - 2(t_{j+1} - t_j)\mathbb{E}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] \\
&\qquad\quad + (t_{j+1} - t_j)^2.
\end{aligned}$$

The fourth moment of a normal random variable with zero mean is three times its variance squared (see Exercise 3.3). Therefore,

$$\mathbb{E}\left[\left(W(t_{j+1}) - W(t_j)\right)^4\right] = 3(t_{j+1} - t_j)^2,$$

$$\begin{aligned}
\mathrm{Var}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] &= 3(t_{j+1} - t_j)^2 - 2(t_{j+1} - t_j)^2 + (t_{j+1} - t_j)^2 \\
&= 2(t_{j+1} - t_j)^2,
\end{aligned} \tag{3.4.7}$$

and

$$\begin{aligned}
\mathrm{Var}(Q_\Pi) &= \sum_{j=0}^{n-1} \mathrm{Var}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] = \sum_{j=0}^{n-1} 2(t_{j+1} - t_j)^2 \\
&\le \sum_{j=0}^{n-1} 2\|\Pi\|(t_{j+1} - t_j) = 2\|\Pi\|T.
\end{aligned}$$

In particular, $\lim_{\|\Pi\|\to 0} \mathrm{Var}(Q_\Pi) = 0$, and we conclude that $\lim_{\|\Pi\|\to 0} Q_\Pi = \mathbb{E}Q_\Pi = T$. □

[2] The convergence we prove is actually *convergence in mean square*, also called L^2*-convergence*. When this convergence takes place, there is a subsequence along which the convergence is *almost sure* (i.e., the convergence takes place for all paths except for a set of paths having probability zero). We shall not dwell on subtle differences among types of convergence of random variables.

Remark 3.4.4. In the proof above, we derived the equations (3.4.6) and (3.4.7):

$$\mathbb{E}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] = t_{j+1} - t_j$$

and

$$\mathrm{Var}\left[\left(W(t_{j+1}) - W(t_j)\right)^2\right] = 2(t_{j+1} - t_j)^2.$$

It is tempting to argue that when $t_{j+1} - t_j$ is small, $(t_{j+1} - t_j)^2$ is *very* small, and therefore $\left(W(t_{j+1}) - W(t_j)\right)^2$, although random, is with high probability near its mean $t_{j+1} - t_j$. We could therefore claim that

$$\left(W(t_{j+1}) - W(t_j)\right)^2 \approx t_{j+1} - t_j. \tag{3.4.8}$$

This approximation is trivially true because, when $t_{j+1} - t_j$ is small, both sides are near zero. It would also be true if we squared the right-hand side, multiplied the right-hand side by 2, or made any of several other significant changes to the right-hand side. In other words, (3.4.8) really has no content. A better way to try to capture what we think is going on is to write

$$\frac{\left(W(t_{j+1}) - W(t_j)\right)^2}{t_{j+1} - t_j} \approx 1 \tag{3.4.9}$$

instead of (3.4.8). However,

$$\frac{\left(W(t_{j+1}) - W(t_j)\right)^2}{t_{j+1} - t_j}$$

is in fact not near 1, regardless of how small we make $t_{j+1} - t_j$. It is the square of the standard normal random variable

$$Y_{j+1} = \frac{W(t_{j+1}) - W(t_j)}{\sqrt{t_{j+1} - t_j}},$$

and its distribution is the same, no matter how small we make $t_{j+1} - t_j$.

To understand better the idea behind Theorem 3.4.3, we choose a large value of n and take $t_j = \frac{jT}{n}$, $j = 0, 1, \ldots, n$. Then $t_{j+1} - t_j = \frac{T}{n}$ for all j and

$$\left(W(t_{j+1}) - W(t_j)\right)^2 = T \cdot \frac{Y_{j+1}^2}{n}.$$

Since the random variables $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed, the Law of Large Numbers implies that $\sum_{j=0}^{n-1} \frac{Y_{j+1}^2}{n}$ converges to the common mean $\mathbb{E}Y_{j+1}^2$ as $n \to \infty$. This mean is 1, and hence $\sum_{j=0}^{n-1} \left(W(t_{j+1}) - W(t_j)\right)^2$ converges to T. Each of the terms $\left(W(t_{j+1}) - W(t_j)\right)^2$ in this sum can be quite different from its mean $t_{j+1} - t_j = \frac{T}{n}$, but when we sum many terms like this, the differences average out to zero.

We write informally

$$dW(t)\,dW(t) = dt, \tag{3.4.10}$$

but this should not be interpreted to mean either (3.4.8) or (3.4.9). It is only when we sum both sides of (3.4.9) and call upon the Law of Large Numbers to cancel errors that we get a correct statement. The statement is that on an interval $[0,T]$, Brownian motion accumulates T units of quadratic variation.

If we compute the quadratic variation of Brownian motion over the time interval $[0,T_1]$, we get $[W,W](T_1) = T_1$. If we compute the quadratic variation over $[0,T_2]$, where $0 < T_1 < T_2$, we get $[W,W](T_2) = T_2$. Therefore, if we partition the interval $[T_1,T_2]$, square the increments of Brownian motion for each of the subintervals in the partition, sum the squared increments, and take the limit as the maximal step size approaches zero, we will get the limit $[W,W](T_2) - [W,W](T_1) = T_2 - T_1$. Brownian motion accumulates $T_2 - T_1$ units of quadratic variation over the interval $[T_1,T_2]$. Since this is true for every interval of time, we conclude that

Brownian motion accumulates quadratic variation at rate one per unit time.

We write (3.4.10) to record this fact. In particular, the dt on the right-hand side of (3.4.10) is multiplied by an understood 1.

As mentioned earlier, the quadratic variation of Brownian motion is the source of volatility in asset prices driven by Brownian motion. We shall eventually scale Brownian motion, sometimes in time- and path-dependent ways, in order to vary the rate at which volatility enters these asset prices. □

Remark 3.4.5. Let $\Pi = \{t_0, t_1, \ldots, t_n\}$ be a partition of $[0,T]$ (i.e., $0 = t_0 < t_1 < \cdots < t_n = T$). In addition to computing the quadratic variation of Brownian motion

$$\lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} \big(W(t_{j+1}) - W(t_j)\big)^2 = T, \tag{3.4.11}$$

we can compute the cross variation of $W(t)$ with t and the quadratic variation of t with itself, which are

$$\lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} \big(W(t_{j+1}) - W(t_j)\big)(t_{j+1} - t_j) = 0, \tag{3.4.12}$$

$$\lim_{\|\Pi\|\to 0} \sum_{j=0}^{n-1} (t_{j+1} - t_j)^2 = 0. \tag{3.4.13}$$

To see that 0 is the limit in (3.4.12), we observe that

$$\Big|\big(W(t_{j+1}) - W(t_j)\big)(t_{j+1} - t_j)\Big| \le \max_{0\le k\le n-1} \big|W(t_{k+1}) - W(k)\big|(t_{j+1} - t_j),$$

and so

$$\left|\sum_{j=0}^{n-1}\big(W(t_{j+1})-W(t_j)\big)(t_{j+1}-t_j)\right| \le \max_{0\le k\le n-1}\left|W(t_{k+1})-W(t_k)\right|\cdot T.$$

Since W is continuous, $\max_{0\le k\le n-1}\left|W(t_{k+1})-W(k)\right|$ has limit zero as $\|\Pi\|$, the length of the longest subinterval, goes to zero. To see that 0 is the limit in (3.4.13), we observe that

$$\sum_{j=0}^{n-1}(t_{j+1}-t_j)^2 \le \max_{0\le k\le n-1}(t_{k+1}-t_k)\cdot\sum_{j=0}^{n-1}(t_{j+1}-t_j) = \|\Pi\|\cdot T,$$

which obviously has limit zero as $\|\Pi\|\to 0$.

