Math 6040
The University of Utah
Mathematical Probability

Davar Khoshnevisan
Department of Mathematics
University of Utah
Salt Lake City, UT 84112–0090
davar@math.utah.edu
http://www.math.utah.edu/~davar

Spring 2002

Contents

Preface (Rough outline)

NOT FOR CIRCULATION

- These notes constitute a one-semester first graduate course in probability theory at the University of Utah. All of the non-optional chapters, and all but a handful of the exercises, have been taught and assigned in a one-semester course. However, it would be possible to offer a slower-paced course based on Chapters 1–7 only.

  In their current state, these are still lecture-notes and not a book; so please keep in mind that the write-up is occasionally too brief, the historical discussions are too inconsistent, and the all-around form is somewhat incomplete.

  The final form of this preface will address the following two questions:

- **Question:** Why probability? Three important addresses: Gnedenko [Gne69], Doob [Doo89], Mumford [Mum00]. All three present very convincing reasons for why study probability in the 60's, 80's and 00's. Interestingly enough, the reasons are different in flavor, but also similar in their essence.

  There is also available a recent document that tries to address the "why probability" question for the forthcoming century. See

  `http://www.math.cornell.edu/~durrett/probrep/probrep.html`.

  Ed Waymire and Phil Protter are in the process of writing a expandedversion of this report for a Siam publication.

- **Question:** Why this book? I have found well in excess of 50 graduate texts in probability! I found that fewer than a handful of them are written for a one-semester treatment. The present notes take a stab at a one-semester curriculum that contains enough probability for many of today's graduate students.

Rather than modeling them after other graduate texts (with personalized embellishments), I have decided to follow a route that is closer in spirit to that of the typical better-thought-out undergraduate probability curriculum. Unlike its undergraduate cousins, however, the present notes are rigorous, and much more importantly, contain recent more sophisticated advances in this and a few neighboring subjects. I hope that you will find these notes unapologetic and highly non-encyclopedic in form.

- *So far many thanks are due to*: Nelson Beebe, Bob Brooks, and Nat Smale. Last but certainly not the least, Irina and Adrian Gushin for their patience and understanding.

- *Special Thanks are due to*: All my former students, esp. Liz Levina, Irina Grabovsky, Pando Son, Jim Turner, and Jun Zhang.

Davar Khoshnevisan
Salt Lake City, UT
December 19, 2002

Part I

Measure Theory Primer

Chapter 1

A Crash-Course in Measure Theory

1 Introduction

Modern probability is spoken in the language of measure theory, and to me this is where the connections between the two theories begin, as well as end. It is for this reason that I have tried to minimize the introductory measure-theoretic discussions that are typically found in probability texts. At the same time, it is difficult to imagine how one can try to understand many of the advances of modern probability theory without first learning the requisite language. As such, we begin these notes with a few brief primer chapters on measure and integration.

The first part of these notes is self-contained, and the motivated student can learn enough measure theory here to use the remainder of the notes successfully. However, the present treatment may be too rapid and perhaps even too sparse for some. My intention is rather to recall some facts, and describe ideas that are needed in developing probability theory. To this I should add that many of the said facts are typically not sufficiently well-stressed in standard books on measure and integration. So it is best not to omit this first part in a first reading.

Perhaps the best way to appreciate our need for using measure theory is to ask, "*What is a random variable?*". After all, no matter how they may be defined, random variables are one of the central objects in probability and many of its applications.

Classically, one thinks of a random variable as the numerical outcome of a random experiment. Moreover, each time we perform our random experiment, we should obtain a "realization" of this random variable that may or may not be the same as the previous realizations. This is far from an exact description, and leads to various inaccuracies not to mention some paradoxes.

The modern viewpoint, in rough terms, is the following: We have an unknown—possibly complicated—function X. Each time we perform our random experiment, we see the evaluation $X(\omega)$ of X at some point ω in the domain of definition of X, where the point (or "realization") ω is selected according to weights (or "probabilities") that are predescribed by a probability measure (or "distribution").

This description is not as complicated as it may seem, and has the appealing property that it can be rigorously introduced. As an example, consider the unit interval $[0, 1]$, and let $\mathrm{P}(E)$ denote the "length" of any $E \subseteq [0, 1]$; more precisely, P is the Lebesgue measure on $[0, 1]$. Now consider the function,

$$X(\omega) := \begin{cases} 1, & \text{if } \omega \in \left[0, \frac{1}{2}\right], \\ 0, & \text{if } \omega \in \left(\frac{1}{2}, 1\right]. \end{cases} \tag{1.1}$$

Note that the P-measure of the set of all $\omega \in [0, 1]$ such that $X(\omega) = 1$ is the length of $\left[0, \frac{1}{2}\right]$ which is $\frac{1}{2}$. This is often written as "$\mathrm{P}\{X = 1\} = \frac{1}{2}$." Likewise, $\mathrm{P}\{X = 0\} = \frac{1}{2}$. Viewed as such, X provides us with a mathematical model for the outcome of a fair coin-toss. For instance, if we observe an ω such that $X(\omega) = 1$, then this describes having tossed heads. Moreover, such an event (i.e., $X = 1$) can happen with probability $\frac{1}{2}$; i.e., for one-half of the ω's (in the sense of measure).

In order to study more complicated random variables, we need to have a much deeper understanding of measures, and measure theory. Having said this, let us begin with a formal description of aspects of the theory of measure.

2 Measure Spaces

Throughout, let Ω be a set that is sometimes referred to as the *sample space.*

Definition 1.1 A collection $\mathfrak{F}$ of subsets of Ω is a *σ-algebra* if: (i) $\Omega \in \mathfrak{F}$; (ii) it is closed under complementation, i.e., if $A \in \mathfrak{F}$ then $A^{\complement} \in \mathfrak{F}$; and (iii) it is closed under countable unions, i.e., if $A_1, A_2, \ldots \in \mathfrak{F}$, then $\cup_{n=1}^{\infty} A_n \in \mathfrak{F}$.

It is an *algebra* if instead of being closed under countable union, it is merely closed under finite unions.

Of course, σ-algebras (respectively, algebras) are also closed under countable (respectively, finite) intersections, and they also contain the empty set. Furthermore, any σ-algebra is obviously an algebra but the converse is false: The collection of all finite unions of subintervals of $[0,1]$ is an algebra but not a σ-algebra.

Example 1.2 $\mathfrak{F} = \{\Omega, \varnothing\}$ is a σ-algebra that is aptly called the *trivial* σ-algebra. The power set of Ω is also a σ-algebra. (Recall that the power set of any set is the collection of all of its subsets.) These are the two extremal examples.

Lemma 1.3 (Hausdorff [Hau27, p. 85]) *If I is any set (denumerable or not), and if $\mathfrak{F}_i$ is a σ-algebra of subsets of Ω for each $i \in I$, then $\cap_{i\in I}\mathfrak{F}_i$ is also such a σ-algebra. Consequently, given any algebra $\mathfrak{A}$, there exists a smallest σ-algebra containing $\mathfrak{A}$.*

Definition 1.4 If $\mathfrak{A}$ is a collection of subsets of Ω, we write $\sigma(\mathfrak{A})$ for the smallest σ-algebra that contains $\mathfrak{A}$; this is the *σ-algebra generated by* $\mathfrak{A}$.

Note that this is a consistent definition, since $\sigma(\mathfrak{A}) = \cap\mathfrak{F}$, where the intersection is taken over all σ-algebras $\mathfrak{F}$ such that $\mathfrak{A} \subseteq \mathfrak{F}$. Note also that this is a nonempty intersection since the power set of Ω is at least one such σ-algebra.

An important class of σ-algebras are introduced in the following.

Definition 1.5 If Ω is a topological space, then the open subsets of Ω generate a σ-algebra $\mathfrak{B}(\Omega)$ that is called the *Borel σ-algebra* of Ω.[1.1] Elements of a σ-algebra $\mathfrak{F}$ are said to be *$\mathfrak{F}$-measurable*, or *measurable with respect to* $\mathfrak{F}$. When it is clear from the context that $\mathfrak{F}$ is the σ-algebra under study, its elements are referred to as *measurable.* If $\mathfrak{F}$ is a σ-algebra of subsets of Ω, a set function $\mu : \mathfrak{F} \to \mathbb{R}_+ \cup \{\infty\}$ is said to be a *measure* on $(\Omega, \mathfrak{F})$ if: (i) $\mu(\varnothing) = 0$; and (ii) given any denumerable collection $A_1, A_2, \ldots$ of disjoint sets,

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{1.2}$$

[1.1] When $\Omega = \mathbb{R}^d$, this is due to Hausdorff [Hau27, pp. 177–181].

I emphasize that by virtue of their definition ($\mu : \mathfrak{F} \to \mathbb{R}_+$ and not $\mathbb{R}$), measures are nonnegative. Of course, real-valued (often called signed) or complex measures can be defined just as easily.

Definition 1.6 Let $\mathfrak{S}$ denote a collection of subsets of Ω, and let μ be a set function on Ω. Then μ is said to be *countably additive* on $\mathfrak{S}$ if for all disjoint sets $A_1, A_2, \ldots$—all in $\mathfrak{S}$—as soon as we have $\cup_n A_n \in \mathfrak{S}$, then (1.2) holds. It is said to be *countably subadditive* on $\mathfrak{S}$ if for all $A_1, A_2, \ldots \in \mathfrak{S}$ such that $\cup_n A_n \in \mathfrak{S}$, then $\mu(\cup_n A_n) \leq \sum_n \mu(A_n)$.

The following is simple but not entirely obvious if you read the definitions carefully.

Lemma 1.7 *Countably additive set functions are countably subadditive.*

Definition 1.8 If $\mathfrak{F}$ is a σ-algebra of subsets of Ω and if μ is a measure on $(\Omega, \mathfrak{F})$, then $(\Omega, \mathfrak{F}, \mu)$ is called a *measure space.*

Listed below are some of the elementary properties of measures:

Lemma 1.9 *If* $(\Omega, \mathfrak{F}, \mu)$ *is a measure space, then:*

(i) (Continuity from below) *If* $A_1 \subseteq A_2 \subseteq \cdots$ *are all measurable, then as* $n \to \infty$ *we have* $\mu(A_n) \uparrow \mu\left(\cup_{m=1}^{\infty} A_m\right)$.

(ii) (Continuity from above) *If* $A_1 \supseteq A_2 \supseteq \cdots$ *are all measurable, and if* $\mu(A_n) < +\infty$ *for some* n*, then as* $n \to \infty$ *we have* $\mu(A_n) \downarrow \mu\left(\cap_{m=1}^{\infty} A_m\right)$.

Definition 1.10 A measure space $(\Omega, \mathfrak{F}, \mu)$ is a *probability space* if $\mu(\Omega) = 1$. In this case, μ is a *probability measure.* If instead $\mu(\Omega) < +\infty$, then μ is a *finite measure.* Finally, it is *σ-finite* if there exist measurable sets $\Omega_1 \subseteq \Omega_2 \subseteq \cdots$ such that $\cup_n \Omega_n = \Omega$ and $\mu(\Omega_n) < +\infty$.

We often denote probability measures as $\mathrm{P}, \mathrm{Q}, \ldots$ rather than $\mu, \nu, \ldots$.

The following result characterizes probability spaces that are based on a finite set Ω.

Lemma 1.11 *Suppose $\Omega = \{\omega_1, \ldots, \omega_n\}$ is a finite set. Then, we can find $p_1, \ldots, p_n \in [0,1]$ such that: (a) $p_1 + \cdots + p_n = 1$, and (b) for all $A \subseteq \Omega$, $\mathrm{P}(A) = \sum_{i:\ \omega_i \in A} p_i$. Conversely, any sequence $p_1, \ldots, p_n \in [0,1]$ that has the property (a) above defines a probability measure P on the power set of Ω via the assignment, $\mathrm{P}(\{\omega_i\}) := p_i$ $(i = 1, \ldots, n)$.*

Definition 1.12 Given any measure space $(\omega, \mathfrak{F})$, and any $x \in \Omega$, we define the *point-mass* δ_x at x to be the probability measure that is defined by setting $\delta_x(A) := \mathbf{1}_A(x)$.

Remark 1.13 In the notation of point-masses, that the probability measure of Lemma 1.11 can be written as $\mathrm{P} = \sum_{j=1}^n \delta_{\omega_j}$ (check!).

3 Lebesgue Measure

Lebesgue measure on $(0,1]$ gives us a way of measuring the length of a subset of $(0,1]$. To construct the Lebesgue measure then, we first define a set function m on half-closed finite intervals of the form $(a,b] \subseteq (0,1]$ that evaluates the length of the said intervals:

$$m\left((a,b]\right) = b - a. \tag{1.3}$$

Let $\mathfrak{A}$ denote the collection of all finite unions of half-closed subintervals of $(0,1]$. It is easy to see that $\mathfrak{A}$ is an algebra. We extend the definition of m to $\mathfrak{A}$ as follows: For all disjoint half-closed intervals $I_1, \ldots, I_n \subseteq (0,1]$,

$$m\left(\bigcup_{i=1}^n I_i\right) := \sum_{i=1}^n m(I_i). \tag{1.4}$$

This defines $m(E)$ for all $E \in \mathfrak{A}$. Of course, we need to insure that this definition is consistent. Consistency follows from induction and the following obvious fact: For all $0 < a < b < c$,

$$m\left((a,c]\right) = m\left((a,b]\right) + m\left((b,c]\right). \tag{1.5}$$

More importantly, we have

Lemma 1.14 *The set function m is countably additive on the algebra $\mathfrak{A}$.*

That is, whenever $A_1, A_2, \ldots$ are disjoint elements of $\mathfrak{A}$ such that $\cup_{n=1}^{\infty} A_n \in \mathfrak{A}$, then $m(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} m(A_n)$. (If there are only finitely many nonempty A_n's, this is an obvious consequence of (1.4).)

Proof Since $\cup_{n=1}^{\infty} A_n$ and $\cup_{n=1}^{N-1} A_n$ are both in $\mathfrak{A}$, so is $\cup_{n=N}^{\infty} A_n$. Moreover,

$$m\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{N-1} m(A_n) + m\left(\bigcup_{n=N}^{\infty} A_n\right). \tag{1.6}$$

It suffices to show that $\lim_{N\to\infty} m(\cup_{n=N}^{\infty} A_n) = 0$. Hence, our goal is this: Given a sequence of sets $B_n \downarrow \varnothing$—all in $\mathfrak{A}$—we wish to show that $m(B_n) \downarrow 0$. Suppose to the contrary that there exists $\varepsilon > 0$ so that for all $n \geq 1$, $m(B_n) \geq \varepsilon$. We will derive a contradiction from this.

Write B_n as a finite union of half-open intervals, viz., $B_n := \cup_{j=1}^{k_n} (a_j^n, b_j^n]$, where $0 \leq a_j^n < b_j^n \leq 1$. Also recall that the B_n's are decreasing. Choose some $\alpha_j^n \in (a_j^n, b_j^n)$ so close to a_j^n that $\alpha_j^n \leq a_j^n + \varepsilon/(2^{n+1}k_n)$. Then the sets B_n and $C_n \subseteq B_n$ are close in measure, where $C_n := \cup_{j=1}^{k_n} [\alpha_j^n, b_j^n]$. Here is a measure of how close they are:

$$m\left(\bigcup_{j=1}^{n} (B_j \setminus C_j)\right) \leq \sum_{j=1}^{n} \sum_{i=1}^{k_n} \left(\alpha_i^j - a_i^j\right) \leq \frac{\varepsilon}{2}. \tag{1.7}$$

In particular, $m(C_n) \geq m(B_n) - (\varepsilon/2) \geq (\varepsilon/2)$. But the C_n's are closed bounded and nonempty. If we knew that they were also decreasing, this and the Heine–Borel property of $[0,1]$ would together imply that $\cap_n C_n \neq \varnothing$, which cannot be since $C_n \subseteq B_n$ and $B_n \downarrow \varnothing$. This would give us the desired contradiction. Unfortunately, the C_n's need not be decreasing. So instead consider $D_n := \cap_{j=1}^{n} C_j$. The D_n's are closed, bounded, and decreasing. It suffices to show that they are nonempty. But this is easy: Note that

$$\begin{aligned} B_n &= D_n \cup \left(B_n \cap D_n^{\complement}\right) \subseteq D_n \cup \left(\bigcup_{j=1}^{n} B_j \cap D_n^{\complement}\right) \\ &= D_n \cup \bigcup_{j=1}^{n} (B_j \setminus C_j), \end{aligned} \tag{1.8}$$

since the B_n's are decreasing. Therefore, $\varepsilon \leq m(B_n) \leq m(D_n) + \varepsilon/2$, thanks to (1.7). This shows that $m(D_n) \geq \varepsilon/2$, so that $D_n \neq \varnothing$, and this completes the proof of countable additivity. The rest of the proof is smooth sailing (try it!). □

Lemma 1.14 and the following result immediately extends the domain of the definition of m to $\sigma(\mathfrak{A})$; the latter is obviously equal to $\mathfrak{B}((0,1])$.

Theorem 1.15 (Carathéodory Extension [Car48]) *Suppose Ω is a set, and $\mathfrak{A}$ denotes an algebra of subsets of Ω. Then, given a countably additive set function μ on $\mathfrak{A}$, there exists a measure $\bar{\mu}$ on $(\Omega, \sigma(\mathfrak{A}))$ such that for all $E \in \mathfrak{A}$, $\mu(E) = \bar{\mu}(E)$. Furthermore, if $\mu(\Omega) < +\infty$, then the extension $\bar{\mu}$ of μ is unique, and $\bar{\mu}$ is a finite measure with $\bar{\mu}(\Omega) = \mu(\Omega)$.*

This result is proved in §5 at the end of this chapter. Note that the method of this section also yields the Lebesgue measure on $\mathbb{R}$, $[0,1]$, etc.

To construct the Lebesgue measure on $(0,1]^d$ where $d \geq 1$, we proceed as in the case $d = 1$, except start by defining the measure of a *hypercube* $\prod_{j=1}^d (a_j, b_j] := \{x \in (0,1]^d : a_j < x_j \leq b_j\}$ as $\prod_{j=1}^d (b_j - a_j)$. Since the collection of all finite unions of hypercubes is an algebra that generates $\mathfrak{B}((0,1]^d)$, we then appeal—as in the one-dimensional case—to the Carathéodory extension theorem to construct the Lebesgue measure on $\mathfrak{B}((0,1]^d)$. Further extensions to $[0,1]^d, \mathbb{R}^d$, etc. are made similarly. It is also possible to construct Lebesgue measure on $\mathfrak{B}(\mathbb{R}^d)$ as a *product measure*; stay tuned!

Once we have Theorem 1.15, we can easily construct many measures on the Borel–measurable subsets of $\mathbb{R}^d$ as the following shows. This example will be greatly generalized in the next chapter where we introduce the abstract integral.

Theorem 1.16 *Suppose $f : \mathbb{R}^d \to \mathbb{R}_+$ is a continuous function such that the Riemann integral $\int_{\mathbb{R}^d} f(x)\,dx$ equals 1. Given $a, b \in \mathbb{R}^d$ with $a_j \leq b_j$ for all $j \leq d$, consider the hypercube $\mathcal{C}_{a,b} := (a_1, b_1] \times \cdots \times (a_d, b_d]$, and define*

$$\mu(\mathcal{C}_{a,b}) := \int_{\mathcal{C}_{a,b}} f(x)\,dx. \tag{1.9}$$

Then μ uniquely extends to a probability measure on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$, and f is called the probability density function of μ.

Proof (Sketch) Let $\mathfrak{A}$ denote the collection of all finite unions of disjoint hypercubes of the type mentioned. Then it is easy to see that $\mathfrak{A}$ is an algebra of subsets of $\mathbb{R}^d$. If $A \in \mathfrak{A}$, then we can write $A = \cup_{i=1}^m \mathcal{C}_{a_i,b_i}$, and define

$$\mu\left(\bigcup_{i=1}^m \mathcal{C}_{a_i,b_i}\right) := \sum_{i=1}^m \mu(\mathcal{C}_{a_i,b_i}). \tag{1.10}$$

One can check that this is a consistent definition. This is an ugly task but it is not too hard to do. Now we proceed as we did when constructing the Lebesgue measure in order to show that μ is countably additive on $\mathfrak{A}$. Finally, we appeal to Theorem 1.15 to finish. □

Below are some examples of measures that are important in applications. I refer to these measures as distributions.

Example 1.17[The Uniform Distribution] Suppose $X \in \mathfrak{B}(\mathbb{R}^n)$ has finite and positive n-dimensional Lebesgue measure $m(X)$. Then, the *uniform* distribution ν is the measure defined by $\nu(A) := m(A \cap X) \div m(X)$ for all $A \in \mathfrak{B}(\mathbb{R}^n)$.

Example 1.18[The Exponential Distribution] Given a number $\lambda > 0$, $f(x) := \lambda \exp(-\lambda x)$ defines a probability density function on $(\mathbb{R}_+, \mathfrak{B}(\mathbb{R}_+))$, and the corresponding measure is the *exponential* distribution with parameter λ.

Example 1.19[The Normal—or Gaussian—Distribution] Given two constants $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$, the *normal (or Gaussian)* distribution with parameters μ and σ corresponds to the density function

$$f(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \qquad \forall x \in \mathbb{R}. \tag{1.11}$$

When $\mu = 0$ and $\sigma = 1$, this is the so-called *standard normal* distribution. To streamline formulas, we can also define the normal distribution with parameters μ and $\sigma = 0$ as the distribution that corresponds to the probability measure δ_μ; i.e., the point-mass at μ. This last case describes the archetypal *degenerate normal distribution* since, in this case, $X \equiv \mu$, a.s.

Example 1.20[The Normal—or Gaussian—Distribution in $\mathbb{R}^n$] Given a column vector of constants $\mu \in \mathbb{R}^n$, and an $(n \times n)$ symmetric invertible matrix Q, the *normal (or Gaussian)* distribution with parameters μ and Q corresponds to the following density function: For all $x \in \mathbb{R}^n$,

$$f(x) := \frac{1}{(2\pi)^{n/2}\sqrt{\det(Q)}} \exp\left(-\frac{1}{2}(x-\mu)\cdot Q^{-1}(x-\mu)\right). \tag{1.12}$$

4 Completion

In the previous section we constructed the Lebesgue measure on $\mathfrak{B}((0,1]^d)$ (say), and this is good enough for most people's needs. However, one can extend the definition of m slightly further by defining $m(E)$ for a slightly larger class of sets E.

Definition 1.21 Given a measure space $(\Omega, \mathfrak{F}, \mu)$, a measurable set E is *null* if $\mu(E) = 0$. The σ-algebra $\mathfrak{F}$ is said to be *complete* if all subsets of null sets are themselves (measurable) and null. When $\mathfrak{F}$ is complete, we also say that $(\Omega, \mathfrak{F}, \mu)$ is complete.

We can always insure completeness, viz.,

Theorem 1.22 *Given a measure space $(\Omega, \mathfrak{F}, \mu)$, there exists a complete σ-algebra $\mathfrak{F}' \supseteq \mathfrak{F}$, and a measure μ' on $(\Omega, \mathfrak{F}')$ such that on $\mathfrak{F}$, μ and μ' agree.*

Definition 1.23 The above measure space $(\Omega, \mathfrak{F}', \mu')$ is the *completion* of $(\Omega, \mathfrak{F}, \mu)$.

Proof of Theorem 1.22 (Sketch) Let $A \triangle B := (A \cap B^{\complement}) \cup (A^{\complement} \cap B)$ denote the set difference of A and B, and for any two sets A and B define

$$\mathfrak{F}' := \{A \subseteq \Omega : \ \exists B, N \in \mathfrak{F} \text{ so that } \mu(N) = 0 \text{ and } A \triangle B \subseteq N\}. \tag{1.13}$$

In words, we construct $\mathfrak{F}'$ by adding in all of the subsets of null sets of $\mathfrak{F}$, and declaring them null.

Step 1. $\mathfrak{F}'$ is a σ-algebra.
Since $A^{\complement} \triangle B^{\complement} = A \triangle B$, $\mathfrak{F}'$ is closed under complementation. If $A_1, A_2, \ldots \in$

$\mathfrak{F}'$, then we can find $B_1, B_2, \ldots \in \mathfrak{F}$ and null sets $N_1, N_2, \ldots$ such that $A_i \triangle B_i \subseteq N_i$ for all $i \geq 1$. But $\cup_i A_i \triangle \cup_i B_i = \cup_i (A_i \triangle B_i) \subseteq \cup_i N_i$, and the latter is null, thanks to countable subadditivity. Thus, $\mathfrak{F}'$ is a σ-algebra.

Step 2. The measure μ'.

For any $A \in \mathfrak{F}'$ define $\mu'(A) := \mu(B)$, where $B \in \mathfrak{F}$ is a set such that for a null set $N \in \mathfrak{F}$, $A \triangle B \subseteq N$. It is not hard to see that this is well defined; i.e., it does not depend on the representation (B, N) of A. Clearly, $\mu' = \mu$ on $\mathfrak{F}$; we need to show that μ' is a measure on $(\Omega, \mathfrak{F}')$. The only interesting portion is countable additivity.

Step 3. Countable Additivity.

Suppose $A_1, A_2, \ldots \in \mathfrak{F}'$ are disjoint. Find $B_i, N_i \in \mathfrak{F}$ as before and note that whenever $j \leq i$, then

$$B_{i+1} \cap B_j \subseteq (A_{i+1} \cup N_*) \cap (A_j \cup N_*) = N_* \tag{1.14}$$

where $N_* := \cup_i N_i$ is a null set. Define $C_1 := B_1$ and iteratively define $C_{i+1} := B_{i+1} \backslash (C_1 \cup \cdots \cup C_i)$. The C_i's are disjoint, and thanks to the previous display, $B_{i+1} \cap C_i \subseteq N_*$; in particular, $\mu'(B_{i+1}) = \mu'(B_{i+1} \setminus C_i) = \mu'(C_{i+1})$. Since $B_1 = C_1$, this shows that for all i, $\mu(B_i) = \mu(C_i)$; we have used the fact that $\mu' = \mu$ on $\mathfrak{F}$. Because the C_j's are disjoint, and since $\cup_j C_j = \cup_j B_j$, we obtain $\sum_j \mu(B_j) = \mu(\cup_j B_j)$. In other words, $\mu'(\cup_j A_j) = \sum_j \mu'(A_j)$ and our task is done. □

If m denotes the Lebesgue measure on $(0,1]^d$, then we can complete $\big((0,1]^d, \mathfrak{B}((0,1]^d)), m\big)$ to obtain the probability space $\big((0,1]^d, \mathfrak{L}((0,1]^d), \lambda\big)$, where $\mathfrak{L}((0,1]^d)$ denotes the completion of $\mathfrak{B}((0,1]^d)$ and is the collection of all *Lebesgue measurable sets* in $(0,1]^d$. Likewise, we could define $\mathfrak{L}([0,1]^d)$, $\mathfrak{L}(\mathbb{R}^d)$, etc. We have now defined the Lebesgue measure $\lambda(E)$ of $E \subset \mathbb{R}^d$ for a large class of sets E. Exercise 1.5 shows that one cannot define $\lambda(E)$ for all $E \subset \mathbb{R}^d$ and preserve the all-important translation-invariance of the Lebesgue measure.

5 Proof of Carathéodory's Extension Theorem

The proof of Theorem 1.15 is somewhat long, and relies on a set of ingenious ideas that are also useful elsewhere. Throughout, Ω is a set, and $\mathfrak{A}$ is an algebra of subsets of Ω.

Definition 1.24 A collection of subsets of Ω is a *monotone class* if it is closed under increasing unions and decreasing countable intersections.

Lemma 1.25 *An arbitrary intersection of monotone classes is a monotone class. In particular, there exists a smallest monotone class containing* $\mathfrak{A}$.

Definition 1.26 The smallest monotone class that contains $\mathfrak{A}$ is written as $\mathrm{mc}(\mathfrak{A})$, and is called the monotone class *generated* by $\mathfrak{A}$.

The following result is of paramount use in measure theory:

Theorem 1.27 (The Monotone Class Theorem) *Any monotone class that contains* $\mathfrak{A}$ *also contains* $\sigma(\mathfrak{A})$. *In other words,* $\mathrm{mc}(\mathfrak{A}) = \sigma(\mathfrak{A})$.

Before proving this, let us use it to prove the uniqueness assertion of Carathédory's extension theorem.

Proof of Theorem 1.15 (Uniqueness) Suppose there were two extensions $\bar{\mu}$ and ν. Clearly, the collection $\mathfrak{C} := \{E \in \sigma(\mathfrak{A}) : \ \nu(E) = \bar{\mu}(E)\}$ is a monotone class that contains $\mathfrak{A}$. Thus, $\mathfrak{C} = \sigma(\mathfrak{A})$, which is another way of saying that ν and $\bar{\mu}$ agree on $\sigma(\mathfrak{A})$. □

Proof of Theorem 1.27 Since $\sigma(\mathfrak{A})$ is a monotone class, $\sigma(\mathfrak{A}) \supseteq \mathrm{mc}(\mathfrak{A})$, and it suffices to show that $\mathrm{mc}(\mathfrak{A}) \supseteq \sigma(\mathfrak{A})$; the proof is nonconstructive. First, note that the following are monotone classes:

$$\begin{aligned} \mathfrak{C}_1 &:= \left\{ E \in \sigma(\mathfrak{A}) : \ E^{\complement} \in \mathrm{mc}(\mathfrak{A}) \right\}, \\ \mathfrak{C}_2 &:= \{ E \in \sigma(\mathfrak{A}) : \ \forall F \in \sigma(\mathfrak{A}), \ E \cup F \in \mathrm{mc}(\mathfrak{A}) \}. \end{aligned} \tag{1.15}$$

Since $\mathfrak{C}_1$ is a monotone class that contains $\mathfrak{A}$, we have $\mathfrak{C}_1 \supseteq \mathrm{mc}(\mathfrak{A})$, and this means that $\mathrm{mc}(\mathfrak{A})$ is closed under complementation. If we also knew that $\mathfrak{A} \subseteq \mathfrak{C}_2$, then we could deduce that $\mathfrak{C}_2 \supseteq \mathrm{mc}(\mathfrak{A})$, which would imply that $\mathrm{mc}(\mathfrak{A})$ is closed under finite unions and is therefore a σ-algebra. This would show that $\mathrm{mc}(\mathfrak{A}) \supseteq \sigma(\mathfrak{A})$ and complete our proof. However, proving that $\mathfrak{A} \subseteq \mathfrak{C}_2$ requires one more idea. Consider

$$\mathfrak{C}_3 := \{ E \in \sigma(\mathfrak{A}) : \ \forall F \in \mathfrak{A}, \ E \cup F \in \mathrm{mc}(\mathfrak{A}) \}. \tag{1.16}$$

Since it is a monotone class that contains $\mathfrak{A}$, $\mathfrak{C}_3 \supseteq \mathrm{mc}(\mathfrak{A})$ and in particular, $\mathfrak{C}_3 \supseteq \mathfrak{A}$. By reversing the roles of E and F in the definition of $\mathfrak{C}_2$, we can see that $\mathfrak{C}_2 \supseteq \mathfrak{A}$ as well, and this is what we needed to prove. □

Proof of Theorem 1.15 (Existence; Optional) Producing a completely rigorous proof takes too much effort, so I will outline a proof instead. The main idea is to try and prove more. Namely, we will define, in a natural way, $\bar{\mu}(E)$ for all $E \subseteq \Omega$. This defines a set function $\bar{\mu}$ on the power set $\mathfrak{P}$ or Ω which maybe too big a σ-algebra in the sense that $\bar{\mu}$ may fail to be countably additive on $\mathfrak{P}$. However, it will be countably additive on $\sigma(\mathfrak{A})$. Now, we fill in more details.

For all $E \subseteq \Omega$, define

$$\bar{\mu}(E) := \inf \left\{ \sum_{n=1}^{\infty} \mu(E_n) : \ \forall j \geq 1, \ E_j \in \mathfrak{A} \text{ and } E \subseteq \bigcup_{n=1}^{\infty} E_n \right\}. \tag{1.17}$$

In other words, the infimum is taken over all sequences $E_1, E_2, \ldots$ in $\mathfrak{A}$ that cover E. This ought to be a natural and appealing extension of μ. The proof proceeds in three steps.

Step 1. Countable Subadditivity of $\bar{\mu}$.
First, one shows that $\bar{\mu}$ is countably subadditive on $\mathfrak{P}$. (In the jargon of measure theory, $\bar{\mu}$ is an *outer measure* on $(\Omega, \mathfrak{P})$.) Indeed, we wish to show that for any $A_1, A_2, \ldots$, all subsets of Ω, $\bar{\mu}(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \bar{\mu}(A_n)$. To this end, consider any collection $(A_{j,n})$ of elements of $\mathfrak{A}$ such that for all n, $A_n \subseteq \cup_{j=1}^{\infty} A_{j,n}$. By the definition of $\bar{\mu}$, $\bar{\mu}(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \mu(A_{j,n})$. On the other hand, once more using the definition of $\bar{\mu}$, we see that for any $\varepsilon > 0$, we could choose the $A_{j,n}$'s such that $\sum_{j=1}^{\infty} \mu(A_{j,n}) \leq \varepsilon 2^{-n} + \bar{\mu}(A_n)$. This yields, $\bar{\mu}(\cup_{n=1}^{\infty} A_n) \leq \varepsilon + \sum_{n=1}^{\infty} \bar{\mu}(A_n)$, which is the desired countable subadditivity of $\bar{\mu}$, since $\varepsilon > 0$ is arbitrary.

Step 2. $\bar{\mu}$ extends μ.
Next one shows that $\bar{\mu}$ and μ agree on $\mathfrak{A}$ so that $\bar{\mu}$ is indeed an extension of μ. Since for all $E \in \mathfrak{A}$, $\bar{\mu}(E) \leq \mu(E)$, we seek to prove the converse inequality. Consider any collection $E_1, E_2, \ldots$ of elements of $\mathfrak{A}$ that covers E. For any $\varepsilon > 0$, we can arrange things so that $\sum_{n=1}^{\infty} \mu(E_n) \leq \bar{\mu}(E) + \varepsilon$. Since μ is countably additive on $\mathfrak{A}$, $\mu(E) \leq \mu(\cup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mu(E_n) \leq \bar{\mu}(E) + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, Step 2 is completed.

Step 3. Countable Additivity.
We now complete our proof by showing that the restriction of $\bar{\mu}$ to $\sigma(\mathfrak{A})$ is countably additive. Thanks to Step 1, it suffices to show that for all disjoint $A_1, A_2, \ldots$, all in $\sigma(\mathfrak{A})$, $\sum_{n=1}^{\infty} \bar{\mu}(A_n) \leq \bar{\mu}(\cup_{n=1}^{\infty} A_n)$. With this in mind, consider

$$\mathfrak{M} := \left\{ E \subseteq \Omega : \ \forall F \in \mathfrak{A}, \ \bar{\mu}(E) = \bar{\mu}\left(E \cap F\right) + \bar{\mu}\left(E \cap F^{\complement}\right) \right\}. \tag{1.18}$$

According to Step 2, $\mathfrak{M}$ contains $\mathfrak{A}$. Thus, thanks to the monotone class theorem, if $\mathfrak{M}$ were a monotone class, then $\sigma(\mathfrak{A}) \subseteq \mathfrak{M}$. This shows that $\bar{\mu}$ is finitely additive on $\sigma(\mathfrak{A})$, which in turn implies that for any $N \geq 1$, $\bar{\mu}(\cup_{n=1}^{\infty} A_n) \geq \bar{\mu}\left(\cup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} \bar{\mu}(A_n)$, since the A_n's were disjoint. Step 3—whence the Carathédory extension theorem—follows from this upon letting $N \uparrow \infty$. Owing to Step 1, it suffices to show that the following is a monotone class:

$$\left\{E \subseteq \Omega : \ \forall F \in \mathfrak{A}, \ \bar{\mu}(E) \geq \bar{\mu}\left(E \cap F\right) + \bar{\mu}\left(E \cap F^{\complement}\right)\right\}. \tag{1.19}$$

This is proved by appealing to similar covering arguments that we used in Steps 1 and 2. □

6 Exercises

Exercise 1.1 Prove Lemma 1.3.

Exercise 1.2 Prove Lemma 1.9.
(Hint: Any countable union can be expressed as a disjoint countable union.)

Exercise 1.3 Prove Lemma 1.11.

Exercise 1.4 Prove Lemma 1.25.

Exercise 1.5 Prove that Lebesgue measure is translation invariant, i.e., the measure of $x + A$ is the same as the measure of A provided $x + A$ and A are Borel measurable. Here, $x+A := \{x+y;\ y \in A\}$. Furthermore, if m_α denotes the Lebesgue measure on $([0,\alpha], \mathfrak{B}([0,\alpha]))$ for a given $\alpha > 0$, prove that for all measurable $A \subseteq [0,\alpha]$, $\alpha^{-1}A$ is in $\mathfrak{B}([0,1])$, and $m_\alpha(A) = \alpha m_1(\alpha^{-1}A)$. In other words, prove that Lebesgue measure is also scale invariant.

Exercise 1.6 Suppose $(\Omega, \mathfrak{F}, \mu)$ is a measure space. Let Ω' be a set, and let $\mathfrak{F}'$ denote a σ-algebra of subsets of Ω'. If $f : \Omega \to \Omega'$ is measurable, show that $\mu \circ f^{-1}$ is a measure on $(\Omega', \mathfrak{F}')$, where

$$\mu \circ f^{-1}(A) := \mu\left(\{\omega \in \Omega : \ f(\omega) \in A\}\right). \tag{1.20}$$

Exercise 1.7 In this exercise we construct a set in the circle S^1 that is not Borel measurable. As usual, we can think of S^1 as a subset of $\mathbb{C}$:

$$S^1 := \left\{ e^{i\theta} : \ \theta \in (0, 2\pi] \right\}. \tag{1.21}$$

1. Given any $z = e^{i\alpha}, w = e^{i\beta} \in S^1$, we write $z \sim w$ if $\alpha - \beta$ is a rational number. Show that this defines an equivalence relation on S^1.

2. Use the axiom of choice to construct a set Λ whose elements are one from each $\sim$-equivalence class of S^1.

3. For any rational $\alpha \in (0, 2\pi]$, define $\Lambda_\alpha := e^{i\alpha}\Lambda$ denote the rotation of Λ by angle α, and check that if $\alpha, \beta \in (0, 2\pi] \cap \mathbb{Q}$ are distinct, then $\Lambda_\alpha \cap \Lambda_\beta = \varnothing$.

4. Let μ denote the Lebesgue on $(S^1, \mathfrak{B}(S^1))$, and show that $\mu(\Lambda)$ is not defined. Lebesgue measure on S^1 is defined as $m \circ f^{-1}$, where m is the Lebesgue measure on $(0, 2\pi]$, and $f : (0, 2\pi] \to S^1$ is an isometry; cf. (1.20) for the definition of $m \circ f^{-1}$.
   (Hint: Note that $S^1 = \cup_{\alpha \in (0,2\pi] \cap \mathbb{Q}} \Lambda_\alpha$ is a countable disjoint union.)

5. Conclude that Λ is not Borel measurable.

Chapter 2

A Crash-Course in Integration

1 Introduction

We are ready to define nearly-household terms such as "random variables," "expectation," "standard deviation," and "correlation." To give a brief preview of what we are about to see, let me mention:[2.1]

- A random variable X is a measurable function.
- The expectation $\mathrm{E}\{X\}$ is the integral $\int X\,d\mathrm{P}$ of the function X with respect to the underlying probability measure P.
- The standard deviation and the correlation are the expectations of certain types of random variables.

Thus, in this chapter I will describe measurable functions, as well as the abstract integral $\int X\,d\mathrm{P}$, together with many of its salient features. Throughout, $(\Omega, \mathfrak{F}, \mu)$ denotes a a measure space.

2 Measurable Functions

Definition 2.1 A function $f : \Omega \to \mathbb{R}^n$ is (Borel) *measurable* if for all $E \in \mathfrak{B}(\mathbb{R}^n)$, $f^{-1}(E) \in \mathfrak{F}$. Measurable functions on probability spaces are often referred to as *random variables,* and written as $X, Y, \ldots$ instead of $f, g \ldots$. In the context of probability spaces, measurable sets are often referred to as *events.*

[2.1] This viewpoint is due to Fréchet [Fré30].

Since $f^{-1}(E) := \{\omega \in \Omega : f(\omega) \in E\}$, f is measurable (equivalently, f is a random variable) if and only if the preimages of measurable sets under f are themselves measurable.

Example 2.2 An important example of a measurable function is the *indicator* function of a measurable set. Indeed, suppose $A \in \mathfrak{F}$ and define the *indicator of* A as the function,

$$\mathbf{1}_A(\omega) := \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \in A^{\complement}. \end{cases} \tag{2.1}$$

We have already seen such functions in (1.1).

Note that quite generally, $\mathbf{1}_A : \Omega \to \mathbb{R}$, and consider $\mathbf{1}_A^{-1}(E)$ for a measurable $A \subseteq \mathbb{R}$. It is easy to see that

$$\mathbf{1}_A^{-1}(E) = \begin{cases} A, & \text{if } 1 \in E \text{ but } 0 \in E^{\complement}, \\ A^{\complement}, & \text{if } 0 \in E \text{ but } 1 \in E^{\complement}, \\ \Omega, & \text{if } 0, 1 \in E, \\ \varnothing, & \text{if } 0, 1 \in E^{\complement}. \end{cases} \tag{2.2}$$

Since $A \in \mathfrak{F}$, it follows that $\mathbf{1}_A$ is measurable.

Checking the measurability of a function can be a painful chore. The following alleviates some of the pain most of the time.

Lemma 2.3 *Suppose that $\mathfrak{A}$ is an algebra of subsets of $\mathfrak{B}(\mathbb{R}^n)$, and that for all $A \in \mathfrak{A}$, $f^{-1}(A) \in \mathfrak{F}$. Then, $f : \Omega \to \mathbb{R}^n$ is measurable.*

Proof Since $\{A \in \mathfrak{A} : f^{-1}(A) \in \mathfrak{F}\}$ is a monotone class, the lemma follows from the monotone class theorem (Theorem 1.27). □

Let us use the above to produce some measurable functions next.

Lemma 2.4 *Suppose $f, f_1, f_2, \ldots : \Omega \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^m$.*

(i) If $g : \mathbb{R}^n \to \mathbb{R}^m$ is continuous, then it is measurable.

(ii) If f, f_1, f_2 are measurable, then so are αf for any $\alpha \in \mathbb{R}$, $f_1 + f_2$, and $f_1 \times f_2$.

(iii) If $f_1, f_2, \ldots$ are measurable, then so are $\sup_n f_n$*,* $\inf_n f_n$*,* $\limsup_n f_n$*, and* $\liminf_n f_n$*.*

(iv) If g and f are measurable, then so is their composition $g \circ f = g(f)$.

Proof I will prove this to remind you of the flavor of the subject.

If $g : \mathbb{R}^n \to \mathbb{R}^m$ is continuous, then for all open sets $G \subseteq \mathbb{R}^m$, $g^{-1}(G)$ is open (and hence Borel measurable) in $\mathbb{R}^n$. Part (i) follows from Lemma 2.3. The functions $g(x) := \alpha x$ and $g(x, y) := x + y$ and $g(x, y) := xy$ are all continuous on the appropriate Euclidean spaces. So if we proved (iv), then (ii) would follow from (i) and (iv). But (iv) is an elementary consequence of the identity: $(g \circ f)^{-1}(A) = f^{-1}(g^{-1}(A))$. It remains to prove (iii).

Let $S(\omega) := \sup_n f_n(\omega)$ and note that for all $x \in \mathbb{R}$, $S^{-1}((-\infty, x]) = \cap_n f_n^{-1}((-\infty, x]) \in \mathfrak{F}$. Consequently, for all reals $x < y$, $S^{-1}((x, y]) = S^{-1}((-\infty, y]) \setminus S^{-1}((\infty, x]) \in \mathfrak{F}$. The collection of finite disjoint unions of sets of the form $(y, x]$ is an algebra that generates $\mathfrak{B}(\mathbb{R})$. Therefore, by Lemma 2.3, $\sup_n f_n$ is measurable. Apply (iv) to $g(x) = -x$ to see that $\inf_n f_n = -\sup_n(-f_n)$ is also measurable. But we have $\limsup_n f_n = \inf_k \sup_{n \geq k} f_n := \inf_k h_k$, where $h_k := \sup_{n \geq k} f_n$. Since denumerably many suprema and infima preserve measurability, $\limsup_n f_n$ is measurable. Finally, the lim inf is measurable since $\liminf_n f_n = -\limsup_n(-f_n)$. □

3 The Abstract Integral

Throughout this part $(\Omega, \mathfrak{F}, \mu)$ denotes a finite measure space unless we explicitly specify that μ is σ-finite.

We now wish to define the integral $\int f \, d\mu$ for measurable functions $f : \Omega \to \mathbb{R}$. Much of what we do here works for σ-finite measure spaces using the following localization method: Find disjoint measurable $K_1, K_2, \ldots$ such that $\cup_n K_n = \Omega$ and $\mu(K_n) < +\infty$. Define μ_n to be the restriction of μ to K_n, i.e., $\mu_n(A) := \mu(K_n \cap A)$ for all $A \in \mathfrak{F}$. It is easy to see that μ_n is a finite measure on $(\Omega, \mathfrak{F})$. Apply the integration theory of this module to μ_n, and combine the integrals: $\int f \, d\mu := \sum_n \int f \, d\mu_n$. For us, the details are not worth the effort. After all, probability measures are finite!

The abstract integral is derived in three steps.

3.1 Elementary and Simple Functions

When f is a nice function, $\int f\,d\mu$ is easy to define. Indeed, suppose $f = c\mathbf{1}_A$ where $A \in \mathfrak{F}$ and $c \in \mathbb{R}$. Such functions are called *elementary functions.* Then, we define $\int f\,d\mu := c\mu(A)$.

More generally, suppose $A_1, \ldots, A_n \in \mathfrak{F}$ are disjoint, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and $f := \sum_{j=1}^n \alpha_j \mathbf{1}_{A_j}$. This is measurable by Lemma 2.4, and such functions are called *simple functions.* For them, we define $\int f\,d\mu := \sum_{j=1}^n \alpha_j \mu(A_j)$. This is well defined; in other words, writing a simple function f in two different ways does not yield two different integrals. One proves this first in the case where f is an elementary function. Indeed, suppose $f = a\mathbf{1}_A = b\mathbf{1}_B + c\mathbf{1}_C$, where B, C are disjoint. It easily follows from this that $a = b = c$ and $A = B \cup C$. Therefore, by the finite additivity of μ, $a\mu(A) = b\mu(B) + c\mu(C)$, which is another way of saying that our integral is well defined in this case. The general case follows from this, the next lemma, and induction.

Lemma 2.5 *If f is a simple function, then so is $|f|$. If $f \geq 0$ pointwise, then $\int f\,d\mu \geq 0$. Furthermore, if f, g are simple functions, then for $a, b \in \mathbb{R}$,*

$$\int (af + bg)\,d\mu = a \int f\,d\mu + b \int f\,d\mu. \tag{2.3}$$

In other words, for simple functions, $f \mapsto \int f\,d\mu$ is a nonnegative linear functional. A consequence of this is that whenever $f \leq g$ are simple functions, then $\int f\,d\mu \leq \int g\,d\mu$. In particular, we also have the following important consequence: $|\int f\,d\mu| \leq \int |f|\,d\mu$.

3.2 Bounded Measurable Functions

Suppose $f : \Omega \to \mathbb{R}$ is bounded and measurable. To define $\int f\,d\mu$ we use the following to approximate f by simple functions.

Lemma 2.6 *If $f : \Omega \to \mathbb{R}$ is bounded and measurable, then we can find simple functions $\underline{f}_n, \overline{f}_n$ ($n = 1, 2, \cdots$) such that $\underline{f}_n \uparrow f$, $\overline{f}_n \downarrow f$, and $\overline{f}_n \leq \underline{f}_n + 2^{-n}$ pointwise.*

By combining this with Lemma 2.5 we can see that

- $\int \underline{f}_n\,d\mu \leq \int \overline{f}_n\,d\mu \leq \int \underline{f}_n\,d\mu + 2^{-n}\mu(\Omega)$, for all $n \geq 1$; and

- $\int f\,d\mu := \lim_n \int \underline{f}_n\,d\mu = \lim_n \int \overline{f}_n\,d\mu$ exists and is finite.

This produces an integral $\int f\,d\mu$ that inherits the properties of $\int \overline{f}_n\,d\mu$ and $\int \underline{f}_n\,d\mu$ described by Lemma 2.5. That is,

Lemma 2.7 *If f is a bounded measurable function, then so is $|f|$. If $f \geq 0$ pointwise, then $\int f\,d\mu \geq 0$. Furthermore, if f, g are bounded and measurable functions, then for $a, b \in \mathbb{R}$,*

$$\int (af + bg)\,d\mu = a \int f\,d\mu + b \int f\,d\mu. \tag{2.4}$$

3.3 The General Case

We now define $\int f\,d\mu$ for any measurable $f : \Omega \to \mathbb{R}_+$ (note: these functions are nonnegative!). To do so, define $f_n := \min(f, n)$, which is measurable thanks to Lemma 2.3. Of course, $0 \leq f_n(\omega) \leq f(\omega)$ and $f_n(\omega) \uparrow f(\omega)$ at every point $\omega \in \Omega$. In particular, $\int f_n\,d\mu$ is an increasing sequence thanks to Lemma 2.7. Its limit (which could be $+\infty$) is denoted by $\int f\,d\mu$, and inherits the properties of the integrals for bounded integrands.

In order to define the most general integral of this type, let us consider an arbitrary measurable function $f : \Omega \to \mathbb{R}$ and write $f = f^+ - f^-$ where $f^+(\omega) := \max\{f(\omega), 0\}$ and $f^-(\omega) := -\max\{-f(\omega), 0\}$. Both $f^\pm$ are measurable (Lemma 2.5), and if $\int |f|\,d\mu < +\infty$, then define $\int f\,d\mu := \int f^+\,d\mu - \int f^-\,d\mu$. This integral has the following properties.

Proposition 2.8 *Let f be a measurable function such that $\int |f|\,d\mu < +\infty$. If $f \geq 0$ pointwise, then $\int f\,d\mu \geq 0$. If g is another measurable function with $\int |g|\,d\mu < +\infty$, then for $a, b \in \mathbb{R}$,*

$$\int (af + bg)\,d\mu = a \int f\,d\mu + b \int g\,d\mu. \tag{2.5}$$

Our arduous construction is over, and gives us an “indefinite integral.” We can get “definite integrals” as follows: For any $A \in \Omega$, define

$$\int_A f\,d\mu := \int f \mathbf{1}_A\,d\mu, \tag{2.6}$$

and this is well-defined as long as $\int_A |f|\,d\mu < +\infty$. In particular, note that $\int_\Omega f\,d\mu = \int f\,d\mu$.

Remark 2.9 Occasionally, we write $\int f(\omega)\,\mu(d\omega)$ in place of $\int f\,d\mu$.

Definition 2.10 We say that f is *integrable* (with respect to the measure μ) if $\int |f|\,d\mu < +\infty$. On occasion, we will write $\int f(\omega)\,\mu(d\omega)$ for the integral $\int f\,d\mu$. This will be useful later when f will have other variables in its definition, and the $\mu(d\omega)$ reminds us to "integrate out the ω variable."

Definition 2.11 When $(\Omega, \mathfrak{F}, \mathrm{P})$ is a probability space, and when $X : \Omega \to \mathbb{R}$ is a random variable, we write $\mathrm{E}\{X\} := \int X\,d\mathrm{P}$ and call this integral the *expectation* of X. When $A \in \mathfrak{F}$, i.e., when A is an event, we may write $\mathrm{E}\{X; A\}$ in place of the more cumbersome $\mathrm{E}\{X\mathbf{1}_A\}$ or $\int_A X\,d\mathrm{P}$.

4 Modes of Convergence

There are many ways in which a function can converge. We will be primarily concerned with the following. Throughout, $(\Omega, \mathfrak{F}, \mu)$ is a measure space, and $f, f_1, f_2, \ldots : \Omega \to \mathbb{R}$ are measurable.

Definition 2.12 We say that f_n converges to f μ-*almost everywhere* (written μ-a.e. or a.e. if it is clear which measure is being referred to) if

$$\mu\left\{\omega \in \Omega : \limsup_{n\to\infty} |f_n(\omega) - f(\omega)| > 0\right\} = 0. \tag{2.7}$$

In order to expedite the notation, we will write $\{f \in A\}$ for $\{\omega \in \Omega : f(\omega) \in A\}$ and $\mu\{f \in A\}$ for $\mu(\{f \in A\})$. In this way, $f_n \to f$ a.e. if and only if $\mu\{f_n \not\to f\} = 0$. In the case $(\Omega, \mathfrak{F}, \mathrm{P})$ is a probability space, and when $X, X_1, X_2, \ldots$ are random variables on this space, then we say that X_n converges to X *almost surely* (written a.s.) in place of almost everywhere.

One also has the following.

Definition 2.13 We say that $f_n \to f$ in $L^p(\mu)$ if $\lim_{n\to\infty} \|f_n - f\|_p = 0$. We say that $f_n \to f$ *in measure* if for all $\varepsilon > 0$, $\lim_{n\to\infty} \mu\{|f_n - f| \geq \varepsilon\} = 0$. When $(\Omega, \mathfrak{F}, \mathrm{P})$ is a probability space, and when $X, X_1, X_2, \ldots$ are random variables on this space, we say that X_n converges to X in *probability* if $\lim_n \mathrm{P}\{|X_n - X| \geq \varepsilon\} = 0$ for all $\varepsilon > 0$. Sometimes we may write this as $X_n \xrightarrow{\mathrm{P}} X$.

Here are how the notions of convergence are related:

Theorem 2.14 *Either almost-everywhere convergence or L^p-convergence implies convergence in measure.*

[We will soon see that the converse is false.] The interesting portion of this result relies on

Theorem 2.15 (Markov's Inequality) *If f is a nonnegative element of $L^1(\mu)$, then for all $\lambda > 0$,*

$$\mu\{f \geq \lambda\} \leq \frac{1}{\lambda}\int_{\{f\geq\lambda\}} f\,d\mu \leq \frac{1}{\lambda}\|f\|_1. \tag{2.8}$$

Proof Notice that $\|f\|_1 = \int f\,d\mu \geq \int_{\{f\geq\lambda\}} f\,d\mu \geq \lambda\mu\{f \geq \lambda\}$. Divide by $\lambda > 0$ to finish. □

Corollary 2.16 (Chebyshev's Inequality) [2.2] *For any $p > 0$, $f \in L^p(\mu)$, and $\lambda > 0$,*

$$\mu\{|f| \geq \lambda\} \leq \frac{1}{\lambda^p}\int_{\{f\geq\lambda\}} |f|^p\,d\mu \leq \frac{1}{\lambda^p}\|f\|_p^p. \tag{2.9}$$

Proof Since $\mu\{|f| \geq \lambda\} = \mu\{|f|^p \geq \lambda^p\}$, we can apply Markov's inequality to the function $|f|^p$ to finish. □

Proof of Theorem 2.14 Thanks to Chebyshev's inequality, convergence in $L^p(\mu)$ implies convergence in measure since $\mu\{|f_n - f| \geq \varepsilon\} \leq \varepsilon^{-p}\|f_n - f\|_p^p \to 0$. To prove that a.e.-convergence implies convergence in measure, we need to understand a.e.-convergence better. Indeed, note that $f_n \to f$ if and only if

$$\mu\left(\bigcup_{\ell=1}^{\infty}\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty}\left\{|f_n - f| \geq \frac{1}{\ell}\right\}\right) = 0. \tag{2.10}$$

[2.2] Chebyshev was the first to develop such concentration inequalities; see Chebyshev [Che46, Che67]. Markov, who was a student of Chebyshev at the time, noted the full power of Chebyshev's inequality.

Since μ is continuous from above, this is equivalent to the following: For all $\varepsilon > 0$,

$$\lim_{N\to\infty} \mu\left(\bigcup_{n=N}^{\infty} \{|f_n - f| \geq \varepsilon\}\right) = 0. \tag{2.11}$$

But the above measure is clearly greater than or equal to $\mu\{|f_N - f| \geq \varepsilon\}$, which implies convergence in measure. □

Here are two examples to test the strength of the relations between the various modes of convergence. The first also introduces the Borel–Steinhaus probability space which was a starting-point of modern probability theory.

Example 2.17[The Borel–Steinhaus Space] Define $\Omega := [0, 1]$, $\mathrm{P} :=$ the Lebesgue measure on Ω, and $\mathfrak{F} := \mathfrak{B}(\Omega)$. For all $\omega \in [0, 1]$ define,

$$X_n(\omega) := \begin{cases} n^a, & \text{if } \omega < \frac{1}{n}, \\ 0, & \text{if } \omega \geq \frac{1}{n}, \end{cases} \tag{2.12}$$

where $a > 0$ is fixed. Then $X_n \to 0$ almost surely (in fact for all $\omega \in (0, 1]$), but for $p \geq a^{-1}$, $\|X_n\|_p^p = n^{ap-1}$ is bounded away from 0, so a.s.-convergence does not imply L^p-convergence. The trouble comes from the evident fact that $\sup_n |X_n|$ is not in $L^p(\mathrm{P})$ here. Indeed, if $\sup_n |X_n|^p$ *were* integrable, then a.s.-convergence would imply L^p-convergence thanks to the dominated convergence theorem; (Theorem 2.22).

Example 2.18 Let $(\Omega, \mathfrak{F}, \mathrm{P})$ be the Borel–Steinhaus probability space of the previous example. We will now construct a family of random variables X_n such that $\lim_n X_n(\omega)$ does not exist for any $\omega \in [0, 1]$—in particular, X_n does not converge a.s.—and yet $\lim_n \|X_n\|_p = 0$ for all $p > 0$. This construction may look complicated on paper but is completely apparent once you draw a picture of the X_n's. We define a "triangular array" of functions $f_{i,j}$ ($\forall i \geq 1,\ j \leq 2^{i-1}$) as follows: First let $f_{1,1}(\omega) := 1$ for all $\omega \in (0, 1]$. Then define

$$\begin{aligned} f_{2,1}(\omega) &:= \begin{cases} 2, & \text{if } \omega \in \left(0, \frac{1}{2}\right] \\ 0, & \text{otherwise} \end{cases}, \\ f_{2,2}(\omega) &:= \begin{cases} 0, & \text{if } \omega \in \left(\frac{1}{2}, 1\right] \\ 2, & \text{otherwise} \end{cases}, \dots . \end{aligned} \tag{2.13}$$

In general, for all $i \geq 1$ and $j = 1, \ldots, 2^{i-1}$, we can define a function $f_{i,j}$ by $f_{i,j} := i\mathbf{1}_{((j-1)2^{-i-1}, j2^{-i-1}]}$. Let us enumerate the $f_{i,j}$'s according to the dictionary ordering, and call the resulting relabeling (X_k); i.e., $X_1 := f_{1,1}$, $X_2 := f_{2,1}$, $X_3 := f_{2,2}$, $X_4 := f_{3,1}$, $X_5 := f_{3,2}, \ldots$. It is clear that for all $\omega \in (0,1]$, $\limsup_{k\to\infty} X_k(\omega) = +\infty$, whereas $\liminf_{k\to\infty} X_k(\omega) = 0$. In particular, the random variables $X_1, X_2, \ldots$ do not possess limits at any ω. On the other hand, note that $\|f_{i,j}\|_p = i^p 2^{-(i-1)}$. Consequently, $X_k \to 0$ in $L^p(\mu)$ for all $p > 0$ even though there are no a.s. limits.

5 Lebesgue's Convergence Theorems

Proposition 2.8 expresses two of the most important properties of the abstract integral: (i) Integration is a positive operation (i.e., if $f \geq 0$, then $\int f\,d\mu \geq 0$); and (ii) it is a linear operation (i.e., equation 2.5). We now turn to some of the other important properties of the abstract Lebesgue integral that involve limiting operations. Recall that unless it is stated otherwise, all measures are assumed to be finite.[2.3]

Theorem 2.19 (Bounded Convergence) *Suppose $f_1, f_2, \ldots$ are measurable functions such that $\sup_n |f_n|$ is bounded by some constant K. If $f_n \to f$ in measure, then*

$$\lim_{n\to\infty} \int f_n\,d\mu = \int f\,d\mu. \tag{2.14}$$

Proof Since $|f| \leq K$ pointwise, and since μ is a finite measure, f is also integrable so that the integrals all exist. Now fix an $\varepsilon > 0$ and let $E_n := \{\omega \in \Omega : |f(\omega) - f_n(\omega)| > \varepsilon\}$. Then, by Proposition 2.8,

$$\begin{aligned} \left| \int f\,d\mu - \int f_n\,d\mu \right| &\leq \int |f - f_n|\,d\mu \\ &= \int_{E_n^{\complement}} |f - f_n|\,d\mu + \int_{E_n} |f - f_n|\,d\mu \\ &\leq \varepsilon\mu(\Omega) + 2K\mu(E_n). \end{aligned} \tag{2.15}$$

[2.3] Much of the material of this section was developed by Lebesgue [Leb10]. Notable exceptions are the monotone convergence theorem (Theorem 2.21) of Levi [Lev06], and Fatou's lemma (Theorem 2.20) of Fatou [Fat06].

As $n \to \infty$, $\mu(E_n) \to 0$. We can then let $\varepsilon \downarrow 0$ to finish. □

Theorem 2.20 (Fatou's Lemma) *If μ is σ-finite, then for any sequence of integrable nonnegative functions, $f_1, f_2, \ldots$,*

$$\int \liminf_{n\to\infty} f_n \, d\mu \leq \liminf_{n\to\infty} \int f_n \, d\mu. \tag{2.16}$$

Proof I first prove this under the additional condition that $\mu(\Omega) < +\infty$. Let $g_n := \inf_{j\geq n} f_j$ and note that as $n \to \infty$, we have $g_n \uparrow f := \liminf_k f_k$ pointwise. In particular, for any constant $K > 0$, $(f \wedge K - g_n \wedge K)$ is a bounded measurable function that converges to 0 pointwise as $n \to \infty$. Thus, by the bounded convergence theorem (Theorem 2.19), and since $g_n \leq f_n$,

$$\liminf_{n\to\infty} \int f_n \, d\mu \geq \lim_n \int (g_n \wedge K) \, d\mu = \int (f \wedge K) \, d\mu. \tag{2.17}$$

It suffices to show that

$$\lim_{K\uparrow\infty} \int (f \wedge K) \, d\mu = \int f \, d\mu. \tag{2.18}$$

Now for any $\varepsilon > 0$, find a simple function S such that: (i) $0 \leq S \leq f$ pointwise; (ii) there exists $C > 0$ such that $S(\omega) \leq C$; and (iii) $\int S \, d\mu \geq \int f \, d\mu - \varepsilon$. Now $\int (f \wedge K) \, d\mu \geq \int (S \wedge K) \, d\mu = \int S \, d\mu \geq \int f \, d\mu - \varepsilon$ if $K > C$. This proves (2.18) and hence the result in case μ is finite. In the general σ-finite case, for any $\varepsilon > 0$, we can find a measurable set Γ such that $\mu(\Gamma) < \infty$, and $\int_\Gamma f \, d\mu \geq \int f \, d\mu - \varepsilon$. What we showed in the finite case shows that $\liminf_n \int_\Gamma f_n \, d\mu \geq \int_\Gamma f \, d\mu \geq \int f \, d\mu - \varepsilon$. Since $\varepsilon > 0$ is arbitrary, this completes our proof. □

Theorem 2.21 (Monotone Convergence) *If μ is a σ-finite measure, $f_n \uparrow f$ are all nonnegative and measurable, and f is integrable $[d\mu]$, then $\int f_n \, d\mu \uparrow \int f \, d\mu$.*

Proof We have $\int f_n \, d\mu \leq \int f \, d\mu$, and Fatou's lemma does the rest. □

Theorem 2.22 (Dominated Convergence) *Suppose μ is σ-finite, and $f_1, f_2, \ldots$ is a sequence of measurable functions such that $\sup_m |f_m|$ is integrable. Then, $\lim_n \int f_n \, d\mu = \int \lim_n f_n \, d\mu$ provided that $\lim_n f_n$ exists.*

Proof Define $g_n := \sup_{j \geq n} f_j$, and note that as $n \to \infty$, $g_n \downarrow f := \lim_k f_k$. Since $(g_n - f)$ is a sequence of nonnegative measurable functions that convergence down to 0, the monotone convergence theorem implies that $\int (g_n - f) \, d\mu \to 0$. Because $g_n \geq f_n$, this shows that

$$\limsup_{n \to \infty} \int f_n \, d\mu \leq \int f \, d\mu. \tag{2.19}$$

For the converse, define $h_n := \inf_{j \geq n} f_j$ and note that $(f - h_n) \downarrow 0$. Apply the monotone convergence theorem once more to obtain $\int f \, d\mu = \lim_n \int h_n \, d\mu \leq \liminf_n \int f_n \, d\mu$. Together with the preceding display, this does the job. □

6 L^p-Spaces

Throughout this section $(\Omega, \mathfrak{F}, \mathrm{P})$ denotes a probability space.

We can define for all $p \in (0, \infty)$ and all random variables $X : \Omega \to \mathbb{R}$,

$$\|X\|_p := \left(\mathrm{E}\{|X|^p\}\right)^{\frac{1}{p}}, \tag{2.20}$$

provided that the integral exists; i.e., that $|X|^p$ is P-integrable.

Definition 2.23 The space $L^p(\mathrm{P})$ is the collection of all random variables $X : \Omega \to \mathbb{R}$ that are p-times P-integrable. More precisely, these are the random variables X such that $\|X\|_p < +\infty$.

Remark 2.24 More generally, if $(\Omega, \mathfrak{F}, \mu)$ is a σ-finite measure space, then $L^p(\mu)$ will denote the collection of all measurable functions $f : \Omega \to \mathbb{R}$ such that $\|f\|_p < +\infty$. Occasionally, I write $\|f\|_{L^p(\mu)}$ and $L^p(\Omega, \mathfrak{F}, \mu)$ respectively in place of $\|f\|_p$ and $L^p(\mu)$ to emphasize that the underlying (possibly σ-finite) measure space is $(\Omega, \mathfrak{F}, \mu)$.

Next I list some of the elementary properties of L^p-spaces. Note that the following properties do not rely on the finiteness of μ.

Theorem 2.25 *If μ is a σ-finite measure, then:*

(i) *For all $a \in \mathbb{R}$ and $f \in L^p(\mu)$, $\|af\|_p = |a| \cdot \|f\|_p$, and whenever $f, g \in L^p(\mu)$, so is $f+g$. In particular, $L^p(\mu)$ is a linear space.*

(ii) (Hölder's Inequality) *Suppose $p > 1$ and define its so-called conjugate q by $p^{-1}+q^{-1}=1$. Then, $\|fg\|_1 \le \|f\|_p \cdot \|g\|_q$, provided that $f \in L^p(\mu)$ and $g \in L^q(\mu)$.*

(iii) (Minkowski's Inequality) *If $f, g \in L^p(\mu)$ for some $p \ge 1$, then $\|f+g\|_p \le \|f\|_p + \|g\|_p$.*

Proof It is clear that $\|af\|_p = |a| \cdot \|f\|_p$. On the other hand, note that for $x, y \in \mathbb{R}$, $|x+y|^p \le 2^p\{|x|^p+|y|^p\}$. Consequently, $\|f+g\|_p^p \le 2^p\{\|f\|_p^p+\|g\|_p^p\}$, and this proves part *(i)*.

Hölder's inequality holds trivially if $\|f\|_p$ or $\|g\|_p$ are equal to 0. Thus, we can assume without loss of generality that $\|f\|_p, \|g\|_p > 0$. Let $\phi(x) := p^{-1}x^p + q^{-1}y^q - xy$ $(x \ge 0)$, where $y \ge 0$ is fixed, and observe that ϕ is minimized at $x = y^{q/p}$ and the minimum value is 0. In other words, $xy \le p^{-1}x^p + q^{-1}y^q$. Replace x and y by $F(\omega) := |f(\omega)|/\|f\|_p$ and $G(\omega) := |g(\omega)|/\|g\|_q$ respectively, and integrate to obtain

$$\frac{\|fg\|_1}{\|f\|_p \cdot \|g\|_q} = \int |FG| \, d\mu \le \frac{\|F\|_p}{p} + \frac{\|G\|_q}{q} = \frac{1}{p} + \frac{1}{q} = 1. \tag{2.21}$$

Minkowski's inequality follows from Hölder's inequality as follows: Since $|x+y|^p \le |x| \cdot |x+y|^{p-1} + |y| \cdot |x+y|^{p-1}$, by Hölder's inequality, for any $\alpha > 1$,

$$\begin{aligned} \int |f+g|^p \, d\mu &\le \int |f| \cdot |f+g|^{p-1} \, d\mu + \int |g| \cdot |f+g|^{p-1} \, d\mu \\ &\le \{\|f\|_\alpha + \|g\|_\alpha\} \cdot \left(\int |f+g|^{\beta(p-1)} \, d\mu \right)^{1/\beta}, \end{aligned} \tag{2.22}$$

where β is the conjugate to α; i.e., $\alpha^{-1} + \beta^{-1} = 1$. Choose $\alpha = p$ and note that $\beta = q$ solves $\beta(p-1) = p$. This yields

$$\|f+g\|_p^p \le \{\|f\|_p + \|g\|_p\} \cdot \|f+g\|_p^{p-1}. \tag{2.23}$$

If $\|f+g\|_p = 0$, Minkowski's inequality holds trivially; else, we can solve the preceding display to finish. □

An important special case of Hölder's inequality is the following. While it is an immediate consequence (set $p := 2$), it is sufficiently important that it deserves special mention.

Corollary 2.26 (The Cauchy–Bunyakovsky–Schwarz Inequality) *If $f, g \in L^2(\mu)$, then $|\int fg\,d\mu| \le \|f\|_2 \cdot \|g\|_2$.*

Definition 2.27 A function $\psi : \mathbb{R} \to \mathbb{R}$ is *convex* if for all $\lambda \in [0,1]$ and all $x, y \in \mathbb{R}$, $\psi\left(\lambda x + (1-\lambda)y\right) \le \lambda\psi(x) + (1-\lambda)\psi(y)$.

Theorem 2.28 (Jensen's Inequality) *Suppose μ is a probability measure. If $\psi : \mathbb{R} \to \mathbb{R}$ is convex and if $\psi(f)$ and f are integrable, then $\int \psi(f)\,d\mu \ge \psi\left(\int f\,d\mu\right)$.*

Example 2.29 Since $\psi(x) := |x|$ is convex, the above contains the *triangle inequality:* $\int |f|\,d\mu \ge |\int f\,d\mu|$. A second noteworthy example is: $\int e^f\,d\mu \ge e^{\int f\,d\mu}$ since $\psi(x) := e^x$ is convex. These two examples do not require integrability (why?).

Proof Since ψ is convex, there are affine[2.4] functions $\{\psi_z\}_{z\in\mathbb{R}}$ such that

$$\psi(x) = \sup_{z\in\mathbb{R}} \psi_z(x), \qquad \forall x \in \mathbb{R}_+. \tag{2.24}$$

Therefore, by Proposition 2.8,

$$\int \psi(f)\,d\mu \ge \sup_{z\in\mathbb{R}} \int \psi_z(f)\,d\mu = \sup_{z\in\mathbb{R}} \psi_z\left(\int f\,d\mu\right). \tag{2.25}$$

(Here is where we need μ to be a probability measure.) The right-hand side is equal to $\psi\left(\int f\,d\mu\right)$ which yields the result. To complete this proof, we need to verify (2.24). But this too is easy to prove (draw a picture!). For any $\varepsilon > 0$, define

$$D_\varepsilon^+\psi(x) := \frac{\psi(x+\varepsilon) - \psi(x)}{\varepsilon}. \tag{2.26}$$

Since $\lambda\varepsilon + x = \lambda(x+\varepsilon) + (1-\lambda)x$, for all $\lambda \in [0,1]$, $\psi(\lambda\varepsilon + x) \le \lambda\psi(x + \varepsilon) + (1-\lambda)\psi(x)$. Collect terms to deduce that for all $\lambda \in [0,1]$, $D_\varepsilon^+\psi(x) \ge D_{\lambda\varepsilon}^+\psi(x)$. In other words, $\varepsilon \mapsto D_\varepsilon^+\psi(x)$ is increasing, and this means that

[2.4] Recall that h is affine if it is of the form $h(x) = ax + b$.

$D^+\psi(x) := \lim_{\varepsilon\downarrow 0} D^+_\varepsilon\psi(x)$ exists. For each fixed $z \in \mathbb{R}$ and $\varepsilon > 0$, define the affine function

$$\psi^\varepsilon_z(x) := (x - z)D^+_\varepsilon\psi(z) + \psi(z), \qquad \forall x \in \mathbb{R}. \tag{2.27}$$

Then you should check that:

- ψ^ε_z is affine and hence so is $\psi_z(x) := \lim_{\varepsilon\downarrow 0} \psi^\varepsilon_z(x)$.
- $\psi^\varepsilon_z(z) = \psi(z)$ and $\psi^\varepsilon_z(z+\varepsilon) = \psi(z+\varepsilon)$. In particular, $\psi_z(z) = \psi(z)$.
- For all $\delta > 0$, $\psi^\varepsilon_z(x) \le \psi^{\varepsilon+\delta}_z(x)$ whenever $x \ge z$, and $\psi^\varepsilon_z(x) \ge \psi^{\varepsilon+\delta}_z(x)$ whenever $x \le z$.

The last two observations together imply that $\psi^\varepsilon_z(x) \le \psi(x)$ for all $x \notin (z, z+\varepsilon)$. Let $\varepsilon \downarrow 0$ to see that $\psi_z(x) \le \psi(x)$ and $\psi_z(z) = \psi(z)$. This proves (2.24). □

An important consequence of this is that in the case that μ is finite, the L^p-spaces are nested.

Proposition 2.30 (Monotonicity of L^p-Spaces) *If $\mu(\Omega) < +\infty$ and $r > p \ge 1$, then $L^r(\mu) \subseteq L^p(\mu)$. In fact, for all $f \in L^r(\mu)$,*

$$\|f\|_p \le [\mu(\Omega)]^{\frac{1}{p}-\frac{1}{r}} \cdot \|f\|_r. \tag{2.28}$$

Proof We will derive the displayed inequality; the proposition follows readily from that. Since this is a result that only involves the function $|f|$, we can assume without loss of generality that $f \ge 0$. Furthermore, by replacing f by $f \wedge K$, proving (2.28) with f replaced by $f \wedge K$, and then letting $K \uparrow \infty$ (Theorem 2.21 justifies this part), we can assume without loss of generality that f is bounded and hence in $L^v(\mu)$ for all $v > 0$.

Note that for any $s > 1$, the function $\phi(x) = |x|^s$ is convex. Let $s := (r/p)$ and apply Jensen's inequality (Theorem 2.28) to deduce that when μ is a probability measure,

$$\|f\|_p^r = \phi\left(\int |f|^p \, d\mu\right) \le \int \phi\left(|f|^p\right) \, d\mu = \|f\|_r^r, \tag{2.29}$$

which is the desired result. If $\mu(\Omega) > 0$ is finite but not equal to 1, define $\bar{\mu}(A) := \mu(A)/\mu(\Omega)$. This is a probability measure, and according to what we have shown thus far,

$$\left(\int |f|^p \, d\bar{\mu} \right)^{\frac{1}{p}} \leq \left(\int |f|^r \, d\bar{\mu} \right)^{\frac{1}{r}} . \tag{2.30}$$

Solve for μ-integrals to finish. Finally, if $\mu(\Omega) = 0$, then the result holds trivially. □

Fix any $p \geq 1$, and for all $f, g \in L^p(\mu)$, define $d(f, g) := \|f - g\|_p$. According to Minkowski's inequality (Theorem 2.25), d has the properties: (a) $d(f, f) = 0$; (b) $d(f, g) \leq d(f, h) + d(h, g)$; and (c) $d(f, g) = d(g, f)$. In other words, if it were the case that "$d(f, g) = 0 \Longrightarrow f = g$," then $d(\cdot, \cdot)$ would metrize $L^p(\mu)$. However, this latter property generally does not hold, since we could let $g = f\mathbf{1}_A$, where $A \neq \varnothing$ and $\mu(A^\complement) = 0$, to see that $g \neq f$ but $d(f, g) = \left\{ \int_{A^\complement} |f|^p \, d\mu \right\}^{1/p} = 0$ (why?). Nonetheless, if we can identify the elements of $L^p(\mu)$ that are equal to each other outside a null set, then the resulting collection of equivalence classes (endowed with the usual quotient topology and Borel σ-algebra) is indeed a metric space. It is also complete (i.e., every Cauchy sequence converges).

Theorem 2.31 *Let $(\Omega, \mathfrak{F}, \mu)$ denote a σ-finite measure space. For any $f, g \in L^p(\mu)$, write $f \sim g$ if and only if $f = g$, μ-almost everywhere (i.e., $\mu(\{\omega : f(\omega) \neq g(\omega)\}) = 0$).) Then $\sim$ is an equivalence relation on $L^p(\mu)$. Let $[f]$ denote the $\sim$-orbit of f; i.e., $f \in [f]$ if and only if $f \sim g$. Let $\text{Ł}^p(\mu) := \{[f] : f \in L^p(\mu)\}$ and define $\|[f]\|_p := \|f\|_p$. Then, $\text{Ł}^p(\mu)$ is a complete normed linear space. Moreover, $\text{Ł}^2(\mu)$ is a complete Hilbert space.*

Henceforth, we will not distinguish between $L^p(\mu)$ and $\text{Ł}^p(\mu)$. This should not cause any confusions. Also note that $\text{Ł}^p(\mu)$ is the quotient space $L^p(\mu)/\sim$.

Proof The fact that $L^p(\mu)$—and hence $\text{Ł}^p(\mu)$—is a linear space has already been established; cf. Theorem 2.25. As we argued a few paragraphs earlier, $d(f, g) := \|f - g\|_p$ is a norm (now on $\text{Ł}^p(\mu)$) if we know that $d(f, g) = 0 \Rightarrow [f] = [g]$; but this is obvious. To prove completeness, suppose (f_n) is a Cauchy sequence in $L^p(\mu)$. It suffices to show that f_n converges in $L^p(\mu)$. (Translate this to a statement about $[f_n]$'s). Recall that "(f_n) is Cauchy" means that

$\|f_n - f_m\|_p \to 0$ as $n, m \to \infty$. Thus, we can find a subsequence $n_1, n_2, \ldots$ such that $\|f_{n_{k+1}} - f_{n_k}\|_p \leq 2^{-k}$. In particular, $\sum_k \|f_{n_{k+1}} - f_{n_k}\|_p < +\infty$. Thanks to Minkowski's inequality and the monotone convergence theorem (to get Minkowski's inequality to work for an infinite sum),

$$\left\| \sum_{k=1}^{\infty} \left| f_{n_{k+1}} - f_{n_k} \right| \right\|_p < +\infty. \tag{2.31}$$

(Remember that integrals of nonnegative functions are always defined, but could be infinite; this is true even if the function is in places infinite.) In particular, $\sum_k (f_{n_{k+1}} - f_{n_k})$ converges μ-almost everywhere (i.e., for all but a null set of $\omega \in \Omega$.) Let f denote this sum. By Fatou's lemma, $f \in L^p(\mu)$, and by the triangle inequality for L^p-norms (i.e., by Minkowski's inequality), $\|f - f_{n_k}\|_p \leq \sum_{j=k+1}^{\infty} \|f_{n_{j+1}} - f_{n_j}\|_p \to 0$ as $k \to \infty$. Using Minkowski's inequality once more, we see that for any $N, k \geq 1$, $\|f - f_N\|_p \leq \|f - f_{n_k}\|_p + \|f_{n_k} - f_N\|_p$. Let $N \to \infty$ and $k \to \infty$ in this order to see that $f_n \to f$ in $L^p(\mu)$. To finish this proof, we need to show that $L^2(\mu)$ has an inner-product, but this is $\langle f, g \rangle := \int fg \, d\mu$, which thanks to Hölder's inequality, is finite for all $f, g \in L^2(\mu)$. □

7 Exercises

Exercise 2.1 Given any $p_1, \ldots, p_n > 0$ such that $\sum_{\nu=1}^n p_\nu = 1$, and $x_1, \ldots, x_n > 0$, prove that $\prod_{\nu=1}^n x_\nu^{p_\nu} \leq \sum_{\nu=1}^n p_\nu x_\nu$. Prove also that this is a strict inequality unless $x_1 = \cdots = x_n$.
(HINT: Use Jensen's inequality, Theorem 2.28, using the convexity of the exponential function.)

Exercise 2.2 Let $\phi(a) := |a| \div \{1 + |a|\}$, and given any two random variables X and Y, define $d_{\text{P}}(X, Y) := \text{E}\{\phi(X - Y)\}$. Prove that d_{P} metrizes convergence in probability.

Exercise 2.3 Suppose (X_n) and (Y_n) are two sequences of random variables such that $X_n \xrightarrow{\text{P}} X$ and $Y_n \xrightarrow{\text{P}} Y$. Show that whenever f is a continuous function of two variables, then $f(X_n, Y_n) \xrightarrow{\text{P}} f(X, Y)$. This is called *Slutsky's Lemma.*

Exercise 2.4 Suppose $X_1, X_2, \ldots$ are random variables that converge in probability to a random variable X. Prove that for any subsequence (n_k), there exists a further sub-subsequence (n_{k_j}) such that as $j \to \infty$, $X_{n_{k_j}} \to X$ almost surely.

Exercise 2.5 Prove *Chernoff's inequality*: For any nonnegative random variable X and all $\lambda > 0$,

$$\mathrm{P}\left\{X \geq \lambda\right\} \leq \inf_{\xi \geq 0} \exp\left\{-\lambda\xi + \ln \mathrm{E}\left[e^{\xi X}\right]\right\}. \tag{2.32}$$

(Hint: Apply Markov's inequality, Theorem 2.15, to $e^{\xi X}$.)

Exercise 2.6 Suppose $\Omega = [0,1]^d$, $\mathfrak{F} = \mathfrak{B}(\Omega)$, and μ is the Lebesgue measure on $(\Omega, \mathfrak{F})$. Prove that continuous functions are dense in $L^p(\mu)$ for every $p \geq 1$. That is, prove that given $\varepsilon > 0$ and $f \in L^p(\mu)$, we can find a continuous function $g : [0,1]^d \to \mathbb{R}$ such that $\|f - g\|_p \leq \varepsilon$.

Exercise 2.7 Prove the *generalized Hölder inequality:* Given n random variables $X_1, \ldots, X_n$ and $p_1, \ldots, p_n > 1$ that satisfy $\sum_{\ell=1}^n p_\ell^{-1} = 1$,

$$\mathrm{E}\left\{\left|\prod_{\ell=1}^n X_\ell\right|\right\} \leq \prod_{\ell=1}^n \|X_\ell\|_{p_\ell}. \tag{2.33}$$

(Hint: You can use Exercise 2.1 for instance.)

Chapter 3

Product Spaces

1 Introduction

The product space $A_1 \times A_2$ is the collection of all two-tuples (a_1, a_2) where $a_1 \in A_1$ and $a_2 \in A_2$. In like manner, one defines $A_1 \times A_2 \times A_3$, etc. In this way, we can even define infinite-product spaces of the type $A_1 \times A_2 \times \cdots$.

We have two main reasons for studying the measure theory of product spaces. The obvious one is that an understanding of product spaces allows for the construction and analysis of several random variables simultaneously; a theme that is essential to nearly all of probability theory.

Our second reason for learning more about product spaces is less obvious at this point: We will need the so-called Fubini–Tonelli theorem that allows to interchange the order of various multiple-integrals. This is a central fact, and leads to a number of essential computations in mathematics.

2 Finite Products

Given two finite measure spaces $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_2, \mathfrak{F}_2, \mu_2)$, we can define the *product space* $\Omega_1 \times \Omega_2 := \{(\omega_1, \omega_2) : \ \omega_1 \in \Omega_1, \ \omega_2 \in \Omega_2\}$.

Consider the collection $\mathfrak{A}_0 := \{A_1 \times A_2 : \ A_1 \in \mathfrak{F}_1, \ A_2 \in \mathfrak{F}_2\}$. This is closed under finite (in fact arbitrary) intersections, but not under finite unions. For example, let $A_1 = A_2 = [0, 1]$ and $B_1 = B_2 = [1, 2]$ to see that $(A_1 \times A_2) \cup (B_1 \times B_2)$ is not of the form $C_1 \times C_2$ for any C_1 and C_2. So $\mathfrak{A}_0$ is not an algebra. We correct this by adding to $\mathfrak{A}_0$ all finite disjoint unions of elements of $\mathfrak{A}_0$, and call the resulting collection $\mathfrak{A}$.

Lemma 3.1 *$\mathfrak{A}$ is an algebra, and $\mathfrak{F}_1 \times \mathfrak{F}_2 := \sigma(\mathfrak{A}_0) = \sigma(\mathfrak{A})$.*

Define μ on $\mathfrak{A}_0$ as follows:

$$\mu(A_1 \times A_2) := \mu_1(A_1)\mu_2(A_2), \qquad \forall A_1 \in \mathfrak{F}_1,\ A_2 \in \mathfrak{F}_2. \tag{3.1}$$

If $A^1, \dots, A^n \in \mathfrak{A}_0$ are disjoint, we define $\mu(\cup_{i=1}^n A^i) := \sum_{i=1}^n \mu(A^i)$. This defines μ on the algebra $\mathfrak{A}$. This is well defined. Indeed, suppose $\cup_{i=1}^n A^i = \cup_{j=1}^m B^j$ where the A^i's are disjoint and the B^j's are also disjoint. Then, $\cup_{i=1}^n A^i = \cup_{i=1}^n \cup_{j=1}^n (A^i \cap B^j)$ is a disjoint union of nm sets, so $\mu(\cup_{i=1}^n A^i) = \sum_{i=1}^n \sum_{j=1}^m \mu(A^i \cap B^j) = \mu(\cup_{j=1}^m B^j)$ by symmetry.

Theorem 3.2 *There exists a unique measure $\mu_1 \times \mu_2$ on $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \times \mathfrak{F}_2)$ such that on $\mathfrak{A}$, $\mu_1 \times \mu_2 = \mu$.*

Definition 3.3 The measure $\mu_1 \times \mu_2$ is called the *product measure* of μ_1 and μ_2; the space $\Omega_1 \times \Omega_2$ is the corresponding *product space*, and $\mathfrak{F}_1 \times \mathfrak{F}_2$ is the *product σ-algebra*. The measure space $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \times \mathfrak{F}_1, \mu_1 \times \mu_2)$ is the *product measure space*.

Remark 3.4 By induction, we can construct a product measure space $(\prod_{i=1}^n \Omega_i, \prod_{i=1}^n \mathfrak{F}_i, \prod_{i=1}^n \mu_i)$ based on any finite number of measure spaces $(\Omega_i, \mathfrak{F}_i, \mu_i)$, $i = 1, \dots, n$.

Proof of Theorem 3.2 Thanks to Carathéodory's extension theorem (Theorem 1.15), it suffices to show that $(\mu_1 \times \mu_2)$ is countably additive on the algebra $\mathfrak{A}$. This is done in three successive steps.

Step 1. Sections of Measurable Sets are Measurable.
Given any $E \subseteq \Omega_1 \times \Omega_2$ and for all $\omega_2 \in \Omega_2$, define

$$E_{\omega_2} := \{\omega_1 \in \Omega_1 :\ (\omega_1, \omega_2) \in E\}. \tag{3.2}$$

This is the section of E along ω_2. In the first step of the proof we show that if E is measurable (i.e., if $E \in \mathfrak{F}_1 \times \mathfrak{F}_2$), then for any fixed $\omega_2 \in \Omega_2$, E_{ω_2} is measurable too (i.e., $E_{\omega_2} \in \mathfrak{F}_1$): Fix $\omega_2 \in \Omega_2$ and consider the collection $\mathfrak{M} := \{E \in \mathfrak{F}_1 \times \mathfrak{F}_2 :\ E_{\omega_2} \in \mathfrak{F}_1\}$. This is a monotone class that contains $\mathfrak{A}$. By the monotone class theorem (Theorem 1.27), $\mathfrak{M} = \mathfrak{F}_1 \times \mathfrak{F}_2$, which concludes Step 1.

Step 2. Disintegration.
Since E_{ω_2} is measurable, $\mu_1(E_{\omega_2})$ is well defined. We now show that as a

function in $\omega_2 \in \Omega_2$, $\omega_2 \mapsto \mu_1(E_{\omega_2})$ is measurable. First suppose $E \in \mathfrak{A}_0$ so that $E = A_1 \times A_2$ where $A_i \in \mathfrak{F}_i$. Then $E_{\omega_2} = A_1$ if $\omega_2 \in A_2$, and $E_{\omega_2} = \varnothing$ if $\omega_2 \in A_2^{\complement}$. As a result, $\mu_1(E_{\omega_2}) = \mu_1(A_1)\mathbf{1}_{A_2}(\omega_2)$, which is a measurable function of $\omega_2 \in \Omega_2$. Furthermore, $(\mu_1 \times \mu_2)(E) = \mu_1(A_1)\mu_2(A_2)$, so that

$$(\mu_1 \times \mu_2)(E) = \int \mu_1(E_{\omega_2})\, \mu_2(d\omega_2). \tag{3.3}$$

Equation (3.3) is called a *disintegration formula.*

Step 3. Countable Addititivity.
By finite additivity, (3.3) extends to a definition of $\mu_1 \times \mu_2$ on finite disjoint unions of elements of $\mathfrak{A}_0$; i.e., the above holds for all $E \in \mathfrak{A}$. Furthermore, the dominated convergence theorem shows that $\mu_1 \times \mu_2$ is countably additive on the algebra $\mathfrak{A}$.[3.1] Therefore, owing to the Carathéodory extension theorem (Theorem 1.15), $\mu_1 \times \mu_2$ can be extended uniquely to a measure on all of $\mathfrak{F}_1 \times \mathfrak{F}_2$. This proves the theorem. In addition, it shows that (3.3) holds for all $E \in \mathfrak{F}_1 \times \mathfrak{F}_2$. (The fact that $\omega_2 \mapsto \mu_1(E_{\omega_2})$ is measurable is proved implicitly here; why?). □

The following shows that the two possible ways of constructing Lebesgue's measure coincide.

Corollary 3.5 *If m^d designates the Lebesgue measure on the measure space $((0,1]^d, \mathfrak{B}((0,1]^d))$, then $m^d = m^1 \times \cdots \times m^1$ (d times.)*

Proof If $E = (a_1, b_1] \times \cdots \times (a_d, b_d]$ is a d-dimensional hypercube, then $m^d(E) = \prod_{j=1}^d (b_j - a_j) = (m^1 \times \cdots \times m^1)(E)$. By finite additivity, m^d and $(m^1 \times \cdots \times m^1)$ agree on the smallest algebra that contains hypercubes, and by Carathéodory's extension theorem, m^d and $(m^1 \times \cdots \times m^1)$ agree on the σ-algebra generated by hypercubes. □

An important consequence of our development thus far is the following.

Theorem 3.6 (The Fubini–Tonelli Theorem) *If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is product measurable, then for each $\omega_1 \in \Omega_1$, $\omega_2 \mapsto f(\omega_1, \omega_2)$ is $\mathfrak{F}_2$-measurable, and by symmetry, for each $\omega_2 \in \Omega_2$, $\omega_1 \mapsto f(\omega_1, \omega_2)$ is $\mathfrak{F}_1$-measurable. If*

[3.1] To prove this, it is enough to show that if $E^N \in \mathfrak{A}$ satisfy $E^N \downarrow \varnothing$, then $\mu_1 \times \mu_2(E^N) \downarrow 0$. But this follows from (3.3) and the monotone convergence theorem (Theorem 1.27).

in addition $f \in L^1(\mu_1 \times \mu_2)$, then the following are a.e.-finite measurable functions:

$$\omega_1 \mapsto \int f(\omega_1, \omega_2)\, \mu_2(d\omega_2), \quad \omega_2 \mapsto \int f(\omega_1, \omega_2)\, \mu_1(d\omega_1). \tag{3.4}$$

Finally, the following change-of-variables formula is valid:

$$\begin{aligned} \int f\, d(\mu_1 \times \mu_2) &= \int \left(\int f(\omega_1, \omega_2)\, \mu_1(d\omega_1) \right) \mu_2(d\omega_2) \\ &= \int \left(\int f(\omega_1, \omega_2)\, \mu_2(d\omega_2) \right) \mu_1(d\omega_1). \end{aligned} \tag{3.5}$$

Proof (Sketch) If $f = \mathbf{1}_E$ for some $E \in \mathfrak{F}_1 \times \mathfrak{F}_2$, then our disintegration formula (3.3) contains (3.4) and (3.5). In other words, these two equations hold for all elementary functions f. By linearity, they also hold for simple functions. Finally, we take limits to prove the result for every function $f \in L^1(\mu_1 \times \mu_2)$. □

The following is an important corollary of the proof of Fubini–Tonelli's theorem. (To prove it, approximate f from below by simple functions and use the monotone convergence theorem.)

Corollary 3.7 *If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is measurable and nonnegative, then (3.5) holds in the sense that all three double-integrals converge and diverge together and are equal in the convergent case.*

Remark 3.8 Careful examination of the content of the above proof shows us a little more. Namely, that whenever any of the three integrals in (3.5) are finite when f is replaced by $|f|$, then all three are finite where f is replaced by $|f|$, and in this case (3.5) holds.

Corollary 3.9 *The Fubini–Tonelli theorem continues to hold if μ_1 and μ_2 are σ-finite.*

Proof We can assume without loss of generality that $f \geq 0$ (else consider f^+ and f^- separately), and find measurable $K_n \uparrow \Omega_1 \times \Omega_2$ such that $(\mu_1 \times \mu_2)(K_n) < +\infty$. What we have shown thus far implies that everything is fine on K_n. Then take limits using the monotone convergence theorem (Theorem 2.21). □

Fubini-Tonelli's theorem is a deceptively delicate result as the following two examples show.

Example 3.10 Let $(\Omega, \mathfrak{F}, \mathrm{P})$ denote the Borel–Steinhaus probability space; i.e., $\Omega := (0,1]$, $\mathfrak{F} := \mathfrak{B}(\Omega)$, and $\mathrm{P} :=$ the Lebesgue on Ω. For every integer $n \geq 0$, define ψ_n to be the unique piecewise-linear continuous function that is (i) nonnegative everywhere, and 0 outside $(2^{-n-1}, 2^{-n})$; (ii) symmetric about the middle point of $(2^{-n-1}, 2^{-n})$; and (iii) has total area 1. That is,

$$\psi_n(x) := \begin{cases} 2^{2n+4}x - 2^{n+3}, & \text{if } 2^{-n-1} \leq x < 3 \cdot 2^{-n-2}, \\ -2^{2n+4}x + 2^{n+4}, & \text{if } 3 \cdot 2^{-n-2} \leq x < 2^{-n}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

Now define the measurable function f on $\Omega \times \Omega$ as follows:

$$f(x,y) := \sum_{n=0}^{\infty} \Big[\psi_n(x) - \psi_{n+1}(x)\Big]\psi_n(y), \qquad \forall x, y \in \Omega. \tag{3.7}$$

All but one of these terms are zero, so the function f is perfectly well-defined. Moreover, $\int_0^1 \int_0^1 f(x,y)\,dx\,dy = 0 \neq 1 = \int_0^1 \int_0^1 f(x,y)\,dy\,dx$. Thus, Fubini–Tonelli's theorem fails, and the reason is that $\int |f|\,d(\mathrm{P} \times \mathrm{P}) = +\infty$ in this case. (Prove it!)

Example 3.11[Sierpinski [Sie20]] In this example we show that Fubini–Tonelli's theorem need not hold when f is not product-measurable. To do so, we will rely on the axiom of choice, as well as the continuum hypothesis.[3.2] Throughout, $(\Omega, \mathfrak{F}, \mathrm{P})$ denotes the Borel–Steinhaus probability space.

[3.2] While assuming the continuum hypothesis may make some readers uncomfortable, assuming the axiom of choice should not. For example, I remind you that without the axiom of choice one could not even prove the following theorem of G. Cantor: "*A countable union of countable sets is countable.*" In other words, the axiom of choice is well-lodged in the theories of measure and probability alike. On the other hand, it is likely that a suitable "axiom of non-choice" would be enough to develop a theory of probability that is as useful as the existing theory, and yet avoids the nuances of measure theory altogether. In this connection, see the landmark paper of Solovay [Sol70].

First we define S to be the collection of all ordinal numbers less than or equal to c := the first uncountable ordinal. (In logic this is called Hartog's c-section of the ordinal numbers. Note that the existence of c implicitly relies on the axiom of choice.)

Since S has the power of the continuum, according to the continuum hypothesis, we can find a one-to-one map $\phi : [0,1] \to S$. Now consider the set $E := \{(x,y) \in [0,1]^2 : \ \phi(x) < \phi(y)\}$. For any $x \in [0,1]$, consider the x-section of E defined as $E_x := \{y \in [0,1] : \ (x,y) \in E\} = \{y \in [0,1] : \phi(x) < \phi(y)\}$. Since ϕ is one-to-one, E and E_x are nonempty, and moreover E_x^{c} is denumerable. Consequently, E_x is Borel measurable, and $\mathrm{P}(E_x) = 1$ for all $x \in [0,1]$. On the other hand, we can also define the y-section, ${}_yE := \{x \in [0,1] : \ (x,y) \in E\}$, for any $y \in [0,1]$ and note that ${}_yE$ is denumerable, so that ${}_yE \in \mathfrak{F}$ and $\mathrm{P}({}_yE) = 0$. In particular,

$$\int_0^1 \mathrm{P}(E_x)\,\mathrm{P}(dx) = 1 \neq 0 = \int_0^1 \mathrm{P}({}_yE)\,\mathrm{P}(dy). \tag{3.8}$$

That is, there is no disintegration formula (3.3), which means that Fubini–Tonelli's theorem (cf. equation 3.5) does not hold for the bounded function $f(x,y) = \mathbf{1}_E(x,y)$. But P is a probability measure on the Borel subsets of $[0,1]$. Therefore, all bounded measurable functions are P-integrable, and we see that the source of the difficulty is that f is not product-measurable although $x \mapsto \int f(x,y)\,\mathrm{P}(dy)$ and $y \mapsto \int f(x,y)\,\mathrm{P}(dx)$ are measurable (in fact constants).

Remark 3.12 Mattner [Mat99, §2.2] has constructed a Borel set $A \subset \mathbb{R}$ and two σ-finite measures μ_1 and μ_2 on Borel subsets of $\Omega := \mathbb{R}$ such that if we ignore measurability issues, then we would have the following:

$$\begin{aligned}\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \mathbf{1}_A(x+y)\,\mu_1(dx) \right) \mu_2(dy) \\ \neq \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \mathbf{1}_A(x+y)\,\mu_2(dy) \right) \mu_1(dx).\end{aligned} \tag{3.9}$$

This is interesting since (i) it does not rely on the axiom of choice (nor on the continuum hypothesis); and (ii) shows that the "convolution" $y \mapsto \int f(x-y)\,\mu_1(dx)$ need not be measurable with respect to the smallest σ-algebra with respect to which all functions $\{x \mapsto f(x-y);\ y \in \mathbb{R}\}$ are measurable.

3 Infinite Products

So far, the Lebesgue measure on $(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ is essentially our only nontrivial example of a measure. We have also seen that once we have a few nice measures, we can create other interesting product measures, but for the Lebesgue measure, this procedure does not produce anything new since finite products of the Lebesgue measure only yield the Lebesgue measure on the higher-dimensional product space.

We now wish to add to our repertoire of nontrivial measures by defining measures on infinite-product spaces that we take to be $(0, 1]^\infty$ or $\mathbb{R}^\infty$, where for any Ω the set Ω^∞ is defined as the collection of all infinite sequences of the form $(\omega_1, \omega_2, \ldots)$ where $\omega_i \in \Omega$.

> *Warning:* In case you have not seen infinite-product measures before, read this section with care. The notation can be a bit taxing, but this is important material that you will need to know.

In order to construct measures on $(0, 1]^\infty$, or more generally $\mathbb{R}^\infty$, we first need a topology in order to have a Borel σ-algebra $\mathfrak{B}((0, 1]^\infty)$. For this, we need to borrow a bit from general topology.

Definition 3.13 Given a topological set Ω, a set $A \subseteq \Omega^\infty$ is called a *cylinder set*, if it has the form $A = A_1 \times A_2 \times \cdots$ where for all but a finite number of i's, $A_i = \Omega$. A cylinder set $A = A_1 \times A_2 \times \cdots$ is *open* if every A_i is open in Ω. The *product topology* on Ω^∞ is the smallest topology that contains all open cylinder sets. This, in turn, gives us the Borel σ-algebra $\mathfrak{B}(\Omega^\infty)$.

Suppose we wanted to construct the Lebesgue measure on $(0, 1]^\infty$. Note that any cylinder set has a perfectly well-defined Lebesgue measure in the following sense: Let $I_\ell = (0, \ell^{-1}]$ for $\ell = 1, 2, 3$, and $I_\ell = (0, 1]$ for $\ell \geq 4$. Then, $I = I_1 \times I_2 \times I_3 \times I_4 \times \cdots$ is a cylinder set, and no one would deny that the "Lebesgue measure" of I should be $1 \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$, since this is the 3-dimensional Lebesgue measure of the 3-dimensional set $I_1 \times I_2 \times I_3$. It stands to reason that if m denotes the one-dimensional Lebesgue measure, one should be able to define the Lebesgue measure $m^\infty := m \times m \times \cdots$ on $((0, 1]^\infty, \mathfrak{B}((0, 1]^\infty))$ as the (or perhaps *a*) "projective limit" of the n-dimensional Lebesgue measure $m^n := m \times \cdots \times m$ on $((0, 1]^n, \mathfrak{B}((0, 1]^n))$. In fact, we will make this argument rigorous not only for m^∞, but for a large class of other measures as well. To do this, we need some notation for projections.

For all $1 \leq n \leq \infty$, let $\mathcal{I}^n := (0,1]^n$, and $\mathfrak{B}^n := \mathfrak{B}(\mathcal{I}^n)$. If $A = A_1 \times A_2 \times \cdots$ is a subset of $\mathcal{I}^\infty$, then for any integer $n \geq 1$ we can project A onto its first n coordinates as follows: $\pi_n(A) := A_1 \times \cdots \times A_n$. We will use this notation throughout.

Definition 3.14 We say that a set $A = A_1 \times A_2 \times \cdots \in \mathcal{I}^\infty$ is a *cylinder set* if either $A = \varnothing$ or else there exists $n \geq 1$ such that for all $i > n$, $A_i = (0,1]$. We define the *dimension* of a nonempty set A as

$$\dim(A) := \inf\left\{n \geq 1 : \ \forall i > n, \ A_i = (0,1]\right\}, \tag{3.10}$$

where $\inf\varnothing := +\infty$, and define $\dim(\varnothing) := 0$.

Note that $\dim(\varnothing) = 0$ and $\dim(\mathcal{I}^\infty) = +\infty$. Moreover, $A \in \mathcal{I}^\infty$ is a nonempty cylinder set if and only if all but a finite number of its coordinates are equal to $(0,1]$, and all cylinder sets are finite-dimensional. In case $\dim(A) = n$, then $A_n \neq (0,1]$, but $A_{i+n} = (0,1]$ for all $i \geq 1$, and you should think of $A \in \mathcal{I}^\infty$ as the natural embedding (or lifting) of the n-dimensional (in the Euclidean sense) set $\pi_n(A) = A_1 \times \cdots \times A_n \in \mathcal{I}^n$ on to $\mathcal{I}^\infty$.

Definition 3.15 A family $\{(\mathcal{I}^n, \mathfrak{B}(\mathcal{I}^n), \mathrm{P}^n);\ n = 1, 2, \ldots\}$ of probability spaces is *consistent* if for all $n \geq 1$ and all $A_1, \ldots, A_n \in \mathfrak{B}(\mathcal{I})$, $\mathrm{P}^n(A_1 \times \cdots \times A_n) = \mathrm{P}^{n+1}(A_1 \times \cdots A_n \times (0,1])$. We will also say that (P^n) is consistent.

Remark 3.16 Here is an alternative way to think of a consistent family (P^n): Suppose $n := \dim(A)$ is a finite and (strictly) positive integer. Then, for all $m \geq n$, $\mathrm{P}^m(\pi_m(A)) = \mathrm{P}^n(\pi_n(A))$.

The notation is admittedly heavy-handed, but once you understand it, you are ready for the beautiful theorem of A. N. Kolmogorov whose proof is written out later on in §4.

Theorem 3.17 (The Kolmogorov Extension Theorem) *Suppose* (P^n) *is a consistent family of probability measures on each of the spaces* $(\mathcal{I}^n, \mathfrak{B}^n)$*. Then, there exists a unique probability measure* P^∞ *on* $(\mathcal{I}^\infty, \mathfrak{B}^\infty)$ *such that for any finite* n*, and all* n*-dimensional sets* $B \in \mathfrak{B}^\infty$*,* $\mathrm{P}^\infty(B) = \mathrm{P}^n(\pi_n(B))$*.*

The only hard part of the proof is in proving countable additivity, and this is similar to the analogous proof for the Lebesgue measure.

Remark 3.18 One can use Theorem 3.17 to construct the Lebesgue measure on $((0,1]^\infty, \mathfrak{B}((0,1]^\infty))$.

Remark 3.19 One can just as easily prove Theorem 3.17 on the measure space $(\mathbb{R}^\infty, \mathfrak{B}(\mathbb{R}^\infty))$, where $\mathbb{R}^\infty$ is endowed with the product topology.

4 Proof of Kolmogorov's Extension Theorem (Optional)

Here is the strategy of the proof in a nutshell: Let $\mathfrak{A}$ denote the collection of all finite unions of cylinder sets of the form $(a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_k, b_k] \times (0,1] \times (0,1] \times \cdots$ where $0 \le a_i < b_i \le 1$ for all i, and $k \ge 1$. We also add $\varnothing$ and $\mathcal{I}^\infty$ to $\mathfrak{A}$ and, in this way, it follows from the definition of the product topology that $\mathfrak{A}$ is an algebra that generates $\mathfrak{B}^\infty$. Our goal is to construct a countably additive measure on $\mathfrak{A}$ and then appeal to Carathéodory's theorem (Theorem 1.15) to finish.

Our definition of P^∞ is both simple and intuitively appealing: First, define $\mathrm{P}^\infty(\varnothing) = 0$ and $\mathrm{P}^\infty(\mathcal{I}^\infty) := 1$. This takes care of the 0-dimensional element $(\varnothing)$ as well as the infinite-dimensional element $(\mathcal{I}^\infty)$ of $\mathfrak{A}$. Now, suppose $A \in \mathfrak{A}$ is such that $n = \dim(A) < +\infty$. Then we let $\mathrm{P}^\infty(A) := \mathrm{P}^n(\pi_n(A))$. This is well defined (i.e., does not depend on any given representation of A) since P^n is a measure on $(\mathcal{I}^n, \mathfrak{B}^n)$; you should check the details.

Step 1. Finite Additivity.

Let us first check that P^∞ is finitely additive on $\mathfrak{A}$. We want to show that if $A, B \in \mathfrak{A}$ are disjoint, then $\mathrm{P}^\infty(A \cup B) = \mathrm{P}^\infty(A) + \mathrm{P}^\infty(B)$. If $A = \varnothing$ or $\mathcal{I}^\infty$, then $B = A^\complement$, and finite additivity holds trivially from the fact that by definition, $\mathrm{P}^\infty(\varnothing) = 1 - \mathrm{P}^\infty(\mathcal{I}^\infty) = 0$.

If neither A nor B is $(0,1]^\infty$, then $n := \dim(A)$ and $m := \dim(B)$ are non-trivial numerals. Without loss of generality, we may suppose that $n \ge m$, in which case $\mathrm{P}^\infty(A \cup B) = \mathrm{P}^n(\pi_n(A) \cup \pi_n(B)) = \mathrm{P}^n(\pi_n(A)) + \mathrm{P}^n(\pi_n(B))$, since $\pi_n(A) \cap \pi_n(B) = \varnothing$ and P^n is a measure. On the other hand, $\mathrm{P}^n(\pi_n(A)) = \mathrm{P}^\infty(A)$, and since (P^k) is a consistent family, $\mathrm{P}^n(\pi_n(B)) = \mathrm{P}^m(\pi_m(B))$ (cf. Remark 3.16). The latter equals $\mathrm{P}^\infty(B)$, and we have verified finite additivity.