Just as we capture (3.4.11) by writing (3.4.10), we capture (3.4.12) and (3.4.13) by writing

$$dW(t)\,dt = 0, \quad dt\,dt = 0. \tag{3.4.14}$$

□

3.4.3 Volatility of Geometric Brownian Motion

Let α and $\sigma > 0$ be constants, and define the *geometric Brownian motion*

$$S(t) = S(0)\exp\left\{\sigma W(t) + \left(\alpha - \frac{1}{2}\sigma^2\right)t\right\}.$$

This is the asset-price model used in the Black-Scholes-Merton option-pricing formula. Here we show how to use the quadratic variation of Brownian motion to identify the volatility σ from a path of this process.

Let $0 \le T_1 < T_2$ be given, and suppose we observe the geometric Brownian motion $S(t)$ for $T_1 \le t \le T_2$. We may then choose a partition of this interval, $T_1 = t_0 < t_2 < \cdots < t_m = T_2$, and observe "log returns"

$$\log\frac{S(t_{j+1})}{S(t_j)} = \sigma\big(W(t_{j+1})-W(t_j)\big) + \left(\alpha-\frac{1}{2}\sigma^2\right)(t_{j+1}-t_j)$$

over each of the subintervals $[t_j, t_{j+1}]$. The sum of the squares of the log returns, sometimes called the *realized volatility*, is

$$\begin{aligned}
&\sum_{j=0}^{m-1}\left(\log\frac{S(t_{j+1})}{S(t_j)}\right)^2 \\
&= \sigma^2\sum_{j=0}^{m-1}\big(W(t_{j+1})-W(t_j)\big)^2 + \left(\alpha-\frac{1}{2}\sigma^2\right)^2\sum_{j=0}^{m-1}(t_{j+1}-t_j)^2 \\
&\quad +2\sigma\left(\alpha-\frac{1}{2}\sigma^2\right)\sum_{j=0}^{m-1}\big(W(t_{j+1})-W(t_j)\big)(t_{j+1}-t_j).
\end{aligned} \tag{3.4.15}$$

When the maximum step size $\|\Pi\| = \max_{j=0,\ldots,m-1}(t_{j+1} - t_j)$ is small, then the first term on the right-hand side of (3.4.15) is approximately equal to its limit, which is σ^2 times the amount of quadratic variation accumulated by Brownian motion on the interval $[T_1, T_2]$, which is $T_2 - T_1$. The second term on the right-hand side of (3.4.15) is $\left(\alpha - \frac{1}{2}\sigma^2\right)^2$ times the quadratic variation of t, which was shown in Remark 3.4.5 to be zero. The third term on the right-hand side of (3.4.15) is $2\sigma\left(\alpha - \frac{1}{2}\sigma^2\right)$ times the cross variation of $W(t)$ and t, which was also shown in Remark 3.4.5 to be zero. We conclude that when the maximum step size $\|\Pi\|$ is small, the right-hand side of (3.4.15) is approximately equal to $\sigma^2(T_2 - T_1)$, and hence

$$\frac{1}{T_2 - T_1} \sum_{j=0}^{m-1} \left(\log \frac{S(t_{j+1})}{S(t_j)}\right)^2 \approx \sigma^2. \tag{3.4.16}$$

If the asset price $S(t)$ really is a geometric Brownian motion with constant volatility σ, then σ can be identified from price observations by computing the left-hand side of (3.4.16) and taking the square root. In theory, we can make this approximation as accurate as we like by decreasing the step size. In practice, there is a limit to how small the step size can be. Between trades, there is no information about prices, and when a trade takes place, it is sometimes at the bid price and sometimes at the ask price. On small time intervals, the difference in prices due to the bid–ask spread can be as large as the difference due to price fluctuations during the time interval.

3.5 Markov Property

In this section, we show that Brownian motion is a Markov process and discuss its *transition density.*

Theorem 3.5.1. *Let $W(t)$, $t \geq 0$, be a Brownian motion and let $\mathcal{F}(t)$, $t \geq 0$, be a filtration for this Brownian motion (see Definition 3.3.3). Then $W(t)$, $t \geq 0$, is a Markov process.*

PROOF: According to Definition 2.3.6, we must show that whenever $0 \leq s \leq t$ and f is a Borel-measurable function, there is another Borel-measurable function g such that

$$\mathbb{E}\big[f\big(W(t)\big)\big|\mathcal{F}(s)\big] = g\big(W(s)\big). \tag{3.5.1}$$

To do this, we write

$$\mathbb{E}\big[f\big(W(t)\big)\big|\mathcal{F}(s)\big] = \mathbb{E}\big[f\big((W(t) - W(s)) + W(s)\big)\big|\mathcal{F}(s)\big]. \tag{3.5.2}$$

The random variable $W(t) - W(s)$ is independent of $\mathcal{F}(s)$, and the random variable $W(s)$ is $\mathcal{F}(s)$-measurable. This permits us to apply the Independence

Lemma, Lemma 2.3.4. In order to compute the expectation on the right-hand side of (3.5.2), we replace $W(s)$ by a dummy variable x to hold it constant and then take the unconditional expectation of the remaining random variable (i.e., we define $g(x) = \mathbb{E}f\big(W(t) - W(s) + x\big)$). But $W(t) - W(s)$ is normally distributed with mean zero and variance $t - s$. Therefore,

$$g(x) = \frac{1}{\sqrt{2\pi(t-s)}} \int_{-\infty}^{\infty} f(w+x) e^{-\frac{w^2}{2(t-s)}}\, dw. \tag{3.5.3}$$

The Independence Lemma states that if we now take the function $g(x)$ defined by (3.5.3) and replace the dummy variable x by the random variable $W(s)$, then equation (3.5.1) holds. □

We may make the change of variable $\tau = t - s$ and $y = w + x$ in (3.5.3) to obtain

$$g(x) = \frac{1}{\sqrt{2\pi\tau}} \int_{-\infty}^{\infty} f(y) e^{-\frac{(y-x)^2}{2\tau}}\, dy.$$

We define the *transition density* $p(\tau, x, y)$ for Brownian motion to be

$$p(\tau, x, y) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(y-x)^2}{2\tau}},$$

so that we may further rewrite (3.5.3) as

$$g(x) = \int_{-\infty}^{\infty} f(y) p(\tau, x, y)\, dy \tag{3.5.4}$$

and (3.5.1) as

$$\mathbb{E}\big[f\big(W(t)\big)\big|\mathcal{F}(s)\big] = \int_{-\infty}^{\infty} f(y) p(\tau, W(s), y)\, dy. \tag{3.5.5}$$

This equation has the following interpretation. Conditioned on the information in $\mathcal{F}(s)$ (which contains all the information obtained by observing the Brownian motion up to and including time s), the conditional density of $W(t)$ is $p(\tau, W(s), y)$. This is a density in the variable y. This density is normal with mean $W(s)$ and variance $\tau = t - s$. In particular, the only information from $\mathcal{F}(s)$ that is relevant is the value of $W(s)$. The fact that only $W(s)$ is relevant is the essence of the Markov property.