Step 2. Countable Additivity.
Thanks to Carathéodory's extension theorem (Theorem 1.15), P^∞ can be uniquely extended from a countably additive measure on $\mathfrak{A}$ to a countably additive measure on $\sigma(\mathfrak{A}) = \mathfrak{B}^\infty$, and this extension (still written as P^∞) is the probability measure on $(\mathcal{I}^\infty, \mathfrak{B}^\infty)$ that is stated in the theorem. Thus, it suffices to establish the countable additivity of P^∞ on $\mathfrak{A}$. This uses a similar argument as the one used to show that the Lebesgue measure on $(0,1]$ is countably additive on finite unions of intervals of the form $(a,b]$; make certain that you understand the proof of Lemma 1.14 before proceeding with the present proof.

Let $A^1, A^2, \ldots$ denote disjoint sets in $\mathfrak{A}$ such that $\cup_{j=1}^\infty A^j$ is also in $\mathfrak{A}$; we need to verify that $\mathrm{P}^\infty(\cup_{j=1}^\infty A^j) = \sum_{j=1}^\infty \mathrm{P}^\infty(A^j)$. On the other hand, $\cup_{j=1}^\infty A^j = (\cup_{j=1}^N A^j) \cup (\cup_{j=N}^\infty A^j)$, and $\cup_{j=1}^N A^j$ and $\cup_{j=N}^\infty A^j$ are disjoint elements of $\mathfrak{A}$. By Step 1, $\mathrm{P}^\infty\left(\cup_{j=N}^\infty A^j\right) = \sum_{j=1}^N \mathrm{P}^\infty(A^j) + \mathrm{P}^\infty\left(\cup_{j=N}^\infty A^j\right)$. Thus, it suffice to show that whenever $B^N \downarrow \varnothing$—all in $\mathfrak{A}$—then $\mathrm{P}^\infty(B^N) \downarrow 0$. We suppose to the contrary and derive a contradiction. That is, we suppose that there exists $\varepsilon > 0$ such that for all $n \geq 1$, $\mathrm{P}^\infty(B^n) \geq \varepsilon$. These remarks make it clear that $\dim(B^n)$ is strictly positive (i.e., $B^n \neq \varnothing$) and finite (i.e., $B^n \neq \mathcal{I}^\infty$). Henceforth, let $\gamma(n) := \dim(B^n)$, and note that the condition $B^n \downarrow \varnothing$ forces $\gamma(n) \uparrow +\infty$. Thus,

$$B^n = B_1^n \times \cdots \times B_{\gamma(n)}^n \times (0,1] \times (0,1] \times \cdots, \tag{3.11}$$

where $B_m^n := \cup_{j=1}^{k(n,m)} (a_j^{n,m}, b_j^{n,m}]$ $(m \leq \gamma(n))$. Now define C^n to be an approximation from inside to B via closed intervals, viz.,

$$C^n := C_1^n \times \cdots C_{\gamma(n)}^n \times (0,1] \times (0,1] \times \cdots, \tag{3.12}$$

where $C_m^n = \cup_{i=1}^{k(n,m)} [\alpha_i^{n,m}, b_i^{n,m}]$ $(m \leq \gamma(n))$, and the $\alpha_i^{n,m} \in (a_i^{n,m}, b_i^{n,m})$ are so close to the a's that

$$\mathrm{P}^\infty(B^j \setminus C^j) \leq \varepsilon 2^{-j}, \qquad \forall j \geq 1. \tag{3.13}$$

Proof. This can be always be done because $\mathrm{P}^\infty(B^j \setminus C^j)$ is

$$\mathrm{P}^{\gamma(j)}\left(\bigcup_{i=1}^{k(j,1)} \left[a_i^{j,1}, \alpha_i^{j,1}\right] \times \cdots \times \bigcup_{i=1}^{k(j,\gamma(j))} \left[a_i^{j,\gamma(j)}, \alpha_i^{j,\gamma(j)}\right] \right),$$

and $\mathrm{P}^{\gamma(j)}$ is a measure on $(\mathcal{I}^{\gamma(j)}, \mathfrak{B}^{\gamma(j)})$. □

Therefore, thanks to (3.13), $\mathrm{P}^\infty(D^n) \geq (\varepsilon/2)$, where $D^n := \cap_{j=1}^n C^j$ is a sequence of decreasing sets with $D^n \subseteq [0,1]^{\gamma(n)} \times (0,1] \times (0,1] \times \cdots$. Now we argue that $\cap_{n=1}^\infty D^n \neq \varnothing$; since $D^n \subseteq B^n$, this contradicts $B^n \downarrow \varnothing$ and we would be finished.

We know that $D_n \neq \varnothing$ for any finite n since $\mathrm{P}^\infty(D^n) \geq (\varepsilon/2)$. Moreover, we can write $D^n := D_1^n \times D_2^n \times \cdots$, where (a) for all $j > \gamma(n)$, $D_j^n = (0,1]$; and (b) for all $j \leq \gamma(n)$, D_j^n is closed in $[0,1]$. So we can choose for each $n \geq 1$, a point $x^n \in D^n$ of the following form:

$$x^n := \left(x_1^n, x_2^n, \ldots x_{\gamma(n)}^n, \frac{1}{2}, \frac{1}{2}, \cdots \right), \qquad \forall n \geq 1. \tag{3.14}$$

Since $\gamma(n) \uparrow +\infty$, and since $D_1^1 \supseteq D_1^2 \supseteq D_1^3 \supseteq \cdots$ is a decreasing sequence of closed subsets of $[0,1]$, $z_1 := \lim_{\ell\to\infty} x_1^\ell \in \cap_{n=1}^\infty D_1^n$. Similarly, for any $j \geq 1$, $z_j := \lim_{\ell\to\infty} x_j^\ell \in \cap_{n=1}^\infty D_j^n$. In particular, we have found a point $z := (z_1, z_2, z_3 \ldots)$ in $\cap_{n=1}^\infty D^n$. So $\cap_{n=1}^\infty D^n \neq \varnothing$, which is a contradiction.

5 Exercises

Exercise 3.1 Given an uncountable set Ω, let $\mathfrak{F}$ denote the collection of all subsets $A \subseteq \Omega$ such that either A or $A^\complement$ is denumerable.

1. Prove that $\mathfrak{F}$ is a σ-algebra.

2. Define the set function $\mathrm{P} : \mathfrak{F} \to \{0,1\}$ by: $\mathrm{P}(A) := 1$ if A is uncountable, and $\mathrm{P}(A) := 0$ if A is denumerable. Prove that P is a probability measure on $(\Omega, \mathfrak{F})$.

3. Use only the axiom of choice to construct a set Ω, and an $E \subseteq \Omega \times \Omega$ such that for all $x, y \in \Omega$, E_x and $({}_yE)^\complement$ are denumerable, where $E_x := \{y \in \Omega :\ (x,y) \in E\}$ and ${}_yE := \{x \in \Omega :\ (x,y) \in E\}$.

Exercise 3.2 If $\mu_1, \mu_2, \ldots$ are probability measures on $([0,1], \mathfrak{B}([0,1]))$, carefully make sense of the probability measure $\prod_{\ell=1}^\infty \mu_\ell$. Use this to construct the Lebesgue measure m on $[0,1]^\infty$ endowed with its product σ-algebra. Finally, if $1 > a_1 > a_2 > \cdots a_n \downarrow 0$, then prove that

$$m\left(\prod_{\ell=1}^\infty [a_\ell, 1] \right) > 0 \quad \text{if and only if} \quad \sum_{\ell=1}^\infty a_\ell < +\infty. \tag{3.15}$$

We will do much more on this subject when we see the Borel–Cantelli lemma later on in Theorem 5.23.

Exercise 3.3 Consider a set-valued function $\mathbf{X}$ on our probability space $(\Omega, \mathfrak{F}, \mathrm{P})$. Specifically, $\mathbf{X} : \Omega \to \mathfrak{P}(\mathbb{R}^d)$, where $\mathfrak{P}(\mathbb{R}^d)$ denotes the power set of $\mathbb{R}^d$. We say that $\mathbf{X}$ is a *random set* if $(\omega, x) \mapsto \mathbf{1}_{\mathbf{X}(\omega)}(x)$ is product measurable on a σ-finite measure space $(\Omega \times \mathbb{R}^d, \mathfrak{F} \times \mathfrak{B}(\mathbb{R}^d), \nu)$ where ν is a given measure. Prove:

1. If $A \in \mathfrak{B}(\mathbb{R}^d)$, then A and $\mathbf{X} \cap A$ are both random sets.
2. If $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ are random sets, then so are $\mathbf{X}^{\complement}$, $\cap_{n=1}^{\infty} \mathbf{X}_n$, and $\cup_{n=1}^{\infty} \mathbf{X}_n$.
3. If $A \in \mathfrak{B}(\mathbb{R}^d)$ satisfies $\lambda(A) < +\infty$ where λ is a σ-finite measure on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$, then $\lambda(\mathbf{X} \cap A)$ is a finite random variable.
4. Prove that for any $A \in \mathfrak{B}(\mathbb{R}^d)$ such that $\lambda(A) < +\infty$, and all integers $k \geq 1$,
$$\|\lambda(\mathbf{X} \cap A)\|_k^k = \int_A \cdots \int_A \mathrm{P}\left\{x_1 \in \mathbf{X}, \ldots, x_k \in \mathbf{X}\right\} \lambda(dx_1) \cdots \lambda(dx_k), \tag{3.16}$$
where $\| \cdots \|_k$ denotes the $L^k(\mathrm{P})$-norm.
5. Show that $\mathrm{P}\{x \in \mathbf{X}\} = 0$ for λ-almost every $x \in \mathbb{R}^d$ if and only if $\lambda(\mathbf{X}) = 0$, P-a.s.
6. There exists a nonempty random set $\mathbf{X}$ such that for all $x \in \mathbb{R}^d$, $\mathrm{P}\{x \in \mathbf{X}\} = 0$.
(Hint: If X is a random variable, then first prove that $\mathbf{X}(\omega) := \{X(\omega)\}$ defines a random set.)

Chapter 4

The Radon–Nikodým Theorem

1 Introduction

Given two measures μ and ν, one can ask, "*When can we find a function $\pi_\star$ such that for all measurable sets A, $\nu(A) = \int_A \pi_\star \, d\mu$?*". If μ is the Lebesgue measure, then the function $\pi_\star$ is a probability density function, and the prescription $\nu(A) = \int_A \pi_\star \, d\mu$ defines a probability measure ν. A famous example of this is the (standard) normal (or Gaussian) distribution. This is precisely the measure ν when $\pi_\star(x) := (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2\right)$. Later on when studying conditional expectations, we will see a substantially more important consequence of the Radon–Nikodým theorem. However, this discussion will have to wait.

2 The Radon–Nikodým Theorem

Definition 4.1 Given two measures μ and ν on $(\Omega, \mathfrak{F})$, we say that ν is *absolutely continuous* with respect to μ (written $\nu \ll \mu$) if for any $A \in \mathfrak{F}$ such that $\mu(A) = 0$, then also $\nu(A) = 0$.

For instance, suppose that $(\Omega, \mathfrak{F}, \mu)$ is a finite measure space and $f \in L^1(\mu)$. Let $\nu(A) := \int_A f \, d\mu$ and note that ν is a finite measure that is also absolutely continuous with respect to μ. The following states that what we have just seen is the only example of its kind.

Theorem 4.2 (The Radon–Nikodým Theorem) *If $\nu \ll \mu$ are two finite measures on $(\Omega, \mathfrak{F})$, then there exists a nonnegative $\pi_\star \in L^1(\mu)$, such* *that for all bounded measurable functions $f : \Omega \to \mathbb{R}$,*

$$\int f \, d\nu = \int f \pi_\star \, d\mu. \tag{4.1}$$

Furthermore, this $\pi_\star$ is unique up to a μ-null set.

Remark 4.3 The function π is often written as $d\nu/d\mu$ and is referred to as the *Radon–Nikodým derivative* of ν with respect to μ.

Remark 4.4 Here is a natural way to think of Radon–Nikodým derivatives: Suppose μ is a measure on a measure space $(\Omega, \mathfrak{F})$, and suppose $f \in L^1(\mu)$ is nonnegative. Consider the abstract integral:

$$\nu(E) := \int_E f \, d\mu, \qquad \forall E \in \mathfrak{F}. \tag{4.2}$$

Then, ν is a measure also, and $\frac{d\nu}{d\mu} = f$, μ-a.e. Furthermore, if $\|f\|_1 = 1$, then ν is a probability measure, and f is its probability density function. See Examples 1.17–1.20 for some examples of Radon–Nikodým derivatives that arise in probability theory.

This is a simple but deep theorem. Here is a "geometric proof," due to J. von Neumann.

Proof (von Neumann [vN40, Lemma 3.2.3, p. 127]) First, we suppose that ν is *dominated* by μ (written $\nu \leq \mu$); i.e., that for all $A \in \mathfrak{F}$, $\nu(A) \leq \mu(A)$. If so, then we clearly have $\nu \ll \mu$, but domination of measures is a much stronger requirement than their absolute continuity. Nevertheless, once we understand the Radon–Nikodým theorem in the dominated case, the general result will follow easily.

Step 1. The Case $\nu \leq \mu$.

Consider the linear functional $\mathcal{L}(f) := \int f \, d\nu$ that acts on all $f \in L^1(\nu)$. By Jensen's inequality, $|\mathcal{L}(f)|^2 \leq \nu(\Omega) \cdot \int |f|^2 \, d\mu$. Consequently, $\mathcal{L}$ is a bounded linear functional on $L^2(\mu)$. Since the latter is complete (Theorem 2.31), the general theory of Hilbert spaces tells us that $\mathcal{L}$ is obtained by an inner product; i.e., there exists a μ-almost everywhere unique $\pi \in L^2(\mu)$ such that for all $f \in L^2(\mu)$, $\mathcal{L}(f)$ is the $L^2(\mu)$-inner product between f and π; cf. Theorem A.4 below. In other words, there exists some $\pi \in L^2(\mu)$, such that

$$\int f \, d\nu = \int f \pi \, d\mu, \qquad \forall f \in L^2(\mu). \tag{4.3}$$

If we replace f by the indicator of the measurable set $\{\pi \le -\alpha\}$ for $\alpha > 0$, we see that $\mu\{\pi \le -\alpha\} = 0$, for otherwise $\nu\{\pi \le -\alpha\}$ would be < 0. From the right-continuity of μ, it follows that $\pi \ge 0$, μ-almost everywhere. That is, we have derived the theorem for all $f \in L^2(\mu)$. By the monotone convergence theorem, this fact holds for all measurable $f \ge 0$, and the entire theorem follows for dominated measures (with $\pi_\star := \pi$).

Step 2. General ν, μ.

Given any two finite measures ν and μ on $(\Omega, \mathfrak{F})$, it is obvious that $\nu \le (\mu+\nu)$. Therefore, Step 1 extracts a μ-a.e. unique and nonnegative $\pi \in L^2(\mu + \nu)$ such that for all $f \in L^2(\mu+\nu)$, $\int f(1-\pi)\,d\nu = \int f\pi\,d\mu$. Replace f by the indicator of $\{x : \pi(x) \ge 1\}$ to deduce that $\mu\{\pi \ge 1\} = 0$. Consequently, for all $f \in L^2(\mu+\nu)$, $\int_{\{\pi<1\}} f(1-\pi)\,d\nu = \int_{\{\pi<1\}} f\pi\,d\mu$. Moreover, thanks to the monotone convergence theorem, this holds for any measurable $f \ge 0$. So we can replace f by $f(1-\pi)^{-1}\mathbf{1}_{\{\pi<1\}}$ to see that there exists a measurable function $\Pi := \pi(1-\pi)^{-1}\mathbf{1}_{\{\pi<1\}} \ge 0$, such that for any measurable $f \ge 0$,

$$\int_{\{\pi<1\}} f\,d\nu = \int f\Pi\,d\mu. \tag{4.4}$$

In general, one cannot go further and remove the $\{\pi < 1\}$ from the integral. However, if $\nu \ll \mu$, the already-proven fact that $\mu\{\pi \ge 1\}$ is 0 shows that the left-hand side holds with or without $\{\pi < 1\}$, i.e.,

$$\int f\,d\nu = \int f\Pi\,d\mu. \tag{4.5}$$

Once we show that $\Pi \in L^1(\mu)$, the theorem follows with $\pi_\star := \Pi$, but this is easy since we can plug in $f(x) \equiv 1$ in the above. □

Part II

Foundations

Chapter 5

Independence

1 Introduction

Our review/development of measure theory is finally complete, and we begin studying probability theory in earnest. In this chapter we introduce the all-important notion of independence, and use it to prove a precise formulation of the so-called law of large numbers. In rough terms, the latter states that the sample-average of a large random sample is close to the population average. As such, the law of large numbers opens the door to developing statistical means by which one can estimate various parameters of interest. We will also see more subtle and yet equally fundamental applications of independence that are routinely used in other scientific disciplines.

Throughout, let $(\Omega, \mathfrak{F}, \mathrm{P})$ denote a probability space.

2 Random Variables and Distributions

Given a random variable $X : \Omega \to \mathbb{R}$, we can define a set function on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ as follows:

$$\mathrm{P} \circ X^{-1}(E) := \mathrm{P}\{X \in E\}, \qquad \forall E \in \mathfrak{B}(\mathbb{R}). \tag{5.1}$$

The notation is ugly, but motivated by the fact that $\{X \in E\}$ is another way to write $X^{-1}(E)$, so that $\mathrm{P} \circ X^{-1}(E) = \mathrm{P}(X^{-1}(E))$.

Lemma 5.1 $\mathrm{P} \circ X^{-1}$ *is a probability measure on* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$*.*

Definition 5.2 The measure $\mathrm{P}\circ X^{-1}$ is called the *distribution* of the random variable X.

Proof of Lemma 5.1 The proof is straightforward: $\mathrm{P}\circ X^{-1}(\varnothing) = 0$, and $\mathrm{P}\circ X^{-1}$ is countably additive on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, since P is countably additive on $(\Omega, \mathfrak{F})$, and since X is a function. □

The above lemma tells us that to each random variable on the abstract probability space $(\Omega, \mathfrak{F}, \mathrm{P})$, we can associate a real probability space $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathrm{P}\circ X^{-1})$. The converse is also true.

Lemma 5.3 *If μ is a probability measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, then there exists a random variable X whose distribution is μ.*

Proof Define F to be the *distribution function* of μ; i.e.,

$$F(x) := \mu\left((-\infty, x]\right), \qquad \forall x \in \mathbb{R}. \tag{5.2}$$

Note that (i) F is nondecreasing, right-continuous, and has left-limits; (ii) $F(-\infty) = 0$; and (iii) $F(\infty) = 1$. While F need not have an inverse (it would if and only if F is strictly increasing), we can consider its "right-continuous inverse,"

$$F^{-1}(x) := \inf\left\{y : \ F(y) \geq x\right\}, \qquad \forall x \in \mathbb{R}, \tag{5.3}$$

where $\inf\varnothing := \infty$. The key feature of F^{-1} is that for any $a \in [0,1]$ and $x \in \mathbb{R}$,

$$F^{-1}(a) \leq x \text{ if and only if } a \leq F(x). \tag{5.4}$$

Consider the Borel–Steinhaus probability space $(\Omega, \mathfrak{F}, \mathrm{P})$ where $\Omega := [0,1]$, $\mathfrak{F} = \mathfrak{B}(\Omega)$, and $\mathrm{P} :=$ the Lebesgue measure on $(\Omega, \mathfrak{F})$. For all $\omega \in \Omega$ define $X(\omega) := F^{-1}(\omega)$ which is clearly a random variable on $(\Omega, \mathfrak{F}, \mathrm{P})$. Then, by (5.4), $\mathrm{P}\{X \leq x\}$ is the Lebesgue measure of all $\omega \in [0,1]$ such that $\omega \leq F(x)$, and this equals $F(x)$. What this shows is that $\mathrm{P}\{X \in (-\infty, x]\} = \mu((-\infty, x])$. Now proceed, as one does in integration, and prove whenever E is a finite disjoint unions of half-infinite half-closed intervals of the form $(-\infty, x]$, then $\mathrm{P}\{X \in E\} = \mu(E)$. By the monotone class theorem (Theorem 1.27), for all Borel sets $E \subseteq \mathbb{R}$, $\mathrm{P}\{X \in E\} = \mu(E)$. □

Since we have introduced distribution functions, let me mention a classification theorem for them too.

Theorem 5.4 *A function $F : \mathbb{R} \to [0,1]$ is the distribution function of a probability measure if and only if (i) F is nondecreasing and right-continuous; and (ii) $F(-\infty) = 0$ and $F(\infty) = 1$.*

Proof The necessity of (i) and (ii) has already been mentioned without proof; cf. the proof of Lemma 5.3. Indeed, suppose $F(x) = \mu((-\infty, x])$ for a probability measure μ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Then, F is nondecreasing because μ is a measure (specifically, since $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$.) It is right-continuous, thanks to the outer-continuity of the measure μ, and the assertions about $F(\pm\infty)$ are obvious.

Conversely, suppose $F : \mathbb{R} \to [0,1]$ satisfies (i) and (ii). We can the *define* $\mu((a,b]) := F(b) - F(a)$ for all real numbers $a < b$. Extend the definition of μ to finite disjoint unions of intervals of type $(a_i, b_i]$ by setting $\mu(\cup_{i=1}^n (a_i, b_i]) := \sum_{i=1}^n [F(b_i) - F(a_i)]$. It is not too difficult to check that (a) this is well-defined; and (b) μ is countably additive on the algebra of all disjoint finite unions of intervals of the type $(a,b]$. Now we apply Carathéodory's theorem (Theorem 1.15) to extend μ uniquely to a measure on all of $\mathfrak{B}(\mathbb{R})$. It remains to check that this extended μ is a probability measure, but this follows from $\mu(\mathbb{R}) = \lim_n \mu((-\infty, n]) = F(\infty) = 1$, thanks to the inner-continuity of measures. □

Definition 5.5 If $p > 0$, then the *pth moment* of a random variable X is defined as $\mathrm{E}\{X^p\}$ provided that $X \geq 0$, a.s., or $X \in L^p(\mathrm{P})$.

Lemma 5.6 *If $X \geq 0$, a.s. or $X \in L^p(\mathrm{P})$, then $\mathrm{E}\{X^p\} = \int_\Omega X^p \, d\mathrm{P} = \int_{-\infty}^{\infty} x^p \, \mu(dx)$, where μ is the distribution function of X. More generally still, if $h : \mathbb{R} \to \mathbb{R}$ is measurable, then $\mathrm{E}\{h(X)\} = \int_\Omega h(X) \, d\mathrm{P} = \int_{-\infty}^{\infty} h(x) \, \mu(dx)$, provided that the integrals exist.*

Proof For any random variable Y, $\mathrm{E}\{Y\} = \int_\Omega Y(\omega) \, \mathrm{P}(d\omega)$ provided that the integral is well-defined. Apply this to $Y := h(X)$ for the integral representation which involves the integrals on Ω. The second integral representations are the real message of this lemma since they state that in order for us to compute a moment (say), we can compute a real integral as opposed to an abstract integral. First one works with elementary random variable; i.e., when $X = \alpha \mathbf{1}_A$ where $A \in \mathfrak{F}$, and $\alpha \in \mathbb{R}$. Note that the distribution of X is $\mu = \mathrm{P}(A)\delta_\alpha + [1 - \mathrm{P}(A)]\delta_0$, where δ_x denotes the point mass at $\{x\}$. Thus in this case, $\mathrm{E}\{h(X)\} = h(\alpha)\mathrm{P}(A) + h(0)[1 - \mathrm{P}(A)] = \int h(x)\,\mu(dx)$. Next, one checks this formula for a simple random variable X; i.e., one of the form $X = \sum_{i=1}^n \alpha_n \mathbf{1}_{A_i}$, where $A_1, \ldots, A_n \in \mathfrak{F}$ are disjoint, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Then, $\mu = \sum_{i=1}^n \delta_{\alpha_i} \mathrm{P}(A_i) + \delta_0[1 - \mathrm{P}(\cup_{i=1}^n A_i)]$, and $\mathrm{E}\{h(X)\} = \sum_{i=1}^n h(\alpha_i)\mathrm{P}(A_i) + h(0)[1 - \mathrm{P}(A)] = \int_{-\infty}^\infty h(x)\,\mu(dx)$. Then proceed as one does when constructing the abstract integral. □

Remark 5.7 At this point, you should try and compute the pth moments, if well-defined, of the distributions of Examples 1.17–1.19.

Definition 5.8 The *variance* and the *standard deviation* of the random variable X are respectively defined as $\mathrm{Var}(X) := \mathrm{E}[(X - \mathrm{E}\{X\})^2]$ and $\mathrm{SD}(X) := \sqrt{\mathrm{Var}(X)}$. If (X, Y) is a random variable,[5.1] then the *covariance* and *correlation* between X and Y are respectively defined as $\mathrm{Cov}(X, Y) := \mathrm{E}\{(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])\}$, and $\rho(X, Y) := \mathrm{Cov}(X, Y) \div \{\mathrm{SD}(X) \times \mathrm{SD}(Y)\}$.[5.2]

Lemma 5.9 (Computational Formulas) *Provided that they exist,*

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}\{X^2\} - [\mathrm{E}\{X\}]^2 \\ \mathrm{Cov}(X, Y) &= \mathrm{E}\{XY\} - \mathrm{E}\{X\} \cdot \mathrm{E}\{Y\}. \end{aligned} \tag{5.5}$$

Furthermore, if $X \geq 0$, a.s., then for any $p \geq 1$,

$$\mathrm{E}\{X^p\} = p \int_0^\infty \lambda^{p-1} \mathrm{P}\{X \geq \lambda\}\, d\lambda. \tag{5.6}$$

In particular, $\sum_{n=1}^\infty \mathrm{P}\{X \geq n\} \leq \mathrm{E}\{X\} \leq \sum_{n=0}^\infty \mathrm{P}\{X \geq n\}$.

[5.1] Recall that this means only that $(X, Y) : \Omega \to \mathbb{R}^2$ is Borel measurable.

[5.2] Correlation was invented by K. Pearson in 1893.

3 Independent Random Variables

Definition 5.10 The events $E_1, \ldots, E_n$ are *independent*[5.3] if

$$\mathrm{P}(E_1 \cap \cdots \cap E_n) = \prod_{j=1}^{n} \mathrm{P}(E_j). \tag{5.7}$$

The random variables $X_1, \ldots, X_n : \Omega \to \mathbb{R}^d$ are *independent* if for all $A_1, \ldots, A_n \in \mathfrak{B}(\mathbb{R}^d)$, $X_1^{-1}(A_1), \ldots, X_n^{-1}(A_n) (\in \mathfrak{F})$ are independent. Equivalently, $X_1, \ldots, X_n$ are independent if for all measurable $A_1, \ldots, A_n$,

$$\mathrm{P}\left\{X_1 \in A_1, \ldots, X_n \in A_n\right\} = \prod_{j=1}^{n} \mathrm{P}\left\{X_j \in A_j\right\}. \tag{5.8}$$

An arbitrary collection $\{E_i;\ \alpha \in I\}$ of events is *independent* if for all $i_1, \ldots, i_n \in I$, $E_{i_1}, \ldots, E_{i_n}$ are independent. An arbitrary collection $\{X_i;\ i \in I\}$ is *independent* if for all $i_1, \ldots, i_n \in I$, $X_{i_1}, \ldots, X_{i_n}$ are independent random variables. If $(X_i;\ i \in I)$ are independent and identically distributed, then we say that the X_i's are *i.i.d.*

Remark 5.11 A word of caution is in order here. One can construct random variables X_1, X_2, X_3, such whenever $i \neq j$, X_i and X_j are independent, but X_1, X_2, X_3 are not independent.

Lemma 5.12 *An equivalent definition of the independence of random variables is the following: $X_1, \ldots, X_n$ are independent if for all nonnegative measurable functions $\phi_1, \ldots, \phi_n : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{E}\left[\prod_{j=1}^{n} \phi_j(X_j)\right] = \prod_{j=1}^{n} \mathrm{E}\left[\phi_j(X_j)\right]. \tag{5.9}$$

Consequently, if $X_1, \ldots, X_n$ are independent, then so are $h_1(X_1), \ldots, h_n(X_n)$ for any Borel measurable functions $h_1, \ldots, h_n : \mathbb{R} \to \mathbb{R}$.

[5.3] The definition of independence—also known as "*statistical independence*"—is due to de Moivre [dM18].

Proof I will prove only the first assertion; the second is a ready consequence of the first.

When the ϕ_j's are elementary functions, this is the definition of independence. By linearity (in each of the ϕ_j's), the above display remains to hold when the ϕ_j's are simple functions. Take limits to obtain the full result. □

We will next see that assuming independence places severe restrictions on the restrictions on the random variables in question. But first, we need a definition.

Definition 5.13 The σ-algebra generated by $\{X_i; \in i \in I\}$ is the smallest σ-algebra with respect to which all of the X_i's are measurable; it is often written as $\sigma(X_i; i \in I)$. We say that a random variable Y is independent of $\sigma(X_i;\ i \in I)$ when Y is independent of $\{X_i;\ i \in I\}$. Equivalently, we might say that $\sigma(Y)$ and $\sigma(X_i;\ i \in I)$ are independent. The *tail* σ-algebra $\mathfrak{T}$ of random variables $X_1, X_2, \ldots$ is the σ-algebra, $\mathfrak{T} := \cap_{n=1}^{\infty}\sigma(X_n, X_{n+1}, \ldots)$.

The following tell us that our definitions of independence are compatible. Moreover, the last portion implies that in order to prove that two real-valued random variables X and Y are independent, it is necessary (and sufficient) to prove that for all $x, y \in \mathbb{R}$, $\mathrm{P}\{X \le x, Y \le y\} = \mathrm{P}\{X \le x\}\mathrm{P}\{Y \le y\}$.

Lemma 5.14 *Let $\mathbb{X}$ and $\mathbb{Y}$ denote two topological spaces. Then:*

(i) For all random variables $X : \Omega \to \mathbb{X}$,

$$\sigma(X) = \left\{X^{-1}(A) :\ A \in \mathfrak{B}(\mathbb{X})\right\}. \tag{5.10}$$

(ii) If $X_1, X_2, \ldots$ are random variables all taking values in $\mathbb{X}$, then a $\mathbb{Y}$-valued random variable Y is independent of $\sigma(X_1, X_2, \ldots)$ if and only if Y is independent of $(X_1, X_2, \ldots)$.

(iii) If $Y^{-1}(E)$ is independent of $(X_1, X_2, \ldots)^{-1}(F)$ for all $E \in \mathfrak{A}$ and all $F \in \mathfrak{G}$—where $\mathfrak{A}$ and $\mathfrak{G}$ are subalgebras that generate $\mathfrak{B}(\mathbb{Y})$ and $\mathfrak{B}(\mathbb{X}^{\infty})$ respectively—then Y and $(X_1, X_2, \ldots)$ are independent.

The proof of this is relegated to the exercises. Instead, I turn to the following strange consequence of independence.

Theorem 5.15 (The Kolmogorov Zero-One Law) *If $X_1, X_2, \ldots$ are independent random variables, then their tail σ-algebra $\mathfrak{T}$ is trivial in the sense that for all $E \in \mathfrak{T}$, $\mathrm{P}(E) = 0$ or 1. Consequently, any $\mathfrak{T}$-measurable random variable is a constant, a.s.*

Proof Our strategy is to prove that any $E \in \mathfrak{T}$ is independent of itself, so that $\mathrm{P}(E) = \mathrm{P}(E \cap E) = \mathrm{P}(E)\mathrm{P}(E)$: Since $E \in \mathfrak{T}$, it follows that E is independent of $\sigma(X_1, \ldots, X_{n-1})$. Moreover, because this is true for each n, E is independent of the smallest σ-algebra that contains $\cup_n \sigma(X_1, \ldots, X_{n-1})$ (this is written as $\vee_n \sigma(X_1, \ldots, X_{n-1})$). In other words, E is independent of all of the X_i's, and hence of itself.

To conclude this proof, suppose Y is $\mathfrak{T}$-measurable; we intend to prove that there exists a constant c such that $\mathrm{P}\{Y = c\} = 1$. Since $Y = Y^+ - Y^-$, we can assume—without loss of generality—that $Y \geq 0$, a.s. in which case $\mathrm{E}\{Y\}$ exists but could be infinite. By replacing Y by $Y \wedge n$ and letting $n \to \infty$ we can assume, without loss in generality, that Y is also bounded. Let $c := \mathrm{E}\{Y\}$; this is a positive and finite number, and for all $\varepsilon > 0$, $\mathrm{P}\{Y \leq c+\varepsilon\}$ is 0 or 1. But $c = \mathrm{E}\{Y\} \geq \mathrm{E}\{Y; Y > c + \varepsilon\} \geq (c + \varepsilon)\mathrm{P}\{Y > c + \varepsilon\}$. This shows that for all $\varepsilon > 0$, $\mathrm{P}\{Y \leq c + \varepsilon\} = 1$. Hence, $Y \leq c$, a.s. (why?) Similarly, one proves that $Y \geq c$, a.s. and thus $Y = c$, a.s. □

Example 5.16[A Very Weak Law of Large Numbers] Suppose $X_1, X_2, \ldots$ are independent, and define $S_n := X_1 + \cdots + X_n$. Then, the following are almost surely constants: $\limsup_n n^{-1} S_n$ and $\liminf_n n^{-1} S_n$. Furthermore, $\mathrm{P}\{\lim_n n^{-1} S_n \text{ exists}\} = 0$ or 1, and if this probability is 1, then the said limit is also a constant, almost surely.

Our next result states that independent random variables exist.

Theorem 5.17 *If $\mu_1, \mu_2, \ldots$ are probability measures on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$, then there exist independent random variables $X_1, X_2, \ldots$—all on a suitable probability space—such that the distribution of X_i is μ_i for each $i = 1, 2, \ldots$.*

Proof Without loss of too much generality, we will assume that $d = 1$. This is for notational convenience only.

For every integer $n \geq 1$ let $\Omega^n := \mathbb{R}^n$, $\mathfrak{F}^n := \mathfrak{B}(\mathbb{R}^n)$, and $\mu^n := \mu_1 \times \cdots \times \mu_n$. Clearly, (μ^n) is a consistent family of probability measures. By the Kolmogorov extension theorem (Theorem 3.17 and the accompanying

Remark 3.19), there exists a probability measure P on $(\mathbb{R}^\infty, \mathfrak{B}(\mathbb{R}^\infty))$ that extends $\mu^1, \mu^2, \ldots$. Finally, define for all $\omega \in \mathbb{R}^\infty$ and all integers $i \geq 1$, $X_i(\omega) := \omega_i$. Since $X_i^{-1}(E_i) = E_i$, $\mathrm{P}\{X_i \in E\} = \mu_i(E)$, and $\mathrm{P}\{X_1 \in E_1, \ldots, X_n \in E_n\} = \mu^n(X_1^{-1}(E_1) \times \cdots \times X_n^{-1}(E_n)) = \prod_{j=1}^n \mu_j(E_j)$, so that the X_i's are independent and have the asserted distributions. The extension to the case of $\mathbb{R}^d$ if obtained by letting, instead, $\Omega^n := (\mathbb{R}^d)^n$, etc. □

Corollary 5.18 *Independent $L^2(\mathrm{P})$-valued random variables in $\mathbb{R}$ are uncorrelated.*

Proof Suppose X and Y are independent real-valued random variables both in $L^2(\mathrm{P})$. By the Cauchy–Bunyakovsky–Schwarz inequality (Corollary 2.26), $|\mathrm{E}\{XY\}| \leq \|X\|_2 \cdot \|Y\|_2 < \infty$. Moreover, by Theorem 5.17, $\mathrm{E}\{XY\} = \mathrm{E}\{X\} \cdot \mathrm{E}\{Y\}$. Thanks to Lemma 5.9, $\mathrm{Cov}(X, Y) = 0$, which means that $\rho(X, Y) = 0$. □

Let us conclude this section with a result of computational utility whose second portion is a consequence of Corollary 5.18.

Corollary 5.19 *If $X_1, \ldots, X_n$ are uncorrelated real random variables in $L^2(\mathrm{P})$, then $\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{j=1}^n \mathrm{Var}(X_j)$. In particular, this holds if the X_i's are independent.*

Proof We can use induction on n to reduce it to the case $n = 2$. In this case, we can use Lemma 5.9 to deduce that

$$\begin{aligned}\mathrm{Var}(X_1 + X_2) &= \mathrm{E}\left\{(X_1 + X_2)^2\right\} - \left(\mathrm{E}\{X_1 + X_2\}\right)^2 \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) - 2\mathrm{Cov}(X_1, X_2).\end{aligned} \tag{5.11}$$

The result follows from this. □

4 Khintchine's Weak Law of Large Numbers

The following so-called weak law of large numbers (WLLN) is a consequence of Corollary 5.9, and states that sample averages are, with high probability, close to population averages.[5.4] While the weak law is subsumed by the forthcoming strong law of large numbers, it gives us a good opportunity to learn more about the Markov and Chebyshev inequalities, as well as the so-called "truncation method." The latter was invented by A. A. Markov in 1907, and has been in heavy use in probability and analysis since.

Throughout this section, $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) real-valued random variables, and

$$S_n := X_1 + \cdots + X_n, \qquad \forall n \geq 1. \tag{5.12}$$

Theorem 5.20 (The WLLN; Khintchine [Khi29]) *If $X_1, X_2, \ldots$ are i.i.d. real-valued random variables in $L^1(\mathrm{P})$, then*

$$\frac{S_n}{n} \xrightarrow{\mathrm{P}} \mathrm{E}\{X_1\}, \qquad \textit{as } n \to \infty. \tag{5.13}$$

Proof We prove this in two instructive steps.

Step 1. The L^2-Case.

This is a particularly simple result when $X_1, X_2, \ldots$ are assumed to be in $L^2(\mathrm{P})$. Indeed, note that the expectation of $n^{-1}S_n$ is $\mathrm{E}\{X_1\}$. Thus, Chebyshev's inequality (Corollary 2.16) implies that for all $\varepsilon > 0$,

$$\begin{aligned} \mathrm{P}\left\{ \left| \frac{S_n}{n} - \mathrm{E}\{X_1\} \right| \geq \varepsilon \right\} &\leq \frac{1}{\varepsilon^2} \mathrm{Var}\left(\frac{S_n}{n} \right) \\ &= \frac{1}{n^2\varepsilon^2} \mathrm{Var}(S_n) = \frac{1}{n^2\varepsilon^2} \sum_{j=1}^n \mathrm{Var}(X_j). \end{aligned} \tag{5.14}$$

The last equality is a consequence of Corollary 5.19. Since the X_j's are identically distributed, they have the same variance. Therefore,

$$\mathrm{P}\left\{ \left| \frac{S_n}{n} - \mathrm{E}\{X_1\} \right| \geq \varepsilon \right\} \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}, \tag{5.15}$$

and this goes to zero as $n \to \infty$, thus proving the corollary when the X_i's are in $L^2(\mathrm{P})$.

[5.4] The first weak law of large numbers was proved by J. Bernoulli [Ber13] for ±1 random variables. See Adams [Ada74] for a fascinating account of the history of this classical theorem.

Step 2. The General Case.
When the X_i's are assumed only to be in $L^1(\mathrm{P})$, we use a *truncation argument*: For $i \leq n$ and given a fixed but small $\delta \in (0,1)$, define $X_{i,n} := X_i \mathbf{1}_{\{|X_i| \leq \delta n\}}$. For each fixed n, $X_{1,n}, \ldots, X_{n,n}$ are independent identically distributed random variables; furthermore, since $|X_{i,n}| \leq \delta n$ is bounded, they are all in $L^2(\mathrm{P})$. Thus, writing $S_{n,n} := X_{1,n} + \cdots + X_{n,n}$, we have

$$\mathrm{P}\left\{\left|\frac{S_{n,n}}{n} - \mathrm{E}\{X_{1,n}\}\right| \geq \varepsilon\right\} \leq \frac{\mathrm{Var}(X_{1,n})}{n\varepsilon^2} \leq \frac{\mathrm{E}\{X_{1,n}^2\}}{n\varepsilon^2}, \tag{5.16}$$

thanks to Lemma 5.9. On the other hand, $\mathrm{E}\{X_{i,n}^2\} \leq \mathrm{E}\{X_i^2; |X_i| \leq n\delta\} \leq n\delta \|X_1\|_1$. This yields,

$$\mathrm{P}\left\{\left|\frac{S_{n,n}}{n} - \mathrm{E}\{X_{1,n}\}\right| \geq \varepsilon\right\} \leq \frac{\delta}{\varepsilon^2}\|X_1\|_1. \tag{5.17}$$

Now $\mathrm{P}\{X_{i,n} \neq X_i\} = \mathrm{P}\{|X_i| \geq n\delta\} \leq (n\delta)^{-1}\mathrm{E}\{|X_1|; |X_1| \geq n\delta\} := (n\delta)^{-1}\mathcal{E}_{n,\delta}$, thanks to the Markov inequality (Theorem 2.15). Therefore, by finite subadditivity of measures,

$$\begin{aligned} \mathrm{P}\left\{\exists i \leq n: \; X_{i,n} \neq X_i\right\} &= \mathrm{P}\left(\bigcup_{i=1}^{n}\{X_{i,n} \neq X_i\}\right) \\ &\leq \sum_{i=1}^{n} \mathrm{P}\{X_{i,n} \neq X_i\} \leq \frac{\mathcal{E}_{n,\delta}}{\delta}. \end{aligned} \tag{5.18}$$

Since $\|X_1\|_1 < \infty$, the dominated convergence theorem (Theorem 2.22) guarantees that $\lim_n \mathcal{E}_{n,\delta} = 0$. In particular, there exists $N(\delta)$ such that for all $n \geq N(\delta)$, $\mathcal{E}_{n,\delta} \leq \delta^2$. Subsequently, for all $n \geq N(\delta)$,

$$\mathrm{P}\left\{\exists i \leq n: \; X_{i,n} \neq X_i\right\} \leq \delta. \tag{5.19}$$

Finally, by Jensen's inequality (Theorem 2.28), $|\mathrm{E}\{Z\}| \leq \mathrm{E}\{|Z|\}$, so that for all $n \geq N(\delta)$,

$$\begin{aligned} |\mathrm{E}\{X_{1,n}\} - \mathrm{E}\{X_1\}| &= |\mathrm{E}\{X_1; |X_1| > n\delta\}| \\ &\leq \mathrm{E}\left\{|X_1|; |X_1| \geq n\delta\right\} = \mathcal{E}_{n,\delta} \leq \delta^2. \end{aligned} \tag{5.20}$$

Therefore, for all $\varepsilon > \delta^2 > 0$, and $n \geq N(\delta)$,

$$\begin{aligned}
&\left\{ \left| \frac{S_n}{n} - \mathrm{E}\{X_1\} \right| \geq \varepsilon \right\} \\
&\quad \subseteq \left\{ \left| \frac{S_{n,n}}{n} - \mathrm{E}\{X_1\} \right| \geq \varepsilon \right\} \cup \{\exists j \leq n : \ X_{j,n} \neq X_j\} \qquad (5.21) \\
&\quad \subseteq \left\{ \left| \frac{S_{n,n}}{n} - \mathrm{E}\{X_{1,n}\} \right| \geq \varepsilon - \delta^2 \right\} \cup \{\exists j \leq n : \ X_{j,n} \neq X_j\}.
\end{aligned}$$

Consequently, by (5.17) and (5.19), for all $n \geq N(\varepsilon)$,

$$\mathrm{P}\left\{ \left| \frac{S_n}{n} - \mathrm{E}\{X_1\} \right| \geq \varepsilon \right\} \leq \frac{\delta}{(\varepsilon - \delta^2)^2} \|X_1\|_1 + \delta. \qquad (5.22)$$

Let $n \to \infty$ and then $\delta \downarrow 0$ to prove the theorem. □

5 Kolmogorov's Strong Law of Large Numbers

We are ready to state and prove the strong law of large numbers, so named because it is a stronger theorem than the weak law of large numbers (Theorem 5.20). Throughout this section, $X_1, \ldots, X_n$ are i.i.d. (recall that this means **i**ndependent and **i**dentically **d**istributed) random variables taking values in $\mathbb{R}$. We will write, as before, $S_n := X_1 + \cdots + X_n$ for the partial sum process.

Theorem 5.21 (The Kolmogorov Strong Law) *If the X_j's are in $L^1(\mathrm{P})$, then almost surely,*

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathrm{E}\{X_1\}. \qquad (5.23)$$

Conversely, if $\mathrm{P}\left\{\limsup_n n^{-1}|S_n| < +\infty\right\} > 0$, then the X_j's are in $L^1(\mathrm{P})$, and the strong law (5.23) holds.

Remark 5.22 In fact, independence can be relaxed to the weaker pairwise independence.[5.5] This extension was found by N. Etemadi [Ete81] whose proof is of independent interest.

[5.5] This is the case where any two X_i and X_j are independent, although $X_1, X_2, \ldots$ need not be independent.

There are many proofs of this fact; I will discuss the one that I regard as most informative. This proof relies on two key technical results, either of which is interesting in its own right.

Theorem 5.23 (Borel–Cantelli; [Bor09, Can33]) *If $A_1, A_2, \ldots$ are events, and if $\sum_{n=1}^{\infty} \mathrm{P}(A_n) < +\infty$, then $\sum_{n=1}^{\infty} \mathbf{1}_{A_n} < +\infty$, a.s. The converse holds if the A_i's are uncorrelated, in particular, pairwise independent.*

In the case that the A_n's are independent, the asserted necessary and sufficient condition is due to Borel [Bor09]. Cantelli [Can33] found that independence is not needed in the first half of the theorem.

Proof By the monotone convergence theorem (Theorem 2.21),

$$\begin{aligned}\sum_{n=1}^{\infty} \mathrm{P}(A_n) &= \lim_{N\to\infty} \sum_{n=1}^{N} \mathrm{P}(A_n) = \lim_{N\to\infty} \mathrm{E}\left\{\sum_{n=1}^{N} \mathbf{1}_{A_n}\right\} \\ &= \mathrm{E}\left\{\sum_{n=1}^{\infty} \mathbf{1}_{A_n}\right\}.\end{aligned} \tag{5.24}$$

In particular, whenever $\sum_n \mathrm{P}(A_n)$ is finite, so is $\mathrm{E}\{\sum_n \mathbf{1}_{A_n}\}$. But any nonnegative random variable that is in $L^1(\mathrm{P})$ is a.s. finite (why?); thus, $\sum_n \mathbf{1}_{A_n}$ is finite almost surely.

The converse is more interesting: Suppose that $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = +\infty$, and that the A_j's are uncorrelated. Let $Z_N := \sum_{n=1}^{N} \mathbf{1}_{A_n}$ for simplicity, and note that by Lemma 5.9, for any $N \geq 1$,

$$\begin{aligned}\mathrm{Var}(Z_N) &= \sum_{n=1}^{N} \mathrm{Var}\left(\mathbf{1}_{A_n}\right) = \sum_{n=1}^{N} \mathrm{P}(A_n)\left[1 - \mathrm{P}(A_n)\right] \\ &\leq \sum_{n=1}^{N} \mathrm{P}(A_n) = \mathrm{E}\{Z_N\}.\end{aligned} \tag{5.25}$$

We use this, together with Chebyshev's inequality (Corollary 2.16), to see that

$$\mathrm{P}\Big\{\Big|Z_N - \mathrm{E}\{Z_N\}\Big| \geq \varepsilon \mathrm{E}\{Z_N\}\Big\} \leq \frac{\mathrm{Var}(Z_N)}{\varepsilon^2 \left(\mathrm{E}\{Z_N\}\right)^2} \leq \frac{1}{\varepsilon^2 \mathrm{E}\{Z_N\}}. \tag{5.26}$$

In particular, as $N \to \infty$, $\frac{Z_N}{\mathrm{E}\{Z_N\}} \xrightarrow{\mathrm{P}} 1$. (Why?) Since $\sum_{n=1}^{N} \mathrm{P}(A_n) \uparrow \sum_{n=1}^{\infty} \mathrm{P}(A_n)$ and this is assumed to be infinite, we have shown that as

$N \to \infty$, Z_N converges in probability to $+\infty$; i.e., for any $\lambda > 0$ $\lim_{N\to\infty} \mathrm{P}\{|Z_N| > \lambda\} = 1$. Since $\sum_{n=1}^\infty \mathbf{1}_{A_n} \geq Z_N$ for all N, this shows that $\mathrm{P}\{\sum_{n=1}^\infty \mathbf{1}_{A_n} \geq \lambda\} = 1$ for all λ. Hence, $\sum_{n=1}^\infty \mathbf{1}_{A_n} = +\infty$, almost surely. □

We are now prepared to prove the second half of Theorem 5.21.

Proof of Theorem 5.21 (Necessity) For this part we suppose that the X_j's are not in $L^1(\mathrm{P})$; i.e., that $\mathrm{E}\{|X_1|\} = +\infty$. We will then show that this implies that a.s., $\limsup_n n^{-1}|S_n| = +\infty$.

According to (5.6) of Lemma 5.9, for any $k > 0$,

$$\begin{aligned} k^{-1}\mathrm{E}\{|X_1|\} &= \int_0^\infty \mathrm{P}\{|X_1| \geq k\lambda\}\, d\lambda = \sum_{n=1}^\infty \int_{n-1}^n \mathrm{P}\{|X_1| \geq k\lambda\}\, d\lambda \\ &\leq \sum_{n=1}^\infty \mathrm{P}\{|X_1| \geq k(n-1)\} = \sum_{n=0}^\infty \mathrm{P}\{|X_n| \geq kn\}. \end{aligned} \tag{5.27}$$

Since the left-hand side is infinite, by the independence half of the Borel–Cantelli lemma (Theorem 5.23), with probability one, $|X_n| \geq kn$ for infinitely many n's. Since $|S_n| \geq |X_n| - |S_{n-1}|$, then either $|S_{n-1}| \geq \frac{kn}{2}$ infinitely often, or else for all but a finite number of n's, $|S_n| \geq |X_n| - \frac{kn}{2}$. In any event, it follows that with probability one, $|S_n| \geq \frac{kn}{2}$ for infinitely many n's. In particular, there exists a null set $N(k)$ such that for all $\omega \notin N(k)$, $\limsup_n n^{-1}|S_n(\omega)| \geq \frac{k}{2}$. On the other hand, $N := \cup_{k=1}^\infty N(k)$ is a null set (why?), and for all $\omega \notin N$, $\limsup_n n^{-1}|S_n(\omega)| = +\infty$, almost surely, as was to be shown. □

The second part of the proof of the strong law of large numbers is a *maximal L^2-inequality* of A. N. Kolmogorov. Although they may be new to you at this point, maximal inequalities are some of the fundamental facts in probability and analysis.

Theorem 5.24 (The Kolmogorov Maximal Inequality) [5.6] *If $S_n := X_1 + \cdots + X_n$, and if the X_j's are independent (not necessarily i.i.d.) and in $L^2(\mathrm{P})$, then for any $\lambda > 0$, and all $n = 1, 2, \ldots$,*

$$\mathrm{P}\left\{\max_{1\leq k\leq n} \left|S_k - \mathrm{E}\{S_k\}\right| \geq \lambda\right\} \leq \frac{\mathrm{Var}(S_n)}{\lambda^2}. \tag{5.28}$$

[5.6] This result, together with its history, is well-explained in Kolmogorov [Kol33, Kol50] for instance.

Remark 5.25 If $\max_{k\le n}|S_k - \mathrm{E}\{S_k\}|$ were replaced by $|S_n - \mathrm{E}\{S_n\}|$, then the resulting weaker inequality would follow from the Chebyshev inequality (Corollary 2.16).

Proof Without loss of any generality, we may assume that $\mathrm{E}\{X_i\} = 0$, for otherwise, we can consider $X_i - \mathrm{E}\{X_i\}$ in place of X_i. Note that $\mathrm{E}\{S_n\} = n\mathrm{E}\{X_1\} = 0$ in this case.

For any $k \ge 2$, let A_k denote the event that $|S_k| \ge \lambda$ but that $|S_\ell| < \lambda$ for all $\ell < k$. In symbols,

$$A_k := \{|S_k| \ge \lambda\} \cap \bigcap_{\ell=1}^{k-1} \{|S_\ell| < \lambda\}\,. \tag{5.29}$$

Thinking of the index k as "time," A_k denotes the event that the first time the random process $k \mapsto S_k$ leaves $(-\lambda, \lambda)$ occurs at time k. Since the A_k's are disjoint events, and because $S_n^2 \ge 0$,

$$\begin{aligned}\mathrm{E}\{S_n^2\} &\ge \sum_{k=1}^n \mathrm{E}\left\{S_n^2; A_k\right\} = \sum_{k=1}^n \mathrm{E}\left\{(S_n - S_k + S_k)^2; A_k\right\}\\ &\ge 2\sum_{k=1}^n \mathrm{E}\left\{S_k(S_n - S_k); A_k\right\} + \sum_{k=1}^n \mathrm{E}\{S_k^2; A_k\}.\end{aligned} \tag{5.30}$$

For any $\omega \in A_k$, $S_k^2(\omega) \ge \lambda^2$. Hence,

$$\begin{aligned}\mathrm{E}\{S_n^2\} &\ge 2\sum_{k=1}^n \mathrm{E}\left\{S_k(S_n - S_k); A_k\right\} + \lambda^2 \sum_{k=1}^n \mathrm{P}(A_k)\\ &= 2\sum_{k=1}^n \mathrm{E}\left\{S_k(S_n - S_k); A_k\right\} + \lambda^2 \mathrm{P}\left\{\max_{1\le k\le n} |X_k| \ge \lambda\right\}.\end{aligned} \tag{5.31}$$

The event A_k and the random variable S_k both depend (in a measurable way) on $X_1, \ldots, X_k$, whereas $S_n - S_k = X_{k+1} + \cdots + X_n$ is independent of $X_1, \ldots, X_k$. Consequently, $(S_n - S_k)$ is independent of $S_k\mathbf{1}_{A_k}$ (Lemma 5.12), and from this we get $\mathrm{E}\{(S_n - S_k)S_k; A_k\} = \mathrm{E}\{S_n - S_k\} \times \mathrm{E}\{S_k; A_k\} = 0$, since $\mathrm{E}\{S_n\} = \mathrm{E}\{S_k\} = 0$. This proves the result. □

Proof of Theorem 5.21 Throughout, we may assume, without loss of generality, that $\mathrm{E}\{X_1\} = 0$, for otherwise, consider $X_i - \mathrm{E}\{X_i\}$ in place of X_i.

The proof of the strong law simplifies considerably when we assume further that $X_1 \in L^2(\mathrm{P})$. This will be the first step. The second step uses a truncation argument to reduce to a problem about L^2-random variables. Although we do not need the first step to go through the second, it is a simpler setting in which one can introduce two important techniques: Blocking and subsequencing.

Step 1. The L^2-Case.
If the X_j's are in $L^2(\mathrm{P})$, then by the Kolmogorov maximal inequality (Theorem 5.24), for all $n \geq 1$ and $\varepsilon > 0$,

$$\mathrm{P}\left\{\max_{1\leq k\leq n} |S_k| \geq \varepsilon n\right\} \leq \frac{\mathrm{E}\{S_n^2\}}{n^2\varepsilon^2} = \frac{\|X_1\|_2^2}{n\varepsilon^2}, \tag{5.32}$$

since $\mathrm{E}\{S_n^2\} = \mathrm{Var}(S_n) = \sum_{j=1}^n \mathrm{Var}(X_j) = n\mathrm{E}\{X_1^2\}$ (cf. Lemma 5.9). Now replace n by 2^n to see that

$$\sum_{n=1}^\infty \mathrm{P}\left\{\max_{1\leq k\leq 2^n} |S_k| \geq \varepsilon 2^n\right\} \leq \sum_{n=1}^\infty \frac{\|X_1\|_2^2}{2^n\varepsilon^2} < +\infty. \tag{5.33}$$

By the Borel–Cantelli lemma (Theorem 5.23), with probability one, for all but a finite number of n's, $\max_{1\leq k\leq 2^n} |S_k| \leq \varepsilon 2^n$. Now any integer m can be sandwiched between 2^n and 2^{n+1} for some n. Thus, $|S_m| \leq \max_{1\leq k\leq 2^{n+1}} |S_k|$, which is, with probability one, eventually less than $\varepsilon 2^{n+1} \leq 2\varepsilon m$. We have shown that for any fixed $\varepsilon > 0$, there exists a null set $N(\varepsilon)$ such that for all $\omega \notin N(\varepsilon)$, $\limsup_m m^{-1}|S_m(\omega)| \leq \varepsilon$. Let $N := \cup_{\varepsilon\in\mathbb{Q}_+} N(\varepsilon)$, and note that thanks to countable subadditivity of P, N is a null set. Moreover, for all $\omega \notin N$, $\lim_{m\to\infty} m^{-1}|S_m| = 0$, which is the desired result in the L^2-case.

> *Aside.* The point of the preceding proof is that while S_n and S_{n+1} are not close to being independent, the random variables S_{2^n} and $S_{2^{n+1}}$ are. Of course, this is an informal statement, since being "close to independent" is too vague to be meaningful.

Step 2. The L^1-Case.
As in the proof of the weak law of large numbers (Theorem 5.20), we truncate the X_i's. However, the truncation "levels" are chosen more carefully: For all

$i \geq 1$, define $X_i' := X_i \mathbf{1}_{\{|X_i| \leq i\}}$, and let $S_n' := X_1' + \cdots + X_n'$. Since S_i' is a sum of i independent (though not i.i.d.) random variables, by the Kolmogorov maximal inequality (Theorem 5.24), for all $n \geq 1$ and $\varepsilon > 0$,

$$\mathrm{P}\left\{ \max_{1 \leq k \leq n} \left| S_k' - \mathrm{E}\{S_k'\} \right| \geq n\varepsilon \right\} \leq \frac{\mathrm{Var}(S_n')}{n^2 \varepsilon^2}. \tag{5.34}$$

Furthermore,

$$\mathrm{Var}(S_n') = \sum_{j=1}^n \mathrm{Var}(X_j') \leq \sum_{j=1}^n \mathrm{E}\{(X_j')^2\} = \sum_{j=1}^n \mathrm{E}\{X_1^2; |X_1| \leq j\}. \tag{5.35}$$

Thus,

$$\begin{aligned} &\sum_{n=1}^\infty \mathrm{P}\left\{ \max_{1 \leq k \leq 2^n} \left| S_k' - \mathrm{E}\{S_k'\} \right| \geq \varepsilon 2^n \right\} \\ &\quad \leq \sum_{n=1}^\infty \sum_{j=1}^{2^n} \frac{\mathrm{E}\{X_1^2; |X_1| \leq j\}}{4^n \varepsilon^2} = \sum_{j \geq 1} \sum_{n \geq \log_2(j)} \frac{\mathrm{E}\{X_1^2; |X_1| \leq j\}}{4^n \varepsilon^2}. \end{aligned} \tag{5.36}$$

On the other hand, for any $x > 0$, $\sum_{n \geq x} 4^{-n} \leq \frac{4}{3} 4^{-x}$. Thus,

$$\begin{aligned} &\sum_{n=1}^\infty \mathrm{P}\left\{ \max_{1 \leq k \leq 2^n} \left| S_k' - \mathrm{E}\{S_k'\} \right| \geq \varepsilon 2^n \right\} \\ &\quad \leq \frac{4}{3\varepsilon^2} \sum_{j \geq 1} \frac{\mathrm{E}\{X_1^2; |X_1| \leq j\}}{j^2} \leq \frac{4}{3\varepsilon^2} \mathrm{E}\left\{ X_1^2 \sum_{j \geq |X_1|} j^{-2} \right\}. \end{aligned} \tag{5.37}$$

But for any $x > 1$,

$$\sum_{j \geq x} \frac{1}{j^2} \leq \int_{x-1}^\infty \frac{du}{u^2} = (x-1)^{-1} = x^{-1} + \{x(x-1)\}^{-1} \leq \frac{2}{x}. \tag{5.38}$$

On the other hand, if $x \in (0,1]$, then $\sum_{j \geq x} j^{-2} \leq \sum_{j=1}^\infty j^{-2} \leq 2$. In other words, $|X_1| \sum_{j \geq |X_1|} j^{-2} \leq 2\mathbf{1}_{\{|X_1| > 1\}} + 2\mathbf{1}_{\{|X_1| \leq 1\}} = 2$. Thus,

$$\sum_{n=1}^\infty \mathrm{P}\left\{ \max_{1 \leq k \leq 2^n} \left| S_k' - \mathrm{E}\{S_k'\} \right| \geq \varepsilon 2^n \right\} \leq \frac{8}{3\varepsilon^2} \mathrm{E}\{|X_1|\} < +\infty. \tag{5.39}$$

Now use the proof of Step 1 to deduce that $\lim_{n\to\infty} n^{-1}(S'_n - \mathrm{E}\{S'_n\}) = 0$, almost surely. But the fact that X_1 is mean-zero implies that $|\mathrm{E}\{S'_n\}| = |\sum_{j=1}^n \mathrm{E}\{X_1; |X_1| \le j\}| = |\sum_{j=1}^n \mathrm{E}\{X_1; |X_1| > j\}| \le \sum_{j=1}^n \mathrm{E}\{|X_1|; |X_1| > j\}$. In particular, since $\lim_j \mathrm{E}\{|X_1|; |X_1| > j\} = 0$ (cf. Theorem 2.22), we have $\lim_{n\to\infty} n^{-1}\mathrm{E}\{S'_n\} = 0$. It suffices to show that

$$\lim_{n\to\infty} \frac{S_n - S'_n}{n} = 0, \qquad \text{a.s.} \tag{5.40}$$

To prove this, note that

$$\mathrm{E}\left\{\sum_{j=1}^{\infty} \mathbf{1}_{\{|X_j|>j\}}\right\} = \sum_{j=1}^{\infty} \mathrm{P}\{|X_1| \ge j\} \le \mathrm{E}\{|X_1|\} < +\infty; \tag{5.41}$$

cf. Lemma 5.9. Consequently, with probability one, for all but a finite number of j's, $|X_j| \le j$, and this means that $\sup_n |S_n - S'_n| < +\infty$, a.s. (Why?). This proves (5.40), whence the strong law. □

6 Five Applications

I will conclude this chapter by applying several of these ideas to five applied problems. The first three of these topics are generally considered to be fundamental scientific discoveries, and are at the core of the theories of approximation, information, and empirical processes, respectively. Topics four and five are elegant, as well as natural starting-points for learning more about some of the far-reaching discoveries in discrete mathematics and numerical integration respectively.

6.1 The Weierstrass Approximation Theorem

The Weierstrass approximation theorem is one of the fundamental approximation theorems of analysis. It states that every continuous function (on $[0, 1]$, say) can be uniformly approximated to within any given $\varepsilon > 0$ by a polynomial. We now use the proof of the weak law (Theorem 5.20) to prove the said Weierstrass theorem; this proof is due to S. N. Bernstein.

Theorem 5.26 (Bernstein [Ber13]) *Given a continuous $f : [0,1] \to \mathbb{R}$, define the Bernstein polynomial $\mathcal{B}_n f$ by,*

$$\mathcal{B}_n f(x) := \sum_{j=0}^{n} \binom{n}{j} x^j (1-x)^{n-j} f\left(\frac{j}{n}\right), \qquad \forall x \in [0,1]. \tag{5.42}$$

Then, $\mathcal{B}_n f$ is a polynomial of order n, and $\lim_n |\mathcal{B}_n f(p) - f(p)| = 0$ uniformly in $p \in [0,1]$.

Proof We start with a bit of undergraduate probability: If $X_1, X_2, \ldots, X_n$ are i.i.d. with $\mathrm{P}\{X_1 = 0\} = 1 - \mathrm{P}\{X_1 = 1\} = 1 - p$ where $p \in [0,1]$ is a fixed number, then $\mathrm{P}\{S_n = k\} = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, \ldots, n$, and $\mathrm{P}\{S_n = k\} = 0$ otherwise. Here, as before, $S_n := X_1 + \cdots + X_n$, and its distribution is the so-called *binomial distribution*. It has the following statistical significance: If X_j's are the results of success/failure i.i.d. trials with $X_k = 1$ if and only if the kth trial resulted in a success, then S_n denotes the total number of successes in this random experiment, and its distribution is binomial.

Since f is uniformly continuous, for each $\varepsilon > 0$ we can choose $\delta > 0$ such that for all $p, q \in [0,1]$ with $|p - q| \le \delta$, we have $|f(p) - f(q)| \le \varepsilon$. Now write $\mathcal{B}_n f(p) = \mathrm{E}\{f(A_n)\}$, where $A_n := n^{-1} S_n$ denotes the average number of successes. We decompose this as $\mathcal{B}_n f(p) = \mathrm{E}\{f(A_n); |A_n - p| \le \delta\} + \mathrm{E}\{f(A_n); |A_n - p| > \delta\} := T_1 + T_2$. By the Chebyshev inequality (Corollary 2.16), $|T_2| \le K\mathrm{P}\{|A_n - p| \ge \delta\} \le K\delta^{-2}\mathrm{Var}(A_n)$, where $K := \sup_x |f(x)|$. But by Lemma 5.9, $\mathrm{Var}(A_n) = n^{-2}\mathrm{Var}(S_n) = n^{-1}\mathrm{Var}(X_1) = n^{-1}p(1-p) \le (4n)^{-1}$. Consequently, we see that $|T_2| \le K(4n)^{-1}$, uniformly in $p \in [0,1]$. It is also clear that

$$\begin{aligned} \Big| T_1 - f(p) \Big| &\le \Big| T_1 - f(p) \cdot \mathrm{P}\left\{|A_n - p| \le \delta\right\} \Big| + K\mathrm{P}\left\{|A_n - p| \ge \delta\right\} \\ &\le \varepsilon + \frac{K}{4n}, \end{aligned} \tag{5.43}$$

since $|T_1 - f(p) \cdot \mathrm{P}\{|A_n - p| \le \delta\}| = |\mathrm{E}\{f(A_n) - f(p); |A_n - p| \le \delta\}| \le \mathrm{E}\{|f(A_n) - f(p)|; |A_n - p| \le \delta\}$. Since our bounds on both T_1 and T_2 hold uniformly in $p \in [0,1]$, this concludes the proof. □

6.2 The Shannon Entropy Theorem

Our next application of independence is one of the starting-points of the work of Shannon [Sha48, SW49] who discovered various startling connections between the thermodynamical notion of relative entropy and the mathematical theory of communication.

Consider a finite "alphabet" $\mathbb{A} := \{\sigma_1, \ldots, \sigma_m\}$, and the probability measure $\mu := \sum_{j=1}^m p_j \delta_{\sigma_j}$, defined on the power set of $\mathbb{A}$, where δ_x denotes the point mass at $x \in \Omega$. Thus, $\mu(E) = \sum_{j=1}^m \mathbf{1}_E(\sigma_j)p_j$, for all $E \subseteq \mathbb{A}$. Equivalently, if X is a random variable on some probability space $(\Omega, \mathfrak{F}, \mathrm{P})$ with distribution μ, then $\mathrm{P}\{X = \sigma_j\} = p_j$ for all $j = 1, \ldots, m$.

Definition 5.27 Any element σ_i of $\mathbb{A}$ is a *symbol* or *letter*. A *word* $w := (w_1, \ldots, w_n)$ of length n is a vector of n symbols. Let $\mathbb{W}_n$ denote the collection of all words of length n, and define m *counting functions* $C_1^n, \ldots, C_m^n : \mathbb{W}_n \to \mathbb{N}$ by $C_\ell^n(w) = \sum_{k=1}^n \mathbf{1}_{\{\sigma_\ell\}}(w_k)$ $(\ell = 1, \ldots, m,\ w \in \mathbb{W}_n)$.

That is, $C_\ell^n(w)$ is the number of times the symbol σ_ℓ appears in the word w.

Definition 5.28 Fix a sequence $\lambda_1, \lambda_2, \ldots > 0$ such that $\lim_n n^{-1}\lambda_n = 0$. Then define the n-letter word $w \in \mathbb{W}_n$ to be λ-*typical* if for all $\ell = 1, \ldots, m$, we have $|C_\ell^n(w) - np_\ell| \leq \lambda_n$. Otherwise, w is said to be λ-*atypical.*

In other words, a word w is "typical" if the proportion of times that σ_ℓ appears as a symbol in w is p_ℓ give or take a negligible amount.

Theorem 5.29 (Shannon [Sha48]) *Fix any sequence $\lambda_n > 0$ such that: (i) $n^{-1}\lambda_n \to 0$; and (ii) $\limsup_n n\lambda_n^{-2} < 4m^{-1}$. Also define $T_n(\lambda)$ to be the number of λ-typical words of length n. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \log_2 (T_n(\lambda)) = -H(p), \tag{5.44}$$

where $\log_2$ is the base-2 logarithm, and $H(p) := \sum_{\ell=1}^m p_\ell \log_2(p_\ell)$ is the relative entropy of the sequence $p := (p_1, \ldots, p_m)$.

Proof Let $X_1, X_2, \ldots$ denote i.i.d. random variables with distribution μ, all defined on some probability space $(\Omega, \mathfrak{F}, \mathrm{P})$. Write $W_n := (X_1, \ldots, X_n)$—this is a randomly-sampled word of length n—and define for all $w \in \mathbb{W}_n$,

$$\pi_n(w) := \mathrm{P}\{W_n = w\} = \mathrm{P}\{X_1 = w_1, \ldots, X_n = w_n\}. \tag{5.45}$$

This is the probability of ever sampling the word w. But then

$$\begin{aligned}\sum_{\substack{w\in\mathbb{W}_n:\\ w \text{ is } \lambda\text{-atypical}}} \pi_n(w) &= \mathrm{P}\left\{\exists\ell = 1,\dots,m :\ \left|C_\ell^n(W_n) - np_\ell\right| > \lambda_n\right\}\\ &\le \sum_{\ell=1}^m \mathrm{P}\left\{\left|C_\ell^n(W_n) - np_\ell\right| > \lambda_n\right\}.\end{aligned} \tag{5.46}$$

Moreover, $C_\ell^n(W_n) = \sum_{j=1}^n \mathbf{1}_{\{X_j=\sigma_\ell\}}$ is a sum of n i.i.d. random variables with mean p_ℓ and variance $p_\ell(1-p_\ell) \le \frac{1}{4}$. Thus, by the Chebyshev inequality (Corollary 2.16),

$$\limsup_{n\to\infty} \sum_{\substack{w\in\mathbb{W}_n:\\ w \text{ is } \lambda\text{-atypical}}} \pi_n(w) \le \limsup_{n\to\infty} \frac{mn}{4\lambda_n^2} := \delta < 1. \tag{5.47}$$

On the other hand, the independence of the X_js assures us that for any $w \in \mathbb{W}_n$, $\pi_n(w) = \prod_{\ell=1}^m p_\ell^{C_\ell^n(w)}$; equivalently, $\log_2(\pi_n(w)) = \sum_{\ell=1}^m C_\ell^n(w)\log_2(p_\ell)$. Thus, whenever $w \in \mathbb{W}_n$ is λ-typical, then $|n^{-1}\log_2(\pi_n(w)) - H(p)| \le Kn^{-1}\lambda_n$, where $K := -\sum_{\ell=1}^m \log_2(p_\ell)$, and $H(p) := \sum_{\ell=1}^m p_\ell \log_2(p_\ell)$ is the relative entropy of p. Equivalently still, for all λ-typical $w \in \mathbb{W}_n$,

$$2^{nH(p)-K\lambda_n} \le \pi_n(w) \le 2^{nH(p)+K\lambda_n}. \tag{5.48}$$

To complete this proof, note that

$$1 = \sum_{w\ \lambda\text{-typical}} \pi_n(w) + \sum_{w\ \lambda\text{-atypical}} \pi_n(w). \tag{5.49}$$

Thus, by the first inequality in (5.48),

$$1 \ge \sum_{w\ \lambda\text{-typical}} \pi_n(w) \ge 2^{nH(p)-K\lambda_n} T_n(\lambda). \tag{5.50}$$

This implies that $\limsup_n n^{-1}\log_2(T_n(\lambda)) \le -H(p)$. This is half of the result. For the other half, we can combine (5.47) with the second inequality in (5.48). Namely, for all $\varepsilon > 0$, there exists n_ε such that for all $n \ge n_\varepsilon$, $\sum_{w\ \lambda\text{-atypical}} \pi_n(w) \le \delta + \varepsilon$. Hence, by (5.49), for all $n \ge n_\varepsilon$,

$$\sum_{w\ \lambda\text{-atypical}} \pi_n(w) \ge 1 - \delta - \varepsilon. \tag{5.51}$$

If we choose ε such that $1 - \delta - \varepsilon > 0$, then by (5.48), $T_n(\lambda)2^{nH(p)+K\lambda_n} \ge 1 - \delta - \varepsilon$, which has the desired effect. □

6.3 The Glivenko–Cantelli Theorem

In applied data-analysis you begin with the data $X_1, \ldots, X_n$, then "plot it," and then ideally let the data guide your analysis.[5.7] The plotting is a critical step and is often done by way of drawing a "histogram." A histogram is a random discrete probability distribution; it depends on $X_1, \ldots, X_n$, and assigns probability $p_n(x)$ to any point $x \in \mathbb{R}$, where $p_n(x)$ is the fraction of the data that is equal to x. Its cumulative distribution function is called the empirical distribution function, and defined more formally as follows.

Definition 5.30 If $X_1, X_2, \ldots$ denotes a sequence of i.i.d. random variables all with distribution function F, one can form the *empirical distribution function* F_n by

$$F_n(x) := \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k \le x\}}, \qquad \forall x \in \mathbb{R}. \tag{5.52}$$

The following is due to V. I. Glivenko and F. P. Cantelli; in statistical terms, this theorem presents a uniform approximation to an unknown distribution function F, based on a random sample from this distribution.

Theorem 5.31 (Glivenko–Cantelli [Can33, Gli33]) *Almost surely,*

$$\lim_{n\to\infty} \frac{1}{n} \sup_{-\infty<x<\infty} \Big| F_n(x) - F(x) \Big| = 0. \tag{5.53}$$

Proof Let us dispose of all measurability issues first. Since F_n and F are right-continuous, $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{x\in\mathbb{Q}} |F_n(x) - F(x)|$. Being the supremum of denumerably many random variables, it follows that $\sup_x |F_n(x) - F(x)|$ is itself a random variable. Now note that for each fixed x, $nF_n(x)$ is a sum of n i.i.d. random variables with mean $nF(x)$ and variance $nF(x)[1 - F(x)]$. Thus, by the Kolmogorov maximal inequality

[5.7] See the fundamental book of Tukey [Tuk77] for this and more, as well as the history of the subject.

(Theorem 5.24), for any $\varepsilon > 0$, $n \geq 1$, and $x \in \mathbb{R}$,

$$\begin{aligned}
&\mathrm{P}\left\{ \max_{2^n \leq \ell \leq 2^{n+1}} \left| F_\ell(x) - F(x) \right| > \varepsilon \right\} \\
&\quad \leq \mathrm{P}\left\{ \max_{1 \leq \ell \leq 2^{n+1}} \left| \ell F_\ell(x) - \ell F(x) \right| > \varepsilon 2^{n+1} \right\} \\
&\quad \leq \frac{2^{n+1} F(x)\left[1 - F(x)\right]}{\varepsilon^2 4^{n+1}} \leq \frac{1}{\varepsilon^2 2^{n+3}}.
\end{aligned} \tag{5.54}$$

I have used the fact that for any number $z \in [0,1]$, $z(1-z) \leq \frac{1}{4}$. But F is nondecreasing, right-continuous, $F(\infty) = 1$, and $F(-\infty) = 0$. Therefore, we can always find a sequence of values $x_{-m-1} < x_{-m} < \cdots < x_{-1} \leq 0 < x_1 < x_2 < \cdots < x_{m+1}$ such that: (i) $F(x_{-m}) \leq \varepsilon$; (ii) $F(x_m) \geq 1 - \varepsilon$; and (iii) for all $|j| \leq m$, $\sup_{x_{j-1} \leq x < x_j} |F(x) - F(x_{j-1})| \leq \varepsilon$. Moreover,

$$\sum_{n=1}^{\infty} \mathrm{P}\left\{ \max_{|j| \leq m+1} \max_{2^n \leq \ell \leq 2^{n+1}} \left| F_\ell(x_j) - F(x_j) \right| > \varepsilon \right\} \leq \sum_{n=1}^{\infty} \frac{m+1}{\varepsilon^2 2^n} < \infty. \tag{5.55}$$

By the Borel–Cantelli lemma, with probability one,

$$\max_{|j| \leq m+1} \max_{2^n \leq \ell \leq 2^{n+1}} \left| F_\ell(x_j) - F(x_j) \right| \leq \varepsilon, \quad \text{for all but finitely many } n\text{'s}. \tag{5.56}$$

But if $x \in [x_{j-1}, x_j)$, then for all ℓ as above, $F_\ell(x) - \varepsilon \leq F_\ell(x_j) - \varepsilon \leq F(x_j) \leq F(x) \leq F(x_{j-1}) + \varepsilon \leq F_\ell(x_{j-1}) + 2\varepsilon \leq F_\ell(x) + 2\varepsilon$. In other words, this shows that with probability one,

$$\sup_{x_{-m} \leq x \leq x_m} \max_{2^n \leq \ell \leq 2^{n+1}} \left| F_\ell(x) - F(x) \right| \leq 2\varepsilon, \quad \text{for all but finitely many } n\text{'s}. \tag{5.57}$$

On the other hand, if $x > x_m$, then $F(x) \geq F(x_m) \geq 1 - \varepsilon$ and $F_\ell(x) \geq F(x_m) - \varepsilon \geq 1 - 2\varepsilon$. Therefore, for such values of x, $|F_\ell(x) - F(x)| \leq |1 - F(x)| + |1 - F_\ell(x)| \leq 3\varepsilon$. Similarly, if $x < x_{-m}$, $|F(x) - F_\ell(x)| \leq F(x_m) + F_\ell(x_m) \leq 3\varepsilon$. Consequently, with probability one,

$$\sup_{x \in \mathbb{R}} \max_{2^n \leq \ell \leq 2^{n+1}} \left| F_\ell(x) - F(x) \right| \leq 3\varepsilon, \quad \text{for all but finitely many } n\text{'s}. \tag{5.58}$$

Let $N(\varepsilon)$ denote the null set off of which the above property holds. Then, $\cup_{\varepsilon \in \mathbb{Q}_+} N(\varepsilon)$ is a null set off of which the theorem holds. □

6.4 The Erdős Bound on Ramsey Numbers

Let us begin with a definition from graph theory.

Definition 5.32 The *complete graph* K_m on m vertices is a collection of m distinct vertices any two of which are connected by a unique edge. The nth (diagonal) *Ramsey number* R_n is the smallest integer N such that any bichromatic coloring of the edges of K_N yields a $K_n \subseteq K_N$ whose edges are all of the same color.

In other words, if $R_n = N$, then no matter how we color the edges of K_N using only the colors red and blue, then somewhere inside K_N there exists a K_n all of whose edges are either blue or red, and N is the smallest such value.

Ramsey [Ram30] introduced these and other Ramsey numbers to discuss ways of checking the consistency of a logical formula.[5.8] As a key step in his proofs, he proved that for all $n \geq 1$, $R(n) < +\infty$. It is intuitively clear that as $n \to \infty$, $R(n) \to \infty$, and one wants to know how fast this occurs. This question was answered in 1948 by P. Erdős using independent random variables.

Theorem 5.33 (Erdős [Erd48]) *For all $n \geq 3$, $R_n > 2^{n/2}$.*

Remark 5.34 As far as I know, the best known bounds are $An2^{n/2} < R_n < n^{-B}2^{2n}$ for two constants A and B. For instance, the following is likely to be true but has no known proof:

$$\lim_{n\to\infty} \frac{1}{n} \log_2(R_n) \in \left[\tfrac{1}{2}, 2\right] \text{ exists.} \tag{5.59}$$

This is a conjecture of P. Erdős; see Alon and Spencer [AS91, (3), p. 241].

[5.8] For elementary proofs, consult Skolem [Sko33] and Erdős and Szekeres [ES35].

Proof I will show that given any two integers $N \geq n$,

$$\binom{N}{n} 2^{1-\binom{n}{2}} < 1 \quad \Longrightarrow \quad R_n > N. \tag{5.60}$$

If so, we can then apply the above with $N := \lfloor 2^{n/2} \rfloor$, and note that $\binom{N}{n} \leq 2^{n^2/2} \div n!$; thus, $\binom{N}{n} 2^{1-\binom{n}{2}} \leq 2^{n/2} \div n!$, which is strictly less than 1 for all $n \geq 3$. Hence, it suffices to verify the preceding display.

Consider a random coloring of the edges of K_N; i.e., if E_N denotes the edges of K_N, then consider an i.i.d. collection of random variables $\{X_{\mathfrak{e}};\, \mathfrak{e} \in E_N\}$ where $\mathrm{P}\{X_{\mathfrak{e}} = \pm 1\} = \frac{1}{2}$. Then color $\mathfrak{e}$ red if and only if $X_{\mathfrak{e}} = 1$.

The probability that any n given vertices form a monochromatic K_n is $2^{1-\binom{n}{2}}$. Since there are $\binom{N}{n}$ many choices of these n vertices, the probability that there exist n vertices that form a monochromatic K_n is less than or equal to $\binom{N}{n} 2^{1-\binom{n}{2}}$. In other words, there are bichromatic colorings of K_N that yield no monochromatic $K_n \subseteq K_N$; i.e., $R_n > N$. □

6.5 Monte-Carlo Integration

Suppose we were to find estimate the value of some integral $\mathcal{I}_\phi := \int_{[0,1]^n} \phi(x)\, dx$, where $\phi : \mathbb{R}^n \to \mathbb{R}$ is a Lebesgue-integrable function that is so complicated that the integral $\mathcal{I}_\phi$ of interest is not explicitly computable.

One way to proceed is to first pick i.i.d. random variables $X_1, \ldots, X_N$, all chosen according to the uniform measure on $[0,1]^n$; i.e., each X_j is sampled uniformly at random from the hypercube $[0,1]^n$. By definition, for any $j = 1, \ldots, N$, $\mathrm{E}\{\phi(X_j)\} = \mathcal{I}_\phi$. Since $\phi(X_1), \ldots, \phi(X_N)$ are i.i.d. random variables with expectation $\mathcal{I}_\phi$, the Kolmogorov strong law of large numbers (Theorem 5.21) insures that for n large, $N^{-1} \sum_{j=1}^N \phi(X_j)$ is close to the desired integral. More precisely, that with probability one,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \phi(X_j) = \mathcal{I}_\phi. \tag{5.61}$$

This so-called *Monte-Carlo integration* works well when compared to other numerical integration methods when the dimension n is large.

7 Exercises

Exercise 5.1 Prove Lemma 5.9.
(Hint: For (5.6) use the Fubini–Tonelli theorem in conjunction with the fact that for $g \geq 0$ measurable, $\int g(\lambda)\mathrm{P}\{X \geq \lambda\}\, d\lambda = \mathrm{E}\{\int g(\lambda)\mathbf{1}_{[0,X]}(\lambda)\, d\lambda\}$.)

Exercise 5.2 Prove that two real-value random variables X and Y—both defined on the same probability space—are independent if and only if for all $x, y \in \mathbb{R}$, $\mathrm{P}\{X \leq x\ ,\ Y \leq y\} = \mathrm{P}\{X \leq x\}\mathrm{P}\{Y \leq y\}$.

Exercise 5.3 Verify the claim of Remark 5.11 by constructing three random variables X_1, X_2, and X_3, such that $\{X_1, X_2\}$, $\{X_1, X_3\}$, and $\{X_2, X_3\}$ are independent, but $\{X_1, X_2, X_3\}$ are not independent.
(Hint: Consider $X_1 := \pm 2$ with probability $\frac{1}{2}$ each, an independent $X_2 := \pm 1$ with probability $\frac{1}{2}$ each, and $X_3 := X_1 \times X_2$.)

Exercise 5.4 Prove Lemma 5.14.

Exercise 5.5 Independence is a powerful property. Consider a probability space $(\Omega, \mathfrak{F}, \mathrm{P})$ that is rich enough to support countably many independent random variables $X_1, X_2, \ldots \in L^2(\mathrm{P})$.

1. Prove that $X_n' := X_n / \|X_n\|_2$ is a complete orthonormal system on $L^2(\mathrm{P})$; i.e., given any $Y \in L^2(\mathrm{P})$, we have $Y = \sum_{n=1}^{\infty} \mathrm{E}\{X_n' Y\} X_n'$, where the infinite random sum converges in $L^2(\mathrm{P})$.

2. Construct explicitly a probability space on which one cannot construct infinitely many independent random variables.

Exercise 5.6 Prove the *one-series theorem* of Kolmogorov [Kol30]:[5.9] If $X_1, X_2, \ldots$ are independent mean-zero random variables taking values in $\mathbb{R}$, and if $\sum_j \mathrm{E}\{X_j^2\} < +\infty$, then $\sum_j X_j$ converges almost surely. Use this to prove the following beautiful theorem of Steinhaus [Ste30]. The random harmonic series converges with probability one. That is, if $\sigma_1, \sigma_2, \ldots$ are i.i.d. random variables taking the values ± 1 with probability $\frac{1}{2}$ each, then $\sum_j j^{-1}\sigma_j$ converges almost surely.
(Hint: First, show that $S_n := \sum_{j=1}^{n} X_j$ is a Cauchy sequence in $L^2(\mathrm{P})$. Then

[5.9] In fact, this is preceded by the stronger *three-series theorem* of Khintchine and Kolmogorov [KK25] which characterizes precisely when $\sum_j X_j$ converges a.s.

use the Kolmogorov maximal inequality (Theorem 5.24) along a suitable subsequence.)

Exercise 5.7 In $L^4(\mathrm{P})$, the strong law becomes easier to prove. Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with mean zero and variance one, and let $S_n := X_1 + \cdots + X_n$. If in addition $\|X_1\|_4 < +\infty$, then show that there exists a constant A such that for all $n \geq 1$, $\|S_n\|_4 \leq A\sqrt{n}$. Conclude the strong law under the given L^4-assumption from this, the Borel–Cantelli lemma, and Chebyshev's inequality alone. (You may not use the Kolmogorov maximal inequality in this exercise.)

Exercise 5.8 Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables that take values in some topological space $\mathbb{X}$, and that $A \in \mathfrak{B}(\mathbb{X})$ has the property that $\mathrm{P}\{X_1 \in A\} > 0$. Then prove that with probability one, infinitely-many of the X_n's fall in A. Use this to make precise the following amusing claim: If a monkey types ad infinitum and completely "at random," then with probability one, some portion of this infinite monkey-novel contains the entire works of Shakespeare in exactly the way that they were written.

Exercise 5.9 Suppose that X_n is a sequence of i.i.d. exponential random variables with parameter $\lambda > 0$ (Example 1.18). Then prove that with probability one, $\limsup_n X_n = +\infty$ and $\liminf_n X_n = 0$. Improve this to the following: Almost surely,

$$\limsup_{n\to\infty} \frac{X_n}{\ln n} = \frac{1}{\lambda}, \qquad \liminf_{n\to\infty} \frac{\ln X_n}{\ln n} = -1. \tag{5.62}$$

Exercise 5.10 Prove the *Paley–Zygmund inequality* [PZ32]: For any non-negative random variable $Y \in L^2(\mathrm{P})$, and for any $\varepsilon > 0$,

$$\mathrm{P}\left\{Y > \varepsilon \mathrm{E}[Y]\right\} \geq (1-\varepsilon)^2 \frac{\{\mathrm{E}[Y]\}^2}{\mathrm{E}\{Y^2\}}. \tag{5.63}$$

Now suppose that $E_1, E_2, \ldots$ are events such that $\sum_{j=1}^\infty \mathrm{P}(E_j) = +\infty$, and that

$$\gamma := \liminf_{n\to\infty} \frac{\sum_{i=1}^n \sum_{j=1}^n \mathrm{P}\{E_i \cap E_j\}}{\left(\sum_{j=1}^n \mathrm{P}\{E_j\}\right)^2} < +\infty. \tag{5.64}$$

Then, show that $\mathrm{P}\{\sum_j \mathbf{1}_{E_j} = +\infty\} \geq \gamma^{-1} > 0$. Verify that this improves the independence half of the Borel–Cantelli lemma (Theorem 5.23).

(HINT: For the first part, start by writing $\mathrm{E}\{Y\} = \mathrm{E}\{Y; Y \le a\} + \mathrm{E}\{Y; Y > a\}$ for a suitable number a.)

Exercise 5.11 Suppose X is uniformly distributed on $[0,1]$; i.e., its distribution is the Lebesgue measure on $[0,1]$. Write the binary expansion of X; i.e., $X = \sum_{j=1}^{\infty} 2^{-j} X_j$, and then show that $X_1, X_2, \ldots$ are i.i.d. Find their distribution, and show that with probability one, $\lim_n n^{-1} S_n = \frac{1}{2}$, where $S_n := \sum_{j=1}^{n} \mathbf{1}_{\{X_j = 1\}}$ is the total number of 1's in the first n digits of the binary expansion of X. This is the *normal number theorem* of Borel [Bor09].

Exercise 5.12 (Hard) The classical *central limit theorem* (Theorem 6.22 below) states the following: If $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathrm{E}\{X_1\} = 0$ and $\mathrm{Var}(X_1) = 1$, and if $S_n := X_1 + \cdots + X_n$, then for any real $a < b$,

$$\lim_{n\to\infty} \mathrm{P}\left\{a\sqrt{n} \le S_n \le b\sqrt{n}\right\} = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx. \tag{5.65}$$

Use this, without proof, to derive the following:[5.10] As $n \to \infty$,

$$\frac{1}{\ln n} \sum_{j=1}^{n} \frac{\mathbf{1}_{\{a\sqrt{j} \le S_j \le b\sqrt{j}\}}}{j} \xrightarrow{\mathrm{P}} \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx. \tag{5.66}$$

We will prove this in successive steps.

1. If $Z_n := \sum_{j=1}^{n} \mathbf{1}_{\{a\sqrt{j} \le S_j \le b\sqrt{j}\}}$ and $C := (2\pi)^{-1/2} \int_a^b e^{-x^2/2}\, dx$, then show that as $n \to \infty$, $\mathrm{E}\{Z_n\} \sim C \ln n$. Here and throughout, $a_n \sim b_n$ means that $a_n \div b_n \to 1$.

[5.10] It has been shown that if, in addition, there exists a $\varepsilon > 0$ such that $X_1 \in L^{2+\varepsilon}(\mathrm{P})$, then (5.65) holds almost surely. The resulting convergence is called an *almost-sure central limit theorem* (ASCLT); it was discovered independently and at the same time by Brosamler, Fisher, and Schatte [Bro88, Fis87, Sch88].

A prefatory version of the ASCLT was anticipated by Lévy [Lév37, p. 270].

Lacey and Philipp [LP90] devised a proof of the ASCLT that is robust enough that many of the conditions of this exercise—including the independence of the increments—can be weakened. They also prove that, in the ASCLT, the condition $X_1 \in L^2(\mathrm{P})$ suffices.

Berkes et al. [BCH98, BC01] present remarkably general families of ASCLTs. For further references and related results see the survey article of Berkes [Ber98].

2. Prove that when $j > i$,

$$\begin{aligned} &\mathrm{P}\left\{a\sqrt{i} \le S_i \le b\sqrt{i}\ ,\ a\sqrt{j} \le S_j \le b\sqrt{j}\right\} \\ &\quad\le \mathrm{P}\left\{a\sqrt{i} \le S_i \le b\sqrt{i}\right\} \\ &\qquad\times \mathrm{P}\left\{a\sqrt{j} - b\sqrt{i} \le S_{j-i} \le b\sqrt{j} - a\sqrt{i}\right\}. \end{aligned} \tag{5.67}$$

3. Use the previous part to show that we can find a sequence $\delta_n \to 0$, such that

$$\begin{aligned} &\sum\sum_{\substack{1\le i\le n \\ i^2\le j\le n}} \frac{1}{ij}\,\mathrm{P}\left\{a\sqrt{i} \le S_i \le b\sqrt{i}\ ,\ a\sqrt{j} \le S_j \le b\sqrt{j}\right\} \\ &\quad\le \left(\frac{1}{2} + \delta_n\right)\left(\mathrm{E}\{Z_n\}\right)^2, \end{aligned} \tag{5.68}$$

4. Show that there exists a constant A_3 such that for all n large,

$$\begin{aligned} &\sum\sum_{\substack{1\le i\le n \\ i<j\le i^2\le n}} \frac{1}{ij}\,\mathrm{P}\left\{a\sqrt{i} \le S_i \le b\sqrt{i}\ ,\ a\sqrt{j} \le S_j \le b\sqrt{j}\right\} \\ &\le A_3 \ln n. \end{aligned} \tag{5.69}$$

5. (e) Use this to show that the variance of $Z_n \div \ln n$ goes to zero as $n \to \infty$; conclude the proof from this.

(Hint: Part (e) uses (a), (b), and (c). In part (e), use the trivial estimate, $\mathrm{P}(E \cap F) \le \mathrm{P}(E)$ for any two events E and F.)

Chapter 6

Weak Convergence

1 Introduction

Without a doubt, the central limit theorem (CLT) of de Moivre [dM33, dM38, dM56] and Laplace [Lap10] is one of the great discoveries of nineteenth century mathematics.[6.1] To describe it, consider performing a sequence of independent random trials, each of which can lead to either a success or failure. Suppose also that there exists some $p \in (0,1)$ such that with probability p, any given trial succeeds, and with probability $1-p$ it fails. If T_n denotes the total number of successes in n trials, then it is easy to see that

$$\mathrm{P}\{T_n = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \forall k = 0, 1, \ldots, n, \tag{6.1}$$

and 0 if $k \notin \{1, \ldots, n\}$. The de Moivre–Laplace central limit theorem gives an asymptotic evaluation of the distribution of T_n for large n. In precise terms, it states that for any $a < b$,

$$\lim_{n\to\infty} \mathrm{P}\left\{ a < \frac{T_n - np}{\sqrt{np(1-p)}} \leq b \right\} = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx. \tag{6.2}$$

Therefore, in some sense, T_n is approximately normally distributed with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ (why?).

[6.1] For a detailed historical account of the classical central limit theorem, see Adams [Ada74].

The classical proof of the de Moivre–Laplace theorem is combinatorial, and is a tedious application of the de Moivre–Stirling formula,[6.2]

$$\lim_{n\to\infty} \frac{n!}{n^{n+\frac{1}{2}}e^{-n}} = \sqrt{2\pi}. \tag{6.3}$$

This is in itself a simple but tedious exercise.

In this chapter we discuss the modern approach to this and a powerful generalization that involves the notion of weak convergence.[6.3]

2 Weak Convergence

Definition 6.1 Let $\mathbb{X}$ denote a topological space, and suppose μ, μ_1, $\mu_2, \ldots$ are probability (or more generally finite) measures on $(\mathbb{X}, \mathfrak{B}(\mathbb{X}))$. We say that μ_n *converges weakly* to μ—and write $\mu_n \Longrightarrow \mu$—if for all bounded continuous functions $f : \mathbb{X} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu. \tag{6.4}$$

If X_n is an $\mathbb{X}$-valued random variable with distribution μ_n, and if X is an $\mathbb{X}$-valued random variable with distribution μ, then we also say that X_n *converges weakly* to X and write this as $X_n \Longrightarrow X$. This is equivalent to saying that for all bounded continuous functions $f : \mathbb{X} \to \mathbb{R}$, $\lim_n \mathrm{E}\{f(X_n)\} = \mathrm{E}\{f(X)\}$.

The following important characterization of weak convergence on $\mathbb{R}$ is due to P. Lévy.

Theorem 6.2 (Lévy [Lév37]) *Let F be the distribution function of μ—a probability measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$—and F_n that of μ_n—also probability measures on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Then $\mu_n \Longrightarrow \mu$ if and only if $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ at which F is continuous. Equivalently, in terms of random variables, $X_n \Longrightarrow X$ if and only if $\mathrm{P}\{X_n \le x\} \to \mathrm{P}\{X \le x\}$ for all x such that $\mathrm{P}\{X = x\} = 0$.*

[6.2] This is also known as *Stirling's formula*. The original formulation of Stirling's formula is due to A. de Moivre who showed that there exists a constant β such that $\ln n! = \ln\beta + \left(n+\frac{1}{2}\right)\ln n - n + \frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1260n^4} + \cdots$; see de Moivre [dM38]. The displayed Stirling formula (with $\sqrt{2\pi}$ replaced by β) follows readily from this. The contribution of Stirling [Sti30] was in proving the nontrivial fact that in addition $\beta = \sqrt{2\pi}$.

[6.3] Unfortunately, most nonprobabilists call this weak-* convergence; weak convergence typically means something else.

Remark 6.3 To see how the above may be useful to us, consider the setting of the de Moivre–Laplace central limit theorem of (6.2), let $X_n := (T_n - np) \div \sqrt{np(1-p)}$, $X :=$ a normal random variable with $\mu = 0$ and $\sigma = 1$, $F_n :=$ the distribution function of X_n, and $F :=$ the distribution function of X. Observe that: (i) F is continuous everywhere; and (ii) (6.2) asserts that $\lim_n(F_n(b) - F_n(a)) = F(b) - F(a)$. Thanks to the preceding theorem then, (6.2) is saying that $X_n \Longrightarrow X$.

Remark 6.4 At first glance, it may seem strange that if $\mu_n \Longrightarrow \mu$, then $F_n \to F$ only where F is continuous. If so, then perhaps the following simple example will convince you that continuity is unavoidable: Let $X := \pm 1$ with probability $\frac{1}{2}$ each. Next let $X_n(\omega) = -1$ if $X(\omega) = 1$, and $X_n(\omega) := 1 + \frac{1}{n}$ if $X(\omega) = 1$. Then for all bounded continuous functions f, we have $f(X_n) \to f(X)$ a.s. (in fact surely), and hence $\mathrm{E}\{f(X_n)\} \to \mathrm{E}\{f(X)\}$. However, $F(1) = \mathrm{P}\{X \le 1\} = 1$, whereas $F_n(1) = \mathrm{P}\{X_n \le 1\} = \frac{1}{2}$.

In order to prove the preceding theorem, we first need a simple lemma that is really a result about nondecreasing right-continuous functions that have left-limits everywhere. Note that any distribution function is of this type (why?).

Lemma 6.5 *The set* $\{x \in \mathbb{R} : \mathrm{P}\{X = x\} > 0\}$ *is denumerable.*

Thus, in this sense, X_n converges weakly to X if and only if $F_n(x) \to F(x)$ for most values of $x \in \mathbb{R}$.

Proof The set of the x's in question is the same as the set of all x at which F jumps, where F is the distribution function of X; i.e., $F(x) = \mathrm{P}\{X \le x\}$. So we will show that F can only have a denumerable number of jumps. Let J_n denote the collection of all x such that $F(x) - F(x-) \ge n^{-1}$. Clearly, $1 = F(\infty) - F(-\infty) \ge \sum_{x \in J_n}[F(x) - F(x-)] \ge n^{-1}\#(J_n)$, where $\#$ denotes cardinality. Therefore, J_n is finite, and hence $\cup_n J_n$ is denumerable. □

Proof of Theorem 6.2 It should be clear that the statement about $X_n \Longrightarrow X$ is equivalent to the statement about $\mu_n \Longrightarrow \mu$, so we need only to prove the statement about the random variables.

Suppose first that $X_n \Longrightarrow X$. For any fixed $x \in \mathbb{R}$ and $\varepsilon > 0$, it is not hard to find a bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ such that for all $y \in \mathbb{R}$, $f(y) \le \mathbf{1}_{(-\infty,x]}(y) \le f(y + \varepsilon)$. For instance, f could be the piecewise-linear continuous function such that (i) for all $z \le x$, $f(z) = 0$; (ii) for all $z \ge x + \varepsilon$, $f(z) = 1$; and (iii) for all $z \in [x, x+\varepsilon]$, $f(z) = \varepsilon^{-1}(z - x)$. Then, $\mathrm{E}\{f(X_n)\} \le F_n(x) \le \mathrm{E}\{f(X_n + \varepsilon)\}$. Let $n \to \infty$ to deduce from this that

$$\mathrm{E}\{f(X)\} \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le \mathrm{E}\{f(X+\varepsilon)\}. \tag{6.5}$$

Since $f(y) \ge \mathbf{1}_{(-\infty, x-\varepsilon]}(y)$, the left-most term is greater than or equal to $F(x-\varepsilon)$, and a similar reasoning shows that the right-most term is greater than or equal to $F(x+\varepsilon)$. This tells us that if F were continuous at x, then $F_n(x) \to F(x)$, as asserted.

For the converse, suppose that for all continuity-points x of F, $F_n(x) \to F(x)$; we wish to prove that $X_n \Longrightarrow X$. Equivalently, that for all bounded continuous function $f : \mathbb{R} \to \mathbb{R}_+$, $\mathrm{E}\{f(X_n)\} \to \mathrm{E}\{f(X)\}$. (Note that $f \ge 0$ here, but this is not a restriction since otherwise we would consider f^+ and f^- separately.)

For any $\varepsilon, N > 0$, we can find an increasing collection of points $0 := x_0 < x_1 < x_2 < \ldots \in \mathbb{R}$, and write $x_{-i} := -x_i$, such that

(i) $\max_{|i| \le N} \sup_{y \in (x_i, x_{i+1}]} |f(y) - f(x_i)| \le \varepsilon$;

(ii) F is continuous at x_i for all $i \in \mathbb{Z}$;

(iii) $\lim_n x_n = +\infty$ and $\lim_n x_{-n} = -\infty$.

(We need Lemma 6.5 part (ii).) By (i),

$$\begin{aligned}
&\left|\mathrm{E}\left\{f(X_n); |X_n| \le x_N\right\} - \sum_{j=-N}^{N} f(x_j)\left[F_n(x_{j+1}) - F_n(x_j)\right]\right| \\
&\quad = \left|\mathrm{E}\{f(X_n); |X_n| \le x_N\} - \sum_{j=-N}^{N} f(x_j)\mathrm{P}\left\{X_n \in (x_j, x_{j+1}]\right\}\right| \le \varepsilon.
\end{aligned} \tag{6.6}$$

This remains valid if we replace X_n and F_n by X and F, respectively. Since N is a large but fixed number, $\lim_n \sum_{|j| \le N} f(x_j)\left[F_n(x_{j+1}) - F_n(x_j)\right] = \sum_{|j|\le N} f(x_j)\left[F(x_{j+1}) - F(x_j)\right]$. Therefore, the fact that $\varepsilon > 0$ is arbitrary shows us that for all $N > 0$,

$$\lim_{n\to\infty} \mathrm{E}\{f(X_n); |X_n| \le x_N\} = \mathrm{E}\{f(X); |X| \le x_N\}. \tag{6.7}$$

For the remainder terms, let $K := \sup_{y\in\mathbb{R}} f(y)$ to see that

$$\begin{aligned}\limsup_{n\to\infty} \mathrm{E}\left\{f(X_n); |X_n| > x_N\right\} &\le K \limsup_{n\to\infty} \mathrm{P}\{|X_n| > x_N\} \\ &\le K \lim_{n\to\infty} \left[1 - F_n(x_N) + F_n(-x_N)\right] \\ &= K\left[1 - F(x_N) + F(-x_N)\right].\end{aligned} \tag{6.8}$$

But for all $\varepsilon > 0$, there exists $N_0 > 0$ so large that the last term above is $\le \frac{\varepsilon}{2}$ for $N := N_0$. Thus, there exists n_0 so that for all $n \ge n_0$, $\mathrm{E}\{f(X_n); |X_n| > x_{N_0}\} \le K\varepsilon$. On the other hand, $\lim_{N\to\infty} \max_{n\le n_0} \mathrm{E}\{f(X_n); |X_n| > x_N\} \le K \lim_{N\to\infty} \max_{n\le n_0} \mathrm{P}\{|X_n| > x_N\} = 0$, since the maximum is over a finite set. Thus, we can find N_1 such that $\max_{n\le n_0} \mathrm{E}\{f(X_n); |X_n| > x_{N_1}\} \le \varepsilon$. Let $N_2 := N_0 \vee N_1$ to see that for all $n \ge 1$, $\mathrm{E}\{f(X_n); |X_n| > x_{N_2}\} \le K\varepsilon$. Finally, there exists $N_3 > 0$ such that $\mathrm{E}\{f(X); |X| > x_{N_3}\} \le K\varepsilon$. Let $N_4 := N_3 \vee N_2$, and apply (6.7) with $N := N_4$ to deduce that

$$\limsup_{n\to\infty} \left|\mathrm{E}\{f(X_n)\} - \mathrm{E}\{f(X)\}\right| \le K\varepsilon. \tag{6.9}$$

Since $\varepsilon > 0$ is arbitrary, this concludes our proof. □

3 Weak Convergence and Compact-Support Functions

Definition 6.6 If $\mathbb{X}$ is a metric space, then $C_c(\mathbb{X})$ denotes the collection of all continuous functions $f : \mathbb{X} \to \mathbb{R}$ such that f has *compact support*; i.e., there exists a compact set K such that for all $x \notin K$, $f(x) = 0$. In addition, $C_b(\mathbb{X})$ denotes the collection of all bounded continuous functions $f : \mathbb{X} \to \mathbb{R}$.

Recall that in order to prove that μ_n converges weakly to μ we need to verify that for all $f \in C_b(\mathbb{X})$, $\int f\,d\mu_n \to \int f\,d\mu$. Since $C_c(\mathbb{R}^k) \subseteq C_b(\mathbb{R}^k)$, the next result can be viewed as a simplification of this task in the case that $\mathbb{X} = \mathbb{R}^k$ is a Euclidean space.[6.4]

Theorem 6.7 *If $\mu, \mu_1, \mu_2, \ldots$ are probability measures on $(\mathbb{R}^k, \mathfrak{B}(\mathbb{R}^k))$, then $\mu_n \Longrightarrow \mu$ if and only if for all $f \in C_c(\mathbb{R}^k)$, $\lim_n \int f\,d\mu_n = \int f\,d\mu$.*

[6.4] In fact, with a little general topology, one can go far beyond Euclidean spaces.

Proof As I mentioned a little earlier, it is easy to see that every continuous function of compact support is bounded and continuous; i.e., $C_c(\mathbb{R}^k) \subseteq C_b(\mathbb{R}^k)$. Therefore, we need only suppose that $\int g\, d\mu_n \to \int g\, d\mu$ for all $g \in C_c(\mathbb{R}^k)$, and then prove that for all $f \in C_b(\mathbb{R}^k)$, $\int f\, d\mu_n \to \int f\, d\mu$. With this goal in mind, let us choose and fix such an $f \in C_b(\mathbb{R}^k)$; I will prove the theorem in three successive steps. By considering f^+ and f^- separately, I can assume without loss of generality that $f(x) \geq 0$ for all x. This will be done without further mention.

Step 1. The Lower Bound.
For any $p > 0$ choose and fix a function $f_p \in C_c(\mathbb{R})$ such that:

- For all $x \in [-p, p]^k$, $f_p(x) = f(x)$.
- For all $x \notin [-p-1, p+1]^k$, $f_p(x) = 0$.
- For all $x \in \mathbb{R}^k$, $f_p(x) \leq f(x)$.

You should check that f_p exists, and for each $x \in \mathbb{R}^k$, $f_p(x) \geq 0$, while $f_p(x) \uparrow f(x)$ as $p \uparrow \infty$. It follows that

$$\liminf_{n\to\infty} \int f\, d\mu_n \geq \lim_{n\to\infty} \int f_p\, d\mu_n = \int f_p\, d\mu. \tag{6.10}$$

Let $p \uparrow \infty$ and appeal to the dominated convergence theorem (Theorem 2.22) to deduce half of the theorem. Namely,

$$\liminf_{n\to\infty} \int f\, d\mu_n \geq \int f\, d\mu. \tag{6.11}$$

Step 2. A Variant.
In this step I will prove that in (6.11) we could formally replace f by the indicator function of an open k-dimensional hypercube. More precisely, that given any real numbers $a_1 < b_1, \ldots, a_k < b_k$,

$$\liminf_{n\to\infty} \mu_n\left((a_1, b_1) \times \cdots \times (a_k, b_k)\right) \geq \mu\left((a_1, b_1) \times \cdots \times (a_k, b_k)\right). \tag{6.12}$$

To prove this we first find continuous functions $\psi_m \uparrow \mathbf{1}_{(a_1,b_1)\times\cdots\times(a_k,b_k)}$, pointwise (do it!). By definition, given any $m \geq 1$, $\psi_m \in C_c(\mathbb{R}^k)$, and

$$\liminf_{n\to\infty} \mu_n\left((a_1, b_1) \times \cdots \times (a_k, b_k)\right) \geq \lim_{n\to\infty} \int \psi_m\, d\mu_n = \int \psi_m\, d\mu. \tag{6.13}$$

Now let $m \uparrow \infty$ and appeal to the dominated convergence theorem to deduce (6.12).