3.6 First Passage Time Distribution

In Chapter 5 of Volume I, we studied the first passage time for a random walk, first using the optional sampling theorem for martingales to obtain the distribution in Section 5.2 and then rederiving the distribution using the reflection

principle in Section 5.3. Here we develop the first approach; the second is presented in the next section. In Sections 5.2 and 5.3 of Volume I, we observed after deriving the distribution of the first passage time for the symmetric random walk that our answer could easily be modified to obtain the first passage distribution for an asymmetric random walk. In this section, we work only with Brownian motion, the continuous-time counterpart of the symmetric random walk. The case of Brownian motion with drift, the continuous-time counterpart of an asymmetric random walk, is treated in Exercise 3.7. We revisit this problem in Chapter 7, where it is solved using Girsanov's Theorem. The resulting formulas often provide explicit pricing and hedging formulas for exotic options. Examples of the application of these formulas to such options are given in Chapter 7.

Just as we began in Section 5.2 of Volume I with a martingale that had the random walk in the exponential function, we must begin here with a martingale containing Brownian motion in the exponential function. We fix a constant σ. The so-called *exponential martingale* corresponding to σ, which is

$$Z(t) = \exp\left\{\sigma W(t) - \frac{1}{2}\sigma^2 t\right\}, \tag{3.6.1}$$

plays a key role in much of the remainder of this text.

Theorem 3.6.1 (Exponential martingale). *Let $W(t)$, $t \geq 0$, be a Brownian motion with a filtration $\mathcal{F}(t)$, $t \geq 0$, and let σ be a constant. The process $Z(t)$, $t \geq 0$, of (3.6.1) is a martingale.*

PROOF: For $0 \leq s \leq t$, we have

$$\begin{aligned}
&\mathbb{E}\big[Z(t)\big|\mathcal{F}(s)\big] \\
&= \mathbb{E}\left[\exp\left\{\sigma W(t) - \frac{1}{2}\sigma^2 t\right\}\middle|\,\mathcal{F}(s)\right] \\
&= \mathbb{E}\left[\exp\big\{\sigma\big(W(t) - W(s)\big)\big\} \cdot \exp\left\{\sigma W(s) - \frac{1}{2}\sigma^2 t\right\}\middle|\,\mathcal{F}(s)\right] \\
&= \exp\left\{\sigma W(s) - \frac{1}{2}\sigma^2 t\right\} \cdot \mathbb{E}\left[\exp\big\{\sigma\big(W(t) - W(s)\big)\big\}\big|\,\mathcal{F}(s)\right], \qquad (3.6.2)
\end{aligned}$$

where we have used "taking out what is known" (Theorem 2.3.2(ii)) for the last step. We next use "independence" (Theorem 2.3.2(iv)) to write

$$\mathbb{E}\left[\exp\big\{\sigma\big(W(t) - W(s)\big)\big\}\big|\,\mathcal{F}(s)\right] = \mathbb{E}\left[\exp\big\{\sigma\big(W(t) - W(s)\big)\big\}\right].$$

Because $W(t) - W(s)$ is normally distributed with mean zero and variance $t - s$, this expected value is $\exp\left\{\frac{1}{2}\sigma^2(t-s)\right\}$ (see (3.2.13)). Substituting this into (3.6.2), we obtain the martingale property

$$\mathbb{E}\big[Z(t)\big|\mathcal{F}(s)\big] = \exp\left\{\sigma W(s) - \frac{1}{2}\sigma^2 s\right\} = Z(s). \qquad \square$$

Let m be a real number, and define the *first passage time* to level m

$$\tau_m = \min\{t \geq 0; W(t) = m\}. \tag{3.6.3}$$

This is the first time the Brownian motion W reaches the level m. If the Brownian motion never reaches the level m, we set $\tau_m = \infty$. A martingale that is stopped ("frozen" would be a more apt description) at a stopping time is still a martingale and thus must have constant expectation. (The text following Theorem 4.3.2 of Volume I discusses this in more detail.) Because of this fact,

$$1 = Z(0) = \mathbb{E}Z(t \wedge \tau_m) = \mathbb{E}\left[\exp\left\{\sigma W(t \wedge \tau_m) - \frac{1}{2}\sigma^2(t \wedge \tau_m)\right\}\right], \tag{3.6.4}$$

where the notation $t \wedge \tau_m$ denotes the minimum of t and τ_m.

For the next step, we assume that $\sigma > 0$ and $m > 0$. In this case, the Brownian motion is always at or below level m for $t \leq \tau_m$ and so

$$0 \leq \exp\{\sigma W(t \wedge \tau_m)\} \leq e^{\sigma m}. \tag{3.6.5}$$

If $\tau_m < \infty$, the term $\exp\left\{-\frac{1}{2}\sigma^2(t \wedge \tau_m)\right\}$ is equal to $\exp\left\{-\frac{1}{2}\sigma^2\tau_m\right\}$ for large enough t. On the other hand, if $\tau_m = \infty$, then the term $\exp\left\{-\frac{1}{2}\sigma^2(t \wedge \tau_m)\right\}$ is equal to $\exp\left\{-\frac{1}{2}\sigma^2 t\right\}$, and as $t \to \infty$, this converges to zero. We capture these two cases by writing

$$\lim_{t\to\infty} \exp\left\{-\frac{1}{2}\sigma^2(t \wedge \tau_m)\right\} = \mathbb{I}_{\{\tau_m<\infty\}} \exp\left\{-\frac{1}{2}\sigma^2\tau_m\right\},$$

where the notation $\mathbb{I}_{\{\tau_m<\infty\}}$ is used to indicate the random variable that takes the value 1 if $\tau_m < \infty$ and otherwise takes the value zero. If $\tau_m < \infty$, then $\exp\{\sigma W(t \wedge \tau_m)\} = \exp\{\sigma W(\tau_m)\} = e^{\sigma m}$ when t becomes large enough. If $\tau_m = \infty$, then we do not know what happens to $\exp\{\sigma W(t \wedge \tau_m)\}$ as $t \to \infty$, but we at least know that this term is bounded because of (3.6.5). That is enough to ensure that the product of $\exp\{\sigma W(t \wedge \tau_m)\}$ and $\exp\left\{-\frac{1}{2}\sigma^2\tau_m\right\}$ has limit zero in this case. In conclusion, we have

$$\lim_{t\to\infty} \exp\left\{\sigma W(t \wedge \tau_m) - \frac{1}{2}\sigma^2(t \wedge \tau_m)\right\} = \mathbb{I}_{\{\tau_m<\infty\}} \exp\left\{\sigma m - \frac{1}{2}\sigma^2\tau_m\right\}.$$

We can now take the limit in (3.6.4)[3] to obtain

$$1 = \mathbb{E}\left[\mathbb{I}_{\{\tau_m<\infty\}} \exp\left\{\sigma m - \frac{1}{2}\sigma^2\tau_m\right\}\right]$$

or, equivalently,

[3] The interchange of limit and expectation implicit in this step is justified by the Dominated Convergence Theorem, Theorem 1.4.9.

$$\mathbb{E}\left[\mathbb{I}_{\{\tau_m<\infty\}}\exp\left\{-\frac{1}{2}\sigma^2\tau_m\right\}\right] = e^{-\sigma m}. \tag{3.6.6}$$

Equation (3.6.6) holds when m and σ are positive. We may not substitute $\sigma = 0$ into this equation, but since it holds for every positive σ, we may take the limit on both sides as $\sigma \downarrow 0$. This yields[4] $\mathbb{E}\left[\mathbb{I}_{\{\tau_m<\infty\}}\right] = 1$ or, equivalently,

$$\mathbb{P}\{\tau_m < \infty\} = 1. \tag{3.6.7}$$

Because τ_m is finite with probability one (we say τ_m is finite *almost surely*), we may drop the indicator of this event in (3.6.6) to obtain

$$\mathbb{E}\left[\exp\left\{-\frac{1}{2}\sigma^2\tau_m\right\}\right] = e^{-\sigma m}. \tag{3.6.8}$$

We have done the hard work in the proof of the following theorem.