Step 3. The Upper Bound.
For the upper bound, we use f_p from Step 1 and write

$$\begin{aligned}\int f\,d\mu_n &= \int_{[-p,p]^k} f\,d\mu_n + \int_{\mathbb{R}^k\setminus[-p,p]^k} f\,d\mu_n \\ &\le \int f_p\,d\mu_n + \sup_{z\in\mathbb{R}} |f(z)|\cdot\left[1-\mu_n\left([-p,p]^k\right)\right].\end{aligned} \tag{6.14}$$

Now let $n \to \infty$ and appeal to (6.12) to find that

$$\limsup_{n\to\infty} \int f\,d\mu_n \le \int f_p\,d\mu + \sup_{z\in\mathbb{R}} |f(z)|\cdot\left[1-\mu\left((-p,p)^k\right)\right]. \tag{6.15}$$

Let $p \uparrow \infty$ and appeal to the monotone convergence theorem (2.21) to deduce that $\limsup_n \int f\,d\mu_n \le \int f\,d\mu$ and finish the proof. □

4 Harmonic Analysis in One Dimension

Definition 6.8 The *Fourier transform* of a probability measure μ on $\mathbb{R}$ is

$$\widehat{\mu}(t) := \int_{-\infty}^{\infty} e^{itx}\,\mu(dx), \qquad \forall t \in \mathbb{R}, \tag{6.16}$$

where $i := \sqrt{-1}$. This still makes sense if μ is a finite measure, and even if μ is replaced by a Lebesgue-integrable function f as follows: $\widehat{f}(t) := \int_{-\infty}^{\infty} e^{ixt} f(x)\,dx$. In this case, we are identifying the Fourier transform of the function f with that of the measure μ, where $f(x) := \frac{d\mu}{dx}$. If X is a real-valued random variable whose distribution is some probability measure μ, then $\widehat{\mu}$ is also called the *characteristic function* of X and/or μ, and $\widehat{\mu}(t)$ is equal to $\mathrm{E}\{e^{itX}\}$. This is equal to $\mathrm{E}\{\cos(tX)\} + i\mathrm{E}\{\sin(tX)\}$ if you wish to deal only with real integrands.[6.5]

Here are some of the elementary properties of characteristic functions. You should be sure to understand the extent to which the following properties depend on the measure μ's being a probability measure.

[6.5] Unfortunately, in most other areas of mathematics, a characteristic function is our indicator function, and our characteristic function is the Fourier transform!

Lemma 6.9 *If μ is a finite measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, then $\widehat{\mu}$ exists, is uniformly continuou on $\mathbb{R}$, and satisfies the following:*

1. $\sup_{t\in\mathbb{R}} |\widehat{\mu}(t)| = \widehat{\mu}(0) = \mu(\mathbb{R})$*, and* $\widehat{\mu}(-t) = \overline{\widehat{\mu}(t)}$.
2. $\widehat{\mu}$ *is a nonnegative-definite function; i.e., for any* $z_1, \ldots, z_n \in \mathbb{C}$*, and for all* $t_1, \ldots, t_n \in \mathbb{R}$, $\sum_{j=1}^n \sum_{k=1}^n \widehat{\mu}(t_j - t_k) z_j \overline{z_k} \geq 0$.

Proof Without loss of generality, we may assume that μ is a probability measure, for otherwise we can prove the theorem for the probability measure $\nu := \mu \div \mu(\mathbb{R})$, and then multiply through by $\mu(\mathbb{R})$.

Let X be a random variable whose distribution is μ, so that $\widehat{\mu}(t) = \mathrm{E}\{\exp(itX)\}$. This is always defined and bounded, since $|e^{itX}| \leq 1$. To prove uniform continuity, we note that for any $a, b \in \mathbb{R}$, $|e^{ia} - e^{ib}| = |1 - e^{i(a-b)}| \leq |a - b| \wedge 2$ (why?), so that

$$\sup_{|s-t|\leq\delta} |\widehat{\mu}(t) - \widehat{\mu}(s)| \leq \sup_{|s-t|\leq\delta} \mathrm{E}\left\{\left|1 - e^{i(t-s)X}\right|\right\} \leq \mathrm{E}\left\{\delta|X| \wedge 2\right\}. \tag{6.17}$$

By the dominated convergence theorem (Theorem 2.22), as $\delta \downarrow 0$, this goes to 0 and this yields uniform continuity. Part 1 is elementary, and we turn to proving 2:

$$\begin{aligned}\sum_{j=1}^n \sum_{k=1}^n \widehat{\mu}(t_j - t_k) z_j \overline{z_k} &= \sum_{j=1}^n \sum_{k=1}^n \mathrm{E}\left\{e^{i(t_j-t_k)X}\right\} z_j \overline{z_k} \\ &= \mathrm{E}\left\{\left|\sum_{j=1}^n e^{it_jX} z_j\right|^2\right\},\end{aligned} \tag{6.18}$$

which is nonnegative. □

Example 6.10[The Uniform Distribution; Example 1.17] Given two numbers $a < b$, a random variable X is uniformly distributed on (a, b) if its density function is $f(x) = (b{-}a)^{-1}\mathbf{1}_{(a,b)}(x)$. Equivalently, the uniform distribution on (a, b) is the same as the Lebesgue measure on (a, b) normalized to have total mass one. (Check this for intervals first and then proceed.) Its characteristic function is then given by

$$\mathrm{E}\{e^{itX}\} = \frac{e^{itb} - e^{ita}}{it(b-a)}, \qquad \forall t \in \mathbb{R}. \tag{6.19}$$

Example 6.11[The Exponential Distribution; Example 1.18] Given some number $\lambda > 0$, a random variable X is said to have an exponential distribution with parameter λ if its density function is $f(x) = \lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}(x)$. Its characteristic function is

$$\mathrm{E}\{e^{itX}\} = \frac{\lambda}{it - \lambda}, \qquad \forall t \in \mathbb{R}. \tag{6.20}$$

Example 6.12[The Normal Distribution; Example 1.19] Given two numbers $\mu \in \mathbb{R}$ and $\sigma > 0$, a random variable X is said to have an normal distribution with parameters μ and σ if its density function is $f(x) = (2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/(2\sigma^2)\}$. Its characteristic function is

$$\mathrm{E}\{e^{itX}\} = \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right), \qquad \forall t \in \mathbb{R}. \tag{6.21}$$

This remains valid in the degenerate case where $\sigma = 0$ (why?). You should check that $\mathrm{E}\{X\} = \mu$ and $\mathrm{Var}(X) = \sigma^2$, so that we can refer to the distribution of X as normal with mean μ and variance σ^2.

I mention a few "discrete" distributions as well.

Example 6.13[Discrete Distributions] Suppose that $X = x_j$ with probability p_j ($j = 1, 2, \ldots$), where the x_j's are real, and $p_j > 0$ and $\sum_{j=1}^\infty p_j = 1$. Then,

$$\mathrm{E}\{e^{itX}\} = \sum_{j=1}^\infty e^{itx_j} p_j, \qquad \forall t \in \mathbb{R}. \tag{6.22}$$

Here are two noteworthy consequences of this:

Example 6.14[Binomial Distributions] Given a number $p \in (0,1)$ and an integer $n \geq 1$, a random variable X is said to have the binomial distribution with parameters n and p if $\mathrm{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$ and $\mathrm{P}\{X = k\} = 0$ otherwise. According to Example 6.13, for all $t \in \mathbb{R}$,

$$\mathrm{E}\{e^{itX}\} = \left(pe^{it} + 1 - p\right)^n, \tag{6.23}$$

thanks to the binomial formula.

Example 6.15[Poisson Distributions] Given a number $\lambda > 0$, X is said to have the Poisson distribution with parameter λ if $\mathrm{P}\{X = k\} = e^{-\lambda}\lambda^k/k!$ if $k = 0, 1, \ldots$ and otherwise $\mathrm{P}\{X = k\} = 0$. Its characteristic function is given by Example 6.13, and is equal to

$$\mathrm{E}\{e^{itX}\} = \exp\left(-\lambda + \lambda e^{it}\right), \qquad \forall t \in \mathbb{R}. \tag{6.24}$$

5 The Plancherel Theorem

In this section I state and prove a modern variant of an important result of Plancherel ([Pla10, Pla33]). Roughly speaking, it shows us how to reconstruct a distribution from its characteristic function.

In order to state it in a convenient form, I will need to introduce a definition, as well as some notation.

Definition 6.16 Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are measurable. Then, when defined, the *convolution* $f * g$ is the function,

$$f * g(x) := \int_{-\infty}^{\infty} f(x-y)g(y)\,dy. \tag{6.25}$$

Remark 6.17 You should check that convolution is a symmetric operation; i.e., $f * g = g * f$ in the sense that one is defined if and only if the other is, and the identity holds in such a case.

Throughout this section, we define φ_ε denote the density of a mean-zero normal random variable in $\mathbb{R}$ whose variance is ε^2. That is,

$$\varphi_\varepsilon(x) := \frac{e^{-x^2/(2\varepsilon^2)}}{\varepsilon\sqrt{2\pi}}, \qquad \forall x \in \mathbb{R}. \tag{6.26}$$

According to Example 6.12, the Fourier transform of φ_ε has the following neat form:

$$\widehat{\varphi_\varepsilon}(t) = e^{-\frac{1}{2}\varepsilon^2 t^2}, \qquad \forall t \in \mathbb{R}. \tag{6.27}$$

Theorem 6.18 (The Plancherel Theorem) *If μ is a finite measure on $\mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ is Lebesgue-integrable, then for any $\varepsilon > 0$,*

$$\int_{-\infty}^{\infty} f * \varphi_\varepsilon(x)\, \mu(dx) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2} \widehat{f}(t) \overline{\widehat{\mu}(t)}\, dt. \tag{6.28}$$

Consequently, for all $f \in C_c(\mathbb{R})$ whose Fourier transform $\widehat{f}$ is integrable,

$$\int_{-\infty}^{\infty} f\, d\mu = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(t) \overline{\widehat{\mu}(t)}\, dt. \tag{6.29}$$

Remark 6.19

(i) You should check that if f and g are sufficiently well-behaved, then $\widehat{f * g}(t) = \widehat{f}(t)\widehat{g}(t)$. That is, the Fourier transform maps convolutions into products. In particular, the Fourier transform of $\psi := f * \varphi_\varepsilon$ is $e^{-\frac{1}{2}\varepsilon^2 t^2} \widehat{f}(t)$. Therefore, (6.28) states that $\int \psi\, d\mu = (2\pi)^{-1} \int \widehat{\psi}(t)\overline{\widehat{\mu}}(t)\, dt$, which is a little like (6.29). This is far from being an accident, but you will have to learn about this in a text on distribution theory and/or Fourier analysis.

(ii) Equation (6.29) is sometimes referred to as the *Parseval identity*, named after Marc-Antoine Parseval des Chénes for his 1801 discovery of a discrete version of (6.29) in the context of Fourier series.

Proof of Theorem 6.18 In order to prove (6.28), we simply integrate the right-hand side and appeal to the Fubini–Tonelli theorem (Theorem 3.6). Here is how:[6.6]

$$\begin{aligned}
&\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2} \widehat{f}(t) \overline{\widehat{\mu}(t)}\, dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2} \left(\int_{-\infty}^{\infty} f(x) e^{itx}\, dx \right) \left(\int_{-\infty}^{\infty} e^{-ity}\, \mu(dy) \right) dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2} e^{it(x-y)}\, dt \right) \mu(dy)\, f(x)\, dx.
\end{aligned} \tag{6.30}$$

[6.6] Since f is integrable, all of the integrals converge absolutely.

A direct calculation reveals that

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2} e^{it(x-y)}\,dt = \frac{\sqrt{2\pi}}{\varepsilon} e^{-(x-y)^2/(2\varepsilon^2)} = 2\pi\varphi_\varepsilon(x-y). \tag{6.31}$$

(Check by completing the square!) This and another doze of the Fubini–Tonelli theorem together yield (6.28). Equation (6.29) follows from (6.28) and the dominated convergence theorem (Theorem 2.22) once we show the following:[6.7]

$$\lim_{\varepsilon\to 0} \sup_{x\in\mathbb{R}} |f*\varphi_\varepsilon(x) - f(x)| = 0, \qquad \forall f\in C_c(\mathbb{R}). \tag{6.32}$$

To show this, I first note the obvious identity, $f(x) = \int_{-\infty}^{\infty} f(x)\varphi_\varepsilon(y)\,dy$, valid for all $x\in\mathbb{R}$. Then, for any $\varepsilon, \delta > 0$,

$$\begin{aligned} |f*\varphi_\varepsilon(x) - f(x)| &\le \sup_{x\in\mathbb{R}} \int_{-\infty}^{\infty} \varphi_\varepsilon(x-y)\,|f(y)-f(x)|\;dy \\ &\le \Omega f(\delta) \times \sup_{x\in\mathbb{R}} \int_{y:\ |y-x|\le\delta} \varphi_\varepsilon(x-y)\,dy \\ &\quad + 2\sup_{z\in\mathbb{R}} |f(z)| \cdot \sup_{x\in\mathbb{R}} \int_{y:\ |y-x|\ge\delta} \varphi_\varepsilon(x-y)\,dy, \end{aligned} \tag{6.33}$$

where Ωf is the so-called *modulus of continuity* of f; i.e., $\Omega f(\delta) := \sup|f(u) - f(v)|$, and the supremum is taken over all $u, v\in\mathbb{R}$ such that $|u-v|\le\delta$. What we have done so far, and a bit of algebra, together show that

$$|f*\varphi_\varepsilon(x) - f(x)| \le \Omega f(\delta) + 2\sup_{z\in\mathbb{R}}|f(z)| \times \int_{|y|\ge\delta/\varepsilon} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,dy. \tag{6.34}$$

Let $\varepsilon\to 0$, and then $\delta\to 0$, to deduce that

$$\limsup_{\varepsilon\to 0} |f*\varphi_\varepsilon(x) - f(x)| \le \lim_{\delta\to 0} \Omega f(\delta) = 0, \tag{6.35}$$

since any $f\in C_c(\mathbb{R})$ is uniformly continuous. This proves (6.32) and hence the result. □

[6.7] This is a 1900 theorem of the nineteen-year-old L. Fejér; cf. [Tan83].

The Plancherel theorem is one of the deep results of classical analysis, and has a number of profound consequences. I will state a few that we will need. The first states that the characteristic function of a finite measure determines the measure.

Theorem 6.20 (The Uniqueness Theorem) *If μ and ν are two finite measures such that for Lebesgue-almost every $t \in \mathbb{R}$, $\widehat{\mu}(t) = \widehat{\nu}(t)$, then $\mu = \nu$.*

Proof First of all, we apply Plancherel's theorem (Theorem 6.18) and (6.35), together, to deduce that for all $f \in C_c(\mathbb{R})$, $\int f\,d\mu = \int f\,d\nu$. We can choose $f_k \in C_c(\mathbb{R})$ such that $f_k \downarrow \mathbf{1}_{[a,b]}$ (do it!). Therefore, by the monotone convergence theorem (Theorem 2.21), $\mu([a,b]) = \nu([a,b])$.

This implies that μ and ν agree on all finite unions of disjoint closed intervals of the form $[a, b]$. Because the collection of all such intervals generates $\mathfrak{B}(\mathbb{R})$, $\mu = \nu$ on $\mathfrak{B}(\mathbb{R})$. □

Another significant consequence of the Plancherel theorem is the following convergence theorem of Glivenko and Lévy.[6.8]

Theorem 6.21 (Glivenko–Lévy; [Gli36, Lév25]) *If μ, μ_1, $\mu_2, \cdots$ are all probability measures on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that for each $t \in \mathbb{R}$, $\widehat{\mu_n}(t) \to \widehat{\mu}(t)$, then $\mu_n \Longrightarrow \mu$.*

Proof A ccording to Theorem 6.7, it suffices to show that for all $f \in C_c(\mathbb{R})$, $\lim_n \int f\,d\mu_n = \int f\,d\mu$. But thanks to (6.34), given any $\delta > 0$ we can choose $\varepsilon > 0$ such that uniformly for all $x \in \mathbb{R}$, $|f * \varphi_\varepsilon(x) - f(x)| \leq \delta$. So we can apply the triangle inequality twice to see that for any $\varepsilon > 0$,

$$\begin{aligned}
&\left|\int f\,d\mu_n - \int f\,d\mu\right| \\
&\quad \leq 2\delta + \left|\int f * \varphi_\varepsilon\,d\mu_n - \int f * \varphi_\varepsilon\,d\mu\right| \qquad (6.36)\\
&\quad = 2\delta + \frac{1}{2\pi}\left|\int_{-\infty}^{\infty} \widehat{f}(t) e^{-\frac{1}{2}\varepsilon^2 t^2}\left(\widehat{\mu_n}(t) - \widehat{\mu}(t)\right)dt\right|.
\end{aligned}$$

[6.8] As it is stated, this theorem is due to Glivenko [Gli36]. Earlier, Lévy [Lév25] had published the following refinement that I will not prove since we will not need it: *If $L(t) := \lim_n \widehat{\mu_n}(t)$ exists and is continuous in a neighborhood of $t = 0$, then there exists a probability measure μ such that $L = \widehat{\mu}$, and $\mu_n \Longrightarrow \mu$.*

The last line holds by the Plancherel theorem (Theorem 6.18). Since $f \in C_c(\mathbb{R})$, $\widehat{f}$ is uniformly bounded by $\int |f(x)|\,dx < \infty$ (Lemma 6.9). Therefore, thanks to the dominated convergence theorem, $\limsup_{n\to\infty} |\int f\,d\mu_n - \int f\,d\mu| \le 2\delta$. Since $\delta > 0$ is arbitrary, the theorem follows. □

6 The One-Dimensional Central Limit Theorem

We are ready to state and prove the main result of this chapter, that is a cornerstone of classical probability theory:[6.9]

Theorem 6.22 (The CLT) *Suppose $X_1, X_2, \ldots$ are i.i.d. real-valued random variables in $L^2(\mathrm{P})$, and assume that $\mathrm{Var}(X_1) > 0$. Then writing $S_n := X_1 + \cdots + X_n$ as before, it follows that for all real $a < b$,*

$$\begin{aligned}\lim_{n\to\infty} \mathrm{P}\left\{ a < \frac{S_n - \mathrm{E}\{S_n\}}{\mathrm{SD}(S_n)} \le b \right\} \\ &= \lim_{n\to\infty} \mathrm{P}\left\{ a < \frac{S_n - n\mathrm{E}\{X_1\}}{\sqrt{n\mathrm{Var}(X_1)}} \le b \right\} \\ &= \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx. \end{aligned} \tag{6.37}$$

Proof Let $X_j^* := (X_j - \mathrm{E}\{X_j\}) \div \sqrt{\mathrm{Var}(X_j)}$ and $S_n^* := \sum_{j=1}^n X_j^*$. Then, the X_i^*'s have mean zero and variance one. Moreover, $S_n^* = (S_n - n\mathrm{E}\{X_1\}) \div \sqrt{\mathrm{Var}(X_1)}$. In other words, we can assume without loss of generality that the X_j's have mean zero and variance one. This simplifies the assertion to the following: $n^{-1/2}S_n \Longrightarrow Z$, where Z is a normal random variable with mean zero and variance one.

Thanks to the Glivenko–Lévy convergence theorem (Theorem 6.21) and Example 6.12, we need to prove that for all $t \in \mathbb{R}$, $\lim_{n\to\infty} \mathrm{E}\{e^{itS_n/\sqrt{n}}\} = e^{-t^2/2}$. By Taylor's expansion, for all $x \in \mathbb{R}$, $e^{ix} = 1 + ix - \frac{1}{2}x^2 + \mathcal{R}(x)$, where $\mathcal{R}$ is a complex-valued function, and $|\mathcal{R}(x)| \le \frac{1}{6}|x|^3$. If $|x| \le 1$, this is a good

[6.9] This is, in fact, the beginning of a rich and complete theory that you can learn from reading the works of Gnedenko and Kolmogorov [GK68], Lévy [Lév37], and Feller [Fel66].

estimate. On the other hand, the worst estimate of all works well enough when $|x| > 1$, viz., $|e^{ix}-1-ix+\frac{1}{2}x^2| \leq |e^{ix}|+1+|x|+\frac{1}{2}x^2 \leq 1+|x|+\frac{1}{2}x^2 \leq \frac{7}{2}x^2$. Combine terms to obtain the bound:

$$|\mathcal{R}(x)| \leq \frac{7}{2}\left(|x|^3 \wedge x^2\right). \tag{6.38}$$

But by independence (Lemma 5.12) and the identical distribution of the X_j's,

$$\mathrm{E}\left\{e^{itS_n/\sqrt{n}}\right\} = \prod_{j=1}^{n} \mathrm{E}\left\{e^{itX_j/\sqrt{n}}\right\}. \tag{6.39}$$

Therefore, we apply Taylor's expansion once more, and deduce that

$$\begin{aligned}
&\mathrm{E}\left\{e^{itS_n/\sqrt{n}}\right\} \\
&\quad= \left[1 + it\mathrm{E}\left\{\frac{X_1}{\sqrt{n}}\right\} - \frac{t^2}{2}\mathrm{E}\left\{\frac{X_1^2}{n}\right\} + \mathrm{E}\left\{\mathcal{R}\left(it\frac{X_1}{\sqrt{n}}\right)\right\}\right]^n \\
&\quad= \left[1 - \frac{t^2}{2n} + \mathrm{E}\left\{\mathcal{R}\left(it\frac{X_1}{\sqrt{n}}\right)\right\}\right]^n.
\end{aligned} \tag{6.40}$$

The last expectation is bounded as follows:

$$n\left|\mathrm{E}\left\{\mathcal{R}\left(it\frac{X_1}{\sqrt{n}}\right)\right\}\right| \leq \frac{7}{2}\mathrm{E}\left\{\frac{|X_1|^3}{n^{1/2}} \wedge X_1^2\right\} := \delta_n, \tag{6.41}$$

which goes to zero by the dominated convergence theorem (Theorem 2.22). From this we obtain the following for some sequence $\varepsilon_n \to 0$:

$$\lim_{n\to\infty} \mathrm{E}\left\{e^{itS_n/\sqrt{n}}\right\} = \lim_{n\to\infty}\left[1 - \frac{t^2}{2n}(1+\varepsilon_n)\right]^n = e^{-\frac{1}{2}t^2}, \tag{6.42}$$

since by Taylor expansion, for $x \simeq 0$, we have $\ln(1-x) \simeq -x$ (why?). This proves the central limit theorem. □

7 The Multidimensional CLT (Optional)

Now we turn to the case of random variables in $\mathbb{R}^d$. Throughout, $X, X_1, X_2, \ldots$ are i.i.d. random variables that takes values in $\mathbb{R}^d$, and

$S_n := X_1 + \cdots + X_n$. Our discussion is somewhat sketchy since we have already encountered most of the key ideas earlier on in this chapter. Throughout this section, $\|x\|$ denotes the usual Euclidean norm of a variable $x \in \mathbb{R}^d$; i.e., $\|x\|^2 := \sqrt{x_1^2 + \cdots + x_d^2}$, for all $x \in \mathbb{R}^d$.

Definition 6.23 The *characteristic function* of X is the function $f(t) := \mathrm{E}\{\exp(it \cdot X)\}$ $(t \in \mathbb{R}^d)$, where "$\cdot$" denotes the Euclidean inner product. If μ is the distribution of X, then this is also written as $\widehat{\mu}$.

Next consider the d-dimensional mean-zero normal density with covariance matrix ε times the identity:

$$\varphi_\varepsilon(x) := \frac{e^{-\|x\|^2/(2\varepsilon^2)}}{(2\pi\varepsilon^2)^{d/2}}, \qquad \forall x \in \mathbb{R}^d. \tag{6.43}$$

You should check that its characteristic function is

$$\widehat{\varphi}_\varepsilon(t) = e^{-\frac{1}{2}\varepsilon^2\|t\|^2}, \qquad \forall t \in \mathbb{R}^d. \tag{6.44}$$

The following is the simplest analogue of the uniqueness theorem; it is an immediate consequence of Theorem 6.21.

Theorem 6.24 (The Convergence Theorem, $d \geq 1$) *If $\mu, \mu_1, \mu_2, \ldots$ are probability measures on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ such that $\widehat{\mu_n} \to \widehat{\mu}$, then $\mu_n \Longrightarrow \mu$.*

The proof follows the argument that we used to derive Theorem 6.21. This leads to the following rather quickly.

Theorem 6.25 (The CLT in $\mathbb{R}^d$) *If $X_1, X_2, \ldots$ are i.i.d. random variables in $\mathbb{R}^d$ with $\mathrm{E}\{X_1^i\} = \mu_i$ $\mu \in \mathbb{R}^d$, and $\mathrm{Cov}(X_1^i, X_1^j) := Q_{i,j}$ for an invertible $(d \times d)$ matrix Q, then $n^{-1/2}(S_n - n\mu)$ converges weakly to a multi-dimensional Gaussian distribution with mean vector 0 and covariance matrix Q; i.e., for all d-dimensional hypercubes $G := (a_1, b_1] \times \cdots \times (a_d, b_d]$,*

$$\lim_{n\to\infty} \mathrm{P}\left\{\frac{S_n - n\mu}{\sqrt{n}} \in G\right\} = \int_G \frac{e^{-\frac{1}{2}y'Q^{-1}y}}{(2\pi)^{d/2}\sqrt{\det(Q)}}\, dy. \tag{6.45}$$

The following is an important (perhaps the most important) consequence of the development of this chapter:

Theorem 6.26 (The Cramér–Wald Device) *$X_n \Longrightarrow X$ if and only if for all $t \in \mathbb{R}^d$, $t \cdot X_n \Longrightarrow t \cdot X$.*

The point is that the weak convergence of the d-dimensional random variable X_n is boiled down to that of the one-dimensional $t \cdot X_n$, but this needs to be checked for all $t \in \mathbb{R}^d$.

Proof Suppose $X_n \Longrightarrow X$; i.e., for all bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$, $\mathrm{E}\{f(X_n)\} \to \mathrm{E}\{f(X)\}$. Since $g_t(x) := t \cdot x$ is continuous, this implies also that $\mathrm{E}\{f(g_t(X_n))\} \to \mathrm{E}\{f(g_t(X))\}$, which is half of the result. The converse follows from the continuity theorem: Let μ_n and μ denote the distributions of X_n and X, respectively. The condition $t \cdot X_n \Longrightarrow t \cdot X$ is saying that for all $t \in \mathbb{R}^d$, $\widehat{\mu_n}(t) \to \widehat{\mu}(t)$, and the converse follows from Theorem 6.24. □

8 Cramér's Theorem (Optional)

In this section we use characteristic function methods to prove the following striking theorem of Cramér [Cra36].[6.10]

Theorem 6.27 (Cramér's Theorem [Cra36]) *Suppose X_1 and X_2 are independent real-valued random variables such that $X_1 + X_2$ is a standard normal random variable. Then X_1 and X_2 are normal random variables too.*

Remark 6.28 Equivalently, Cramér's theorem states that if μ_1 and μ_2 are probability measures such that $\widehat{\mu_1}(t)\widehat{\mu_2}(t) = e^{-\frac{1}{2}t^2}$, then μ_1 and μ_2 are Gaussian probability measures (why?). Note that Theorem 6.27 remains valid if $X_1 + X_2$ is assumed to have any normal distribution (why?).

The original proof of Cramér's theorem is quite difficult, and I will take a different route that rests on two elementary lemmas: One from complex analysis, and one from probability.

[6.10] In order to read this section you need to know Cauchy's integral formula from undergraduate complex analysis. Zabell [Zab95, p. 487] points out that Cramér's theorem is preceded by the following result of Turing [Tur34, Theorem 3]: *If X_1 and X_2 are independent, and both X_1 and $X_1 + X_2$ are normal, then so is X_2.* Although this is substantially simpler to prove than Cramér's theorem, it was written while Turing was an undergraduate who wished to apply for a Fellowship at King's College, Cambridge. For discussions on the connection of this result to Turing's independent discovery of the central limit theorem of Lindeberg [Lin22] (Exercise 6.9); see Zabell [Zab95].

Recall that an *entire* function is one that is analytic on all of $\mathbb{C}$.

Lemma 6.29 (The Liouville Theorem) *Suppose that $f : \mathbb{C} \to \mathbb{C}$ is an entire function, and there exists an integer $n \geq 0$ such that*

$$\limsup_{|z|\to\infty} \frac{|f(z)|}{|z|^n} < +\infty. \tag{6.46}$$

Then there exist constants $a_0, \ldots, a_n \in \mathbb{C}$ such that $f(z) = \sum_{j=0}^n a_j z^j$.

Remark 6.30 When $n = 1$, condition (6.46) is equivalent to $\sup_z |f(z)| < +\infty$. Hence, in this case, Lemma 6.29 is the standard form of the Liouville theorem of complex analysis; it states that bounded entire functions are constants. The general case is proved by means of a similar argument, as you will see next.

Proof For any $z_0 \in \mathbb{C}$ and $R > 0$, define $\gamma := \{z \in \mathbb{C} : \ |z - z_0| = R\}$, and recall the Cauchy integral formula: For any $n \geq 0$, the nth derivative $f^{(n)}$ is analytic and satisfies

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_\gamma \frac{f(z)}{(z - z_0)^{n+1}}\, dz = \frac{n!}{2\pi i R^n} \int_0^{2\pi} \frac{f\left(z_0 + Re^{i\theta}\right)}{e^{i(n+1)\theta}}\, d\theta. \tag{6.47}$$

Since f is continuous, (6.46) tells us that there exists a constant $A > 0$ such that for all $R > 0$ sufficiently large, and for all $\theta \in [0, 2\pi)$, $|f(z_0 + Re^{i\theta})| \leq AR^n$. In particular, $|f^{(n+1)}(z_0)| \leq (n+1)!AR^{-1}$. Because this holds for all $R > 0$, $f^{(n+1)}(z_0) = 0$ for all $z_0 \in \mathbb{C}$, whence the result. □

The second preliminary lemma is the following probabilistic one.

Lemma 6.31 *If $V \geq 0$ a.s., then for any $a > 0$,*

$$\mathrm{E}\left\{e^{aV}\right\} = 1 + a \int_0^\infty e^{ax} \mathrm{P}\{V \geq x\}\, dx. \tag{6.48}$$

In particular, if $U \geq 0$ a.s. and there exists $r \geq 1$ such that for all $x > 0$, $\mathrm{P}\{V \geq x\} \leq r\mathrm{P}\{U \geq x\}$, then $\mathrm{E}\{\exp(aV)\} \leq r\mathrm{E}\{\exp(aU)\}$ for all $a > 0$.

Proof Because $e^{aV(\omega)} = 1 + a\int_0^\infty \mathbf{1}_{\{V(\omega)\geq x\}} e^{ax}\,dx$ and the integrand is nonnegative, we can take expectations and use Fubini–Tonelli (Theorem 3.6; even without integrability. Why?). This yields (6.48). The second assertion is a ready corollary of the first since $r \geq 1$. □

Proof of Theorem 6.27 Let μ_1 and μ_2 denote the distributions of X_1 and X_2, respectively. Also recall that the condition that X_1+X_2 is standard normal is *equivalent* to $\widehat{\mu_1}(t)\widehat{\mu_2}(t) = \exp(-\frac{1}{2}t^2)$ (Remark 6.28). When defined, $f_k(z) := \mathrm{E}\{e^{zX_k}\}$ $(z \in \mathbb{C})$, so that $\widehat{\mu_k}(t) = f_k(it)$ for all $t \in \mathbb{R}$. You should note that for any $x \in \mathbb{R}$ such that $f_k(x)$ is defined (and is finite), $f_k(x) > 0$. In particular, $\log f_k(z)$ is well-defined when possible, where log denotes the branch of the logarithm that is real on $(0,\infty)$. Our first task is to prove that $\log f_k$ is entire for $k = 1, 2$.

By the dominated convergence theorem (Theorem 2.22), once we prove that for all $c > 0$, $\mathrm{E}\{\exp(c|X_k|)\} < +\infty$, then it follows that f_k is an entire function and $f_k(x) > 0$ for all $x \in \mathbb{R}$ (why?). I will do this for $k = 1$. The case $k = 2$ follows analogously. Throughout, Z designates the sum $X_1 + X_2$, which is a standard normal random variable.

Whenever $X_1 \geq \lambda$ and $X_2 \geq m_1$, it follows that $Z \geq \lambda - m_1$. Since X_1 and X_2 are independent, we get $\mathrm{P}\{Z \geq \lambda - m_1\} \geq \mathrm{P}\{X_1 \geq \lambda\}\mathrm{P}\{X_2 \geq -m_1\}$. Choose and fix $m_1 > 0$ so large that $\mathrm{P}\{X_2 \geq -m_1\} \geq \frac{1}{2}$.[6.11] This yields,

$$\mathrm{P}\{X_1 \geq \lambda\} \leq 2\mathrm{P}\{Z \geq \lambda - m_1\}, \qquad \forall \lambda \in \mathbb{R}. \tag{6.49}$$

Similarly, we can choose $m_2 > 0$ so large that $\mathrm{P}\{X_2 \leq m_2\} \geq \frac{1}{2}$ and obtain

$$\mathrm{P}\{X_1 \leq -\lambda\} \leq 2\mathrm{P}\{Z \leq -\lambda + m_2\}, \qquad \forall \lambda \in \mathbb{R}. \tag{6.50}$$

Finally, let $m := \max(m_1, m_2)$ and combine (6.49) and (6.50) to deduce that

$$\mathrm{P}\{|X_1| \geq \lambda\} \leq 4\mathrm{P}\{|Z| + m \geq \lambda\}, \qquad \forall \lambda > 0. \tag{6.51}$$

Lemma 6.31 ensures that $\mathrm{E}\{\exp(c|X_1|)\} \leq 4e^{cm}\mathrm{E}\{\exp(c|Z|)\}$. But

$$\mathrm{E}\left\{e^{c|Z|}\right\} = \sqrt{\frac{2}{\pi}}\int_0^\infty e^{cw - \frac{1}{2}w^2}\,dw \leq 2e^{\frac{1}{2}c^2}. \tag{6.52}$$

[6.11] The largest such $-m_1$ is called the *median* of X_2.

Therefore, $\mathrm{E}\{\exp(c|X_1|)\} < +\infty$ for all $c \in \mathbb{R}$. This shows that $\log f_1$ is an entire function. Moreover, since $|f_1(z)| \le \mathrm{E}\{\exp(|z| \cdot |X_1|)\}$,

$$|f_1(z)| \le 8 \exp\left(|z|m + \frac{1}{2}|z|^2\right). \tag{6.53}$$

In particular, $\log f_1$ satisfies (6.46) with $n = 2$, and this implies that $f(z) = \exp(a_0 + a_1 z + a_2 z^2)$ for some $a_0, a_1, a_2 \in \mathbb{C}$. Since $f(0) = 1$, $a_0 = 1$. Therefore, in terms of the characteristic function $\widehat{\mu_1}$, we have

$$\widehat{\mu_1}(t) = \exp(a_1 it - a_2 t^2), \qquad \forall t \in \mathbb{R}. \tag{6.54}$$

But (i) $\widehat{\mu_1}(-t) = \overline{\widehat{\mu_1}(t)}$ (Lemma 6.9), and (ii) $|\widehat{\mu_1}(t)| \ge 0$. These two facts and a few lines of calculations together show that a_1 and a_2 are real, and $a_2 \ge 0$. Thus, X_1 is normal with mean a_1 and variance $2a_2$ (Example 6.12). [Recall that $a_2 = 0$ is permissible.] The normality of X_2 follows from a similar argument. □

9 Exercises

Exercise 6.1 If $\mu, \mu_1, \mu_2, \ldots, \mu_n$ is a sequence of probability measures on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$, then show that the following are characteristic functions of probability measures: $\overline{\widehat{\mu}}$, $\operatorname{Re}\widehat{\mu}$, $|\widehat{\mu}|^2$, $\prod_{j=1}^n \widehat{\mu}_j$, and $\sum_{j=1}^n p_j \widehat{\mu}_j$, where $p_1, \ldots, p_n \ge 0$ and $\sum_{j=1}^n p_j = 1$. Also prove that $\overline{\widehat{\mu}(\xi)} = \widehat{\mu}(-\xi)$. Consequently, if μ is a *symmetric measure* (i.e., $\mu(-A) = \mu(A)$ for all $A \in \mathfrak{B}(\mathbb{R}^d)$) then $\widehat{\mu}$ is a real function.

Exercise 6.2 If X has the probability density function $f(x) := (1 - |x|)^+$, then compute the characteristic function of X. Use this and the Plancherel theorem (Theorem 6.18) to show that f itself is the characteristic function of a probability measure. In particular, conclude that there are probability measures that possess a real nonnegative characteristic function that vanishes outside a compact set.

Exercise 6.3 Use the central limit theorem (Theorem 6.22) to derive (6.2).

Exercise 6.4 Prove the following variant of the Plancherel theorem (Theorem 6.18):[6.12] For any $a < b$ and all probability measures μ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$,

$$\begin{aligned}\lim_{\varepsilon \downarrow 0} \frac{1}{2\pi i} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2} \left(\frac{e^{-ita} - e^{-itb}}{t} \right) \widehat{\mu}(t)\, dt \\ = \mu\left((a,b)\right) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\}).\end{aligned} \tag{6.55}$$

Exercise 6.5 Prove the *law of rare events:* If X_n is a binomial random variable (Example 6.14) with parameters n and λn^{-1}, where $\lambda \in (0, n)$ is fixed, then as $n \to \infty$, X_n converges weakly to a Poisson distribution (Example 6.15) with parameter λ.

Exercise 6.6 Suppose f is a probability density function on $\mathbb{R}$; i.e., $f \geq 0$ a.e., and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

1. Apply (6.29) to deduce the *inversion theorem*: Whenever $\widehat{f}$ is integrable, then f is continuous, and

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \widehat{f}(t)\, dt, \qquad \forall x \in \mathbb{R}. \tag{6.56}$$

2. The density function f is said to be of *positive type* if $\widehat{f} \geq 0$ is integrable. Prove that whenever f is of positive type, then for all $x \in \mathbb{R}$, $f(x) \leq f(0)$. In particular, conclude that $f(0) > 0$.

3. If f is of positive type, then prove that $g(x) := \widehat{f}(x)/(2\pi f(0))$ is a probability density function whose characteristic function is $\widehat{g}(t) = f(t)/(2\pi f(0))$.

4. Show that the characteristic function of $g(x) := \frac{1}{2}e^{-|x|}$ is $\widehat{g}(t) = (2\pi)^{-1}(1+t^2)^{-1}$. Conclude that $f(x) := \frac{1}{\pi}(1+x^2)^{-1}$ is a probability density function whose characteristic function is $\widehat{f}(t) = \exp(-|t|)$. The function f defines the so-called *Cauchy distribution.*

[6.12] The convergence of this limit is fairly subtle. For instance, let me mention that in general the simpler-looking $\int \lim_\varepsilon (\cdots)\widehat{\mu}(t)\, dt$ does not exist. The formula of this exercise is a variant of a calculation of Lévy [Lév37, (10), p. 38] who refers to this formula as the *formule de réciprocité* of Fourier.

Exercise 6.7 A probability measure μ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ is said to be *infinitely divisible* if for any $n \geq 1$, there exists a probability measure ν such that $\widehat{\mu} = (\widehat{\nu})^n$.

1. Let X have distribution μ. Show that infinite divisibility is equivalent to the following: For all $n \geq 1$, there exist i.i.d. random variables $X_1, \ldots, X_n$ such that X has the same distribution as $X_1 + \cdots + X_n$.

2. Prove that the normal distribution, and the Poisson distribution are infinitely divisible. So is the probability density $f(x) = \pi^{-1}(1 + x^2)^{-1}$, known as the *Cauchy distribution*.
   (Hint: For the Cauchy, use Exercise 6.6.)

3. Prove that the uniform distribution on $(0, 1)$ is *not* infinitely divisible.

Exercise 6.8 It is not necessary to have identical distributions to have a central limit theory, however the form of such a theorem is inevitably more complicated than Theorem 6.22 as this exercise shows: Let $X_1, X_2, \ldots$ denote independent random variables such that

$$X_j = \begin{cases} j, & \text{with probability } \frac{1}{2j^2}, \\ -j, & \text{with probability } \frac{1}{2j^2}, \\ 1, & \text{with probability } \frac{1}{2} - \frac{1}{4j^2}, \\ -1, & \text{with probability } \frac{1}{2} - \frac{1}{4j^2}. \end{cases} \tag{6.57}$$

Show that if $S_n := \sum_{j=1}^n X_j$, then $(S_n - \mathrm{E}\{S_n\})/\mathrm{Var}(S_n)$ converges weakly to a normal distribution with mean 0 and variance $\sqrt{2/3}$ (not 1).
(Hint: Consider summing the truncated variables, $Y_j := X_j \mathbf{1}_{\{|X_j| \leq 1\}}$.)

Exercise 6.9 Let $X_1, X_2, \ldots$ denote independent $L^2(\mathrm{P})$-random variables in $\mathbb{R}$, and for all n define $s_n^2 := \sum_{j=1}^n \mathrm{Var}(X_j)$. In addition, suppose that $s_n \to \infty$, and that the following *Lindeberg Condition* holds: For all $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^n \mathrm{E}\left\{X_j^2;\ |X_j| > \varepsilon s_n\right\} = 0. \tag{6.58}$$

If $S_n := X_1 + \cdots + X_n$, then prove the *Lindeberg Central Limit Theorem* (cf. Lindeberg [Lin22]): For all $a < b$,

$$\lim_{n\to\infty} \mathrm{P}\left\{a < \frac{S_n - \mathrm{E}\{S_n\}}{s_n} \leq b\right\} = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx. \tag{6.59}$$

Check that the variables of Exercise 6.8 do not satisfy the Lindeberg condition (6.58). (Hint: Consider the truncated random variables, $X_{j,n} := X_j \mathbf{1}_{\{|X_j| \le \varepsilon s_n\}}$.)

Exercise 6.10 Let $\mathsf{e}_1, \ldots, \mathsf{e}_d$ denote the usual basis vectors of $\mathbb{R}^d$; i.e., $\mathsf{e}_1' := (1, 0, \ldots, 0)$, $\mathsf{e}_2' := (0, 1, 0, \ldots, 0)$, etc. Consider i.i.d. random variables $X_1, X_2, \ldots$ such that $\mathrm{P}\{X_1 = \pm \mathsf{e}_j\} = (2d)^{-1}$. Then the random process $S_n := X_1 + \cdots + X_n$ with $S_0 := 0$ is the *simple walk* on $\mathbb{Z}^d$; it starts at zero and moves to each of the neighboring sites in $\mathbb{Z}^d$ with equal probability, and the process continues in this way ad infinitum. Find vectors a_n and constants b_n such that $(S_n - a_n)/b_n$ converges weakly to a nontrivial limit distribution. Compute the latter distribution.

Chapter 7

Martingale Theory

As usual, $(\Omega, \mathfrak{F}, \mathrm{P})$ is the underlying probability space throughout.

1 Conditional Probabilities and Expectations

The simplest definition of conditional probabilities is the one that was handed to us from the nineteenth century: Given any two events A and B such that $\mathrm{P}(B) > 0$, define $\mathrm{P}(A \,|\, B) := \mathrm{P}(A \cap B) \div \mathrm{P}(B)$. This is the "conditional probability of A given B," and represents the odds of the occurance of A given that B is known to have occurred. The corresponding "conditional expectation" is defined as follows: If $Y \in L^1(\mathrm{P})$ and $\mathrm{P}(B) > 0$, then

$$\mathrm{E}\{Y \,|\, B\} = \frac{1}{\mathrm{P}(B)} \mathrm{E}\{Y; B\}. \tag{7.1}$$

Example 7.1 Let (X, Y) denote a discrete random variable taking the values $(x_1, y_1), \ldots, (x_n, y_n)$ (all pairs distinct) with probabilities $p_1, \ldots, p_n$ (all positive). Then, the conditional distribution of X given that we have observed that $Y = y_1$, for instance, is given by the following: $\mathrm{P}\{X = x_\ell \,|\, Y = y_1\} = \mathrm{P}\{X = x_\ell, Y = y_1\} \div \mathrm{P}\{Y = y_1\}$. This is a probability distribution on $\{x_1, \ldots, x_n\}$ (endowed with its power set if you want properly-defined σ-algebras), and thus also leads to conditional expectation: For any function h,

$$\mathrm{E}\{h(X) \,|\, Y = y_1\} = \sum_{\ell=1}^{n} h(x_\ell) \mathrm{P}\{X = x_\ell \,|\, Y = y_1\}. \tag{7.2}$$

For instance, in the above we have (check!):

$$\begin{aligned}\mathrm{E}\{X\} &= \sum_{\ell=1}^{n} x_\ell \mathrm{P}\{X = x_\ell\} \\ \mathrm{E}\{X \,|\, Y = y_1\} &= \sum_{\ell=1}^{n} x_\ell \mathrm{P}\{X = x_\ell \,|\, Y = y_1\}.\end{aligned} \tag{7.3}$$

What if Y has an absolutely continuous distribution? Now the intuitive notion that $Y = y_1$ still makes sense. We can after all observe Y, and it must have *some* value although the event $\{Y = y_1\}$ may have zero probability for all nonrandom choices of y_1.

Similar considerations led A. N. Kolmogorov to a much more general, though somewhat abstract, notion of conditional expectations. To explain Kolmogorov's idea, let us return to the Example 7.1 above.

Example 7.2[Example 7.1; Continued] Fix a function h, and consider the *random variable* $\varphi(Y) := \mathrm{E}\{h(X) \,|\, Y\}$, where φ is described by the assignment $\varphi(y_j) = \mathrm{E}\{h(X) \,|\, Y = y_j\}$ ($j = 1, \ldots, n$), and the latter conditional expectation is defined in (7.2). This function φ has the following property that is known as *Bayes' formula*:[7.1]

$$\mathrm{E}\{h(X)\} = \sum_{j=1}^{n} \varphi(y_j) \mathrm{P}\{Y = y_j\} = \mathrm{E}\{\varphi(Y)\}. \tag{7.4}$$

The starting point of Kolmogorov's idea—in this example—is the observation that given any other function ψ,

$$\begin{aligned}\mathrm{E}\{\psi(Y)\varphi(Y)\} &= \sum_{j=1}^{n} \psi(y_j)\varphi(y_j) \mathrm{P}\{Y = y_j\} \\ &= \sum_{j=1}^{n} \psi(y_j) \mathrm{E}\{h(X); Y = y_j\}.\end{aligned} \tag{7.5}$$

[To prove the last equality, you first derive it for simple functions h, then elementary functions, and then extend, as usual, to general h.] Thus, we can

[7.1] Bayes' formula was published posthumously in Bayes [Bay63]. It was discovered independently in Laplace [Lap12] who was also responsible for popularizing the Bayes formula among mathematicians. This was noted for instance by Poincaré [Poi12].

rewrite this as follows:

$$\mathrm{E}\{\psi(Y)\varphi(Y)\} = \mathrm{E}\left\{\sum_{j=1}^{n} \psi(y_j)h(X)\mathbf{1}_{\{Y=y_j\}}\right\} = \mathrm{E}\left\{\psi(Y)h(X)\right\}. \tag{7.6}$$

Yet another way to think of this is:

$$\mathrm{E}\Big(\xi\mathrm{E}\{h(X)\,|\,Y\}\Big) = \mathrm{E}\{\xi h(X)\}, \tag{7.7}$$

for all random variables ξ that are Borel-measurable functions of Y. That is, you can think of the random variable $\mathrm{E}\{h(X)\,|\,Y\}$ as the projection—in $L^2(\mathrm{P})$—of $h(X)$ onto the σ-algebra generated by Y—more precisely, onto the linear space of all Borel-measurable functions of Y.

The preceding example suggests a way out of the mentioned difficulties with working with conditional expectations: The right notion is that of conditioning with respect to general σ-algebras.

Definition 7.3 Suppose $\mathfrak{G}$ is a sub-σ-algebra of $\mathfrak{F}$, and X is an integrable random variable. Then, the *conditional expectation of* X *given* $\mathfrak{G}$ is defined by the following: $\mathrm{E}\{X\,|\,\mathfrak{G}\}$ is a $\mathfrak{G}$-measurable random variable in $L^1(\mathrm{P})$, and for all bounded $\mathfrak{G}$-measurable random variables ξ, $\mathrm{E}(\xi\mathrm{E}\{X\,|\,\mathfrak{G}\}) = \mathrm{E}\{\xi X\}$. In the case that $\mathfrak{G} = \sigma(Y)$—the σ-algebra generated by Y—for some random variable Y, then $\mathrm{E}\{X\,|\,\sigma(Y)\}$ is also written as $\mathrm{E}\{X\,|\,Y\}$. This makes sense even if Y takes values in $\mathbb{R}^n$ for $n > 1$.

The following shows that our new notion of conditional expectations contains the one from classical probability theory.

Remark 7.4 If $B \in \mathfrak{F}$, and if $0 < \mathrm{P}(B) < 1$, then for all $Y \in L^1(\mathrm{P})$,

$$\mathrm{E}\{Y\,|\,\sigma(B)\} = \begin{cases} \frac{1}{\mathrm{P}(B)}\mathrm{E}\{Y;B\}, & \text{a.s. on } B, \\ \frac{1}{\mathrm{P}(B^{\complement})}\mathrm{E}\{Y;B^{\complement}\}, & \text{a.s. on } B^{\complement}. \end{cases} \tag{7.8}$$

where $\sigma(B) := \{\varnothing, \Omega, B, B^{\complement}\}$ is the *σ-algebra generated by* B. More generally, whenever $B_1, \ldots, B_n$ are disjoint subsets of $\mathfrak{F}$ with positive P-measure, then for almost all ω,

$$\mathrm{E}\Big\{Y\,\Big|\,\sigma\{B_1,\ldots,B_n\}\Big\}(\omega) = \sum_{j=1}^{n} \mathbf{1}_{B_j}(\omega)\frac{\mathrm{E}\{Y;B_j\}}{\mathrm{P}(B_j)}. \tag{7.9}$$

This is valid for all $Y \in L^1(\mathrm{P})$, and the proof involves a direct verification. Indeed, if $A \in \{B_1, \ldots, B_n\}$, then $A = B_j$ for some j, and therefore, we can multiply the right-hand side of the previous display by $\mathbf{1}_A(\omega)$ and take expectations to obtain $\mathrm{E}\{Y; B_j\} = \mathrm{E}\{Y; A\}$, as needed. The general case follows by considering finite (automatically disjoint) unions of the B_j's.

Remark 7.5 Intuitively speaking, you should think of $\mathrm{E}\{X\}$ as the "best" predictor of $X \in L^1(\mathrm{P})$. However, if we are aware of the (partial) information that is contained in $\mathfrak{G}$, then the "best" predictor of X is $\mathrm{E}\{X \,|\, \mathfrak{G}\}$ (and not necessarily $\mathrm{E}\{X\}$ anymore). Given this, try to convince yourself heuristically of the following before reading further: (i) $\mathrm{E}\{X \,|\, \{\varnothing, \Omega\}\} = \mathrm{E}\{X\}$, since knowing the information in the trivial σ-algebra $\{\varnothing, \Omega\}$ amounts to knowing that either nothing happens or something in Ω will (why?). (ii) $\mathrm{E}\{X \,|\, \mathfrak{F}\} = X$, since knowing $\mathfrak{F}$ amounts to knowing everything there is to know. [The said computations will be proved next, but you should try to understand their meaning intuitive before reading further. In particular, convince yourselves that knowing $\mathfrak{F}$ is the same as knowing all $\mathfrak{F}$-measurable random variables including X.]

Theorem 7.6 *If $\mathfrak{G}$ is a sub-σ-algebra of $\mathfrak{F}$ and $X \in L^1(\mathrm{P})$ is real-valued, then $\mathrm{E}\{X \,|\, \mathfrak{G}\}$ always exists and is unique a.s. Furthermore, conditional expectations have the following properties:*

(i) $\mathrm{E}\{\mathrm{E}(X \,|\, \mathfrak{G})\} = \mathrm{E}\{X\}$, $\mathrm{E}\{X \,|\, \mathfrak{F}\} = X$, *a.s., and* $\mathrm{E}(X \,|\, \{\varnothing, \Omega\}) = \mathrm{E}\{X\}$, *a.s.*

(ii) If ξ is $\mathfrak{G}$-measurable, then a.s., $\mathrm{E}\{\xi X \,|\, \mathfrak{G}\} = \xi \mathrm{E}\{X \,|\, \mathfrak{G}\}$, *and if $X \geq 0$, a.s., then with probability one,* $\mathrm{E}\{X \,|\, \mathfrak{G}\} \geq 0$.

(iii) If $X_1, X_2, \ldots, X_n \in L^1(\mathrm{P})$ and $a_1, a_2, \ldots, a_n \in \mathbb{R}$, then

$$\mathrm{E}\left\{ \sum_{j=1}^n a_j X_j \,\middle|\, \mathfrak{G} \right\} = \sum_{j=1}^n a_j \mathrm{E}\{X_j \,|\, \mathfrak{G}\}, \quad \textit{a.s.} \tag{7.10}$$

(iv) (Conditional Jensen's inequality) *If $\psi : \mathbb{R} \to \mathbb{R}$ is convex and if $\psi(X) \in L^1(\mathrm{P})$, then* $\mathrm{E}\{\psi(X) \,|\, \mathfrak{G}\} \geq \psi\left(\mathrm{E}\{X \,|\, \mathfrak{G}\}\right)$, *a.s.*

(v) (Conditional Fatou's Lemma) *If $X_1, X_2, \ldots \in L^1(\mathrm{P})$ are nonnegative a.s., then a.s.,* $\mathrm{E}\{\liminf_n X_n \,|\, \mathfrak{G}\} \leq \liminf_n \mathrm{E}\{X_n \,|\, \mathfrak{G}\}$.

(vi) (Conditional Bounded Convergence Theorem) *If* $X_1, X_2, \ldots$ *are a.s. bounded random variables, and if* $\lim_n X_n$ *exists a.s., then with probability one,* $\mathrm{E}\{\lim_n X_n \,|\, \mathfrak{G}\} = \lim_n \mathrm{E}\{X_n \,|\, \mathfrak{G}\}$.

(vii) (Conditional Monotone Convergence Theorem) *If* $X_n \uparrow X$ *a.s. and* $X \in L^1(\mathrm{P})$*, then with probability one,*

$$\mathrm{E}\left\{\lim_{n\to\infty} X_n \,\middle|\, \mathfrak{G}\right\} = \lim_n \mathrm{E}\{X_n \,|\, \mathfrak{G}\}. \tag{7.11}$$

(viii) (Conditional Dominated Convergence Theorem) *If* $\sup_n |X_n| \in L^1(\mathrm{P})$ *and* $\lim_n X_n$ *exists a.s. then with probability one,* $\mathrm{E}\{\lim_n X_n \,|\, \mathfrak{G}\} = \lim_n \mathrm{E}\{X_n \,|\, \mathfrak{G}\}$.