Theorem 3.6.2. *For $m \in \mathbb{R}$, the first passage time of Brownian motion to level m is finite almost surely, and the Laplace transform of its distribution is given by*

$$\mathbb{E}e^{-\alpha\tau_m} = e^{-|m|\sqrt{2\alpha}} \text{ for all } \alpha > 0. \tag{3.6.9}$$

PROOF: We consider first the case when m is positive. Let α be a positive constant, and set $\sigma = \sqrt{2\alpha}$, so that $\frac{1}{2}\sigma^2 = \alpha$. Then (3.6.8) becomes (3.6.9). If m is negative, then because Brownian motion is symmetric, the first passage times τ_m and $\tau_{|m|}$ have the same distribution. Equation (3.6.9) for negative m follows. □

Remark 3.6.3. Differentiation of (3.6.9) with respect to α results in

$$\mathbb{E}\left[\tau_m e^{-\alpha\tau_m}\right] = \frac{|m|}{\sqrt{2\alpha}}e^{-|m|\sqrt{2\alpha}} \text{ for all } \alpha > 0.$$

Letting $\alpha \downarrow 0$, we obtain $\mathbb{E}\tau_m = \infty$ so long as $m \neq 0$.

3.7 Reflection Principle

3.7.1 Reflection Equality

In this section, we repeat for Brownian motion the reflection principle argument of Section 5.3 of Volume I for the random walk. The reader may wish to review that section before reading this one.

We fix a positive level m and a positive time t. We wish to "count" the Brownian motion paths that reach level m at or before time t (i.e., those paths for which the first passage time τ_m to level m is less than or equal to t). There are two types of such paths: those that reach level m prior to t but at time t are at some level w below m, and those that exceed level m at time t. There are also Brownian motion paths that are exactly at level m at time t, but unlike the case of the random walk in Section 5.3 of Volume I, the probability of this for Brownian motion is zero. We may thus ignore this possibility.

[4] Here we use the Monotone Convergence Theorem, Theorem 1.4.5.

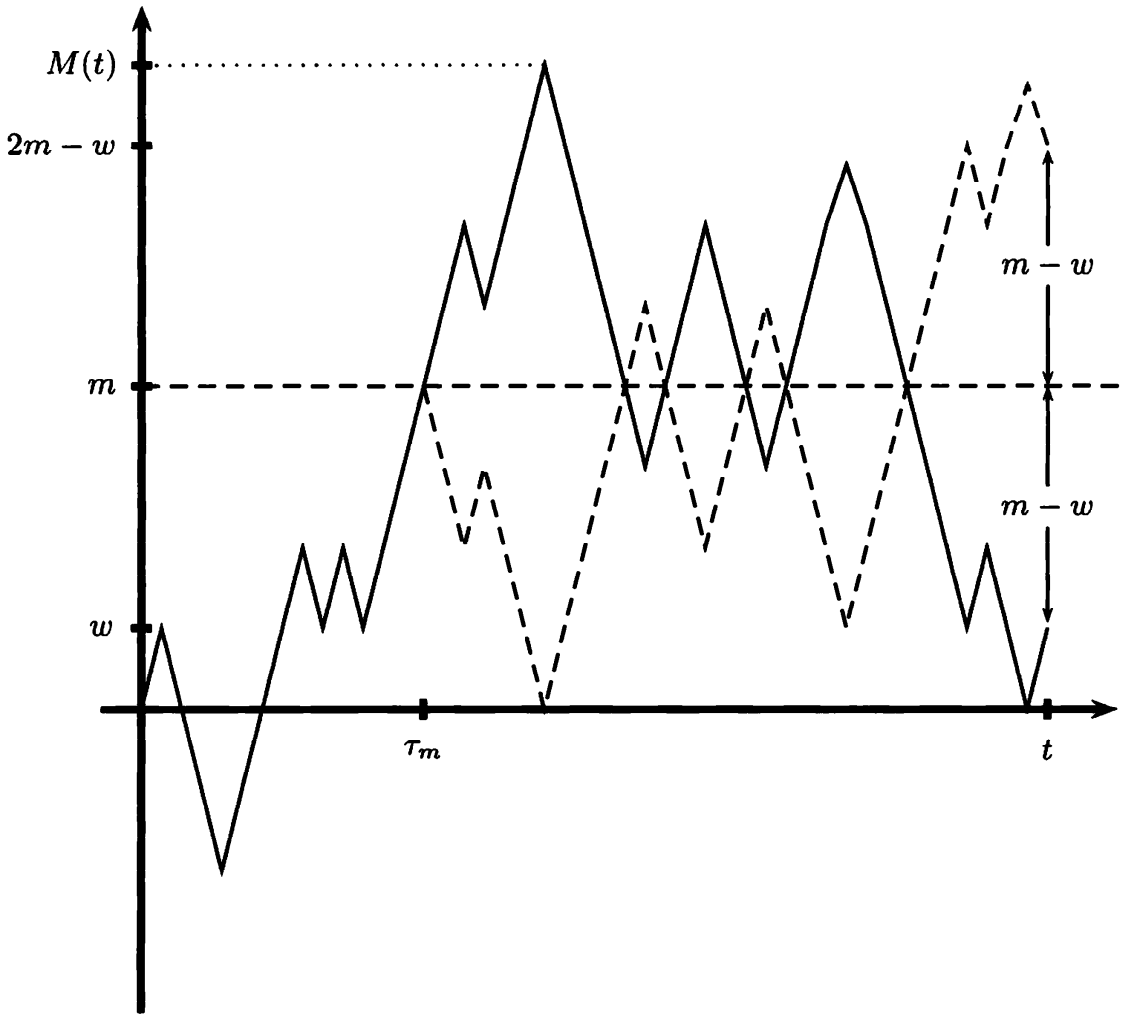

Fig. 3.7.1. Brownian path and reflected path.

As Figure 3.7.1 illustrates, for each Brownian motion path that reaches level m prior to time t but is at a level w below m at time t, there is a "reflected path" that is at level $2m - w$ at time t. This reflected path is constructed by switching the up and down moves of the Brownian motion from time τ_m onward. Of course, the probability that a Brownian motion path ends at exactly w or at exactly $2m - w$ is zero. In order to have nonzero probabilities, we consider the paths that reach level m prior to time t and are *at or below* level w at time t, and we consider their reflections, which are *at or above* $2m - w$ at time t. This leads to the key *reflection equality*

$$\mathbb{P}\{\tau_m \le t, W(t) \le w\} = \mathbb{P}\{W(t) \ge 2m - w\}, \quad w \le m,\ m > 0. \tag{3.7.1}$$

3.7.2 First Passage Time Distribution

We draw two conclusions from (3.7.1). The first is the distribution for the random variable τ_m.