(ix) (Conditional Hölder Inequality) *If for some* $p > 1$, $X \in L^p(\mathrm{P})$ *and* $Y \in L^q(\mathrm{P})$ *where* $p^{-1} + q^{-1} = 1$*, then with probability one,*

$$\left|\mathrm{E}\{XY \,\middle|\, \mathfrak{G}\right\}| \le [\mathrm{E}\{|X|^p \,|\, \mathfrak{G}\}]^{\frac{1}{p}} \cdot [\mathrm{E}\{|Y|^q \,|\, \mathfrak{G}\}]^{\frac{1}{q}}. \tag{7.12}$$

(x) (Conditional Minkowski Inequality) *If for some* $p \ge 1$, $X, Y \in L^p(\mathrm{P})$ *then with probability one,* $[\mathrm{E}\{|X+Y|^p \,|\, \mathfrak{G}\}]^{\frac{1}{p}} \le [\mathrm{E}\{|X|^p \,|\, \mathfrak{G}\}]^{\frac{1}{p}} + [\mathrm{E}\{|Y|^p \,|\, \mathfrak{G}\}]^{\frac{1}{p}}$.

Remark 7.7 This suggests that for almost every $\omega \in \Omega$, $\mathrm{E}\{X \,|\, \mathfrak{G}\}(\omega) = \int_\Omega X(x)\, \mathrm{P}_\omega(dx)$, where for each (or perhaps P-almost all) ω, P_ω is a probability measure. Depending on the topological structure of Ω, this is often the case, but we will not dwell on it. We will however see a weaker version of such a result shortly; cf. Proposition 7.12 below.

Before proving Theorem 7.6, let us state and prove a preliminary technical lemma.

Lemma 7.8 *Suppose that* Z *and* W *are* $\mathfrak{G}$*-measurable random variables, and that for all* $A \in \mathfrak{G}$ $\mathrm{E}\{Z; A\} \le \mathrm{E}\{W; A\}$*. Then* $Z \le W$*, a.s. In particular, if* $\mathrm{E}\{Z; A\} = \mathrm{E}\{W; A\}$ *for all* $A \in \mathfrak{G}$*, then* $Z = W$*, a.s.*

Proof Fix $\varepsilon > 0$ and consider the set $A_\varepsilon := \{\omega \in \Omega : \; Z(\omega) \geq W(\omega) + \varepsilon\}$. Equivalently, we can write $A_\varepsilon = (Z - W)^{-1}([\varepsilon, \infty))$, which also shows that $A_\varepsilon \in \mathfrak{G}$ since $Z - W$ is also $\mathfrak{G}$-measurable. Furthermore,

$$0 \geq \mathrm{E}\{Z; A_\varepsilon\} - \mathrm{E}\{W; A_\varepsilon\} = \mathrm{E}\{Z - W; A_\varepsilon\} \geq \varepsilon \mathrm{P}(A_\varepsilon). \tag{7.13}$$

In particular, $\mathrm{P}(A_\varepsilon) = 0$ for all $\varepsilon > 0$, and $\mathrm{P}(\cup_{\varepsilon \in \mathbb{Q}_+} A_\varepsilon) = \lim_{\varepsilon \to 0} \mathrm{P}(A_\varepsilon) = 0$ (why?). Since $\cup_{\varepsilon \in \mathbb{Q}_+} A_\varepsilon = \{Z > W\}$, this shows that $Z \leq W$, a.s. To prove the second part, we apply the first to (Z, W) in this order to obtain that $Z \leq W$, a.s. But then we can apply the first part to (W, Z) to obtain also that $W \leq Z$, a.s. Together, the last two observations complete the proof. □

Proof of Theorem 7.6 We begin our proof by showing the existence of conditional expectations in the case that $X \geq 0$, a.s. Consider

$$\mathrm{Q}(A) := \mathrm{E}\{X; A\} = \int_A X \, d\mathrm{P}, \qquad \forall A \in \mathfrak{G}. \tag{7.14}$$

Since $X \in L^1(\mathrm{P})$, it follows easily that Q is a finite measure on $(\Omega, \mathfrak{G})$ (why?). It is also easy to see that $\mathrm{Q} \ll \mathrm{P}$, so that by the Radon–Nikodým theorem (Theorem 4.2): (a) $\mathrm{E}\{X \,|\, \mathfrak{G}\} := \frac{d\mathrm{Q}}{d\mathrm{P}}$ exists; (b) it is in $L^1(\Omega, \mathfrak{G}, \mathrm{P})$; (c) it is P-a.s. unique; and (d) it is P-a.s. nonnegative since X is. Being in the said L^1-space implicitly shows that $\mathrm{E}\{X \,|\, \mathfrak{G}\}$ is $\mathfrak{G}$-measurable. Moreover, for all bounded $\mathfrak{G}$-measurable functions ξ, $\mathrm{E}\{\xi X\} = \int \xi \, d\mathrm{Q} = \int \xi \mathrm{E}\{X \,|\, \mathfrak{G}\} \, d\mathrm{P} = \mathrm{E}(\xi \mathrm{E}\{X \,|\, \mathfrak{G}\})$. (If ξ is a simple function, this follows from (7.14); then proceed by checking this for elementary functions, and finally take limits as usual.) This proves the existence and uniqueness of $\mathrm{E}\{X \,|\, \mathfrak{G}\}$ when $X \geq 0$, a.s. In general, we can *define* $\mathrm{E}\{X \,|\, \mathfrak{G}\} := \mathrm{E}\{X^+ \,|\, \mathfrak{G}\} - \mathrm{E}\{X^- \,|\, \mathfrak{G}\}$.

For the first portion of (ii), we first work with $\xi = c\mathbf{1}_A$ where $A \in \mathfrak{G}$ and $c \in \mathbb{R}$, and show that $\mathrm{E}\{Xc\mathbf{1}_A \,|\, \mathfrak{G}\} = c\mathbf{1}_A \mathrm{E}\{X \,|\, \mathfrak{G}\}$, a.s. Equivalently, we need to show that for all $B \in \mathfrak{G}$, $\mathrm{E}\{c\mathbf{1}_A X; B\} = \mathrm{E}\{c\mathbf{1}_A \mathrm{E}(X \,|\, \mathfrak{G}); B\}$; cf. Lemma 7.8. But this is equivalent to $\mathrm{E}\{X; A \cap B\} = \mathrm{E}\{\mathrm{E}(X \,|\, \mathfrak{G}); A \cap B\}$, which holds obviously since $A \cap B \in \mathfrak{G}$. This proves (ii) in the case that ξ is elementary. It is not hard to see how this argument works equally well when ξ is simple. To prove the first portion of (ii) for general ξ, we can assume without loss of generality that $\xi \geq 0$, a.s., for otherwise we could consider ξ^+ and ξ^- separately. Now choose simple $\mathfrak{G}$-measurable random variables $\xi_n \uparrow \xi$. What we have shown so far implies that for all $A \in \mathfrak{G}$, $\mathrm{E}\{\xi_n X; A\} = \mathrm{E}\{\xi_n \mathrm{E}(X \,|\, \mathfrak{G}); A\}$. Let $n \uparrow \infty$ and use the monotone convergence theorem to deduce that for all $A \in \mathfrak{G}$, $\mathrm{E}\{\xi X; A\} = \mathrm{E}\{\xi \mathrm{E}(X \mid \mathfrak{G}); A\}$. An appeal to Lemma 7.8 completes the proof of the first portion of (ii).

Now we proceed to verify (i). First of all, X and $\mathrm{E}\{X \mid \mathfrak{G}\}$ have the same expectation since $\mathrm{E}\{X\} = \mathrm{E}\{X; \Omega\} = \mathrm{E}\{\mathrm{E}(X \mid \mathfrak{G}); \Omega\}$. The fact that the conditional expectation of X given the trivial σ-algebra $\{\varnothing, \mathfrak{F}\}$ equals $\mathrm{E}\{X\}$ (the unconditional expectation) follows from (7.14) for then $\mathrm{Q}(\varnothing) = 0$ and $\mathrm{Q}(\Omega) = \mathrm{E}\{X\}$ together define Q. To finish proving (i), we need to verify that $\mathrm{E}\{X \mid \mathfrak{F}\} = X$. We know that for all bounded random variables ξ,

$$\mathrm{E}\Big\{\xi \Big[\mathrm{E}(X \mid \mathfrak{F}) - X\Big]\Big\} = 0. \tag{7.15}$$

Consider $\xi :=$ the sign of $(\mathrm{E}\{X \mid \mathfrak{F}\} - X)$ to deduce from the preceding display that $\|\mathrm{E}(X \mid \mathfrak{F}) - X\|_1 = 0$, thus proving (i).

It remains to prove (iii) with $n = 2$, since the remaining properties are proved in the same manner as their unconditional counterparts were, and the proofs only rely on (i)–(iii). Define for $j = 1, 2$, $\mathrm{Q}_j(A) := \mathrm{E}\{X_j; A\}$, and $\mathrm{Q}'(A) := \mathrm{E}\{a_1 X_1 + a_2 X_2; A\}$ for all $A \in \mathfrak{G}$. Then, $\mathrm{Q}'(A) = a_1 \mathrm{Q}_1(A) + a_2 \mathrm{Q}_2(A)$, and for all bounded ξ,

$$\begin{aligned}\int \xi \, d\mathrm{Q}' &= a_1 \int \xi \, d\mathrm{Q}_1 + a_2 \int \xi \, d\mathrm{Q}_2 \\ &= a_1 \mathrm{E}\{\xi \mathrm{E}(X_1 \mid \mathfrak{G})\} + a_2 \mathrm{E}\{\xi \mathrm{E}(X_2 \mid \mathfrak{G})\} \\ &= \mathrm{E}\left\{\xi \Big[a_1 \mathrm{E}(X_1 \mid \mathfrak{G}) + a_2 \mathrm{E}(X_2 \mid \mathfrak{G})\Big]\right\}.\end{aligned} \tag{7.16}$$

On the other hand, $\int \xi \, d\mathrm{Q}' = \mathrm{E}\{\xi \mathrm{E}(a_1 X_1 + a_2 X_2 \mid \mathfrak{G})\}$, and the result follows. □

Theorem 7.6 describes some of the elementary properties of conditional expectations. The following describes two more properties that are quite useful.

Theorem 7.9 *If* $X \in L^1(\mathrm{P})$, *and if* $\mathfrak{F}_1 \subseteq \mathfrak{F}_2$ *are both sub-*σ*-algebras of* $\mathfrak{F}$, *then with probability one,*

$$\mathrm{E}\Big\{\mathrm{E}(X \mid \mathfrak{F}_1) \Big| \mathfrak{F}_2\Big\} = \mathrm{E}\Big\{\mathrm{E}(X \mid \mathfrak{F}_2) \Big| \mathfrak{F}_1\Big\} = \mathrm{E}\{X \mid \mathfrak{F}_1\}. \tag{7.17}$$

If X *is independent of* $\mathfrak{F}_1$, *then* $\mathrm{E}\{X \mid \mathfrak{F}_1\} = \mathrm{E}\{X\}$.

The preceding has implicitly relied on the following definition:

Definition 7.10 Two σ-algebras $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are *independent* if for all $A_i \in \mathfrak{F}_i$, $\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\mathrm{P}(A_2)$. A random variable X is independent of a σ-algebra $\mathfrak{F}$ if $\sigma(X)$ and $\mathfrak{F}$ are independent. That is, for all bounded measurable functions $h : \mathbb{R} \to \mathbb{R}$, and for all $A \in \mathfrak{F}$, $\mathrm{E}\{h(X); A\} = \mathrm{E}\{h(X)\} \cdot \mathrm{P}(A)$.

Remark 7.11 Equation (7.17) is known as the *towering property of conditional expectations*. Roughly speaking, it states that one can only use the least amount of available information to make a good prediction on the value of X. One can construct examples where the conditional expectation of $\mathrm{E}\{X \,|\, \mathfrak{F}_1\}$ given $\mathfrak{F}_2$ is not the same as the conditional expectation of $\mathrm{E}\{X \,|\, \mathfrak{F}_2\}$ given $\mathfrak{F}_1$. Thus, the condition $\mathfrak{F}_1 \subseteq \mathfrak{F}_2$ is not to be taken lightly.

Proof If $A \in \mathfrak{F}_1$, then $A \in \mathfrak{F}_2$, and by Theorem 7.6, $\mathrm{E}\{\mathrm{E}(X \,|\, \mathfrak{F}_2); A\} = \mathrm{E}\{\mathrm{E}(X\mathbf{1}_A \,|\, \mathfrak{F}_2)\}$. This equals $\mathrm{E}\{X; A\}$ thanks to the definition of conditional expectations. Consequently, $\mathrm{E}\{\mathrm{E}(X \,|\, \mathfrak{F}_2) \,|\, \mathfrak{F}_1\} = \mathrm{E}\{X \,|\, \mathfrak{F}_1\}$, a.s. On the other hand, $\mathrm{E}\{X \,|\, \mathfrak{F}_1\}$ is $\mathfrak{F}_1$- and hence $\mathfrak{F}_2$-measurable. Thus, by Theorem 7.6, $\mathrm{E}\{\mathrm{E}(X \,|\, \mathfrak{F}_1) \,|\, \mathfrak{F}_2\} = \mathrm{E}\{X \,|\, \mathfrak{F}_1\} \times \mathrm{E}\{1 \,|\, \mathfrak{F}_2\} = \mathrm{E}\{X \,|\, \mathfrak{F}_1\}$, a.s. (Why does the conditional expectation of 1 equal 1?) This proves the first portion.

For the final portion let me note that whenever $A \in \mathfrak{F}_1$, then $\mathrm{E}\{X; A\} = \mathrm{E}\{X\}\mathrm{P}(A)$, which is equal to $\mathrm{E}\{\mathrm{E}(X); A\}$. □

Conditional probabilities follow readily from conditional expectations via the assignment,

$$\mathrm{P}\{A \,|\, \mathfrak{G}\} := \mathrm{E}\{\mathbf{1}_A \,|\, \mathfrak{G}\}, \qquad \forall A \in \mathfrak{F}. \tag{7.18}$$

Their salient properties are not hard to derive, and are listed in the following:

Proposition 7.12 *For any sub-σ-algebra $\mathfrak{G} \subseteq \mathfrak{F}$, the following holds:*

(i) $\mathrm{P}\{\varnothing \,|\, \mathfrak{G}\} = 0$, *a.s., and for all* $A \in \mathfrak{F}$, $\mathrm{P}\{A \,|\, \mathfrak{G}\} = 1 - \mathrm{P}\{A^{\complement} \,|\, \mathfrak{G}\}$, *a.s.*

(ii) For any disjoint measurable $A_1, A_2, \ldots$ there exists a null set outside of which, $\mathrm{P}\{\cup_{i=1}^{\infty} A_i \,|\, \mathfrak{G}\} = \sum_{i=1}^{\infty} \mathrm{P}\{A_i \,|\, \mathfrak{G}\}$.

2 Filtrations, Semimartingales, and Stopping Times

Definition 7.13 A *stochastic process* (or a random process, or a process) is a collection of random variables. If $\mathfrak{F}_1 \subseteq \mathfrak{F}_2 \subseteq \cdots$ are sub-σ-algebras of $\mathfrak{F}$, then $\{\mathfrak{F}_i;\, i \geq 1\}$ is a *filtration.* A process $X_1, X_2, \ldots$ is *adapted* to a filtration $\mathfrak{F}_1, \mathfrak{F}_2, \ldots$ if for every $n \geq 1$, X_n is $\mathfrak{F}_n$-measurable.

The best way to construct such things is to start with a random process of your choice, call it $X_1, X_2, \ldots$, and *defining* $\mathfrak{F}_n := \sigma(X_1, \ldots, X_n)$; i.e., $\mathfrak{F}_n$ is the σ-algebra generated by X_n. Clearly, $\mathfrak{F}_n \subseteq \mathfrak{F}_{n+1}$, and by definition $\{X_n;\, n \geq 1\}$ is adapted to $\{\mathfrak{F}_n;\, n \geq 1\}$. Suppose, in addition, that $X_n \in L^1(\mathrm{P})$ for all $n \geq 1$, and think of $X_1, X_2, \ldots$ as a random process that evolves in (discrete) time. Then, a sensible prediction of the value of the process at time $n+1$ given the values of the process by time n is $\mathrm{E}\{X_{n+1} \,|\, \mathfrak{F}_n\}$. We say that $X := \{X_n;\, n \geq 1\}$ is a martingale if this predicted value is X_n. In this way, you should convince yourself that fair games are martingales, and in a sense, the converse is also true.

Definition 7.14 A stochastic process $X := \{X_n;\, n \geq 1\}$ is a *submartingale* with respect to a filtration $\mathfrak{F} := \{\mathfrak{F}_n;\, n \geq 1\}$ if:

(i) X is adapted to $\mathfrak{F}$.

(ii) $X_n \in L^1(\mathrm{P})$ for all $n \geq 1$.

(iii) For each $n \geq 1$, $\mathrm{E}\{X_{n+1} \,|\, \mathfrak{F}_n\} \geq X_n$, a.s.

X is said to be a *supermartingale* if $-X$ is a submartingale. It is a *martingale* if it is both a sub- and a supermartingale; it is a *semimartingale* if it can be written as $X_n = Y_n + Z_n$ where Y_n is a martingale and Z_n is a bounded variation process; i.e., $Z_n = U_n - V_n$ where $U_1 \leq U_2 \leq \cdots$ and $V_1 \leq V_2 \leq \cdots$ are integrable, adapted processes.

Here are a few examples of martingales.

Example 7.15[Independent Sums] Suppose that a fair game is repeatedly played, each time independently from other times. Suppose also that each game results in ±1 dollar for the gambler. One way to model this is to let $X_1, X_2, \ldots$ be i.i.d. random variables with the values ± 1 with probability one-half each. Then, the gambler's gains, after k games, is $S_k := X_1 + \cdots X_k$. By independence, $\mathrm{E}\{X_k \,|\, X_1, \ldots, X_{k-1}\} = \mathrm{E}\{X_k\} = 0$. This implies that S is a martingale with respect to the filtration $\mathfrak{F}_1, \mathfrak{F}_2, \ldots$, where $\mathfrak{F}_n = \sigma(X_1, \ldots, X_n)$. More generally still, if $S_n = X_1 + \cdots + X_n$ where the X_j's are independent (not necessarily i.i.d.) and mean-zero , then S is a martingale with respect to $\mathfrak{F}$. You should check that $\mathfrak{F}_n$ is also equal to $\sigma(S_1, \ldots, S_n)$.

For our second class of examples, we need a definition.

Definition 7.16 A stochastic process $A_1, A_2, \ldots$ is *previsible* with respect to a given filtration $\mathfrak{F}_n$ if for every $n \geq 1$, A_n is $\mathfrak{F}_{n-1}$-measurable, where $\mathfrak{F}_0$ is always the trivial σ-algebra.

Example 7.17[Martingale Transforms] Suppose S is a martingale with respect to some filtration $\mathfrak{F}_n$, and consider the process Y defined as

$$Y_n := y_0 + \sum_{j=1}^{n} A_j(S_j - S_{j-1}), \qquad \forall n \geq 0, \tag{7.19}$$

where $S_0 := 0$, y_0 is a constant, and A is a previsible process with respect to $\mathfrak{F}_n := \sigma(X_1, \ldots, X_n)$ with $\mathfrak{F}_0 := \{\varnothing, \Omega\}$ denoting the trivial σ-algebra. The process Y is called the *martingale transform* of S, and we next argue that Y is itself a martingale.

Since it is clear that Y is adapted, let us check the martingale property: When $n = 0$, $\mathrm{E}\{Y_{n+1} \,|\, \mathfrak{F}_n\} = \mathrm{E}\{Y_1\} = y_0 = Y_0$. When $n \geq 1$, we can write $Y_{n+1} = Y_n + A_{n+1}(S_{n+1} - S_n)$ to see that $\mathrm{E}\{Y_{n+1} \,|\, \mathfrak{F}_n\} = Y_n + \mathrm{E}\{A_{n+1}(S_{n+1} - S_n) \,|\, \mathfrak{F}_n\}$, a.s. But A_{n+1} is $\mathfrak{F}_n$-measurable. So by Theorem 7.6, with probability one, $\mathrm{E}\{Y_{n+1} \,|\, \mathfrak{F}_n\} = Y_n + A_{n+1}\mathrm{E}\{S_{n+1} - S_n \,|\, \mathfrak{F}_n\}$, and the last term equals 0 a.s. thanks to the martingale property of S. This proves that Y is a martingale.

Note that the martingale transform of (7.19) has the equivalent definition, $Y_{n+1} - Y_n = A_{n+1}(S_{n+1} - S_n)$ for $n \geq 0$, where $Y_0 := y_0$. You should think of this, informally, as the discrete analogue of a stochastic differential identity of the type, $dY = A\,dS$.

Example 7.18[Doob Martingales] Suppose $Y \in L^1(\mathrm{P})$, and let $\mathfrak{F}_1, \mathfrak{F}_2, \ldots$ denote a filtration. Then, $X_n := \mathrm{E}\{Y \,|\, \mathfrak{F}_n\}$ defines a martingale (check!) that is called a *Doob martingale.*

Lemma 7.19 *If X is a submartingale with respect to a filtration $\mathfrak{F}$, then it is also a submartingale with respect to the filtration generated by X itself. That is, for all n, $\mathrm{E}\{X_{n+1} \,|\, X_1, \ldots, X_n\} \geq X_n$, a.s.*

Proof For all n and all $A \in \mathfrak{B}(\mathbb{R})$, $X_n^{-1}(A) \in \mathfrak{F}_n$. Because $\sigma(X_1, \ldots, X_n)$ is the smallest σ-algebra that contains $X_n^{-1}(A)$ for all $A \in \mathfrak{B}(\mathbb{R})$, it follows that $\sigma(X_1, \ldots, X_n) \subseteq \mathfrak{F}_n$ for all n. Consequently, by the towering property of conditional expectations (Theorem 7.9), a.s.,

$$\begin{aligned} \mathrm{E}\{X_{n+1} \,|\, X_1, \ldots, X_n\} &= \mathrm{E}\Big\{\mathrm{E}(X_{n+1} \,|\, \mathfrak{F}_n)\Big| X_1, \ldots, X_n\Big\} \\ &\geq \mathrm{E}\{X_n \,|\, X_1, \ldots, X_n\} = X_n. \end{aligned} \tag{7.20}$$

The last equality is a consequence of Theorem 7.6. □

Lemma 7.20 *If X is a martingale and ψ is convex, then $\psi(X)$ is a submartingale provided that $\psi(X_n) \in L^1(P)$ for all n. If X is a submartingale and ψ is a nondecreasing convex function, and if $\psi(X_n) \in L^1(\mathrm{P})$ for all n, then $\psi(X)$ is a submartingale.*

Proof By the conditional form of Jensen's inequality (Theorem 7.6), with probability one, $\mathrm{E}\{\psi(X_{n+1}) \,|\, \mathfrak{F}_n\} \geq \psi(\mathrm{E}\{X_{n+1} \,|\, \mathfrak{F}_n\})$. This holds for any process X and any convex function ψ as long as $\psi(X_n) \in L^1(\mathrm{P})$.

Now if X is a martingale, then $\psi(\mathrm{E}\{X_{n+1} \,|\, \mathfrak{F}_n\})$ equals $\psi(X_n)$, a.s., and this proves the submartingale property of $\psi(X)$. If in addition ψ is increasing but X is a submartingale, then by the submartingale property of X, and by the fact that ψ is nondecreasing, $\psi(\mathrm{E}\{X_{n+1} \,|\, \mathfrak{F}_n\}) \geq \psi(X_n)$, a.s. which is the desired result. □

Remark 7.21 In particular, whenever X is a martingale, X^+, $|X|^p$, and e^X are submartingales provided that they are integrable at each time n. If X is a submartingale, then X^+ and e^X are also submartingale, provided integrability. However, one can construct a submartingale whose absolute value is not a submartingale; e.g., consider $X_k := -\frac{1}{k}$.

Remark 7.22 An equivalent formulation of (iii) above is that for all k, n, $\mathrm{E}\{X_{n+k} \,|\, \mathfrak{F}_n\} \geq X_n$, a.s. To prove this, we use the towering property of conditional expectations (Theorem 7.9) as follows: Almost surely, $\mathrm{E}\{X_{n+k} \,|\, \mathfrak{F}_n\} = \mathrm{E}(\mathrm{E}\{X_{n+k} \,|\, \mathfrak{F}_{n+k-1}\} \,|\, \mathfrak{F}_n) = \mathrm{E}(X_{n+k-1} \,|\, \mathfrak{F}_n)$, and proceed by induction.

The definition of semimartingales is motivated by the following fact whose proof is at least as interesting as the fact itself:

Theorem 7.23 (Doob Decomposition) *Any submartingale X can be written as $X_n = Y_n + Z_n$ where Y is a martingale, and Z is a nonnegative adapted a.s.-increasing process with $Z_n \in L^1(\mathrm{P})$ for all n. In particular, sub- and supermartingales are semimartingales, and any semimartingale can be written as the difference of a sub- and a supermartingale.*

Proof We can write $X_n = X_1 + \sum_{j=2}^n d_j$ where $d_j := X_j - X_{j-1}$. This can be expanded as $X_n = Y_n + Z_n$, where $Y_1 := X_1$, for all $n \geq 2$, $Y_n := X_1 + \sum_{j=2}^n (d_j - \mathrm{E}\{d_j \,|\, \mathfrak{F}_{j-1}\})$, and $Z_n := \sum_{j=2}^n \mathrm{E}\{d_j \,|\, \mathfrak{F}_{j-1}\}$. Since for $n \geq 2$, $Y_n = Y_{n-1} + (d_n - \mathrm{E}\{d_n \,|\, \mathfrak{F}_{n-1}\})$, Y is a martingale. Moreover, the properties of Z follows from $\mathrm{E}\{d_j \,|\, \mathfrak{F}_{j-1}\} \geq 0$ which is another way to write the submartingale property. □

This is one of many decomposition theorems for semimartingales. Next is another one, due to Krickeberg [Kri63, Satz 33, p. 131] (see also the English translation [Kri65, Theorem 33, p. 144]). Before introducing it however, we need a brief definition.

Definition 7.24 A stochastic process $X_1, X_2, \ldots$ is said to be *bounded in* $L^p(\mathrm{P})$ if $\sup_n \|X_n\|_p < +\infty$.

Theorem 7.25 (Krickeberg Decomposition) *Suppose X_n is a submartingale that is also bounded in $L^1(\mathrm{P})$. Then we can write $X_n = Y_n - Z_n$ where Y_n is a martingale, and Z_n is a nonnegative supermartingale.*

Proof If $m \geq n$, then $\mathrm{E}\{X_m \,|\, \mathfrak{F}_n\} \leq \mathrm{E}\{X_{m+1} \,|\, \mathfrak{F}_n\}$; therefore, $Y_n := \lim_{m\to\infty} \mathrm{E}\{X_m \,|\, \mathfrak{F}_n\}$ exists a.s. as an increasing limit. Note that Y is an adapted process, and $Y_n \geq X_n$. Moreover, thanks to the monotone convergence theorem (Theorem 2.21), $\mathrm{E}\{Y_n\} = \lim_m \mathrm{E}\{X_m\} = \sup_m \mathrm{E}\{X_m\}$, which is finite since X is bounded in $L^1(\mathrm{P})$. Finally, by the towering property of conditional expectations (Theorem 7.9), and by the conditional form of the monotone convergence theorem (Theorem 7.6),

$$\begin{aligned} \mathrm{E}\{Y_{n+1} \,|\, \mathfrak{F}_n\} &= \lim_{m\to\infty} \mathrm{E}\left\{ \mathrm{E}(X_m \,|\, \mathfrak{F}_n) \,\middle|\, \mathfrak{F}_{n+1} \right\} \\ &= \lim_{m\to\infty} \mathrm{E}\{X_m \,|\, \mathfrak{F}_n\} = Y_n, \qquad \text{a.s.} \end{aligned} \tag{7.21}$$

This shows that Y_n is a martingale, and $Z_n := Y_n - X_n \geq 0$. On the other hand, it is not hard to see that Z_n is a supermartingale: Since Y_n is a martingale and X_n is a submartingale,

$$\begin{aligned} \mathrm{E}\{Z_{n+1} \mid \mathfrak{F}_n\} &= \mathrm{E}\{Y_{n+1} \mid \mathfrak{F}_n\} - \mathrm{E}\{X_{n+1} \mid \mathfrak{F}_n\} \\ &\leq Y_n - X_n = Z_n, \qquad \text{a.s.} \end{aligned} \tag{7.22}$$

This completes our proof. □

Remark 7.26 One of the implications of Doob's decomposition is that any submartingale X is bounded below by some martingale. The Krickeberg decomposition implies a powerful converse to this: Every L^1-bounded submartingale is also bounded above by a martingale.

Remark 7.27 Note that in the Krickeberg decomposition, the processes Y and Z are also bounded in $L^1(\mathrm{P})$. Here is a proof: $\mathrm{E}\{|Y_n|\} \leq \mathrm{E}\{|X_n|\} + \mathrm{E}\{Z_n\}$. The first term is bounded in n, so it suffices to show that $\mathrm{E}\{Z_n\}$ is bounded in n. But then $\mathrm{E}\{Z_n\} = \mathrm{E}\{Y_n\} - \mathrm{E}\{X_n\} \leq \mathrm{E}\{Y_n\} + \mathrm{E}\{|X_n|\}$. Since Y is a martingale, $\mathrm{E}\{Y_n\} = \mathrm{E}\{Y_1\} \leq \mathrm{E}\{|Y_1|\} < \infty$. This proves the L^1-boundedness of Y and Z both.

Thinking of the the index n of a process $X_1, X_2, \ldots$ as time, we will be interested in a certain family of random times that are introduced next.

Definition 7.28 A *stopping time* (with respect to the filtration $\mathfrak{F}$) is a random variable $T : \Omega \mapsto \mathbb{N} \cup \{\infty\}$ such that for any $k \in \mathbb{N}$, $\{T = k\} \in \mathfrak{F}_k$. This is equivalent to saying that for all $k \in \mathbb{N}$, $\{T \leq k\} \in \mathfrak{F}_k$.

You should think of $\mathfrak{F}_k$ as the total amount of information available by time k; e.g., if we know $\mathfrak{F}_k$, then we know whether or not any $A \in \mathfrak{F}_k$ has occurred by time k. With this in mind, the above can be interpreted as saying that T is a stopping time if and only if we only need to know the state of things by time k to measurably decide whether or not $T \leq k$.

Example 7.29 Nonrandom times are stopping times (check!). Next suppose X_n is a stochastic process that is adapted to a filtration $\mathfrak{F}_n$. If $A \in \mathfrak{B}(\mathbb{R})$, then $T(\omega) := \inf\{n \geq 1 : \; X_n(\omega) \in A\}$ is a stopping time provided that we define $\inf \varnothing := \infty$. Indeed for each $k \geq 2$,

$$\{T = k\} = \bigcap_{j=1}^{k-1} \{X_j \notin A\} \cap \{X_k \in A\}\,. \tag{7.23}$$

On the other hand, when $k = 1$, $\{T = 1\} = \{X_1 \in A\}$. In any event, we see that always $\{T = k\} \in \mathfrak{F}_k$. The random variable is the first time the process X enters the set A. Likewise, one shows that for any k, the kth time that X enters A is a stopping time. This example is generic in the following sense: Suppose T is a stopping time with respect to a filtration. Then there exists an adapted process X such that $T := \inf\{j \geq 1 : \ X_j = 1\}$, where $\inf \varnothing := \infty$. The receipt for X is simple; namely, $X_j := \mathbf{1}_{\{T=j\}}$. You should check the remaining details.

Remark 7.30 The previous example shows that the first time that a process enters a Borel set is a stopping time. However, not all random times are stopping times. For instance, consider $L(\omega) := \sup\{n \geq 1 : \ X_n(\omega) \in A\}$, where $\sup \varnothing := 0$, and A is a Borel set. Thus, L is the last time X enters A, and $\{L = k\}$ is the event that for all $j \geq k$, X_j is not in A; i.e., $\{L = k\} = \cap_{j=k}^{\infty} \{X_j \notin A\}$. This is in $\mathfrak{F}_k$ if and only if $X_k, X_{k+1}, \ldots$ are all Borel functions of $X_1, \ldots, X_k$; a property that does not generally hold. (For example, consider the case when the X_n's are independent.)

Lemma 7.31 *If $T_1, T_2, \ldots, T_n$ are a finite number of stopping times, then $T_1 + \cdots + T_n$, $\min_{1 \leq j \leq n} T_j$, and $\max_{1 \leq j \leq n} T_j$ are stopping times.*

Given a finite (or an a.s.-finite) stopping time T (with respect to a given underlying filtration of course), consider

$$\mathfrak{F}_T := \{A \in \mathfrak{F} : \ \forall k \geq 1, \ A \cap \{T \leq k\} \in \mathfrak{F}_k\}. \tag{7.24}$$

It should be recognized that the "T" in the notation $\mathfrak{F}_T$is meant only to remind us of the relation of the collection $\mathfrak{F}_T$ to the stopping time T. (It is not the case that $\mathfrak{F}_T$ is a function of $T(\omega)$, for instance.)

Lemma 7.32 *If $S \leq T$ (a.s.) are (a.s.-finite) stopping time, then $\mathfrak{F}_S \subseteq \mathfrak{F}_T$; moreover, $\mathfrak{F}_T$ is a σ-algebra. In addition, for any $Y \in L^1(\mathrm{P})$ and all $n \geq 1$, $\mathrm{E}\{Y \,|\, \mathfrak{F}_T\}\mathbf{1}_{\{T=n\}} = \mathrm{E}\{Y \,|\, \mathfrak{F}_n\}\mathbf{1}_{\{T=n\}}$, a.s. Finally, if X is adapted to $\mathfrak{F}$, then the random variable X_T is $\mathfrak{F}_T$-measurable, where $X_T(\omega)$ is defined to be $X_{T(\omega)}(\omega)$ for all $\omega \in \Omega$.*

The following theorem is due to J. L. Doob [Doo53, Doo71], while the present formulation is due to G. A. Hunt [Hun66]. It is our first important result on semimartingales.

Theorem 7.33 (Doob–Hunt Optional Stopping Theorem) *If S and T are a.s.-bounded stopping times such that $S \le T$ a.s., and if X is a submartingale, then with probability one, $\mathrm{E}\{X_T \,|\, \mathfrak{F}_S\} \ge X_S$. If X is a supermartingale, then this inequality holds but in the other direction; if X is martingale, then the inequality is replaced with equality.*

Remark 7.34 This result has the following interpretation in terms of a fair game. Suppose $X_1, X_2, \ldots$ are i.i.d. mean-zero random variables so that $S_n := X_1 + \cdots + X_n$ can be thought of as the reward (or loss) at time n in a fair game. In particular, for any n, $\mathrm{E}\{S_n\} = 0$, which means that one cannot expect to win at any nonrandom time n. The optional stopping theorem states that in fact there is no winning "previsible strategy" (i.e., one that does not depend on the "future" outcomes in order to predict what happens "next"). In other words, when playing fair games, there is no free lunch unless you are clairvoyant.

Proof It suffices to consider the submartingale case.

We can find a nonrandom $K > 0$ such that with probability one, $S \le T \le K$. Now the trick is to write things in terms of the "(sub-)martingale differences," $d_n := X_n - X_{n-1}$, where we can define $X_0 := 0$ in order to have compact notation. Equivalently, $X_n = \sum_{j=1}^n d_j$, and we can deduce from this that a.s., $X_T = \sum_{j=1}^K d_j \mathbf{1}_{\{j \le T\}}$, and a similar expression holds for X_S. Therefore, for all $A \in \mathfrak{F}_S$,

$$\begin{aligned}
\mathrm{E}\{X_T - X_S; A\} &= \sum_{j=1}^K \mathrm{E}\left[d_j \mathbf{1}_{\{S<j\le T\}\cap A}\right] \\
&= \sum_{j=1}^K \mathrm{E}\left[\mathrm{E}\left\{d_j \mathbf{1}_{\{S<j\le T\}\cap A} \,\middle|\, \mathfrak{F}_{j-1}\right\}\right] \qquad (7.25)\\
&= \sum_{j=1}^K \mathrm{E}\Big[\mathrm{E}\{d_j \,|\, \mathfrak{F}_{j-1}\} \mathbf{1}_{\{S<j\le T\}\cap A}\Big],
\end{aligned}$$

since $\{T \ge j\} = \{T \le j-1\}^{\complement} \in \mathfrak{F}_{j-1}$, and $\{S < j\} \cap A = \{S \le j-1\} \cap A \in \mathfrak{F}_{j-1}$ by the definition of $\mathfrak{F}_S$, so that $A \cap \{S < j \le T\} \in \mathfrak{F}_{j-1}$. By the definition of a submartingale, $\mathrm{E}\{d_j \,|\, \mathfrak{F}_{j-1}\} \ge 0$ almost surely. This implies that $\mathrm{E}\{X_T; A\} \ge \mathrm{E}\{X_S; A\}$, and this is equivalent to the desired result (why?).

Our next result follows immediately from the preceding one, but is an important fact that deserves special mention.

Corollary 7.35 *If T is a stopping time with respect to a filtration $\mathfrak{F}$, and if X is a submartingale (respectively, supermartingale or martingale) with respect to $\mathfrak{F}$, then $n \mapsto X_{T\wedge n}$ is a submartingale (respectively, supermartingale or martingale) with respect to $n \mapsto \mathfrak{F}_{T\wedge n}$.*

3 Gambler's Ruin Formula

We are ready to take a side-tour, and use the optional stopping theorem to have a starting look at random walks.

A *random walk* is simply the process that is obtained by successively summing i.i.d. random variables (in any dimension). In symbols, we have the following:

Definition 7.36 If $X_1, X_2, \ldots$ are i.i.d. random variables in $\mathbb{R}^m$, then the process $n \mapsto S_n$ is a *random walk* (in m dimensions) where $S_n := X_1 + \cdots + X_n$.

In other words, after centering, every L^1-random walk is a martingale. Of course, since martingales are defined to be one-dimensional processes here, our next result too is one-dimensional.

Lemma 7.37 *If $S_n := X_1 + \cdots + X_n$ defines a random walk in one dimension, and if $X_1 \in L^1(\mathrm{P})$ has mean $\mu := \mathrm{E}(X_1)$, then $n \mapsto S_n - n\mu$ is a mean-zero martingale. If in addition $X_1 \in L^2(\mathrm{P})$, $\mu = 0$, and $\sigma^2 := \mathrm{Var}(X_1)$, then $n \mapsto S_n^2 - n\sigma^2$ is a mean-zero martingale.*

Nearest-neighborhood walks are a very natural class of random walks that are defined as follows:

Definition 7.38 A random walk $S_n := X_1 + \cdots + X_n$ is called a *nearest-neighborhood walk* if with probability one, $X_1 \in \{-1, +1\}$; i.e., if at all times $n = 1, 2, \ldots$, we have $S_n = S_{n-1} \pm 1$ almost surely.

In other words, S_n is a nearest-neighborhood walk if there exists $p \in [0, 1]$ such that $\mathrm{P}\{X_1 = 1\} = p = 1 - \mathrm{P}\{X_1 = -1\}$. The case $p = \frac{1}{2}$ is particularly special and has its own name.

Definition 7.39 When $\mathrm{P}\{X_1 = 1\} = \mathrm{P}\{X_1 = -1\} = \frac{1}{2}$, S_n is called the *simple walk*.

We can think of a nearest-neighborhood walk S_n as the amount of money won (lost if negative) in n independent plays of a games, where in each play one wins and loses a dollar with probabilities p and $1-p$ respectively. Then the simple walk corresponds to the fortune-process of the gambler in the case that the game is fair.

Suppose that the gambler is playing against the house, there is a maximum house-limit of h dollars, and the gambler's resources amount to a total of g dollars. Then consider the first time that either the house or the gambler stops the play; i.e.,

$$T := \inf\left\{ j \geq 1 : \; S_j = -g \text{ or } h \right\}, \tag{7.26}$$

where $T(\omega) = \inf \varnothing := +\infty$ amounts to the statement that for the particular realization ω of the game, it is played indefinitely.

Lemma 7.40 *With probability one, $T < +\infty$.*

Proof I will prove this first in the case $p = \frac{1}{2}$. In this case, $n^{-1/2}S_n$ converges weakly to a standard normal (Theorem 6.22; see also equation 6.2). In particular, for all $\lambda > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left\{ \frac{S_n}{\sqrt{n}} \geq \lambda \right\} = \int_\lambda^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx, \tag{7.27}$$

which is positive. On the other hand, by the right-continuity of measures,

$$\begin{aligned} \mathrm{P}\left\{ \limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} \geq \lambda \right\} &= \mathrm{P}\left(\bigcap_{m=1}^\infty \left\{ \sup_{n\geq m} \frac{S_m}{\sqrt{m}} \geq \lambda \right\} \right) \\ &= \lim_{m\to\infty} \mathrm{P}\left\{ \sup_{n\geq m} \frac{S_m}{\sqrt{m}} \geq \lambda \right\}. \end{aligned} \tag{7.28}$$

Now if $m^{-1/2}S_m \geq \lambda$, then certainly $\sup_{n\geq m} n^{-1/2}S_n \geq \lambda$. In other words, the numerical value of the above display is greater than the display preceding it. In particular, for all $\lambda > 0$,

$$\mathrm{P}\left\{ \limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} \geq \lambda \right\} \geq \int_\lambda^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx > 0. \tag{7.29}$$

On the other hand, thanks to the Kolmogorov zero-one law (Theorem 5.15), $c := \limsup_n n^{-1/2} S_n$ is a.s. a constant; thus, $c = +\infty$. (This is greatly refined by the law of the iterated logarithm; cf. Theorem 7.48 below). Therefore, when $p = \frac{1}{2}$ there must exist a random sequence n_k along which $S_{n_k} \to \infty$. Since $|S_n - S_{n-1}| = 1$, this shows that $S_M = h$ for some random M. In particular, $T \leq M < \infty$, a.s.

When $p \neq \frac{1}{2}$, we appeal to the strong law of large numbers to see that a.s., $\lim_{n\to\infty} n^{-1} S_n = 2p - 1$. In the case $p < \frac{1}{2}$, we have shown that $S_n \to -\infty$ a.s., and when $p > \frac{1}{2}$, then $S_n \to \infty$. In either case, S_n must cross g or $-h$ at some random but finite time. This concludes the proof. □

The "gambler's ruin" problem asks for the probability that the gambler is ruined in terms of the parameter p; i.e.,

$$\mathrm{P}_{\mathsf{ruin}}(p) := \mathrm{P}\{S_T = -g\}. \tag{7.30}$$

The following is our first important application of the optional stopping theorem (Corollary 7.35), and gives a formula for the ruin probability. Surprisingly, the form of $\mathrm{P}_{\mathsf{ruin}}$ depends on the probability $p = \mathrm{P}\{X_1 = 1\}$.

Theorem 7.41 (Gambler's Ruin Formula) $\mathrm{P}_{\mathsf{ruin}}(\frac{1}{2}) = h \div (g + h)$. *On the other hand, if* $p \neq \frac{1}{2}$, *then* $\mathrm{P}_{\mathsf{ruin}}(p) = (\zeta^{h+g} - \zeta^g) \div (\zeta^{h+g} - 1)$, *with* $\zeta := (1-p) \div p$.

Remark 7.42 Note that $\lim_{p\to\frac{1}{2}} \mathrm{P}_{\mathsf{ruin}}(p) = \mathrm{P}_{\mathsf{ruin}}(\frac{1}{2})$.

Proof I will begin with the case $p = \frac{1}{2}$: By the optional stopping theorem, and by Lemma 7.37, $n \mapsto S_{T\wedge n}$ is a mean-zero martingale so that $\mathrm{E}\{S_{T\wedge n}\} = 0$ for all n. It is also a.s. bounded since $\sup_n |S_{T\wedge n}| \leq \max(g, h) < \infty$. Therefore, by the dominated convergence theorem (Theorem 2.22), and using the fact that a.s. $T < \infty$ (Lemma 7.40), we deduce that $\mathrm{E}\{S_T\} = \mathrm{E}\{\lim_n S_{T\wedge n}\} = \lim_n = \mathrm{E}\{S_{T\wedge n}\} = 0$. But $\omega \mapsto S_T(\omega)$ is a simple function so that $0 = \mathrm{E}\{S_T\} = -g \times \mathrm{P}_{\mathsf{ruin}}(\frac{1}{2}) + h \times (1 - \mathrm{P}_{\mathsf{ruin}}(\frac{1}{2}))$. Solve to obtain the expression for the ruin probability in the case $p = \frac{1}{2}$.

When $p \neq \frac{1}{2}$, we have to only find a suitable bounded martingale and then follow the preceding argument. But it is not hard to check that the following is one such candidate:

$$\zeta^{S_n} \text{ is a bounded mean-one martingale, where } \zeta := \frac{1-p}{p}. \tag{7.31}$$

A similar reasoning as the one used in the $p = \frac{1}{2}$ case shows that $\mathrm{E}\{\zeta^{S_T}\} = 1$, and the result follows readily from this. □

4 Doob's Maximal Inequality, and Pointwise Convergence

We now use the optional stopping theorem to prove our second maximal inequality of these notes.

Theorem 7.43 (Doob's Inequality; [Doo40]) *If $X_1, X_2, X_3, \ldots$ is a submartingale, then for all $\lambda > 0$ and $n \geq 1$,*

$$\begin{aligned} \mathrm{P}\left\{\max_{1\le j\le n} X_j \geq \lambda\right\} &\leq \frac{1}{\lambda}\mathrm{E}\left\{X_n; \max_{1\le j\le n} X_j \geq \lambda\right\} \leq \frac{1}{\lambda}\mathrm{E}\{X_n^+\}, \\ \mathrm{P}\left\{\min_{1\le j\le n} X_j \leq -\lambda\right\} &\leq \frac{1}{\lambda}\left(\mathrm{E}\{X_n^+\} - \mathrm{E}\{X_1\}\right), \qquad (7.32) \\ \mathrm{P}\left\{\max_{1\le j\le n} |X_j| \geq \lambda\right\} &\leq \frac{1}{\lambda}\left(2\mathrm{E}\{|X_n|\} - \mathrm{E}\{X_1\}\right). \end{aligned}$$

Remark 7.44 The last assertion can be improved if X is a martingale, for then $|X|$ is a submartingale (Lemma 7.20). Hence, we apply the first equation in (7.32) to get $\mathrm{P}\{\max_{j\geq n}|X_j| \geq \lambda\} \leq \lambda^{-1}\mathrm{E}\{|X_n|\}$.

It should also be recognized that Theorem 7.43 contains an extension of Kolmogorov's L^2-maximal inequality (Theorem 5.24). To see this let $\xi_1, \xi_2, \ldots$ denote a sequence of mean-zero i.i.d. random variables with $\xi_1 \in L^p(\mathrm{P})$ for some $p > 1$, and consider the random walk $S_n := \xi_1 + \cdots + \xi_n$.

In light of Lemma 7.37, $n \mapsto S_n$ is a mean-zero martingale, and so $n \mapsto |S_n|^p$ is a nonnegative supermartingale for any $p \geq 1$ (Lemma 7.20). To this we apply Theorem 7.43 with $X_n := |S_n|^p$, and deduce that for all $\lambda > 0$ and $n \geq 1$,

$$\mathrm{P}\left\{\max_{1\le j\le n} |S_j| \geq \lambda\right\} \leq \frac{\|S_n\|_p^p}{\lambda^p}. \qquad (7.33)$$

In particular, if $p = 2$ then this and the easy fact that $\|S_n\|_2^2 = n\|X_1\|_2^2$ (Lemma 5.9) together yield Kolmogorov's L^2-maximal inequality.

More generally still, if X is a submartingale, then so is $|X|^p$, for any $p \geq 1$ (Lemma 7.20). Consequently, for all $n \geq 1$ and $\lambda > 0$, $\mathrm{P}\{\max_{j\leq n} X_j \geq \lambda\} \leq \lambda^{-p}\mathrm{E}\{X_n; \max_{j\leq n} X_j^p \geq \lambda\} \leq \lambda^{-p}\mathrm{E}\{X_n^p\}$.

Proof of Theorem 7.43 If we let $\inf\varnothing := +\infty$, then T is a stopping time where $T := \inf\{j \geq 1 : X_j \geq \lambda\}$. Furthermore, $\{T \leq n\} = \{\max_{j\leq n} X_j \geq \lambda\}$, and by the submartingale property,

$$\begin{aligned}\mathrm{E}\{X_n^+\} &\geq \mathrm{E}\left\{X_n^+;\ T \leq n\right\} \geq \mathrm{E}\left\{X_n;\ T \leq n\right\}\\ &= \sum_{j=1}^n \mathrm{E}\left\{X_n;\ T = j\right\} \geq \sum_{j=1}^n \mathrm{E}\left\{X_j;\ T = j\right\}.\end{aligned} \tag{7.34}$$

I have used the fact that $\{T = j\} \in \mathfrak{F}_j$. On the other hand, whenever $T(\omega) = j$, then $X_j(\omega) = X_{T(\omega)}(\omega) \geq \lambda$. Thus, $\sum_{j\leq n} \mathrm{E}\{X_j;\ T = j\} \geq \lambda\mathrm{P}\{T \leq n\}$, which proves the first part of (7.32).

To prove the second portion of (7.32), let $\tau := \inf\{1 \leq j \leq n :\ X_j \leq -\lambda\}$ where $\inf\varnothing := \infty$. By the optional stopping theorem (Theorem 7.33), $\mathrm{E}\{X_1\} \leq \mathrm{E}\{X_{\tau\wedge n}\}$. Since $X_\tau \leq -\lambda$ on $\{\tau < +\infty\}$, we have

$$\begin{aligned}\mathrm{E}\{X_1\} \leq \mathrm{E}\{X_{\tau\wedge n}\} &= \mathrm{E}\{X_\tau; \tau \leq n\} + \mathrm{E}\{X_n; \tau > n\}\\ &\leq -\lambda\mathrm{P}\{\tau \leq n\} + \mathrm{E}\{X_n^+\}.\end{aligned} \tag{7.35}$$

Since $\{\tau \leq n\} = \{\min_{j\leq n} X_j \leq -\lambda\}$, this prove the second portion of (7.32). Adding the two parts of (7.32) yields $\lambda\mathrm{P}\{\max_{j\leq n} |X_j| \geq \lambda\} \leq 2\mathrm{E}\{X_n^+\} - \mathrm{E}\{X_1\} \leq 2\mathrm{E}\{|X_n|\} - \mathrm{E}\{X_1\}$. □

This implies the following fundamental theorem of J. L. Doob; it is known as the *martingale convergence theorem*:

Theorem 7.45 ([Doo40]) *A submartingale X converges a.s. if either (i) X is bounded in $L^1(\mathrm{P})$; or (ii) X is nonpositive a.s. In either case, the limiting random variable* $\lim_n X_n$ *is finite a.s.*

Proof (Isaac [Isa65]) As for the Kolmogorov strong law (Theorem 5.21), I first prove things in the L^2-case. Then I truncate down to $L^1(\mathrm{P})$. With this in mind, the proof is divided into four easy steps.

Step 1. The Nonnegative L^2-Bounded Case.
The easiest case is when X is nonnegative and bounded in $L^2(\mathrm{P})$. In this

case, note that for all $n, k \geq 1$,

$$\begin{aligned}\|X_{n+k} - X_n\|_2^2 &= \|X_{n+k}\|_2^2 + \|X_n\|_2^2 - 2\mathrm{E}\{X_{n+k}X_n\} \\ &= \|X_{n+k}\|_2^2 + \|X_n\|_2^2 - 2\mathrm{E}\left\{\mathrm{E}(X_{n+k}\,|\,\mathfrak{F}_n)X_n\right\} \\ &\leq \|X_{n+k}\|_2^2 - \|X_n\|_2^2.\end{aligned} \tag{7.36}$$

On the other hand, by Lemma 7.20, X^2 is a submartingale since $X_n \geq 0$ for all n. Therefore, $\|X_n\|_2 \uparrow \sup_m \|X_m\|_2$ as n increases without bound. This and the preceding display together show that $\{X_n\}$ is a Cauchy sequence in $L^2(\mathrm{P})$, and so it converges in $L^2(\mathrm{P})$. Let X_∞ to be the $L^2(\mathrm{P})$-limit of X_n, and find $n_k \uparrow \infty$ such that $\|X_\infty - X_{n_k}\|_2 \leq 2^{-k}$. By Chebyshev's inequality (Corollary 2.16), for all $\varepsilon > 0$, $\sum_k \mathrm{P}\{|X_\infty - X_{n_k}| \geq \varepsilon\} \leq \varepsilon^{-2}\sum_k 4^{-k} < +\infty$. Thus, by the Borel–Cantelli lemma, $\lim_{k\to\infty} X_{n_k} = X_\infty$, a.s. On the other hand, $\|X_{n_{k+1}} - X_{n_k}\|_1 \leq \|X_{n_{k+1}} - X_{n_k}\|_2 \leq \|X_\infty - X_{n_k}\|_2 + \|X_\infty - X_{n_{k+1}}\|_2 \leq 2^{-k} + 2^{-k+1} = 3\cdot 2^{-k}$. Therefore, Theorem 7.43 shows us that for any $\varepsilon > 0$,

$$\begin{aligned}\sum_k \mathrm{P}\left\{\max_{n_k \leq j \leq n_{k+1}} |X_j - X_{n_k}| \geq \varepsilon\right\} &\leq \frac{2}{\varepsilon}\sum_{k=1}^\infty \|X_{n_{k+1}} - X_{n_k}\|_1 \\ &\leq \frac{6}{\varepsilon}\sum_{k=1}^\infty 2^{-k} < +\infty.\end{aligned} \tag{7.37}$$

I have used the fact that $\{X_{n+j} - X_n; j \geq 0\}$ is a submartingale for each fixed n with respect to the filtration $\{\mathfrak{F}_{j+n}; j \geq 0\}$, and that this submartingale starts at 0. Therefore, by the Borel–Cantelli lemma,

$$\lim_{k\to\infty} \max_{n_k \leq j \leq n_{k+1}} |X_j - X_{n_k}| = 0, \qquad \text{a.s.} \tag{7.38}$$

Since $X_{n_k} \to X$ a.s., this shows that $\lim_{m\to\infty} X_m = X_\infty$ a.s. Since $X_\infty \in L^2(\mathrm{P})$, it is a.s. finite.

Step 2. The Nonpositive Case.