Theorem 3.7.1. *For all $m \neq 0$, the random variable τ_m has cumulative distribution function*

$$\mathbb{P}\{\tau_m \le t\} = \frac{2}{\sqrt{2\pi}} \int_{\frac{|m|}{\sqrt{t}}}^{\infty} e^{-\frac{y^2}{2}}\, dy, \quad t \ge 0, \tag{3.7.2}$$

and density

$$f_{\tau_m}(t) = \frac{d}{dt}\mathbb{P}\{\tau_m \le t\} = \frac{|m|}{t\sqrt{2\pi t}} e^{-\frac{m^2}{2t}}, \quad t \ge 0. \tag{3.7.3}$$

PROOF: We first consider the case $m > 0$. We substitute $w = m$ into the reflection formula (3.7.1) to obtain

$$\mathbb{P}\{\tau_m \le t, W(t) \le m\} = \mathbb{P}\{W(t) \ge m\}.$$

On the other hand, if $W(t) \ge m$, then we are guaranteed that $\tau_m \le t$. In other words,

$$\mathbb{P}\{\tau_m \le t, W(t) \ge m\} = \mathbb{P}\{W(t) \ge m\}.$$

Adding these two equations, we obtain the cumulative distribution function for τ_m:

$$\begin{aligned}\mathbb{P}\{\tau_m \le t\} &= \mathbb{P}\{\tau_m \le t, W(t) \le m\} + \mathbb{P}\{\tau_m \le t, W(t) \ge m\} \\ &= 2\mathbb{P}\{W(t) \ge m\} \;=\; \frac{2}{\sqrt{2\pi t}} \int_m^{\infty} e^{-\frac{x^2}{2t}}\, dx.\end{aligned}$$

We make the change of variable $y = \frac{x}{\sqrt{t}}$ in the integral, and this leads to (3.7.2) when m is positive. If m is negative, then τ_m and $\tau_{|m|}$ have the same distribution, and (3.7.2) provides the cumulative distribution function of the latter. Finally, (3.7.3) is obtained by differentiating (3.7.2) with respect to t. □

Remark 3.7.2. From (3.7.3), we see that

$$\mathbb{E}e^{-\alpha\tau_m} = \int_0^{\infty} e^{-\alpha m} f_{\tau_m}(t)\, dt = \int_0^{\infty} \frac{|m|}{t\sqrt{2\pi t}} e^{-\alpha m - \frac{m^2}{2t}}\, dt \text{ for all } \alpha > 0. \tag{3.7.4}$$

Theorem 3.6.2 provides the apparently different Laplace transform formula (3.6.9). These two formulas are in fact the same, and the steps needed to verify this are provided in Exercise 3.9. □

3.7.3 Distribution of Brownian Motion and Its Maximum

We define the *maximum to date* for Brownian motion to be

$$M(t) = \max_{0 \le s \le t} W(s). \tag{3.7.5}$$

This stochastic process is used in pricing barrier options. For the value of t in Figure 3.7.1, the random variable $M(t)$ is indicated. For positive m, we have

$M(t) \geq m$ if and only if $\tau_m \leq t$. This observation permits us to rewrite the reflection equality (3.7.1) as

$$\mathbb{P}\{M(t) \geq m, W(t) \leq w\} = \mathbb{P}\{W(t) \geq 2m - w\}, \quad w \leq m,\ m > 0. \tag{3.7.6}$$

From this, we can obtain the joint distribution of $W(t)$ and $M(t)$.

Theorem 3.7.3. *For $t > 0$, the joint density of $\big(M(t), W(t)\big)$ is*

$$f_{M(t),W(t)}(m, w) = \frac{2(2m - w)}{t\sqrt{2\pi t}} e^{-\frac{(2m-w)^2}{2t}}, \quad w \leq m,\ m > 0. \tag{3.7.7}$$

PROOF: Because

$$\mathbb{P}\{M(t) \geq m, W(t) \leq w\} = \int_m^\infty \int_{-\infty}^w f_{M(t),W(t)}(x, y)\, dy\, dx$$

and

$$\mathbb{P}\{W(t) \geq 2m - w\} = \frac{1}{\sqrt{2\pi t}} \int_{2m-w}^\infty e^{-\frac{z^2}{2t}}\, dz,$$

we have from (3.7.6) that

$$\int_m^\infty \int_{-\infty}^w f_{M(t),W(t)}(x, y)\, dy\, dx = \frac{1}{\sqrt{2\pi t}} \int_{2m-w}^\infty e^{-\frac{z^2}{2t}}\, dz.$$

We differentiate first with respect to m to obtain

$$-\int_{-\infty}^w f_{M(t),W(t)}(m, y)\, dy = -\frac{2}{\sqrt{2\pi t}} e^{-\frac{(2m-w)^2}{2t}}.$$

We next differentiate with respect to w to see that

$$-f_{M(t),W(t)}(m, w) = -\frac{2(2m - w)}{t\sqrt{2\pi t}} e^{-\frac{(2m-w)^2}{2t}}.$$

This is (3.7.7). □

When simulating Brownian motion to price exotic options, it is often convenient to first simulate the value of the Brownian motion at some time $T > 0$ and then simulate the maximum of the Brownian motion between times 0 and t. This second step requires that we know the distribution of the maximum of the Brownian motion $M(t)$ on $[0, t]$ conditioned on the value of $W(t)$. This conditional distribution is provided by the following corollary.

Corollary 3.7.4. *The conditional distribution of $M(t)$ given $W(t) = w$ is*

$$f_{M(t)|W(t)}(m|w) = \frac{2(2m - w)}{t} e^{-\frac{2m(m-w)}{t}}, \quad w \leq m, m > 0.$$

PROOF: The conditional density is the joint density divided by the marginal density of the conditioning random variable. The conditional density we seek here is

$$\begin{aligned} f_{M(t)|W(t)}(m|w) &= \frac{f_{M(t),W(t)}(m,w)}{f_{W(t)}(w)} \\ &= \frac{2(2m-w)}{t\sqrt{2\pi t}} \cdot \sqrt{2\pi t}\, e^{-\frac{(2m-w)^2}{2t} + \frac{w^2}{2t}} \\ &= \frac{2(2m-w)}{t} e^{-\frac{2m(m-w)}{t}}. \end{aligned}$$

□

3.8 Summary

Brownian motion is a continuous stochastic process $W(t)$, $t \geq 0$, that has independent, normally distributed increments. In this text, we adopt the convention that Brownian motion starts at zero at time zero, although one could add a constant a to our Brownian motion and obtain a "Brownian motion starting at a". For either Brownian motion starting at 0 or Brownian motion starting at a, if $0 = t_0 < t_1 < \cdots < t_m$, then the increments

$$W(t_1) - W(t_0),\ W(t_2) - W(t_1), \ldots, W(t_m) - W(t_{m-1})$$

are independent and normally distributed with

$$\mathbb{E}\big[W(t_{i+1}) - W(t_i)\big] = 0, \quad \mathrm{Var}\big[W(t_{i+1}) - W(t_i)\big] = t_{i+1} - t_i.$$

This is Definition 3.3.1. Associated with Brownian motion there is a filtration $\mathcal{F}(t)$, $t \geq 0$, such that for each $t \geq 0$ and $u \geq t$, $W(t)$ is $\mathcal{F}(t)$-measurable and $W(u) - W(t)$ is independent of $\mathcal{F}(t)$.