If $X_n \leq 0$ is a submartingale, then e^{X_n} is a bounded nonnegative submartingale (Lemma 7.20). Thanks to Step 1, $\lim_n e^{X_n}$ exists and is finite a.s.

Step 3. The Nonnegative L^1-Bounded Case.

If X_n is a nonnegative submartingale that is bounded in $L^1(\mathrm{P})$, then thanks to the Krickeberg decomposition (Theorem 7.25), we can write $X_n = Y_n - Z_n$ where Y_n is a nonnegative martingale, and Z_n is a nonnegative supermartingale. By Step 2, $\lim_n Y_n$ and $\lim_n Z_n$ exist and are finite a.s. This shows that $\lim_n X_n$ exists and is finite a.s.

Step 4. The L^1-Bounded Case.
If X_n is an L^1-bounded submartingale, then we can write it as $Y_n^+ - Y_n^- - Z_n$, where Y is a martingale, and Z is a nonnegative supermartingale. On the other hand, Y^+ and Y^- are nonnegative submartingales (Lemma 7.20). Thus, thanks to Steps 2 and 3, $\lim_n Y_n^+$, $\lim_n Y_n^-$, and $\lim_n Z_n$ all exists and are finite a.s. This completes the proof. □

5 Four Applications

Martingale theory provides us with a powerful set of analytical tools and, as such, it is not surprising that it has made an impact on a tremendous number of diverse mathematical problems. In keeping with the unwritten tradition of these notes, I will mention a few applications. More examples can be found in the exercises, as well as in the general bibliography at the end of these notes.

5.1 Kolmogorov's Strong Law

For our first application of martingale theory, I present the martingale proof of the Kolmogorov strong law of large numbers (Theorem 5.21) that is due to Doob [Doo49].

Let $X_1, X_2, \ldots$ be i.i.d. real random variables in $L^1(\mathrm{P})$, and define $S_n := X_1 + \cdots + X_n$. Recall that in this case, $\lim_n n^{-1} S_n = \mathrm{E}\{X_1\}$, a.s.

Let $\mathfrak{F}_n$ denote the σ-algebra generated by $\{S_n, S_{n+1}, S_{n+2}, \ldots\}$, and note that the time-reversed $\mathfrak{F}_n$'s are a filtration; i.e., $\mathfrak{F}_1 \supseteq \mathfrak{F}_2 \supseteq \cdots$. The following states that $n^{-1}S_n$ is a time-reversed martingale with respect to this time-reversed filtration.

Lemma 7.46 (Doob [Doo49]) *For all $n \geq 1$, $\mathrm{E}\{X_1 \,|\, \mathfrak{F}_n\} = \frac{1}{n} S_n$, a.s.*

Proof I first show that

$$\mathrm{E}\{X_1 \,|\, S_n\} = \frac{1}{n} S_n, \qquad \text{a.s.} \tag{7.39}$$

To do this I prove that

$$\mathrm{E}\{X_k \,|\, S_n\} = \mathrm{E}\{X_1 \,|\, S_n\}, \qquad \text{a.s., } \forall k = 1, \ldots, n. \tag{7.40}$$

This implies (7.39), since we can sum the above from $k = 1$ to n to deduce that almost surely,

$$n\mathrm{E}\{X_1 \,|\, S_n\} = \mathrm{E}\left\{ \sum_{k=1}^{n} X_k \,\middle|\, S_n \right\} = \mathrm{E}\{S_n \,|\, S_n\} = S_n. \tag{7.41}$$

Thus, it suffices to prove (7.40). This follows from an "exchangeability argument." Namely, that the distribution of the random variable $(X_1, \ldots, X_n)$ is the same as that of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ for any permutation π of $\{1, \ldots, n\}$.[7.2] The said "exchangeability" implies that for any bounded measurable function g, and for all $k \leq n$, $\mathrm{E}\{X_1 g(S_n)\} = \mathrm{E}\{X_k g(S_n)\}$ (why?). Equation (7.40), and hence (7.39), follows immediately from this and the definition of conditional expectations. We conclude our proof by showing that $\mathrm{E}\{X_1 \,|\, \mathfrak{F}_n\} = \mathrm{E}\{X_1 \,|\, S_n\}$, a.s. Since $\mathfrak{F}_n = \sigma\{S_n, X_{n+1}, X_{n+2}, \ldots\}$ (why?), it suffices to show that for all finite integers k, and all $B_0, B_1, \ldots, B_k \in \mathfrak{B}(\mathbb{R})$,

$$\begin{aligned} &\mathrm{E}\left\{ \mathrm{E}(X_1 \,|\, \mathfrak{F}_n) \prod_{\ell=0}^{k} \mathbf{1}_{\{S_n \in B_0, X_{n+1} \in B_1, \ldots, X_{n+k} \in B_k\}} \right\} \\ &\quad = \mathrm{E}\left\{ \mathrm{E}(X_1 \,|\, S_n) \prod_{\ell=0}^{k} \mathbf{1}_{\{S_n \in B_0, X_{n+1} \in B_1, \ldots, X_{n+k} \in B_k\}} \right\}. \end{aligned} \tag{7.42}$$

Indeed, this and a monotone class argument together show that for all $A \in \mathfrak{F}_n$, $\mathrm{E}\{\mathrm{E}(X_1 \,|\, \mathfrak{F}_n); A\} = \mathrm{E}\{\mathrm{E}(X_1 \,|\, S_n); A\}$; i.e., $\mathrm{E}\{X_1 \,|\, \mathfrak{F}_n\} = \mathrm{E}\{X_1 \,|\, S_n\}$, a.s. This would then complete our proof.

Since the product in (7.42) is $\mathfrak{F}_n$-measurable. Moreover, the left-hand side is equal to

$$\begin{aligned} &\mathrm{E}\left\{ X_1 \prod_{\ell=0}^{k} \mathbf{1}_{\{S_n \in B_0, X_{n+1} \in B_1, \ldots, X_{n+k} \in B_k\}} \right\} \\ &\quad = \mathrm{E}\left\{X_1; S_n \in B_0\right\} \times \mathrm{E}\left\{ \prod_{\ell=0}^{k} \mathbf{1}_{\{X_{n+1} \in B_1, \ldots, X_{n+k} \in B_k\}} \right\}, \end{aligned} \tag{7.43}$$

[7.2] For example, you will need to show that for all bounded measurable functions ϕ of two variables, $\mathrm{E}\{\phi(X_1, X_2)\} = \mathrm{E}\{\phi(X_2, X_1)\}$. This is essentially obvious if $\phi(x, y) = \phi_1(x)\phi_2(y)$, since $\mathrm{E}\{\phi_1(X_1)\phi_2(X_2)\} = \int \phi_1 \, d\mu \cdot \int \phi_2 \, d\mu$, where μ is the distribution of X_1 (equivalently, that of X_2). Then proceed by appealing to a monotone-class argument.

thanks to the independence of $\{X_1, \cdots, X_n\}$ and $\sigma\{X_{n+1}, X_{n+2}, \ldots\}$. A similar argument shows that the right-hand side of (7.42) is equal to the same number (check this!). This concludes our proof. □

Kolmogorov's strong law of large numbers follows at once from the previous lemma used in conjunction with the Kolmogorov zero-one law (Theorem 5.15), and the following convergence theorem for Doob-type time-reversed martingales. [Recall that a Doob martingale is of the form $\mathrm{E}\{f \,|\, \mathfrak{A}_n\}$ where $f \in L^1(\mathrm{P})$, and $\mathfrak{A}_n$ is a filtration.]

Theorem 7.47 (Doob [Doo40]) *If $Y \in L^1(\mathrm{P})$ and $\mathfrak{F}_n$ is any time-reversed filtration, then $Z_n := \mathrm{E}\{Y \,|\, \mathfrak{F}_n\}$ converges almost surely and in $L^1(\mathrm{P})$ to $\mathrm{E}\{Y \,|\, \mathfrak{F}_\infty\}$ where $\mathfrak{F}_\infty := \cap_n \mathfrak{F}_n$.*

Proof First of all, we note that since $\mathfrak{F}_n \subseteq \mathfrak{F}_{n-1}$, the process Z is a time-reversed Doob-type martingale in the sense that Z_n is $\mathfrak{F}_n$-measurable, and $\mathrm{E}\{Z_{n-1} \,|\, \mathfrak{F}_n\} = Z_n$, a.s. In particular, the process $\{Z_{-n}; n = 1, 2, \cdots\}$ is an L^1-bounded martingale indexed by $-\mathbb{N} := \{-1, -2, \cdots\}$. Thanks to Theorem 7.45, $Z_\infty := \lim_n Z_n$ exists and is finite a.s. This defines $Z_\infty(\omega)$ for almost all ω. We can extend Z_∞ to a function defined for all ω in a number of ways. The quickest way is to redefine $Z_\infty(\omega) := \limsup_n Z_n(\omega)$ for all $\omega \in \Omega$. In this way, Z_∞ is $\mathfrak{F}_\infty$-measurable.

Next we prove L^1-convergence. If the random variable Y is bounded a.s., then the L^1-convergence of Z_n to Z_∞ follows from the bounded convergence theorem (Theorem 2.19).

If Y is not bounded, then for each $\varepsilon > 0$ we can find a bounded Y^ε such that $\|Y - Y^\varepsilon\|_1 \leq \varepsilon$. Let $Z_n^\varepsilon := \mathrm{E}\{Y^\varepsilon \,|\, \mathfrak{F}_n\}$ and note that $Z_\infty^\varepsilon := \lim_n Z_n^\varepsilon$ exists a.s. and in $L^1(\mathrm{P})$. On the other hand,

$$\|Z_n - Z_\infty\|_1 \leq \|Z_n - Z_n^\varepsilon\|_1 + \|Z_n^\varepsilon - Z_\infty^\varepsilon\|_1 + \|Z_\infty^\varepsilon - Z_\infty\|_1. \tag{7.44}$$

We estimate each term separately. As for the first term, $\|Z_n - Z_n^\varepsilon\|_1 = \mathrm{E}\{|\mathrm{E}(Y - Y^\varepsilon \,|\, \mathfrak{F}_n)|\}$. Thanks to the conditional Jensen's inequality (Theorem 7.6), $\|Z_n - Z_n^\varepsilon\|_1 \leq \mathrm{E}\{\mathrm{E}(|Y - Y^\varepsilon| \,|\, \mathfrak{F}_n)\}$, which equals $\mathrm{E}\{|Y - Y^\varepsilon|\} = \|Y - Y^\varepsilon\|_1 \leq \varepsilon$ thanks to (i) of Theorem 7.6. This inequality can be used in conjunction with Fatou's lemma (Theorem 2.20) to show that $\|Z_\infty - Z_\infty^\varepsilon\|_1 \leq \liminf_{n\to\infty} \|Z_n - Z_n^\varepsilon\|_1 \leq \varepsilon$. Thus, we have shown that $\limsup_n \|Z_n - Z_\infty\|_1 \leq 2\varepsilon$ for all $\varepsilon > 0$. This proves the L^1-convergence.

In order to show that $Z_\infty = \mathrm{E}\{Y \,|\, \mathfrak{F}_\infty\}$ a.s., we can note that for all $A \in \mathfrak{F}_\infty$, $\mathrm{E}\{Y; A\} = \mathrm{E}\{Z_n; A\}$. This follows because $\mathfrak{F}_n$ contains $\mathfrak{F}_\infty$. Then the L^1-convergence of Z_n to Z_∞ proves that $\mathrm{E}\{Y; A\} = \mathrm{E}\{Z_\infty; A\}$ for all $A \in \mathfrak{F}_\infty$. This verifies that $Z_\infty = \mathrm{E}\{Y \,|\, \mathfrak{F}_\infty\}$ a.s., since Z_∞ is $\mathfrak{F}_\infty$-measurable. □

5.2 The Khintchine LIL

Suppose $X_1, X_2, \ldots$ are i.i.d. random variables taking the value ±1 with probability $\frac{1}{2}$ each. Then, by the Kolmogorov strong law (Theorem 5.21), $\lim_{n\to\infty} \frac{1}{n} S_n = 0$ where $S_n := X_1 + \cdots + X_n$ is the partial sum process based on the X_n's. This problem was heavily popularized in the context of the normal number theorem of Borel [Bor09]; cf. Exercise 5.11.

One might then ask about the correct asymptotic size of S_n. An application of the central limit theorem shows that as $n \to \infty$, $n^{-1/2} S_n$ converges weakly to a standard Gaussian (i.e., where the parameters μ and σ^2 are 0 and 1, respectively). However, this is not saying much about the random variables S_n so much as their distributions. Hausdorff [?] proved that $n^{1/2}$ is roughly the correct order of magnitude; he did this by showing that for all $\rho < \frac{1}{2}$, $S_n = o(n^{-\rho})$ almost surely. This was refined by Hardy and Littlewood [?] who showed that $|S_n| = O((n \ln n)^{-1/2})$ almost surely, and later by Khintchine [?] who showed that $|S_n| = O((n \ln\ln n)^{-1/2})$ almost surely.[7.3]Finally, Khintchine [Khi24] showed the following *law of the iterated logarithm* (LIL):

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \ln\ln n}} = -\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \ln\ln n}} = 1, \quad \text{a.s.} \tag{7.45}$$

Two decades later, Hartman and Wintner [HW41] derived a highly nontrivial extension of this LIL and settled an old conjecture of A. N. Kolmogorov by proving that Khintchine's LIL holds quite generally. Namely, they proved the following result.

Theorem 7.48 (The LIL; Hartman and Wintner [HW41]) *If $X_1, X_2, \ldots$ are i.i.d. random variables in $L^2(\mathrm{P})$, then*

$$\limsup_{n\to\infty} \frac{S_n - \mathrm{E}\{S_n\}}{\sqrt{2n \ln\ln n}} = \|X_1\|_2, \quad a.s. \tag{7.46}$$

[7.3] I have used the following "little-o/big-O" notation invented in 1894 by the number-theorist P. Bachmann: $s_n = o(a_n)$ iff $\lim_n a_n^{-1} s_n = 0$; $s_n = O(a_n)$ iff $\limsup_n a_n^{-1} s_n < +\infty$.

At this level of generality, there are no simple-to-describe proofs of this fact. However, the LIL is not exceedingly difficult to understand in the case that the X's are Gaussian random variables.

Proof of Theorem 7.48 for Gaussian Increments Without loss of generality, we may assume that $\mathrm{E}\{X_1\} = 0$ and $\mathrm{E}\{X_1^2\} = 1$, for otherwise we can consider the random variables $X_\ell^\star := (X_\ell - \mathrm{E}\{X_\ell\}) \div \|X_\ell\|_2$ (why?) together with their partial sums $S_n^\star := X_1^\star + \cdots + X_n^\star$. In other words, it suffices to assume that X_i's are standard Gaussians. With this in mind, the proof is divided into a few easy steps. For simplicity, I will write $\Lambda := \lim_m S_m \div \sqrt{2m \ln\ln m}$; thanks to the Kolmogorov 0–1 law, Λ is a.s. a constant. So out task is to show that $\Lambda = 1$.

Step 1. A Large Deviations Estimate.
Keeping in mind that $X_1, X_2, \ldots$ are i.i.d. standard Gaussians, we wish to show in this first step that for all integers $n \geq 1$ and all real numbers $t > 0$,

$$\mathrm{P}\left\{\max_{1\leq j\leq n} S_j \geq nt\right\} \leq e^{-\frac{1}{2}nt^2}. \tag{7.47}$$

We start by reciting without proof a calculation that you should check yourself: For all $t > 0$,

$$\mathrm{E}\left\{e^{tX_1}\right\} = e^{\frac{1}{2}t^2}. \tag{7.48}$$

Consequently, if $\max_{j\leq n} S_j$ is replaced by S_n, (7.47) follows from Exercise 2.5.

Next let us fix a $t > 0$ and define

$$M_n := \exp\left(tS_n - \frac{t^2 n}{2}\right). \tag{7.49}$$

Let $\mathfrak{F}_n$ denote the filtration generated by $X_1, \ldots, X_n$, and note that $M_{n+1} = M_n \exp(tX_{n+1} - \frac{1}{2}t^2)$. Thus,

$$\begin{aligned}\mathrm{E}\left\{M_{n+1}\,\middle|\,\mathfrak{F}_n\right\} &= M_n \mathrm{E}\left\{\exp\left(tX_{n+1} - \frac{t^2}{2}\right)\,\middle|\,\mathfrak{F}_n\right\}\\ &= M_n \mathrm{E}\left\{\exp\left(tX_{n+1} - \frac{t^2}{2}\right)\right\}, \quad \text{a.s.}\end{aligned} \tag{7.50}$$

The first equality follows from the fact that M_n is $\mathfrak{F}_n$-measurable, and the second from the fact that X_{n+1} is independent of $\mathfrak{F}_n$. This and (7.48) together show that M_n is a nonnegative mean-one martingale. Moreover,

$$\begin{aligned} \mathrm{P}\left\{\max_{1\le j\le n} S_j \ge nt\right\} &\le \mathrm{P}\left\{\exists j \le n:\ tS_j - \frac{jt^2}{2} \ge \frac{nt^2}{2}\right\} \\ &= \mathrm{P}\left\{\max_{1\le j\le n} M_j \ge e^{\frac{1}{2}nt^2}\right\}. \end{aligned} \tag{7.51}$$

Doob's maximal inequality (Theorem 7.43) then implies (7.47).

Step 2. The Upper Bound.

Let $\theta > 1$ be fixed, and define $\theta_k := \lfloor \theta^k \rfloor$ $(k = 1, 2, \ldots)$. We can apply (7.47) with $n := \theta_k$ and $t := \sqrt{2cn^{-1}\ln\ln n}$ to see that for all $c > 1$ and all K sufficiently large,

$$\sum_{k\ge K} \mathrm{P}\left\{\max_{1\le j\le \theta_k} S_j \ge \sqrt{2c\theta_k \ln\ln\theta_k}\right\} \le \sum_{k\ge K} (\ln\theta_k)^{-c} < +\infty. \tag{7.52}$$

Thanks to the Borel–Cantelli lemma (Theorem 5.23), with probability one there exists a random variable k_0 such that for all $k \ge k_0$, $\max_{j\le\theta_k} S_j \le \sqrt{2c\theta_k\ln\ln\theta_k}$. For all m larger than θ_{k_0}, we can find $k \ge k_0$ such that $\theta_k \le m \le \theta_{k+1}$. Thus,

$$S_m \le \max_{j\le\theta_{k+1}} S_j \le \sqrt{2c\theta_{k+1}\ln\ln\theta_{k+1}} \le \Theta_m\sqrt{2c\theta m\ln\ln m}, \tag{7.53}$$

where $\Theta_m \to 1$ as $m \to \infty$. In other words, $\Lambda \le \sqrt{c\theta}$ for all $c, \theta > 1$,which shows that $\Lambda \le 1$. This is one-half of the LIL in the case of standard Gaussian increments. Moreover, by also applying the above to the Gaussian partial-sum process $(-S_n)$, we obtain

$$\limsup_{m\to\infty} \frac{|S_m|}{\sqrt{2m\ln\ln m}} \le 1. \tag{7.54}$$

Step 3. A Lower Estimate.

I first prove the following bound: There exists a constant $C > 0$ such that for all λ sufficiently large,

$$\mathrm{P}\{X_1 \ge \lambda\} \ge C\lambda^{-1}e^{-\frac{1}{2}\lambda^2}. \tag{7.55}$$

Indeed by the L'Hôpital rule of calculus,[7.4]

$$\mathrm{P}\left\{X_1 \geq \lambda\right\} = \int_\lambda^\infty \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\,dx \sim \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{1}{2}\lambda^2}, \qquad \text{as } \lambda \to \infty, \tag{7.56}$$

where $f(\lambda) \sim g(\lambda)$ means that $\lim_{\lambda\to\infty} f(\lambda) \div g(\lambda) = 1$. Equation (7.55) readily follows from this.

Step 4. The Lower Bound.
Once more fix some $\theta > 1$, $c \in (0,1)$, and define $\theta_k := \lfloor \theta^k \rfloor$. Consider the events

$$E_k := \left\{ S_{\theta_{k+1}} - S_{\theta_k} \geq \sqrt{2c(\theta_{k+1} - \theta_k)\ln\ln\theta_k} \right\}. \tag{7.57}$$

Since E_k depends on $X_{\theta_k+1}, \ldots, X_{\theta_{k+1}}$, it follows that $E_1, E_2, \ldots$ are independent events. Moreover, since $S_m/\sqrt{m}$ has a standard Gaussian distribution for any m,

$$\begin{aligned} \mathrm{P}(E_k) &= \mathrm{P}\left\{ S_{\theta_{k+1}-\theta_k} \geq \sqrt{2c(\theta_{k+1} - \theta_k)\ln\ln\theta_k} \right\} \\ &= \mathrm{P}\left\{ X_1 \geq \sqrt{2c\ln\ln\theta_k} \right\} \geq \frac{C}{\sqrt{2c\ln\ln\theta_k}(\ln\theta_k)^c}. \end{aligned} \tag{7.58}$$

The last inequality holds for all k large, thanks to (7.55). Therefore, for any $\theta > 1$ and $c \in (0,1)$, $\sum_k \mathrm{P}(E_k) = +\infty$. By the second (independent) part of the Borel–Cantelli Lemma (Theorem 5.23), a.s. infinitely many of the E_k's occur; more precisely, almost all $\omega \in \Omega$ is in infinitely many of the E_k's. Thus,

$$\limsup_{k\to\infty} \frac{S_{\theta_{k+1}} - S_{\theta_k}}{\sqrt{2c(\theta_{k+1} - \theta_k)\ln\ln\theta_k}} \geq 1, \quad \text{a.s.} \tag{7.59}$$

Since $\theta_{k+1} - \theta_k \sim \theta_{k+1}(1 - \theta^{-1})$,

$$\limsup_{k\to\infty} \frac{S_{\theta_{k+1}} - S_{\theta_k}}{\sqrt{2\theta_{k+1}\ln\ln\theta_k}} \geq \sqrt{c\left(1 - \frac{1}{\theta}\right)}, \quad \text{a.s.} \tag{7.60}$$

[7.4] Much more is known. For instance, Laplace [Lap05, pp. 490–493] derived the following remarkable continued fraction expansion:

$$\mathrm{P}\left\{X_1 \geq \lambda\right\} = \frac{e^{-\frac{1}{2}\lambda^2}}{\lambda\sqrt{2\pi}} \div \left[1 + \lambda^2 \div \left(1 + 2\lambda^2 \div \left\{1 + 3\lambda^2 \div \cdots\right\}\right)\right], \quad \forall\lambda > 0.$$

Thanks to this, (7.54), and the fact that $\theta_{k+1} \sim \theta \cdot \theta_k$, we can deduce the following:

$$\begin{aligned}\Lambda &\geq \limsup_{k\to\infty} \frac{S_{\theta_{k+1}}}{\sqrt{2\theta_{k+1}\ln\ln\theta_{k+1}}} = \limsup_{k\to\infty} \frac{S_{\theta_{k+1}}}{\sqrt{2\theta_{k+1}\ln\ln\theta_{k}}} \\ &\geq \limsup_{k\to\infty} \frac{S_{\theta_{k+1}} - S_{\theta_k}}{\sqrt{2\theta_{k+1}\ln\ln\theta_{k}}} - \limsup_{k\to\infty} \frac{|S_{\theta_k}|}{\sqrt{2\theta_{k+1}\ln\ln\theta_{k}}} \\ &\geq \sqrt{c\left(1-\frac{1}{\theta}\right)} - \sqrt{\frac{1}{\theta}}, \quad \text{a.s.}\end{aligned} \tag{7.61}$$

Since this is true for all $c, \theta > 1$, we see that $\Lambda \geq 1$, as was to be proved. □

5.3 The Lebesgue Differentiation Theorem

Given a continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{\delta\downarrow 0} \frac{1}{\delta}\int_{\omega}^{\omega+\delta} f(y)\, dy = f(\omega), \tag{7.62}$$

uniformly for all ω in any given compact set. The proof is not too difficult: Suppose we are interested in $\omega \in [0,1]$ to be concrete. Then, for any $\varepsilon > 0$, there exists $\delta \in (0,1)$ such that whenever $\omega, y \in [0, 1+\delta]$ satisfy $|\omega - y| \leq \delta$, then $|f(\omega) - f(y)| \leq \varepsilon$. Consequently,

$$\left|\frac{1}{\delta}\int_{\omega}^{\omega+\delta} f(y)\, dy - f(\omega)\right| \leq \frac{1}{\delta}\int_{\omega}^{\omega+\delta} |f(y) - f(\omega)|\, dy \leq \varepsilon. \tag{7.63}$$

This verifies (7.62). There is a surprising extension of this, due to H. Lebesgue, that holds for all integrable functions.[7.5]

Theorem 7.49 (Lebesgue Differentiation) *Given the Borel–Steinhaus probability space* $([0,1), \mathfrak{B}([0,1)), \mathrm{P})$*, if* $f \in L^1(\mathrm{P})$*, then (7.62) holds for Lebesgue almost all* $\omega \in [0,1)$*.*

[7.5] This is also known as the *Lebesgue density theorem.* In rough terms, it states that there are many functions that are a.e.-derivatives: At least one for each element of $L^1(dx)$. There is an interesting counterpart to this that states that "most" continuous functions are nowhere-differentiable. See Banach [Ban31] and Mazurkiewicz [Maz31].

Just as we did for the Kolmogorov strong law (Theorem 5.21) and Doob's martingale convergence theorem (Theorem 7.45) we need a maximal inequality, this time for the following function $\mathcal{M}f$ that is known as the *Hardy-Littlewood maximal function*; cf. Hardy and Littlewood [HL30, Theorem 17]. Extend f periodically to a function on $[0,2)$, and define,

$$\mathcal{M}f(\omega) := \sup_{\delta\in(0,1)} \frac{1}{\delta}\int_\omega^{\omega+\delta} |f(y)|\,dy, \qquad \forall \omega\in[0,1). \tag{7.64}$$

Theorem 7.50 *For all $\lambda > 0$, $p \geq 1$, and $f \in L^p(\mathrm{P})$,*

$$\mathrm{P}\left\{\mathcal{M}f \geq \lambda\right\} \leq \left(\frac{4}{\lambda}\right)^p \|f\|_p. \tag{7.65}$$

I first prove the Lebesgue differentiation theorem assuming the preceding maximal function inequality. We will then verify Theorem 7.50.

Proof of Theorem 7.49 For notational convenience, let me first define the "averaging operators" $\mathcal{A}_\delta$ as follows:

$$\mathcal{A}_\delta(f)(\omega) = \mathcal{A}_\delta f(\omega) := \frac{1}{\delta}\int_\omega^{\omega+\delta} f(y)\,dy, \qquad \forall\omega\in[0,1),\ f\in L^1(\mathrm{P}). \tag{7.66}$$

Thus, we have the following pointwise equality : $\mathcal{M}f = \sup_{\delta\in(0,1)}\mathcal{A}_\delta(|f|)$.

We have already seen that for any continuous function $g : [0,1) \to \mathbb{R}$ (extended periodically to $[0,2)$), $\lim_{\delta\downarrow 0}\mathcal{A}_\delta g = g$, uniformly. On the other hand, continuous functions are dense in $L^1(\mathrm{P})$ (Exercise 2.6). Therefore, for every $n \geq 1$, we can find a continuous function g_n such that $\|g_n - f\|_1 \leq n^{-1}$. With this in mind, we can note that thanks to the triangle inequality, the following inequality holds pointwise (i.e., for every ω).

$$\begin{aligned}\limsup_{\delta\downarrow 0}|\mathcal{A}_\delta f - f| &\leq \lim_{\delta\downarrow 0}|\mathcal{A}_\delta g_n - g_n| + \limsup_{\delta\downarrow 0}|\mathcal{A}_\delta g_n - \mathcal{A}_\delta f| \\ &\quad + |g_n - f| \\ &= \limsup_{\delta\downarrow 0}|\mathcal{A}_\delta g_n - \mathcal{A}_\delta f| + |g_n - f|.\end{aligned} \tag{7.67}$$

Consequently, if the left-hand side were greater than λ, then one of the two terms on the right-most side must be at least $\lambda/2$ (triangle inequality); i.e.,

$$\begin{aligned} \mathrm{P}\left\{\limsup_{\delta\downarrow 0} |\mathcal{A}_\delta f - f| \geq \lambda\right\} \leq \mathrm{P}&\left\{\limsup_{\delta\downarrow 0} |\mathcal{A}_\delta g_n - \mathcal{A}_\delta f| \geq \frac{\lambda}{2}\right\} \\ &+ \mathrm{P}\left\{|g_n - f| \geq \frac{\lambda}{2}\right\}. \end{aligned} \tag{7.68}$$

Because of the inequality $|\mathcal{A}_\delta g_n - \mathcal{A}_\delta f| \leq \mathcal{A}_\delta(|g_n - f|)$, Theorem 7.50 (with $p = 1$) shows that the first term in the right-hand side of (7.68) is bounded above by $8\lambda^{-1}\|g_n - f\|_1 \leq 8\lambda^{-1}n^{-1}$. We can also apply Chebyshev's inequality (Corollary 2.16) to deduce that the second term is at most $2\lambda^{-1}\|g_n - f\|_1 \leq 2\lambda^{-1}n^{-1}$. In other words, the left-hand side of (7.68) is at most $10\lambda^{-1}n^{-1}$ for all n. Since the left-hand side of (7.68) is independent of n, it must equal zero for all $\lambda > 0$. This proves the Lebesgue differentiation theorem (why?). □

Proof of Theorem 7.50 By replacing f with $|f|$, we can assume without any loss in generality that $f \geq 0$. Now consider $\mathfrak{F}_n^0$ that is defined as the collection of all dyadic intervals of the form $D(j;n) := [j2^{-n}, (j+1)2^{-n})$ where $j = 0, \ldots, 2^n - 1$ and $n \geq 1$, and define $\mathfrak{F}_n$ to be the σ-algebra generated by $\mathfrak{F}_n^0$. Since every element of $\mathfrak{F}_n^0$ is a union of two of the elements of $\mathfrak{F}_{n+1}^0$, it follows that $\mathfrak{F}_n \subseteq \mathfrak{F}_{n+1}$; i.e., $\mathfrak{F}_n$ is a filtration.

Next, let us view the function f as a random variable, and compute $\mathrm{E}\{f \,|\, \mathfrak{F}_n\}$ by using Remark 7.4. It follows that for almost all $\omega \in [0,1)$, if $\omega \in D(j;n)$, then $\mathrm{E}\{f \,|\, \mathfrak{F}_n\}$ equals $\mathrm{E}\{f \,|\, D(j;n)\}$ in the classical sense; i.e., for almost all $\omega \in [0,1)$,

$$\begin{aligned} \mathrm{E}\{f \,|\, \mathfrak{F}_n\}(\omega) &= \sum_{j=0}^{2^n-1} \mathbf{1}_{D(j;n)}(\omega) \frac{1}{\mathrm{P}(D(j;n))} \int_{D(j;n)} f(y)\, dy \\ &= \sum_{j=0}^{2^n-1} \mathbf{1}_{D(j;n)}(\omega) 2^n \int_{D(j;n)} f(y)\, dy. \end{aligned} \tag{7.69}$$

If $\delta \in (0,1)$, then there exists $n \geq 0$ such that $2^{-n-1} \leq \delta \leq 2^{-n}$. For this value of n, we can find $j \in \{0, \ldots, 2^n - 1\}$ such that $\omega \in D(j;n)$. There are two cases to consider:

Case 1. If j is even, then there we can find a unique $k \in \{0, \ldots, 2^{n-1}-1\}$ such that $j = 2k$ and $k2^{-(n-1)} = j2^{-n} \leq \omega + \delta < (j+1)2^{-n} \leq (k+1)2^{-(n-1)}$. In other words, in this case $[\omega, \omega+\delta] \subseteq D(k; n-1)$ and $\omega \in D(j; n) = [k2^{-(n-1)}, (k+\frac{1}{2})2^{-(n-1)})$. Since $f \geq 0$, we have a.s.,

$$\begin{aligned} &\sum_{\substack{j=0 \\ j \text{ even}}}^{2^n-1} \int_\omega^{\omega+\delta} f(y)\,dy \cdot \mathbf{1}_{D(j;n)}(\omega) \\ &\quad \leq \sum_{k=0}^{2^{n-1}-1} \int_{D(k;n-1)} f(y)\,dy \cdot \mathbf{1}_{\left[k2^{-(n-1)},(k+\frac{1}{2})2^{-(n-1)}\right)}(\omega). \end{aligned} \tag{7.70}$$

When $n = 0$ or $n = 1$, the above holds tautologically if we define $\mathfrak{F}_{-1} := \mathfrak{F}_0 := \{\varnothing, \Omega\}$. This is because $\mathrm{E}\{f \mid \{\varnothing, \Omega\}\} = \int_0^1 f(y)\,dy$ and f is nonnegative. (Note that $\{\mathfrak{F}_n;\ n \geq -1\}$ is still a filtration.)

Case 2. If j is odd, then there we can find $k \in \{0, \ldots, 2^{n-1}-1\}$ such that $j = 2k+1$ and $k2^{-(n-1)} \leq j2^{-n} \leq \omega + \delta < (j+1)2^{-n} = (k+1)2^{-(n-1)}$. In other words, in this case, ω is in $[(k+\frac{1}{2})2^{-(n-1)}, (k+1)2^{-n})$, and $[\omega, \omega+\delta] \subseteq D(k; n-1)$. Consequently, with probability one,

$$\begin{aligned} &\sum_{\substack{j=0 \\ j \text{ odd}}}^{2^n-1} \int_\omega^{\omega+\delta} f(y)\,dy \cdot \mathbf{1}_{D(j;n)}(\omega) \\ &\quad \leq \sum_{k=0}^{2^{n-1}-1} \int_{D(k;n-1)} f(y)\,dy \cdot \mathbf{1}_{\left[(k+\frac{1}{2})2^{-(n-1)},(k+1)2^{-n}\right)}(\omega). \end{aligned} \tag{7.71}$$

We can add (7.70) and (7.71), and appeal to (7.69), to deduce that for almost all $\omega \in [0,1)$,

$$\begin{aligned} \frac{1}{\delta}\int_\omega^{\omega+\delta} f(y)\,dy &\leq \frac{1}{\delta}\sum_{k=0}^{2^{n-1}-1} \int_{D(k;n-1)} f(y)\,dy \cdot \mathbf{1}_{D(k;n-1)}(\omega) \\ &= \frac{2^{-n+1}}{\delta}\mathrm{E}\{f \mid \mathfrak{F}_{n-1}\} \leq 4\mathrm{E}\{f \mid \mathfrak{F}_{n-1}\}, \end{aligned} \tag{7.72}$$

because $\delta \geq 2^{-n-1}$. In particular, we can take the supremum over all $n \geq 0$ (equivalently, all $\delta \in (0,1)$) to see that a.s., $\mathcal{M}f \leq 4\sup_{n\geq 0} X_n$ where X

is the Doob martingale $X_{n-1} := \mathrm{E}\{f \mid \mathfrak{F}_{n-1}\}$. The theorem follows from applying Doob's maximal inequality (Theorem 7.43) to the submartingale X_{n-1}^p; cf. Lemma 7.20. □

Let me mention also the following corollary of Theorem 7.50 that is essentially due to Hardy and Littlewood [HL30, Theorem 17]. This result has a number of interesting consequences in real and harmonic analysis, but this is a discussion that is better suited elsewhere.

Corollary 7.51 *If $p > 1$ and $f \in L^p(\mathrm{P})$, then*

$$\int_0^1 |\mathcal{M}f(t)|^p \, dt \le \left(\frac{4p}{p-1}\right)^p \int_0^1 |f(t)|^p \, dt. \tag{7.73}$$

5.4 Option-Pricing in Discrete Time

We now take a look at an application of martingale theory to the mathematics of securities in finance.[7.6] In this example, we consider the oversimplified case where there is only one type of stock, and the value of this stock changes at times $n = 1, 2, 3, \ldots, N$. You start with y_0 dollars at time 0, and during the time-period $(n, n+1)$, you can look at the performance of this stock up to time n, and based on this information, you may decide to buy A_{n+1}-many share, where a negative A_{n+1} means selling A_{n+1}-many shares. If S_n denotes the value of the stock at time n, we simplify the model further by assuming that $|S_{n+1} - S_n| = 1$. That is, the stock value fluctuates by exactly one unit at each time-step, and the stock-value is updated precisely at time n for every $n = 1, 2, \ldots$. The only unexplained variable is the ending time N; this is the so-called *time to maturity* that will be explained later. Now we can place things in a more precise framework.

Let Ω denote the collection of all possible $\omega := (\omega_1, \ldots, \omega_N)$ where every ω_j takes the values ± 1. Intuitively, $\omega_j = 1$ if and only if the value of our stock went up by 1 dollar at time j. Thus, $\omega_j = -1$ means that the stock went down by a dollar, and Ω is the collection of all theoretically possible stock movements.

Define the functions $S_1, \ldots, S_N$ by $S_0(\omega) := 0$, and $S_n(\omega_1, \ldots, \omega_n) := \omega_1 + \cdots + \omega_n$ $(1 \le n \le N)$. We may slightly abuse the notation and also

[7.6] This section is based on the discussions of Baxter and Rennie [BR96, Chapter 2] and Williams [Wil91, Section 15.2].

write $S_n(\omega)$ for $S_n(\omega_1, \ldots, \omega_n)$. although it is clear that S_n only depends on $\omega_1, \ldots, \omega_n$. In this way, $S_n(\omega)$ represents the value of the stock at time n, and corresponds to the stock movements $\omega_1, \ldots, \omega_n$. During the time-interval $(n, n+1)$, you may look at $\omega_1, \ldots, \omega_n$, choose a number $A_{n+1}(\omega) := A_{n+1}(\omega_1, \ldots, \omega_n)$ that might depend on $\omega_1, \ldots, \omega_n$, and buy $A_{n+1}(\omega)$-many shares. Therefore, if your starting fortune at time 0 is y_0, then your fortune at time n is

$$Y_n(\omega) = Y_n(\omega_1, \ldots, \omega_n) := y_0 + \sum_{j=1}^{n} A_j(\omega) \left[S_j(\omega) - S_{j-1}(\omega) \right], \tag{7.74}$$

as n ranges from 1, to N. The sequence $A_1(\omega), \ldots, A_N(\omega)$ is your investment *strategy*, and recall that it depends on the stock movements $\omega_1, \ldots, \omega_N$ in a "previsible manner;" i.e., for each n, $A_n(\omega)$ depends only on $\omega_1, \ldots, \omega_n$.

In simple terms, a *European call option* is a gamble wherein you buy the option to buy the stock at a given price C—the *strike or exercise price*—at time N.

Now suppose that you have the option to call at C dollars. If it happens that $S_N(\omega) > C$, then you have gained $(S_N(\omega) - C)$ dollars. This is because you can buy the stock at C dollars, and then instantaneously sell the stock at $S_N(\omega)$. On the other hand, if $S_N(\omega) \leq C$, then it is unwise (if "$S_N < C$", and immaterial if "$S_N = C$") to buy at C, and you gain nothing. Therefore, no matter what happens, the value of your option at time N is $(S_N(\omega) - C)^+$. An important question that needs to be settled is this:

$$\text{"\textit{What is the fair price for a call at } C\text{?}"}. \tag{7.75}$$

This was answered by Black and Scholes [BS73], and the connections to probability theory were discovered later by Harrison and Kreps [HK79] and Harrison and Pliska [HP81]. To explain the solution to (7.75), we need a brief definition from finance.

Definition 7.52 A strategy A is a *hedging strategy* if:

(i) Using A does not lead you to bankruptcy; i.e., $Y_n(\omega) \geq 0$ for all $n = 1, \ldots, N$.

(ii) Y attains the value of the stock at time N; i.e., $Y_N(\omega) = (S_N(\omega) - C)^+$.

Of course any strategy A is also previsible.

In terms of our model, we can think of the notion of y_0 as the "fair price of a given option" if, starting with y_0 dollars, we can find a hedging/investment strategy that yields the value of the said option at time N, *no matter how the stock values behave.*

The solution of Black and Scholes [BS73]—transcribed to the present simplified setting—depends on first making $(\Omega, \mathfrak{P}(\Omega))$ into a probability space, where $\mathfrak{P}(\Omega)$ is the power set of Ω. Define the probability measure P so that $X_j(\omega) := \omega_j$ are i.i.d. taking the values ± 1 with probability $\frac{1}{2}$ each. In words, under the measure P, the stock values fluctuate at random but in a fair manner. Another, yet equivalent, way to define P is as the product measure:

$$\mathrm{P}(d\omega) := \mathrm{Q}(d\omega_1) \cdots \mathrm{Q}(d\omega_N), \qquad \forall \omega \in \Omega, \tag{7.76}$$

where $\mathrm{Q}(\{1\}) = \mathrm{Q}(\{-1\}) = \frac{1}{2}$. Using this probability space $(\Omega, \mathfrak{P}(\Omega), \mathrm{P})$, the functions $A_1, A_2, \ldots,$ $S_1, S_2, \ldots$, and $Y_1, Y_2, \ldots$ are stochastic processes, and we can present the so-called Black–Scholes formula for the fair price y_0 of a European option.

Theorem 7.53 (The Black–Scholes Formula) *There is a hedging strategy iff* $y_0 = \mathrm{E}\{(S_N - C)^+\}$.

Proof (First Part) We first prove Theorem 7.53 assuming that a hedging strategy A exists. If so, then the process Y_n defined in (7.74) is a martingale; cf. Exercise 7.17.[7.7] Moreover, by the definition of a hedging strategy, $Y_n \geq 0$ for all n, and $Y_N = (S_N - C)^+$ a.s. (in fact for all ω). On the other hand, martingales have a constant mean; i.e., $\mathrm{E}\{Y_N\} = \mathrm{E}\{Y_1\} = y_0$, thanks to (7.74). Therefore, we have shown that $y_0 = \mathrm{E}\{(S_N - C)^+\}$ as desired. □

In order to prove the second—more important—half, we need the following fact.

Theorem 7.54 (Martingale Representations) *In* $(\Omega, \mathfrak{P}(\Omega), \mathrm{P})$*, the process S is a mean-zero martingale. Moreover, if M is any other martingale, then it is a martingale transform of S; i.e., there exists a previsible process H such that* $M_n = \mathrm{E}\{M_1\} + \sum_{j=1}^n H_j(S_j - S_{j-1}),\ \forall n = 1, \ldots, N$.

[7.7] We need only check that $\mathrm{E}\{|Y_n|\} < +\infty$, but this is elementary since $(\Omega, \mathfrak{P}(\Omega), \mathrm{P})$ is a finite probability space.

Proof Since Lemma 7.37 proves that S is a mean-zero martingale, we can concentrate on proving that M is a martingale transform.

Since M is adapted, M_n is a function of $\omega_1, \dots, \omega_n$ only; i.e., by abusing the notation slightly, $M_n(\omega) = M_n(\omega_1, \dots, \omega_n)$ for all $\omega \in \Omega$. Now the martingale property states that $\mathrm{E}\{M_{n+1} \,|\, \mathfrak{F}_n\} = M_n$, a.s. On the other hand, thanks to the independence of the ω_j's, for all bounded $\phi_1, \dots, \phi_n$, where ϕ_j is a function of ω_j only,

$$\begin{aligned}
&\int_\Omega \prod_{j=1}^{n} \phi_j(\omega_j) M_{n+1}(\omega_1, \dots, \omega_n, \omega_{n+1})\, \mathrm{P}(d\omega) \\
&\quad = \frac{1}{2} \int_\Omega \prod_{j=1}^{n} \phi_j(\omega_j) M_{n+1}(\omega_1, \dots, \omega_n, -1)\, \mathrm{Q}(d\omega_1) \cdots \mathrm{Q}(d\omega_n) \\
&\qquad + \frac{1}{2} \int_\Omega \prod_{j=1}^{n} \phi_j(\omega_j) M_{n+1}(\omega_1, \dots, \omega_n, 1)\, \mathrm{Q}(d\omega_1) \cdots \mathrm{Q}(d\omega_n).
\end{aligned} \tag{7.77}$$

Define

$$N_n(\omega) := \frac{1}{2} M_{n+1}(\omega_1, \dots, \omega_n, 1) + \frac{1}{2} M_{n+1}(\omega_1, \dots, \omega_n, -1). \tag{7.78}$$

Then N_n is $\mathfrak{F}_n$-measurable, and $\mathrm{E}\{\prod_j \phi_j \cdot M_{n+1}\} = \mathrm{E}\{\prod_j \phi_j \cdot N_n\}$. This and the martingale property of M together show that[7.8]

$$\begin{aligned}
M_n(\omega) &= \mathrm{E}\left\{M_{n+1} \,|\, \mathfrak{F}_n\right\}(\omega) \\
&= \frac{1}{2} M_{n+1}(\omega_1, \dots, \omega_n, 1) + \frac{1}{2} M_{n+1}(\omega_1, \dots, \omega_n, -1),
\end{aligned} \tag{7.79}$$

for almost all $\omega \in \Omega$. In fact, since Ω is finite, and since P assigns positive measure to each ω_j, the preceding equality must hold for all ω. Moreover, since $\mathfrak{F}_0 = \{\varnothing, \Omega\}$, the preceding discussion continues to hold for $n = 0$ if we define $M_0 := \mathrm{E}\{M_1\}$. Since $M_n(\omega) = \frac{1}{2} M_n(\omega) + \frac{1}{2} M_n(\omega)$, the following holds for all $0 \le n \le N-1$ and all $\omega \in \Omega$:

$$M_{n+1}(\omega_1, \dots, \omega_n, 1) - M_n(\omega) = M_n(\omega) - M_{n+1}(\omega_1, \dots, \omega_n, -1). \tag{7.80}$$

[7.8] While this calculation is intuitively clear, you should prove its validity by first checking it for M_{n+1} of the form, $M_{n+1}(\omega_1, \dots, \omega_{n+1}) = \prod_{j=1}^{n+1} h_j(\omega_j)$, and then appealing to a monotone class argument.

We can apply this as follows:

$$\begin{aligned} M_{n+1}(\omega) - M_0 &= \sum_{j=0}^{n} \left(M_{j+1}(\omega) - M_j(\omega)\right) \\ &= \sum_{j=0}^{n} \Big[\left(M_{j+1}(\omega) - M_j(\omega)\right) \mathbf{1}_{\{1\}}(\omega_{j+1}) \\ &\quad + \left(M_{j+1}(\omega) - M_j(\omega)\right) \mathbf{1}_{\{-1\}}(\omega_{j+1})\Big] \qquad (7.81) \\ &= \sum_{j=0}^{n} \left(M_{j+1}(\omega_1, \dots, \omega_j, 1) - M_j(\omega)\right) \left[\mathbf{1}_{\{1\}}(\omega_{j+1}) - \mathbf{1}_{\{-1\}}(\omega_{j+1})\right] \\ &= \sum_{j=0}^{n} \left(M_{j+1}(\omega_1, \dots, \omega_j, 1) - M_j(\omega)\right) \left[S_{j+1}(\omega) - S_j(\omega)\right]. \end{aligned}$$

Define

$$H_{j+1}(\omega) := M_{j+1}(\omega_1, \dots, \omega_j, 1) - M_j(\omega). \qquad (7.82)$$

Since $H_{j+1}(\omega)$ is a function of $\omega_1, \dots, \omega_j$, H is a previsible process, and the theorem follows. □

We are ready to prove the second half of the Black–Scholes formula.

Proof of Theorem 7.53 (Second Half) Note that the stochastic process $Y_n := \mathrm{E}\{(S_N - C)^+ \,|\, \mathfrak{F}_n\}$ $(0 \le n \le N)$ is a nonnegative Doob martingale and has the property that $Y_N = (S_N - C)^+$ almost surely, and hence for all ω (why?). Thanks to the martingale representation theorem (Theorem 7.54), we can find a previsible process A such that $Y_n = \mathrm{E}\{Y_1\} + \sum_{j=1}^{n-1} A_j(S_j - S_{j-1})$. Thus, as soon as we can prove that $A_j(\omega) \ge 0$ for all ω, it follows that A is a hedging strategy with $y_0 := \mathrm{E}\{Y_1\}$. By the martingale property, $\mathrm{E}\{Y_1\} = \mathrm{E}\{Y_2\} = \cdots = \mathrm{E}\{Y_N\}$, which implies that $y_0 = \mathrm{E}\{(S_N - C)^+\}$ and proves the theorem. Thus, it suffices to prove that for all n, $A_n \ge 0$ almost surely (why?).[7.9]

Recall from (7.82) that $A_{n+1}(\omega) = Y_{n+1}(\omega_1, \dots, \omega_n, 1) - Y_n(\omega)$. Therefore, it remains to show that a.s.,

$$Y_{n+1}(\omega_1, \dots, \omega_n, 1) \ge Y_n(\omega_1, \dots, \omega_n). \qquad (7.83)$$

[7.9] Negative investments are in fact allowed in the marketplace: If $A_n(\omega) \le 0$ for some n and ω, then for that ω, we may be *selling short.* This means that we may be selling stocks that we do not own, hoping that when the clock strikes n, we will earn enough to pay our debts.

But we can write

$$\begin{aligned} Y_n(\omega) &= \mathrm{E}\left\{(S_N - C)^+ \,\middle|\, \mathfrak{F}_n\right\}(\omega) \\ &= \mathrm{E}\left\{(S_N - S_n + S_n - C)^+ \,\middle|\, \mathfrak{F}_n\right\}(\omega) \\ &\le \mathrm{E}\left\{(S_N - S_{n+1} + 1 + S_n - C)^+ \,\middle|\, \mathfrak{F}_n\right\}(\omega), \end{aligned} \tag{7.84}$$

since the function $x \mapsto x^+$ is nondecreasing, and $S_N(\omega) - S_n(\omega) = S_N(\omega) - S_{n+1}(\omega) + \omega_{n+1} \le S_N(\omega) - S_{n+1}(\omega) + 1$. A similar calculation is made for $Y_{n+1}(\omega)$, viz.,

$$Y_{n+1}(\omega) = \mathrm{E}\left\{(S_N - S_{n+1} + S_{n+1} - C)^+ \,\middle|\, \mathfrak{F}_{n+1}\right\}(\omega). \tag{7.85}$$

Since $S_{n+1}(\omega) = S_n(\omega) + \omega_{n+1}$, then almost surely,

$$\begin{aligned} &Y_{n+1}(\omega_1, \dots, \omega_n, 1) \\ &\quad = \mathrm{E}\left\{(S_N - S_{n+1} + 1 + S_n - C)^+ \,\middle|\, \mathfrak{F}_{n+1}\right\}(\omega_1, \dots, \omega_n, 1). \end{aligned} \tag{7.86}$$

In light of (7.82) and the above, it suffices to show that

$$\begin{aligned} &\mathrm{E}\left\{(S_N - S_{n+1} + 1 + S_n - C)^+ \,\middle|\, \mathfrak{F}_{n+1}\right\}(\omega) \\ &\quad = \mathrm{E}\left\{(S_N - S_{n+1} + 1 + S_n - C)^+ \,\middle|\, \mathfrak{F}_n\right\}(\omega), \quad \text{a.s.}, \end{aligned} \tag{7.87}$$

for then the right hand side is a.s. a function of $(\omega_1, \dots, \omega_n)$. But this is not too hard to do. Since $S_N - S_{n+1}$ is independent of $\mathfrak{F}_{n+1}$ (and hence also of $\mathfrak{F}_n$), and since S_n is $\mathfrak{F}_n$-measurable, for any two bounded functions $\phi_1, \phi_2 : \mathbb{Z} \to \mathbb{R}$, two applications of (ii) of Theorem 7.6 reveal that a.s.,

$$\begin{aligned} \mathrm{E}\left\{\phi_1(S_N - S_{n+1})\phi_2(S_n) \,\middle|\, \mathfrak{F}_{n+1}\right\} &= \phi_2(S_n)\mathrm{E}\left\{\phi_1(S_N - S_{n+1}) \,\middle|\, \mathfrak{F}_{n+1}\right\} \\ &= \phi_2(S_n)\mathrm{E}\left\{\phi_1(S_N - S_{n+1})\right\} \\ &= \phi_2(S_n)\mathrm{E}\left\{\phi_1(S_N - S_{n+1}) \,\middle|\, \mathfrak{F}_n\right\} \\ &= \mathrm{E}\left\{\phi_1(S_N - S_{n+1})\,\phi_2(S_n) \,\middle|\, \mathfrak{F}_n\right\}. \end{aligned} \tag{7.88}$$

By a monotone class argument, for any integrable $\phi : \mathbb{Z}^2 \to \mathbb{R}$,

$$\mathrm{E}\left\{\phi(S_N - S_{n+1}, S_n) \,\middle|\, \mathfrak{F}_{n+1}\right\} = \mathrm{E}\left\{\phi(S_N - S_{n+1}, S_n) \,\middle|\, \mathfrak{F}_n\right\}. \tag{7.89}$$

In particular, apply this with $\phi(x, y) := (x + 1 + y - C)^+$ to deduce (7.87) and conclude the proof. □

6 Exercises

Exercise 7.1 We say that $X_n \in L^1(\mathrm{P})$ converges to $X \in L^1(\mathrm{P})$ *weakly in* $L^1(\mathrm{P})$ if for all bounded random variables Z, $\lim_{n\to\infty} \mathrm{E}\{X_n Z\} = \mathrm{E}\{XZ\}$. Show that $X_n \to X$ weakly in $L^1(\mathrm{P})$ if for any sub-σ-algebra $\mathfrak{G} \subseteq \mathfrak{F}$, $\mathrm{E}\{X_n \,|\, \mathfrak{G}\}$ converges to $\mathrm{E}\{X \,|\, \mathfrak{G}\}$ weakly in $L^1(\mathrm{P})$.

Exercise 7.2 Consider a random variable X that equals $2, 1, -1$ with probability $\frac{1}{3}$ each, and let $Y := \mathrm{sgn}(X)$ be the sign of X, while $Z := |X|$ is the modulus of X.

1. (a) Prove that in the sense of elementary probability, $\mathrm{E}\{X \,|\, Z = 1\} = 0$, and $\mathrm{E}\{X \,|\, Z = 2\} = 2$. From this conclude that $\mathrm{E}\{X \,|\, Z\} = 2\mathbf{1}_{\{Z=2\}}$, a.s.

2. (b) Using similar arguments, prove that $\mathrm{E}\{X \,|\, Y\} = \frac{3}{2}\mathbf{1}_{\{Y=1\}} - \mathbf{1}_{\{Y=-1\}}$, a.s.