Brownian motion is both a martingale and a Markov process. Its transition density is

$$p(\tau, x, y) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(y-x)^2}{2\tau}}.$$

This is the density in the variable y for the random variable $W(s+\tau)$ given that $W(s) = x$.

A profound property of Brownian motion is that it accumulates quadratic variation at rate one per unit time (Theorem 3.4.3). If we choose a time interval $[T_1, T_2]$, choose partition points $T_1 = t_0 < t_1 < \cdots < t_m = T_2$, and compute $\sum_{j=0}^{m-1} \big(W(t_{j+1}) - W(t_j)\big)^2$, we get an answer that depends on the path along which the computation is done. However, if we let the number of partition points approach infinity and the length of the longest subinterval $t_{j+1} - t_j$ approach zero, this quantity has limit $T_2 - T_1$, the length of the interval over which the quadratic variation is being computed. We write $dW(t)\, dW(t) = dt$ to symbolize the fact that the amount of quadratic

variation Brownian motion accumulates in an interval is equal to the length of the interval, *regardless of the path along which we do the computation.*

If we compute $\sum_{j=0}^{m-1}\big(W(t_{j+1})-W(t_j)\big)(t_{j+1}-t_j)$ or $\sum_{j=1}^{m-1}(t_{j+1}-t_j)^2$ and pass to the limit, we get zero (Remark 3.4.5). We symbolize this by writing $dW(t)\,dt = dt\,dt = 0$.

The first passage time of Brownian motion,

$$\tau_m = \min\{t \geq 0; W(t) = m\},$$

is the first time the Brownian motion reaches the level m. For $m \neq 0$, we have $\mathbb{P}\{\tau_m < \infty\} = 1$ (equation (3.6.7)) (i.e., the Brownian motion eventually reaches every nonzero level), but $\mathbb{E}\tau_m = \infty$ (Remark 3.6.3). The random variable τ_m is a stopping time, has density (Theorem 3.7.1)

$$f_{\tau_m}(t) = \frac{|m|}{t\sqrt{2\pi t}},$$

and this density has Laplace transform (Theorem 3.6.2; see also Exercise 3.9)

$$\mathbb{E}e^{-\alpha\tau_m} = e^{-|m|\sqrt{2\alpha}} \text{ for all } \alpha > 0.$$

The reflection principle used to determine the density $f_{\tau_m}(t)$ can also be used to determine the joint density of $W(t)$ and its maximum to date $M(t) = \max_{0\leq s\leq t} W(s)$. This joint density is (Theorem 3.7.3)

$$f_{M(t),W(t)}(m,w) = \frac{2(2m-w)}{t\sqrt{2\pi t}}e^{-\frac{(2m-w)^2}{2t}}, \quad w \leq m,\ m > 0.$$

3.9 Notes

In 1828, Robert Brown observed irregular movement of pollen suspended in water. This motion is now known to be caused by the buffeting of the pollen by water molecules, as explained by Einstein [62]. Bachelier [6] used Brownian motion (not geometric Brownian motion) as a model of stock prices, even though Brownian motion can take negative values. Lévy [107], [108] discovered many of the nonintuitive properties of Brownian motion. The first mathematically rigorous construction of Brownian motion is credited to Wiener [159], [160], and Brownian motion is sometimes called the *Wiener process.*

Brownian motion and its properties are presented in numerous texts, including Billingsley [10]. The development in these notes is a summary of that found in Karatzas and Shreve [101]. The properties of Brownian motion and many formulas useful for pricing exotic options are developed in Borodin and Salminen [18].

Convergence of discrete-time and/or discrete-state models to continuous-time models, a topic touched upon in Section 3.2.7, is treated by Amin and Khanna [3], Cox, Ross and Rubinstein [42], Duffie and Protter [60], and Willinger and Taqqu [162], among others.

3.10 Exercises

Exercise 3.1. According to Definition 3.3.3(iii), for $0 \leq t < u$, the Brownian motion increment $W(u) - W(t)$ is independent of the σ-algebra $\mathcal{F}(t)$. Use this property and property (i) of that definition to show that, for $0 \leq t < u_1 < u_2$, the increment $W(u_2) - W(u_1)$ is also independent of $\mathcal{F}(t)$.

Exercise 3.2. Let $W(t)$, $t \geq 0$, be a Brownian motion, and let $\mathcal{F}(t)$, $t \geq 0$, be a filtration for this Brownian motion. Show that $W^2(t) - t$ is a martingale. (Hint: For $0 \leq s \leq t$, write $W^2(t)$ as $(W(t) - W(s))^2 + 2W(t)W(s) - W^2(s)$.)

Exercise 3.3 (Normal kurtosis). The *kurtosis* of a random variable is defined to be the ratio of its fourth central moment to the square of its variance. For a normal random variable, the kurtosis is 3. This fact was used to obtain (3.4.7). This exercise verifies this fact.

Let X be a normal random variable with mean μ, so that $X - \mu$ has mean zero. Let the variance of X, which is also the variance of $X - \mu$, be σ^2. In (3.2.13), we computed the moment-generating function of $X - \mu$ to be $\varphi(u) = \mathbb{E}e^{u(X-\mu)} = e^{\frac{1}{2}u^2\sigma^2}$, where u is a real variable. Differentiating this function with respect to u, we obtain

$$\varphi'(u) = \mathbb{E}\left[(X-\mu)e^{u(X-\mu)}\right] = \sigma^2 u e^{\frac{1}{2}\sigma^2 u^2}$$

and, in particular, $\varphi'(0) = \mathbb{E}(X - \mu) = 0$. Differentiating again, we obtain

$$\varphi''(u) = \mathbb{E}\left[(X-\mu)^2 e^{u(X-\mu)}\right] = \left(\sigma^2 + \sigma^4 u^2\right) e^{\frac{1}{2}\sigma^2 u^2}$$

and, in particular, $\varphi''(0) = \mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$. Differentiate two more times and obtain the normal kurtosis formula $\mathbb{E}\left[(X-\mu)^4\right] = 3\sigma^4$.

Exercise 3.4 (Other variations of Brownian motion). Theorem 3.4.3 asserts that if T is a positive number and we choose a partition Π with points $0 = t_0 < t_1 < t_2 < \cdots < t_n = T$, then as the number n of partition points approaches infinity and the length of the longest subinterval $\|\Pi\|$ approaches zero, the sample quadratic variation

$$\sum_{j=0}^{n-1} \big(W(t_{j+1}) - W(t_j)\big)^2$$

approaches T for almost every path of the Brownian motion W. In Remark 3.4.5, we further showed that $\sum_{j=0}^{n-1} \big(W(t_{j+1}) - W(t_j)\big)(t_{j+1} - t_j)$ and $\sum_{j=0}^{n-1}(t_{j+1} - t_j)^2$ have limit zero. We summarize these facts by the multiplication rules

$$dW(t)\,dW(t) = dt, \quad dW(t)\,dt = 0, \quad dt\,dt = 0. \tag{3.10.1}$$

(i) Show that as the number m of partition points approaches infinity and the length of the longest subinterval approaches zero, the sample first variation

$$\sum_{j=0}^{n-1} \left|W(t_{j+1}) - W(t_j)\right|$$

approaches ∞ for almost every path of the Brownian motion W. (Hint:

$$\sum_{j=0}^{n-1} \left(W(t_{j+1}) - W(t_j)\right)^2$$

$$\leq \max_{0\leq k\leq n-1} \left|W(t_{k+1}) - W(t_k)\right| \cdot \sum_{j=0}^{n-1} \left|W(t_{j+1}) - W(t_j)\right|.)$$

(ii) Show that as the number n of partition points approaches infinity and the length of the longest subinterval approaches zero, the sample cubic variation

$$\sum_{j=0}^{n-1} \left|W(t_{j+1}) - W(t_j)\right|^3$$

approaches zero for almost every path of the Brownian motion W.

Exercise 3.5 (Black-Scholes-Merton formula). Let the interest rate r and the volatility $\sigma > 0$ be constant. Let

$$S(t) = S(0)e^{(r-\frac{1}{2}\sigma^2)t+\sigma W(t)}$$

be a geometric Brownian motion with mean rate of return r, where the initial stock price $S(0)$ is positive. Let K be a positive constant. Show that, for $T > 0$,

$$\mathbb{E}\left[e^{-rT}\left(S(T) - K\right)^+\right] = S(0)N\left(d_+(T, S(0))\right) - Ke^{-rT}N\left(d_-(T, S(0))\right),$$

where

$$d_\pm(T, S(0)) = \frac{1}{\sigma\sqrt{T}}\left[\log\frac{S(0)}{K} + \left(r \pm \frac{\sigma^2}{2}\right)T\right],$$

and N is the cumulative standard normal distribution function

$$N(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-\frac{1}{2}z^2}\,dz = \frac{1}{\sqrt{2\pi}}\int_{-y}^{\infty} e^{-\frac{1}{2}z^2}\,dz.$$

Exercise 3.6. Let $W(t)$ be a Brownian motion and let $\mathcal{F}(t)$, $t \geq 0$, be an associated filtration.

(i) For $\mu \in \mathbb{R}$, consider the *Brownian motion with drift* μ:

$$X(t) = \mu t + W(t).$$

Show that for any Borel-measurable function $f(y)$, and for any $0 \le s < t$, the function

$$g(x) = \frac{1}{\sqrt{2\pi(t-s)}} \int_{-\infty}^{\infty} f(y) \exp\left\{-\frac{(y-x-\mu(t-s))^2}{2(t-s)}\right\} dy$$

satisfies $\mathbb{E}\big[f\big(X(t)\big)\big|\mathcal{F}(s)\big] = g\big(X(s)\big)$, and hence X has the Markov property. We may rewrite $g(x)$ as $g(x) = \int_{-\infty}^{\infty} f(y)p(\tau, x, y)\, dy$, where $\tau = t - s$ and

$$p(\tau, x, y) = \frac{1}{\sqrt{2\pi\tau}} \exp\left\{-\frac{(y-x-\mu\tau)^2}{2\tau}\right\}$$

is the *transition density* for Brownian motion with drift μ.

(ii) For $\nu \in \mathbb{R}$ and $\sigma > 0$, consider the *geometric Brownian motion*

$$S(t) = S(0)e^{\sigma W(t) + \nu t}.$$

Set $\tau = t - s$ and

$$p(\tau, x, y) = \frac{1}{\sigma y\sqrt{2\pi\tau}} \exp\left\{-\frac{\left(\log\frac{y}{x} - \nu\tau\right)^2}{2\sigma^2\tau}\right\}.$$

Show that for any Borel-measurable function $f(y)$ and for any $0 \le s < t$ the function $g(x) = \int_0^{\infty} h(y)p(\tau, x, y)\, dy$ satisfies $\mathbb{E}\big[f\big(S(t)\big)\big|\mathcal{F}(s)\big] = g\big(S(s)\big)$ and hence S has the Markov property and $p(\tau, x, y)$ is its transition density.

Exercise 3.7. Theorem 3.6.2 provides the Laplace transform of the density of the first passage time for Brownian motion. This problem derives the analogous formula for Brownian motions with drift. Let W be a Brownian motion. Fix $m > 0$ and $\mu \in \mathbb{R}$. For $0 \le t < \infty$, define

$$\begin{aligned} X(t) &= \mu t + W(t), \\ \tau_m &= \min\{t \ge 0; X(t) = m\}. \end{aligned}$$

As usual, we set $\tau_m = \infty$ if $X(t)$ never reaches the level m. Let σ be a positive number and set

$$Z(t) = \exp\left\{\sigma X(t) - \left(\sigma\mu + \frac{1}{2}\sigma^2\right)t\right\}.$$

(i) Show that $Z(t)$, $t \ge 0$, is a martingale.

(ii) Use (i) to conclude that

$$\mathbb{E}\left[\exp\left\{\sigma X(t \wedge \tau_m) - \left(\sigma\mu + \frac{1}{2}\sigma^2\right)(t \wedge \tau_m)\right\}\right] = 1, \quad t \ge 0.$$

(iii) Now suppose $\mu \geq 0$. Show that, for $\sigma > 0$,

$$\mathbb{E}\left[\exp\left\{\sigma m - \left(\sigma\mu + \frac{1}{2}\sigma^2\right)\tau_m\right\}\mathbb{I}_{\{\tau_m<\infty\}}\right] = 1.$$

Use this fact to show $\mathbb{P}\{\tau_m < \infty\} = 1$ and to obtain the Laplace transform

$$\mathbb{E}e^{-\alpha\tau_m} = e^{m\mu - m\sqrt{2\alpha+\mu^2}} \text{ for all } \alpha > 0.$$

(iv) Show that if $\mu > 0$, then $\mathbb{E}\tau_m < \infty$. Obtain a formula for $\mathbb{E}\tau_m$. (Hint: Differentiate the formula in (iii) with respect to α.)

(v) Now suppose $\mu < 0$. Show that, for $\sigma > -2\mu$,

$$\mathbb{E}\left[\exp\left\{\sigma m - \left(\sigma\mu + \frac{1}{2}\sigma^2\right)\tau_m\right\}\mathbb{I}_{\{\tau_m<\infty\}}\right] = 1.$$

Use this fact to show that $\mathbb{P}\{\tau_m < \infty\} = e^{-2x|\mu|}$, which is strictly less than one, and to obtain the Laplace transform

$$\mathbb{E}e^{-\alpha\tau_m} = e^{m\mu - m\sqrt{2\alpha+\mu^2}} \text{ for all } \alpha > 0.$$

Exercise 3.8. This problem presents the convergence of the distribution of stock prices in a sequence of binomial models to the distribution of geometric Brownian motion. In contrast to the analysis of Subsection 3.2.7, here we allow the interest rate to be different from zero.

Let $\sigma > 0$ and $r \geq 0$ be given. For each positive integer n, we consider a binomial model taking n steps per unit time. In this model, the interest rate per period is $\frac{r}{n}$, the up factor is $u_n = e^{\sigma/\sqrt{n}}$, and the down factor is $d_n = e^{-\sigma/\sqrt{n}}$. The risk-neutral probabilities are then

$$\tilde{p}_n = \frac{\frac{r}{n} + 1 - e^{-\sigma/\sqrt{n}}}{e^{\sigma/\sqrt{n}} - e^{-\sigma/\sqrt{n}}}, \quad \tilde{q}_n = \frac{e^{\sigma/\sqrt{n}} - \frac{r}{n} - 1}{e^{\sigma/\sqrt{n}} - e^{-\sigma/\sqrt{n}}}.$$

Let t be an arbitrary positive rational number, and for each positive integer n for which nt is an integer, define

$$M_{nt,n} = \sum_{k=1}^{nt} X_{k,n},$$

where $X_{1,n}, \ldots, X_{n,n}$ are independent, identically distributed random variables with

$$\widetilde{\mathbb{P}}\{X_{k,n} = 1\} = \tilde{p}_n, \quad \widetilde{\mathbb{P}}\{X_{k,n} = -1\} = \tilde{q}_n, \ k = 1, \ldots, n.$$

The stock price at time t in this binomial model, which is the result of nt steps from the initial time, is given by (see (3.2.15) for a similar equation)

$$\begin{aligned} S_n(t) &= S(0)u_n^{\frac{1}{2}(nt+M_{nt,n})}d_n^{\frac{1}{2}(nt-M_{nt,n})} \\ &= S(0)\exp\left\{\frac{\sigma}{2\sqrt{n}}(nt+M_{nT,n})\right\}\exp\left\{-\frac{\sigma}{2\sqrt{n}}(nt-M_{nt,n})\right\} \\ &= S(0)\exp\left\{\frac{\sigma}{\sqrt{n}}M_{nt,n}\right\}. \end{aligned}$$

This problem shows that as $n \to \infty$, the distribution of the sequence of random variables $\frac{\sigma}{\sqrt{n}}M_{nt,n}$ appearing in the exponent above converges to the normal distribution with mean $(r - \frac{1}{2}\sigma^2)t$ and variance $\sigma^2 t$. Therefore, the limiting distribution of $S_n(t)$ is the same as the distribution of the geometric Brownian motion $S(0)\exp\{\sigma W(t) + (r - \frac{1}{2}\sigma)t\}$ at time t.

(i) Show that the moment-generating function $\varphi_n(u)$ of $\frac{1}{\sqrt{n}}M_{nt,n}$ is given by

$$\varphi_n(u) = \left[e^{\frac{u}{\sqrt{n}}}\left(\frac{\frac{r}{n}+1-e^{-\sigma/\sqrt{n}}}{e^{\sigma/\sqrt{n}}-e^{-\sigma/\sqrt{n}}}\right) - e^{-\frac{u}{\sqrt{n}}}\left(\frac{\frac{r}{n}+1-e^{\sigma/\sqrt{n}}}{e^{\sigma/\sqrt{n}}-e^{-\sigma/\sqrt{n}}}\right)\right]^{nt}.$$

(ii) We want to compute

$$\lim_{n\to\infty}\varphi_n(u) = \lim_{x\downarrow 0}\varphi_{\frac{1}{x^2}}(u),$$

where we have made the change of variable $x = \frac{1}{\sqrt{n}}$. To do this, we will compute $\log\varphi_{\frac{1}{x^2}}(u)$ and then take the limit as $x \downarrow 0$. Show that

$$\log\varphi_{\frac{1}{x^2}}(u) = \frac{t}{x^2}\log\left[\frac{(rx^2+1)\sinh ux + \sinh(\sigma-u)x}{\sinh\sigma x}\right]$$

(the definitions are $\sinh z = \frac{e^z-e^{-z}}{2}$, $\cosh z = \frac{e^z+e^{-z}}{2}$), and use the formula

$$\sinh(A-B) = \sinh A\cosh B - \cosh A\sinh B$$

to rewrite this as

$$\log\varphi_{\frac{1}{x^2}}(u) = \frac{t}{x^2}\log\left[\cosh ux + \frac{(rx^2+1-\cosh\sigma x)\sinh ux}{\sinh\sigma x}\right].$$

(iii) Use the Taylor series expansions

$$\cosh z = 1 + \frac{1}{2}z^2 + O(z^4),\ \sinh z = z + O(z^3),$$

to show that

$$\begin{aligned} &\cosh ux + \frac{(rx^2+1-\cosh\sigma x)\sinh ux}{\sinh\sigma x} \\ &\qquad = 1 + \frac{1}{2}u^2x^2 + \frac{rux^2}{\sigma} - \frac{1}{2}ux^2\sigma + O(x^4). \end{aligned} \tag{3.10.2}$$

The notation $O(x^j)$ is used to represent terms of the order x^j.

(iv) Use the Taylor series expansion $\log(1+x) = x + O(x^2)$ to compute $\lim_{x \downarrow 0} \log \varphi_{\frac{1}{x^2}}(u)$. Now explain how you know that the limiting distribution for $\frac{\sigma}{\sqrt{n}} M_{nt,n}$ is normal with mean $(r - \frac{1}{2}\sigma^2)t$ and variance $\sigma^2 t$.

Exercise 3.9 (Laplace transform of first passage density). The solution to this problem is long and technical. It is included for the sake of completeness, but the reader may safely skip it.

Let $m > 0$ be given, and define

$$f(t,m) = \frac{m}{t\sqrt{2\pi t}} \exp\left\{-\frac{m^2}{2t}\right\}.$$

According to (3.7.3) in Theorem 3.7.1, $f(t,m)$ is the density in the variable t of the first passage time $\tau_m = \min\{t \geq 0; W(t) = m\}$, where W is a Brownian motion without drift. Let

$$g(\alpha, m) = \int_0^\infty e^{-\alpha t} f(t,m)\, dt, \quad \alpha > 0,$$

be the Laplace transform of the density $f(t,m)$. This problem verifies that $g(\alpha, m) = e^{-m\sqrt{2\alpha}}$, which is the formula derived in Theorem 3.6.2.

(i) For $k \geq 1$, define

$$a_k(m) = \frac{1}{\sqrt{2\pi}} \int_0^\infty t^{-k/2} \exp\left\{-\alpha t - \frac{m^2}{2t}\right\} dt,$$

so $g(\alpha, m) = m a_3(m)$. Show that

$$g_m(\alpha, m) = a_3(m) - m^2 a_5(m),$$
$$g_{mm}(\alpha, m) = -3m a_5(m) + m^3 a_7(m).$$

(ii) Use integration by parts to show that

$$a_5(m) = -\frac{2\alpha}{3} a_3(m) + \frac{m^2}{3} a_7(m).$$

(iii) Use (i) and (ii) to show that g satisfies the second-order ordinary differential equation

$$g_{mm}(\alpha, m) = 2\alpha g(\alpha, m).$$

(iv) The general solution to a second-order ordinary differential equation of the form

$$a y''(m) + b y'(m) + c y(m) = 0$$

is

$$y(m) = A_1 e^{\lambda_1 m} + A_2 e^{\lambda_2 m},$$

where λ_1 and λ_2 are roots of the *characteristic equation*

$$a\lambda^2 + b\lambda + c = 0.$$

Here we are assuming that these roots are distinct. Find the general solution of the equation in (iii) when $\alpha > 0$. This solution has two undetermined parameters A_1 and A_2, and these may depend on α.

(v) Derive the bound

$$g(\alpha, m) \leq \frac{m}{\sqrt{2\pi}} \int_0^m \sqrt{\frac{m}{t}}\, t^{-3/2} \exp\left\{-\frac{m^2}{2t}\right\} dt + \frac{1}{\sqrt{2\pi m}} \int_m^\infty e^{-\alpha t}\, dt$$

and use it to show that, for every $\alpha > 0$,

$$\lim_{m \to \infty} g(\alpha, m) = 0.$$

Use this fact to determine one of the parameters in the general solution to the equation in (iii).

(vi) Using first the change of variable $s = t/m^2$ and then the change of variable $y = 1/\sqrt{s}$, show that

$$\lim_{m \downarrow 0} g(\alpha, m) = 1.$$

Use this fact to determine the other parameter in the general solution to the equation in (iii).

This page intentionally left blank