3. (c) Show that with probability one, $\mathrm{E}\{\mathrm{E}(X \,|\, Y) \,|\, Z\} \neq \mathrm{E}\{\mathrm{E}(X \,|\, Z) \,|\, Y\}$. In particular, there exist random variables U, V, W, such that with positive probability $\mathrm{E}\{\mathrm{E}(U \,|\, V) \,|\, W\} \neq \mathrm{E}\{\mathrm{E}(U \,|\, W) \,|\, V\}$.

Exercise 7.3 Let $([0,1], \mathfrak{B}([0,1]), \mathrm{P})$ denote the Borel–Steinhaus probability space, and consider $X(\omega) := \omega$ for all $\omega \in [0,1]$ so that X is uniformly distributed on $[0,1]$. Compute $\mathrm{E}\{X \,|\, \mathfrak{B}([0, \frac{1}{2}])\}$, and compare this to $\mathrm{E}\{X \,|\, \sigma([0, \frac{1}{2}])\}$. Convince yourselves that the latter is related to $\mathrm{E}\{X \,|\, X \geq \frac{1}{2}\}$ of classical theory. This is due to J. Turner.

Exercise 7.4 Suppose that $Y \in L^1(\mathrm{P})$ is real-valued, and that X is a random variable that takes values in $\mathbb{R}^n$. Then prove that there exists a Borel measurable function b such that $\mathrm{E}\{Y \,|\, X\} = b(X)$, almost surely.
(Hint: First do this in the case that X is simple, then elementary, and then somehow take “limits.”)

Exercise 7.5 Verify Proposition 7.12.

Exercise 7.6 Carefully prove Lemmas 7.31 and 7.32. In addition, construct an example that shows that the difference of two stopping times, even if nonnegative, need not be a stopping time.
(Hint: For Lemma 7.32 first observe that $\mathfrak{F}_T$ is an algebra. Then consider the collection of all $A \in \mathfrak{F}_T$ such that $A^{\complement} \in \mathfrak{F}_T$.)

Exercise 7.7 Prove Lemma 7.37.

Exercise 7.8 Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathrm{P}\{X_1 = 1\} = 1 - \mathrm{P}\{X_1 = -1\} = p \neq \frac{1}{2}$, and prove (7.31). Also compute $\mathrm{E}\{T\}$ in the case $p = \frac{1}{2}$.
(HINT: For the last portion, start by carefully proving that $\mathrm{E}\{S_T^2\} = \mathrm{E}\{T\}$.)

Exercise 7.9 Suppose ξ and ζ are a.s.-nonnegative random variables such that for all $a > 0$,

$$\mathrm{P}\{\xi > a\} \leq \frac{1}{a}\mathrm{E}\{\zeta; \xi \geq a\}. \tag{7.90}$$

Prove then that for all $p > 1$, $\|\xi\|_p \leq (\frac{p}{p-1})\|\zeta\|_p$. Use this show the *strong L^p-inequality of Doob:* If X is a nonnegative submartingale, then as long as $X_n \in L^p(\mathrm{P})$ for all $n \geq 1$ and some $p > 1$,

$$\mathrm{E}\left\{\max_{1\leq j\leq n} X_j^p\right\} \leq \left(\frac{p}{p-1}\right)^p \mathrm{E}\{X_n^p\}. \tag{7.91}$$

Use this to prove Corollary 7.51.

Exercise 7.10 Suppose that $X_1, X_2, \ldots$ are independent mean-zero random variables in $L^2(\mathrm{P})$, and that they are bounded; i.e., that there exists a constant B such that almost surely, $|X_n| \leq B$ for all n. If $S_n := X_1 + \cdots + X_n$ denotes the corresponding partial-sum process, then prove that for all $\lambda > 0$ and $n \geq 1$,

$$\mathrm{P}\left\{\max_{1\leq j\leq n} |S_j| \leq \lambda\right\} \leq \frac{(B+\lambda)^2}{\mathrm{Var}(S_n)}. \tag{7.92}$$

This is from Khintchine and Kolmogorov [KK25].
(HINT: Apply Theorem 7.33 to $M_{n\wedge T}$ where $T := \inf\{j : |S_j| > \lambda\}$ with $\inf \varnothing := \infty$, and $M_n := S_n^2 - \mathrm{Var}(S_n)$, and note that on $\{T < +\infty\}$, $\sup_n |S_{T\wedge n}| \leq B + \lambda$.)

Exercise 7.11 Refine the martingale convergence theorem by showing that whenever X is bounded in L^p (i.e., $\sup_n \|X_n\|_p < +\infty$) for some $p > 1$, then $\lim_n X_n$ exists also in $L^p(\mathrm{P})$.
(HINT: Use Exercise 7.9.)

Exercise 7.12 Let $\gamma_1, \gamma_2, \ldots$ denote a sequence of i.i.d. random variables with $\mathrm{P}\{\gamma_1 = 0\} = \mathrm{P}\{\gamma_1 = 1\} = \frac{1}{2}$. Consider the stochastic process X, where

$X_1 := 1$, and for all $n \geq 2$, $X_n := 2X_{n-1}\gamma_n$. This is the mathematical model for the *double-or-nothing* strategy in a fair game: You start with 1 dollar. At each step, you bet twice your net worth if you won in the previous step, and 1 dollar otherwise. If all these steps are independent from one another, then your net worth at step n is X_n.

Prove that X is an L^1-bounded martingale that does not converge in $L^1(\mathrm{P})$. [In particular, Exercise 7.11 fails when $p = 1$.] Compute the almost sure limit $\lim_n X_n$. (What is the gambling interpretation of all this? This is known as the *St.-Petersburg Paradox*: While double-or-nothing will lead to an eventual win with probability one, you never expect to win.)

Exercise 7.13 Suppose X_n is a submartingale with bounded increments; i.e., there exists a nonrandom finite constant B such that almost surely, $|X_n - X_{n-1}| \leq B$ for all $n \geq 2$. Then prove that $\lim_n X_n$ exists a.s. on the set $\{\sup_m |X_m| < +\infty\}$.
(Hint: For any $\lambda > 0$, let $T := \inf\{j : |X_j| \geq \lambda\}$ and first argue that $X_{n\wedge T}$ is a bounded submartingale.)

Exercise 7.14 Suppose that $X_1, X_2, \ldots$ are i.i.d. with $\mathrm{P}\{X_1 = \pm 1\} = \frac{1}{2}$. As before, let $S_n := X_1 + \cdots + X_n$.

1. Prove that for all $n \geq 1$, and $t \in \mathbb{R}$, $\mathrm{E}\{e^{tS_n}\} \leq e^{\frac{1}{2}nt^2}$.

2. Use this to prove that for all $n \geq 1$, and $t > 0$,
$$\mathrm{P}\left\{\max_{1\leq j\leq n} |S_j| \geq nt\right\} \leq 2e^{-\frac{1}{2}nt^2}. \tag{7.93}$$

3. Use the above to prove the following half of the LIL (Theorem 7.48) for ±1 random variables: With probability one,
$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\ln\ln n}} \leq 1. \tag{7.94}$$

4. If $Y_1, Y_2, \ldots$ are i.i.d. with $\mathrm{P}\{Y_1 = 0\} = \mathrm{P}\{Y_1 = 1\} = \frac{1}{2}$, and if $T_n := Y_1 + \cdots + Y_n$, then prove that with probability one,
$$\limsup_{n\to\infty} \frac{T_n - \frac{n}{2}}{\sqrt{2n\ln\ln n}} \leq \frac{1}{2}. \tag{7.95}$$

   Check that this is one-half of the LIL (Theorem 7.48) for T_n.

(HINT: For part 1, you may need to use Taylor's expansion of $\cosh(\lambda)$.)

Exercise 7.15 (Hard) We continue with the previous exercise in order to prove the other half of the LIL (Theorem 7.48) for ±1 random variables; i.e., that with probability one,

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\ln\ln n}} \geq 1. \tag{7.96}$$

The following argument was devised by de Acosta [dA83, Lemma 2.4].

1. Prove that (7.96) follows once we prove that for any $c > 0$,
$$\liminf_{n\to\infty} \frac{1}{\ln\ln n} \ln \mathrm{P}\left\{S_n \geq \sqrt{2cn\ln\ln n}\right\} \geq -c. \tag{7.97}$$

2. To derive (7.97), choose $p_n \to \infty$ such that n divides p_n, and first prove that
$$\mathrm{P}\left\{S_n \geq \sqrt{2cn\ln\ln n}\right\} \geq \left(\mathrm{P}\left\{S_{p_n} \geq \frac{p_n}{n}\sqrt{2cn\ln\ln n}\right\}\right)^{n/p_n}. \tag{7.98}$$

3. Use the central limit theorem (Theorem 6.22) and the preceding with $p_n \sim \alpha n \div (\ln\ln n)$ to derive (7.97) and hence conclude the proof of the LIL (Theorem 7.48) for ±1 random variables.

(HINT: For part 2 first write $S_n = S_{p_n} + (S_{2p_n} - S_{p_n}) + \cdots + (S_n - S_{(n-1)p_n/n})$, and note that each summand has the same distribution as S_{p_n}. Next observe that if each of these n/p_n terms is greater than $p_n\lambda/n$, then $S_n \geq \lambda$. Finally choose λ judiciously. For part 3 optimize over the choice of α; you may also need (7.48) at this stage.)

Part III

A Glimpse at Brownian Motion

Chapter 8

The Wiener Process: A Mathematical Theory of Brownian Motion

1 A Brief Historical Review

Brown [Bro28] noted empirically that the grains of pollen in water undergo erratic motion, but he had no scientific explanation for this phenomenon. This observation went largely unnoticed in the scientific literature.

Later on, Bachelier [Bac00] published his doctoral dissertation—under the supervision of H. Poincaré—on the mathematics of the stock market. In this work, Bachelier proposed a stochastic process that today is called the Brownian motion. Unfortunately, at that time, the measure-theoretic foundations of probability had not yet been cast. Therefore, Bachelier's work was not considered entirely rigorous, and for this reason, until quite recently, his ideas was appreciated.

Einstein [Ein05] returned to the observation of R. Brown, and proposed a theory for molecular motion that was based on Bachelier's Brownian motion, although Einstein arrived at the Brownian motion independently and apparently unaware of the work of Bachelier [Bac00]. He then used this theory to compute a very good estimate for Avagadro's constant. Einstein's theory was based on the assumption that the Brownian motion process exists; an assumption that was finally verified by Wiener [Wie23]. In the present context, the contributions of von Smoluchowski [vS18] and Perrin [Per03] are also noteworthy.[8.1]

The said postulates on Brownian motion are as follows: Brownian motion $\{W(t);\ t \geq 0\}$ is a random function of t (:= "time") such that

(P-a) $W(0) = 0$, and for any given time $t > 0$, the distribution of $W(t)$ is normal with mean zero and variance t.

(P-b) For any $0 < s < t$, $W(t) - W(s)$ is independent of $\{W(u);\, 0 \leq u \leq s\}$, Think of s as the current time, to see that this condition is saying: Given the present value, the future is independent of the past; i.e., W satisfies the Markov property in continuous time.

(P-c) The random variable $W(t) - W(s)$ has the same distribution as $W(t-s)$. That is, Brownian motion has stationary increments.

(P-d) The random path $t \mapsto W(t)$ is continuous with probability one.

Remark 8.1 One can also have a Brownian motion B that starts at an arbitrary point $x \in \mathbb{R}$ by defining $B(t) := x + W(t)$ where W is a Brownian motion started at the origin. Check that B has all the properties of W except that $B(t)$ has mean x for all $t \geq 0$, and $B(0) = x$. Unless stated to the contrary, our Brownian motions always start at the origin.

So why is this a postulate and not a fact? The sticky point is part (P-d); namely that the Brownian path must be continuous almost surely. In fact, Lévy [Lév37, Théoréme 54.2, page 181], has shown that if, in (P-a), you replace the normal distribution by any other, then either there is no stochastic process that satisfies (P-a)–(P-c), or else (P-d) fails to hold![8.2]

In summary, while the predictions of physics were correct, a rigorous understanding of this phenomenon required the in-depth undertaking of N. Wiener. Since Wiener's work, Brownian motion has been studied by multitudes of mathematicians. At best, these notes might whet your appetite to learn more about this elegant theory.

[8.1] The phrase "Avagadro's constant" is due to Perrin.

[8.2] Incidentally, some of these so-called pure-jump Lévy processes are now being used in diverse applications such as mathematical ecology, economics, geology, and mathematical finance. It now seems as if the normal distribution is not favored by all aspects of nature. To learn more about this work of Lévy and much more, see the books of Bertoin [Ber96] and Sato [Sat99].

2 Normal Random Variables and Gaussian Processes

Let us temporarily leave aside the question of the existence of Brownian motion, and take a detour on normal distributions, random variables, and processes. Before proceeding further, you may wish to recall Examples 1.19, 1.20, and 6.12, where normal random variables and their characteristic functions have been introduced.

2.1 Normal Random Variables

This section's definition of a normal random variable is a suitable generalization that are sometimes called (possibly) degenerate normal variables.

Definition 8.2 An $\mathbb{R}$-valued random variable Y is said to be *centered* if $Y \in L^1(\mathrm{P})$, and $\mathrm{E}\{Y\} = 0$. An $\mathbb{R}^n$-valued random variable $Y := (Y_1, \ldots, Y_n)$ is said to be centered if each Y_i is. If, in addition, $Y_i \in L^2(\mathrm{P})$ for all $i = 1, \ldots, n$, then the *covariance matrix* $Q = (Q_{i,j})$ of Y is the matrix whose (i,j)th entry is the covariance of Y_i and Y_j; i.e., $Q_{i,j} := \mathrm{E}\{Y_i Y_j\}$.

Suppose that $X := (X_1, \ldots, X_n)$ denotes a centered n-dimensional random variable in $L^2(P)$. Let $\alpha \in \mathbb{R}^n$ denote a constant vector, and note that $\alpha \cdot X := \alpha_1 X_1 + \cdots + \alpha_n X_n$ is a centered $\mathbb{R}$-valued random variable in $L^2(\mathrm{P})$ whose variance is computed as follows:

$$\mathrm{Var}(\alpha \cdot X) = \mathrm{E}\left\{(\alpha \cdot X)^2\right\} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \mathrm{E}\{X_i X_j\} \alpha_j = \alpha \cdot Q\alpha, \tag{8.1}$$

where $Q = (Q_{i,j})$ is the covariance matrix of X. Since the variance of any random variable is nonnegative, we have the following.

Lemma 8.3 *If Q denotes the covariance matrix of a centered $L^2(\mathrm{P})$-valued random variable $X := (X_1, \ldots, X_n)$ in $\mathbb{R}^n$, then Q is a symmetric nonnegative-definite matrix. Moreover, the diagonal terms of Q are given by $Q_{j,j} = \mathrm{Var}(X_j)$.*

Definition 8.4 An $\mathbb{R}^n$-valued random variable $X := (X_1, \ldots, X_n)$ is *centered normal* (or centered Gaussian) if for all $\alpha \in \mathbb{R}^n$,

$$\mathrm{E}\left\{e^{i\alpha \cdot X}\right\} = e^{-\frac{1}{2}\alpha \cdot Q\alpha}, \tag{8.2}$$

where Q is a symmetric nonnegative-definite real matrix. The matrix A is called the *covariance matrix* of X.

We have seen in Lemma 8.3 that covariance matrices are symmetric and nonnegative-definite. The converse is also true. This is a mere consequence of the following. Among other things it shows that the matrix Q that was used to define a centered normal random variable is indeed the covariance matrix of X, so that our definitions are not inconsistent.

Theorem 8.5 *If Q is any symmetric nonnegative-definite n-by-n matrix of real numbers, then there exists a centered normal random variable $X := (X_1, \dots, X_n)$ whose covariance matrix is Q. Moreover, if Q is nonsingular, then the distribution of X is absolutely continuous with respect to the n-dimensional Lebesgue measure and has the density of Example 1.20; i.e.,*

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(Q)}} \exp\left(-\frac{1}{2} x \cdot Q^{-1} x\right), \qquad \forall x \in \mathbb{R}^n. \tag{8.3}$$

Finally, an $\mathbb{R}^n$-valued random variable $X := (X_1, \dots, X_n)$ is centered Gaussian if and only if all linear combinations of the X_j's are centered Gaussian random variables in $\mathbb{R}$; i.e., iff for all $\alpha \in \mathbb{R}^n$, $\alpha \cdot X$ is a mean-zero normal random variable.

Proof (Optional) Let $\lambda_1, \dots, \lambda_n$ denote the n eigenvalues of Q; we know that the λ_j's are real and nonnegative. If we let $v_1, \dots, v_n$ denote the orthonormal (column) eigenvectors corresponding to $\lambda_1, \dots, \lambda_n$, then the n-by-n matrix $P := (v_1, \dots, v_n)$ is orthogonal; i.e., $P'P$ is equal to the n-by-n identity matrix. Moreover, we can write $Q = P'\Lambda P$, where Λ is the diagonal matrix of the eigenvalues,

$$\Lambda := \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}. \tag{8.4}$$

Next, let $Z_1, \dots, Z_n$ denote n independent standard normal random variables ($\mathbb{R}$-valued). (They exist on some probability space, thanks to Example 5.17.) It is not difficult to see that $Z := (Z_1, \dots, Z_n)$ is a centered $\mathbb{R}^n$-valued random variable whose covariance is the identity matrix. Having this define

$$X := P'\Lambda^{\frac{1}{2}}Z, \quad \text{where} \quad \Lambda^{\frac{1}{2}} := \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix}. \tag{8.5}$$

Since $X := (X_1, \ldots, X_n)$ is a linear combination of centered random variables, it too is a centered $\mathbb{R}^n$-valued random variable. Now for any $\alpha \in \mathbb{R}^n$,

$$\begin{aligned} \alpha \cdot X = \alpha \cdot P'\Lambda^{\frac{1}{2}}Z &= \sum_{l=1}^n \alpha_l \left(P'\Lambda^{\frac{1}{2}}Z\right)_l \\ &= \sum_{k=1}^n Z_k \left[\sum_{l=1}^n \alpha_l \left(P'\Lambda^{\frac{1}{2}}\right)_{l,k}\right] := \sum_{k=1}^n Z_k A_k. \end{aligned} \tag{8.6}$$

Therefore, by independence (Lemma 5.12),

$$\mathrm{E}\left\{e^{i\alpha\cdot X}\right\} = \prod_{k=1}^n \mathrm{E}\left\{e^{iZ_kA_k}\right\} = \exp\left(-\frac{1}{2}\sum_{k=1}^n A_k^2\right). \tag{8.7}$$

The last equality uses also the explicit computation of the characteristic function of a standard normal variable; see Example 6.12. Now

$$\begin{aligned} A_k^2 = \left[\sum_{l=1}^n \alpha_l \left(P'\Lambda^{\frac{1}{2}}\right)_{l,k}\right]^2 &= \sum_{l=1}^n \sum_{j=1}^n \alpha_l\alpha_j \left(P'\Lambda^{\frac{1}{2}}\right)_{l,k} \left(P'\Lambda^{\frac{1}{2}}\right)_{j,k} \\ = \sum_{l=1}^n \sum_{j=1}^n \alpha_l\alpha_j \left(P'\Lambda^{\frac{1}{2}}\right)_{l,k} \left(\Lambda^{\frac{1}{2}}P\right)_{k,j}. \end{aligned} \tag{8.8}$$

Therefore,

$$\sum_{k=1}^n A_k^2 = \sum_{l=1}^n \sum_{j=1}^n \alpha_l\alpha_j P'\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}P = \sum_{l=1}^n \sum_{j=1}^n \alpha_l\alpha_j Q = \alpha \cdot Q\alpha. \tag{8.9}$$

In other words, we have constructed a centered Gaussian process X that has covariance matrix Q. To check that Q is indeed the matrix of the covariances of X, note that

$$\mathrm{E}\{X_iX_j\} = \sum_{l=1}^n \sum_{k=1}^n \mathrm{E}\left\{\left(P'\Lambda^{\frac{1}{2}}\right)_{i,k} Z_k \times \left(P'\Lambda^{\frac{1}{2}}\right)_{j,l} Z_l\right\}. \tag{8.10}$$

But $\mathrm{E}\{Z_l Z_k\}$ is equal to zero unless $k = l$ in which case it equals one. Therefore,

$$\begin{aligned}\mathrm{E}\{X_i X_j\} &= \sum_{k=1}^{n} \left(P'\Lambda^{\frac{1}{2}}\right)_{i,k} \times \left(P'\Lambda^{\frac{1}{2}}\right)_{j,k} \\ &= \sum_{k=1}^{n} \left(P'\Lambda^{\frac{1}{2}}\right)_{i,k} \times \left(\Lambda^{\frac{1}{2}}P\right)_{k,j} = Q_{i,j},\end{aligned} \tag{8.11}$$

as asserted.

Next we suppose that Q is nonsingular, and propose to show that $\widehat{f}(t) = \exp(-\frac{1}{2}t'Qt)$ for all $t \in \mathbb{R}^n$. In other words, we intend to prove that

$$\int_{\mathbb{R}^n} e^{it\cdot x} f(x)\, dx = e^{-\frac{1}{2}t\cdot Qt}, \qquad \forall t \in \mathbb{R}^n. \tag{8.12}$$

The left-hand side equals

$$\int_{\mathbb{R}^n} e^{it\cdot x} f(x)\, dx = \frac{1}{\sqrt{(2\pi)^n \det(Q)}} \int_{\mathbb{R}^n} e^{it\cdot x - \frac{1}{2}x'Q^{-1}x}\, dx. \tag{8.13}$$

We can write $Q = P'\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}P$, and appeal to fact that $P' = P^{-1}$ to obtain $Q^{-1} = P'\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}P$, where $\Lambda^{-\frac{1}{2}}$ is the inverse of $\Lambda^{\frac{1}{2}}$; i.e., it is a diagonal matrix with diagonal elements $\lambda_1^{-\frac{1}{2}}, \ldots, \lambda_n^{-\frac{1}{2}}$. We can now change variables in the integral of the preceding display ($y := \Lambda^{-\frac{1}{2}}Px$). Note that $x = P'\Lambda^{\frac{1}{2}}y$, and $x \cdot Q^{-1}x = \|y\|^2$—the square of the Euclidean norm of y. Finally, a Jacobian calculation reveals that $dx = \det(P')\det(\Lambda^{\frac{1}{2}})dy$. Since $P' = P^{-1}$, the determinant of P equals 1. Therefore, $dx = \sqrt{\det(P'\Lambda P)}\, dy = \sqrt{\det(Q)}\, dy$. Putting these calculations together leads us to

$$\begin{aligned}\int_{\mathbb{R}^n} e^{ix\cdot t} f(x)\, dx &= (2\pi)^{-n/2} \int_{\mathbb{R}^n} \exp\left(it \cdot P'\Lambda^{\frac{1}{2}}y - \frac{1}{2}\|y\|^2\right) dy \\ &= \prod_{\ell=1}^{n} (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(it_\ell \left[P'\Lambda^{\frac{1}{2}}y\right]_\ell - \frac{1}{2}y_\ell^2\right) dy_\ell\end{aligned} \tag{8.14}$$

I have appealed to Fubini–Tonelli (Theorem 3.6; how?). Each of the one-dimensional integrals can be computed painlessly by completing the square, and this will show that the end right-hand side equals $\exp(-\frac{1}{2}t \cdot Qt)$ as asserted (check!).

To complete this proof, we derive the assertion about linear combinations. First suppose $\alpha \cdot X$ is a centered Gaussian variable in $\mathbb{R}$. We have already seen that its variance must then $\alpha \cdot Q\alpha$ where Q is the covariance matrix of X. In particular, thanks to Example 6.12, $\mathrm{E}\{\exp(i\alpha \cdot X)\} = \exp(-\frac{1}{2}\alpha \cdot Q\alpha)$. Thus, if $\alpha \cdot X$ is a mean-zero normal variable in $\mathbb{R}$ for all $\alpha \in \mathbb{R}^n$, then X is centered Gaussian. The converse is proved similarly: If X is a centered Gaussian variable in $\mathbb{R}^n$, fix any $\alpha \in \mathbb{R}^n$ and define $Y := \alpha \cdot X$ to see that for all $t \in \mathbb{R}$, $\mathrm{E}\{\exp(itY)\} = \exp(-\frac{1}{2}t^2\sigma^2)$ where $\sigma^2 = \alpha \cdot \alpha \geq 0$ (nonnegative-definiteness of Q), and therefore $Y = \alpha \cdot X$ is a mean-zero normal random variable with variance σ^2. □

Remark 8.6 We have seen that the covariance matrix of the centered normal random variable $X := (X_1, \ldots, X_n)$ determines its distribution. To this I will add in passing that there are somewhat strange constructions of two centered normal variables X_1 and X_2 such that (X_1, X_2) is not normal even though the covariance matrix of (X_1, X_2) always exists; cf. Exercise 8.3. This shows that in general the normality of $(X_1, \ldots, X_n)$ is a stronger property than the normality of each of the X_j's.

There is an important corollary of this development that states that for normal random vectors independence and uncorrelatedness are one and the same.[8.3]

Corollary 8.7 *Consider a centered* $\mathbb{R}^{n+m}$*-valued normal random variable* $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ *such that* $\mathrm{E}\{X_iY_j\} = 0$ *for all* $i = 1, \ldots, n$ *and* $j = 1, \ldots, m$. *Then* $(X_1, \ldots, X_n)$ *and* $(Y_1, \ldots, Y_m)$ *are independent.*

Proof (Optional) It is clear that $(X_1, \ldots, X_n)$ is a centered $\mathbb{R}^n$-valued normal random variable whose covariance matrix A is described by $A_{i,j} = \mathrm{E}\{X_iX_j\}$. Similarly, $(Y_1, \ldots, Y_m)$ is a centered $\mathbb{R}^m$-valued normal random variable whose covariance matrix B is $B_{i,j} = \mathrm{E}\{Y_iY_j\}$. Then it is easy to check that the covariance matrix of $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$, where the zeros are square matrices whose entries are all zeroes. This proves

[8.3] Suppose U is uniformly distributed on $[0, 1]$, and define $X := \cos(2\pi U)$ and $Y := \sin(2\pi U)$. Then $\mathrm{E}\{X\} = \mathrm{E}\{Y\} = 0 = \mathrm{E}\{XY\}$, so that X and Y are centered, as well as uncorrelated. However, X and Y are not independent since $X^2 + Y^2 = 1$. (Why is this enough to show lack of independence?)

that for all $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$,

$$\mathrm{E}\left\{e^{ia\cdot X+ib\cdot Y}\right\} = \exp\left(-\frac{1}{2}a\cdot Aa\right)\exp\left(-\frac{1}{2}b\cdot Bb\right). \tag{8.15}$$

(Why?) Exercise 8.1 finishes the proof. □

2.2 Brownian Motion as a Centered Gaussian Processes

Definition 8.8 We say that a real-valued process X is *centered Gaussian* if for any $0 \le t_1, t_2, t_3, \ldots, t_k$, $(X(t_1), \ldots, X(t_k))$ is a centered normal random variable in $\mathbb{R}^k$. The function $Q(s,t) := \mathrm{E}\{X(s)X(t)\}$ is the corresponding *covariance function.*

Now let us assume that we know that Brownian motion exists, and derive some of its elementary properties.

Theorem 8.9 *If $W := \{W(t) : \ t \ge 0\}$ denotes a Brownian motion. Then W is a centered Gaussian process with covariance function $Q(s,t) = s \wedge t$. Moreover, any a.s.-continuous centered Gaussian process whose covariance function is Q and starts at 0 is a Brownian motion.*[8.4] *Furthermore:*

(i) (Symmetry) *The process $-W$ is a Brownian motion.*

(ii) (Scaling) *For any $c > 0$, the process $X(t) := c^{-1/2}W(ct)$ is a Brownian motion.*

(iii) (Time Inversion) *The process $Z(t) := tW(\frac{1}{t})$ is also a Brownian motion.*

(iv) (Time Reversion) *Given any $T > 0$, the process $R(t) := W(T) - W(T-t)$ is a Brownian motion indexed by $t \in [0,T]$.*

(v) (Quadratic Variation) *The process W has "quadratic variation" in the following sense: For each $t > 0$, as $n \to \infty$,*

$$V_n(t) := \sum_{j=0}^{n-1}\left[W\left(\left(\frac{j+1}{n}\right)t\right) - W\left(\left(\frac{j}{n}\right)t\right)\right]^2 \xrightarrow{\mathrm{P}} t. \tag{8.16}$$

[8.4] The same theorem remains to hold without the a.s.-continuity assumption.

(vi) (The Markov Property) *For any $T > 0$, the process $t \mapsto W(t+T) - W(T)$ is a Brownian motion, and is independent of $\sigma\{W(r);\ 0 \le r \le T\}$; i.e., for any $t_1, t_2, \ldots, t_k \ge 0$ and all $r_1, \ldots, r_m \le T$, the vector $(W(t_j+T) - W(T); 1 \le j \le k)$ is independent of $(W(r_l); 1 \le l \le m)$.*

The following is a ready consequence that will be important for us later on.

Corollary 8.10 *Brownian motion is a continuous-time martingale; i.e., for all $t \ge s \ge 0$, $\mathrm{E}\{W(t) \,|\, \mathfrak{F}_s\} = W(s)$, a.s., where $\mathfrak{F}_s$ is the σ-algebra generated by $\{W(u);\ 0 \le u \le s\}$.*

Proof Since $W(t) - W(s)$ has mean zero, and is independent of $\mathfrak{F}_s$, $\mathrm{E}\{W(t) - W(s) \,|\, \mathfrak{F}_s\} = 0$, a.s. The result follows from the obvious fact that $W(s)$ is $\mathfrak{F}_s$-measurable. □

Remark 8.11 I wish to point out three interesting consequences of Theorem 8.9:

1. Since $W(t+T) = [W(t+T) - W(T)] + W(T)$, the Markov property tells us that given the values of W before time T, the "post-T" process $t \mapsto W(t+T)$ is an independent Brownian motion that starts at $W(T)$. In particular, the dependence of the post-T process on the past is local since it only depends on the last value $W(T)$.

2. There can be many different Brownian motions on the same probability space. For instance, note that $\{W(t); 0 \le t \le 1\}$ and $\{W(1-t) - W(t); 0 \le t \le 1\}$ are two different Brownian motions. However, they are not the same process. For instance, evaluate both at time $t = 1$ and compare.

3. If W has a bounded variation with probability $\delta > 0$, then $V_n(t)$ would have to converge to 0 with probability at least δ. This follows from $V_n(t) \le \sup_{u,v} |W(u) - W(v)| \cdot \sum_{j=0}^{n-1} |W((j+1)t/n) - W(jt/n)|$ and the a.s. uniform continuity of W for $t \in [0,1]$. Here, $\sup_{u,v}$ denotes the supremum over all $u, v \in [0,1]$ such that $|v - u| \le n^{-1}$. Consequently, this shows that Brownian motion a.s. has unbounded variation. This will be refined later on (Theorem 8.16).

Proof First, let us find the covariance function of W, assuming that it is indeed a centered Gaussian process: If $t \geq s \geq 0$, then we can appeal to the fact that $W(t) - W(s)$ is a mean-zero random variable that independent of $W(s)$, we can deduce that

$$\begin{aligned} Q(s,t) &= \mathrm{E}\left\{W(s)W(t)\right\} = \mathrm{E}\left\{\big[W(t) - W(s) + W(s)\big]W(s)\right\} \\ &= \mathrm{E}\left\{|W(s)|^2\right\} = s. \end{aligned} \tag{8.17}$$

In other words for all $s, t \geq 0$, $Q(s,t) = s \wedge t$. Next we will prove that W is a centered Gaussian process.

By the independence of the increments of W, for all $0 = t_0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$\begin{aligned} &\mathrm{E}\left\{\exp\left(i\sum_{k=1}^{n} \alpha_k \left[W(t_k) - W(t_{k-1})\right]\right)\right\} \\ &\quad = \prod_{k=1}^{n} \mathrm{E}\left\{e^{i\alpha_k [W(t_k) - W(t_{k-1})]}\right\}. \end{aligned} \tag{8.18}$$

On the other hand, $W(t_k) - W(t_{k-1})$ is a mean-zero normal random variable with variance $t_k - t_{k-1}$; its characteristic function is computed in Example 6.12, and this leads to

$$\begin{aligned} &\mathrm{E}\left\{\exp\left(i\sum_{k=1}^{n} \alpha_k \left[W(t_k) - W(t_{k-1})\right]\right)\right\} \\ &\quad = \prod_{k=1}^{n} \exp\left(-\frac{1}{2}\alpha_k^2 \left[t_k - t_{k-1}\right]\right) = e^{-\frac{1}{2}\alpha \cdot M\alpha}, \end{aligned} \tag{8.19}$$

where the matrix M is described by $M_{i,j} = 0$ if $i \neq j$, and $M_{j,j} = t_j - t_{j-1}$. In other words, the vector $(W(t_j) - W(t_{j-1});\ 1 \leq j \leq n)$ is a centered normal random variable in $\mathbb{R}^n$. On the other hand, for any $\beta := (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$, $\sum_{k=1}^{n} \beta_k W(t_k) = \sum_{k=1}^{n} \alpha_k [W(t_k) - W(t_{k-1})]$ where $\beta_k := \alpha_k + \cdots + \alpha_n$. This proves that $\sum_{k=1}^{n} \beta_k W(t_k)$ is a centered normal random variable in $\mathbb{R}$, which amounts to the fact that W is a centered Gaussian process.

Next we prove that if $G := \{G(t);\ t \geq 0\}$ is an a.s.-continuous centered Gaussian process with covariance function Q, and if $G(0) = 0$, then G is a Brownian motion. It suffices to show that whenever $t > s$, $G(t) - G(s)$ is independent of $\{G(u);\ 0 \leq u \leq s\}$, since the remaining conditions of

Brownian motion are easily-verified for G (check!). To prove this assertion, we fix $0 \leq u_1 \leq \cdots \leq u_k \leq s$ and prove that $G(t) - G(s)$ is independent of $(G(u_1), \ldots, G(u_n))$. However, the distribution of the $(n+1)$-dimensional random vector $(G(t)-G(s), G(u_1), \ldots, G(u_n))$ is the same as that of $(W(t)-W(s), W(u_1), \ldots, W(u_n))$, since everything reduces to the same calculations involving the function Q. This proves that G is a Brownian motion.

Parts (i)–(iv) follow from direct covariance computations. I will work out one of them (part iv say), and leave the rest to you (check them!).

For any $s \leq t$,

$$\begin{aligned}
&\mathrm{E}\left\{R(s)R(t)\right\} \\
&\quad = \mathrm{E}\left\{\left(W(T) - W(T-s)\right)\left(W(T) - W(T-t)\right)\right\} \\
&\quad = \mathrm{E}\left\{(W(T))^2\right\} - \mathrm{E}\left\{W(T)W(T-s)\right\} \\
&\qquad - \mathrm{E}\left\{W(T)W(T-t)\right\} + \mathrm{E}\left\{W(T-s)W(T-t)\right\}.
\end{aligned} \tag{8.20}$$

The above equals $s = s \wedge t$, since by the first portion of this proof, for any $u, v \geq 0$, $\mathrm{E}\{W(u)W(v)\} = u \wedge v$.

In order to prove (v), we first note that

$$\mathrm{E}\left\{V_n(t)\right\} = \sum_{j=0}^{n-1} \mathrm{E}\left\{\left[W\left(\left(\frac{j+1}{n}\right)t\right) - W\left(\left(\frac{j}{n}\right)t\right)\right]^2\right\} = t. \tag{8.21}$$

Therefore, thanks to Chebyshev's inequality (Corollary 2.16), it remains to show that $\mathrm{Var}(V_n(t)) \to 0$. To show this, we compute $\|V_n(t)\|_2^2$ first. Write $d_j := [W((j+1)t/n) - W(jt/n)]^2$ for brevity. Then, since d_j and d_k are independent when $j \neq k$, and since $\mathrm{E}\{d_\ell\} = t/n$ for all ℓ, we have

$$\begin{aligned}
\mathrm{E}\left\{(V_n(t))^2\right\} &= \sum_{k=0}^{n-1} \mathrm{E}\{d_k^2\} + 2 \sum\sum_{0 \leq j < k \leq n-1} \left(\frac{t}{n}\right)^2 \\
&= \sum_{k=0}^{n-1} \mathrm{E}\{d_k^2\} + t^2 - \frac{t^2}{n}.
\end{aligned} \tag{8.22}$$

On the other hand, d_k is the square of a real-valued mean-zero normal random variable with variance t/n. Therefore, d_k has the same distribution as $(t/n)Z$ and $\mathrm{E}\{d_k^2\} = (t/n)^2 \mathrm{E}\{Z^4\}$, where Z is a standard normal random variable in $\mathbb{R}$ (why?). In particular,

$$\mathrm{Var}\left(V_n(t)\right) = \frac{t^2}{n}\left[\mathrm{E}\{Z^4\} - 1\right] \to 0, \tag{8.23}$$

as desired.[8.5] We complete our proof by deriving the Markov property.

Note that $t \mapsto W(t+T) - W(T)$ is a continuous centered Gaussian process. Therefore, we need to verify first that it is Brownian motion, and next that it is independent of $\sigma\{W(r);\ 0 \le r \le T\}$. For the former, suppose $0 \le s \le t$, and notice that

$$\begin{aligned}
&\mathrm{E}\left\{\Big(W(s+T) - W(T)\Big)\Big(W(t+T) - W(T)\Big)\right\} \\
&\quad = \mathrm{E}\{W(s+T)W(t+T)\} - \mathrm{E}\{W(T)W(t+T)\} \\
&\qquad - \mathrm{E}\{W(T)W(s+T)\} + \mathrm{E}\left\{(W(T))^2\right\} \\
&\quad = (s+T) - T - T + (T) = s = s \wedge t.
\end{aligned} \tag{8.24}$$

In particular, $t \mapsto W(t+T) - W(T)$ is a Brownian motion. The assertion about independence is proved similarly: If $s \le T$, then

$$\begin{aligned}
&\mathrm{E}\left\{W(s)\left(W(T+t) - W(T)\right)\right\} \\
&\quad = \mathrm{E}\left\{W(s)W(T+t)\right\} - \mathrm{E}\left\{W(s)W(T)\right\} = s - s = 0.
\end{aligned} \tag{8.25}$$

Corollary 8.7 proves the independence of $t \mapsto W(t+T) - W(T)$ from $\sigma\{W(r); 0 \le r \le T\}$. □

3 Wiener's Construction: Brownian Motion on $[0,1)$

The big problem at this point is to show the existence of Brownian motion. Once we have this, then we can proceed to study various aspects of Brownian motion as we started doing in the previous section. The first step is a simple reduction of this project: If we can show the existence of Brownian motion indexed by $[0,1)$, then we have a general existence result.[8.6] In more precise term, we have

[8.5] In fact, $\mathrm{E}\{Z^4\}$ can be computed as follows:

$$\mathrm{E}\{Z^4\} = \int_{-\infty}^{\infty} \frac{x^4 e^{-x^2/2}}{\sqrt{2\pi}}\, dx = \sqrt{\frac{2}{\pi}} \int_0^\infty x^4 e^{-x^2/2}\, dx = 3.$$

I have used the fact that $\Gamma(\frac{5}{2}) = \frac{3}{4}\sqrt{\pi}$ (check this!), which is a classical result of Stirling [Sti30].

[8.6] Brownian motion indexed by $[0,1)$ has the same properties as Brownian motion, however we only define the process $W(t)$ for $t \in [0,1)$. Likewise, we can discuss Brownian

Lemma 8.12 *Suppose $B_0, B_1, B_2, \ldots$ are independent Brownian motions indexed by $[0,1)$. Then, the following recursive definition is a Brownian motion (indexed by $[0,\infty)$): $W(t) := B_0(t)$ if $t \in [0,1)$. If $t \in [1,2)$, then $W(t) := B_0(1) + B_1(t-1)$; more generally, whenever $t \in [j, j+1)$ for some $j \geq 1$, then*

$$W(t) := \sum_{k=0}^{j-1} B_k(1) + B_j(t-j). \tag{8.26}$$

Remark 8.13 Suppose we know that Brownian motion indexed by $[0,1]$ exists (cf. Theorem 8.15 below). Then you should be able to use Theorem 5.17 to prove that on some probability space we can construct independent Brownian motions $B_0, B_1, B_2, \ldots$.

Remark 8.14 Conversely, suppose that Brownian motion exists. Then we can construct countably many independent Brownian motions indexed by $[0,1)$ as follows: Define $B_j(t) := W(t+j) - W(j)$ if $t \in [j, j+1)$. Indeed, thanks to the Markov property (Theorem 8.9), $B_0, B_1, \ldots$ are independent Brownian motions, each of which is indexed by $[0,1)$.

Proof (Optional) Since the above is a finite sum for any finite $t \geq 0$, the process W is a continuous centered Gaussian process. It remains to check its covariance: If $s \leq t$, then either we can find $j \geq 0$ such that $j \leq s \leq t < j+1$, or $j \leq s \leq j+1 \leq \ell \leq t < \ell+1$ for some $\ell > j$. In the first case,

$$\begin{aligned} &\mathrm{E}\left\{W(s)W(t)\right\} \\ &\quad = \mathrm{E}\left\{\left(\sum_{k=0}^{j-1} B_k(1) + B_j(t-j)\right)\left(\sum_{k=0}^{j-1} B_k(1) + B_j(s-j)\right)\right\} \\ &\quad = \mathrm{E}\left\{\sum_{k=0}^{j-1} \left(B_k(1)\right)^2\right\} + \mathrm{E}\left\{(B_j(s-j)B_j(t-j))\right\} \\ &\quad = j + (s-j) = s = s \wedge t. \end{aligned} \tag{8.27}$$

(Why?) In the second case, one obtains the same final answer, and I will leave the calculations to you. In any event, W has the correct covariance

motion indexed by any interval, and more generally still, any measurable set $T \subseteq [0,\infty)$. When T is "fractal-like," the latter objects are quite interesting still, and appear in Peres et al. [LPX02].

function, and is therefore a Brownian motion. □

The simplest construction of Brownian motion indexed by $[0, 1)$ (in fact, $[0, 1]$) is the following minor variant of the original construction of Norbert Wiener. Throughout, let $X_0, X_1, X_2, \cdots$ denotes a sequence of i.i.d. normal random variables with mean zero and variance one; the existence of this sequence is guaranteed by Theorem 5.17. Then, formally speaking, the Brownian motion $\{W(t); 0 \le t \le 1\}$ is the limit of the following sequence:

$$W_n(t) = tX_0 + \frac{\sqrt{2}}{\pi}\sum_{j=1}^{n} \frac{\sin(j\pi t)}{j} X_j, \qquad 0 \le t \le 1, \ n = 1, 2, \ldots. \tag{8.28}$$

Of course, once we have the existence of Brownian motion indexed by $[0, 1)$, Lemma 8.12 proves the existence of Brownian motion. To properly prove existence, we need to first complete the probability space. Informally, this means that we declare all subsets of null sets measurable and null; this can always be done at no cost (Theorem 1.22).

Theorem 8.15 (Wiener [Wie23]) *If the underlying probability space is complete, then with probability one, $W(t) := \lim_{n\to\infty} W_{2^n}(t)$ exists, and the convergence holds uniformly for all $t \in [0, 1]$. Moreover, the process W is a Brownian motion indexed by $[0, 1]$.*

Proof I will split the proof into three steps.

Step 1. Uniform Convergence.
For $n = 1, 2, \ldots$ and $t \ge 0$ define $S_n(t) := \sum_{k=1}^{n} k^{-1}\sin(k\pi t)X_k$, so that $W_n(t) = tX_0 + \sqrt{2}\pi^{-1}S_n(t)$. Stated more carefully, we have two processes $S_n(t, \omega)$ and $W_n(t, \omega)$, and as always we do not show the dependence on ω.

We will show that S_{2^n} forms a Cauchy sequence a.s. and in $L^2(\mathrm{P})$, uniformly in $t \in [0, 1]$. Note that

$$[S_{2^{n+1}}(t) - S_{2^n}(t)]^2 = \left(\sum_{j=2^n+1}^{2^{n+1}} \frac{\sin(j\pi t)}{j} X_j\right)^2 \le \left|\sum_{j=2^n+1}^{2^{n+1}} \frac{e^{ij\pi t}}{j} X_j\right|^2. \tag{8.29}$$

Therefore,

$$\begin{aligned}[S_{2^{n+1}}(t) - S_{2^n}(t)]^2 &\le \sum_{j=2^n+1}^{2^{n+1}} \sum_{k=2^n+1}^{2^{n+1}} \frac{e^{i(j-k)\pi t}}{jk} X_j X_k \\ &= \sum_{k=2^n+1}^{2^{n+1}} \frac{X_j^2}{j^2} + 2 \sum_{l=1}^{2^n-1} \sum_{k=2^n+1}^{2^{n+1}-l} \frac{e^{il\pi t}}{k(l+k)} X_k X_{l+k} \qquad (8.30) \\ &\le \sum_{k=2^n+1}^{2^{n+1}} \frac{X_j^2}{j^2} + 2 \sum_{l=1}^{2^n-1} \left| \sum_{k=2^n+1}^{2^{n+1}-l} \frac{X_k X_{l+k}}{k(k+l)} \right|.\end{aligned}$$

The right-hand side is independent of $t \ge 0$, so we can take expectations and apply Minkowski's inequality (Theorem 2.25) to obtain

$$\begin{aligned}\left\| \sup_{t\ge 0} |S_{2^{n+1}}(t) - S_{2^n}(t)| \right\|_2^2 &\le \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{j^2} + 2 \sum_{l=1}^{2^n-1} \left\| \sum_{k=2^n+1}^{2^{n+1}-l} \frac{X_k X_{l+k}}{k(k+l)} \right\|_2 \\ &= \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{j^2} + 2 \sum_{l=1}^{2^n-1} \sqrt{\sum_{k=2^n+1}^{2^{n+1}-l} \left\| \frac{X_k X_{l+k}}{k(k+l)} \right\|_2^2}.\end{aligned} \qquad (8.31)$$

In the last step we have used the proof of Lemma 5.9. The final squared-$L^2(\mathrm{P})$-norm is easily seen to be equal to $k^{-2}(k+l)^{-2}$. On the other hand, by monotonicity, $\sum_{k=2^n+1}^{2^{n+1}} k^{-2} \le 2^{-n}$, and $\sum_{l=1}^{2^n-1}(l+2^n)^{-1} \le 1$. Therefore,

$$\mathrm{E}\left\{ \sup_{t\ge 0} |S_{2^{n+1}}(t) - S_{2^n}(t)|^2 \right\} \le 2^{-n} + 2 \cdot 2^{-n/2}. \qquad (8.32)$$

In particular, with probability one, $\sum_n \sup_{t\ge 0} |S_{2^{n+1}}(t) - S_{2^n}(t)| < +\infty$, which shows that as $n \to \infty$, $S_{2^n}(t)$ converges uniformly in $t \ge 0$ to the limiting random process $S_\infty(t) := \sum_{j=1}^\infty \sin(j\pi t) j^{-1} X_j$. In particular, $W(t) := \lim_{n\to\infty} W_{2^n}(t)$ exists uniformly in $t \ge 0$ almost surely, and this concludes Step 1.

Step 2. Continuity and Distributional Properties.
The random map $t \mapsto W_{2^n}(t)$ defined in (8.28) is obviously continuous. Being an a.s. uniform limit of continuous functions, it follows that $t \mapsto W(t)$ is a.s. continuous; see Step 3 below for a technical note on this issue. Moreover, since W_{2^n} is a mean-zero Gaussian process, then so is W (Exercise 8.2). Since $W(0) = 0$, it suffices to show that

$$\mathrm{E}\left\{ |W(t) - W(s)|^2 \right\} = t - s, \qquad \forall 0 \le s \le t. \qquad (8.33)$$

(Why?) But the independence of the X's—together with Lemma 5.9—yields

$$\begin{aligned}\mathrm{E}\left\{|W(t)-W(s)|^2\right\} &= (t-s)^2 + \frac{2}{\pi^2}\mathrm{E}\left\{(S_\infty(t)-S_\infty(s))^2\right\} \\ &= (t-s)^2 + \frac{2}{\pi^2}\sum_{j=1}^{\infty}\left(\frac{\sin(j\pi t)-\sin(j\pi s)}{j}\right)^2 \\ &= (t-s)^2 + \frac{2}{\pi^2}\sum_{j=1}^{\infty}\left|\frac{1}{2}\int_{-\pi}^{\pi} f(x)e^{ijx}\,dx\right|^2,\end{aligned} \tag{8.34}$$

where $f(x) := \mathbf{1}_{[\pi s,\pi t]}(x) + \mathbf{1}_{[-\pi t,-\pi s]}(x)$ for $x \in [-\pi,\pi]$. Define $\phi_n(x) := (2\pi)^{-1/2}\exp(inx)$ for $x \in [-\pi,\pi]$, and $n = 0, \pm 1, \pm 2, \ldots$. Then,

$$\begin{aligned}\mathrm{E}\left\{|W(t)-W(s)|^2\right\} &= (t-s)^2 + \frac{1}{\pi}\sum_{j=1}^{\infty}\left|\int_{-\pi}^{\pi} f(x)\phi_j(x)\,dx\right|^2 \\ &= (t-s)^2 + \frac{1}{2\pi}\sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty}\left|\int_{-\pi}^{\pi} f(x)\phi_j(x)\,dx\right|^2 \\ &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\left|\int_{-\pi}^{\pi} f(x)\phi_j(x)\,dx\right|^2.\end{aligned} \tag{8.35}$$

By the Riesz–Fischer theorem (Theorem B.1), the right-hand side is equal to $(2\pi)^{-1}\int_{-\pi}^{\pi}|f(x)|^2\,dx = (t-s)$. This yields (8.33).

Step 3. Technical Wrap-Up.

There are one or two subtle loose ends that we now address. That is, the limit $W(t) := \lim_{n\to\infty} W_{2^n}(t)$ is known to exist (and holds uniformly over all $t \in [0,1]$) only with probability one. However, we may not yet have a process W, since $W(t,\omega)$ is undefined for all ω at which that $\lim_n W_n$ does not exist uniformly. Therefore, we *define* $W(t) := \limsup_{n\to\infty} W_{2^n}(t)$. This is a well-defined measurable process. Moreover, with probability one, it defines a continuous function. The remainder of the calculations of Step 2 goes through, since by redefining a random variable on a set of measure zero, we do not change its distribution (unscramble this!). Finally, the completion is needed to ensure that the event C that W is continuous is measurable; in Step 1, we showed that $C^{\complement}$ is a subset of a null set. Since the underlying probability is complete, it too is null. □

4 Nowhere-Differentiability

We are in position to prove the following striking theorem. Throughout, W denotes a Brownian motion.

Theorem 8.16 (Paley et al. [PWZ33]) *If the underlying probability space is complete, then W is nowhere-differentiable, a.s.*

Proof For any $\lambda > 0$ and $n \geq 1$, consider the event

$$E_\lambda^n := \left\{ \exists s \in [0,1] : \ |W(s) - W(t)| \leq \lambda 2^{-n}, \forall t \in s \pm 2^{-n} \right\}. \tag{8.36}$$

(Why is this measurable?) We intend to show that $\sum_n \mathrm{P}(E_\lambda^n) < +\infty$ no matter the value of $\lambda > 0$. Indeed, suppose there exists $s \in [0,1]$ such that for all t within 2^{-n} of s, $|W(s) - W(t)| \leq \lambda 2^{-n}$. Now there must exist a (random) $j = 0, \ldots, 2^n - 1$ such that $s \in D(j;n) := [j2^{-n}, (j+1)2^{-n}]$. Thus, for all $t \in D(j;n)$, $|W(s) - W(t)| \leq \lambda 2^{-n}$. By the triangle inequality, we can deduce that for all $u, v \in D(j;n)$, $|W(u) - W(v)| \leq 2\lambda 2^{-n} = \lambda 2^{-n+1}$. Subdivide $D(j;n)$ into four smaller dyadic intervals, and note that the successive differences in the values of W (at the endpoints of the subdivided intervals) are at most $\lambda 2^{-n+1}$. In other words, $\max_{0 \leq \ell \leq 3} |\Delta_{j,\ell}^n| \leq \lambda 2^{-n+1}$, where

$$\Delta_{j,\ell}^n := W\left(j2^{-n} + (\ell+1)2^{-(n+2)}\right) - W\left(j2^{-n} + \ell 2^{-(n+2)}\right). \tag{8.37}$$

So far, we have worked for any fixed ω. Now we obtain a probability estimate: We have shown that

$$\begin{aligned} \mathrm{P}\left(E_\lambda^n\right) &\leq \mathrm{P}\left\{ \exists j = 0, \ldots, 2^n - 1 : \max_{0 \leq \ell \leq 3} \left|\Delta_{j,\ell}^n\right| \leq \lambda 2^{-n+1} \right\} \\ &\leq \sum_{j=0}^{2^n-1} \mathrm{P}\left\{ \max_{0 \leq \ell \leq 3} \left|\Delta_{j,\ell}^n\right| \leq \lambda 2^{-n+1} \right\} \\ &= \sum_{j=0}^{2^n-1} \prod_{\ell=0}^{3} \mathrm{P}\left\{ \left|\Delta_{j,\ell}^n\right| \leq \lambda 2^{-n+1} \right\}, \end{aligned} \tag{8.38}$$

thanks to the independent-increments property of Brownian motion (P-b). On the other hand, by the stationary-increments property of W, $\Delta_{j,\ell}^n$ has a normal distribution with mean zero and variance $2^{-(n+2)}$ (P-a and P-c). Thus, for any $\beta > 0$,

$$\mathrm{P}\left\{\left|\Delta_{j,\ell}^n\right| \le \beta\right\} = \int_{-\beta 2^{(n+2)/2}}^{\beta 2^{(n+2)/2}} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \le \beta 2^{(n+2)/2}. \tag{8.39}$$

(Check this!) Apply this with $\beta := \lambda 2^{-n+1}$ to deduce that $\mathrm{P}(E_\lambda^n) \le \sum_{j=0}^{2^n-1}\prod_{\ell=0}^{3}(4\lambda 2^{-n/2}) = 256\lambda^4 2^{-n}$. In particular, $\sum_n \mathrm{P}(E_\lambda^n) < \infty$, as promised earlier. By the Borel–Cantelli lemma (Theorem 5.23), for any $\lambda > 0$, the following holds with probability one: For all but a finite number of n's,

$$\inf_{0\le s\le 1}\sup_{|t-s|\le 2^{-n}} \frac{|W(s)-W(t)|}{|s-t|} \ge \inf_{0\le s\le 1}\sup_{|t-s|\le 2^{-n}} \frac{|W(s)-W(t)|}{2^{-n}}, \tag{8.40}$$

which is at least λ. Thus, if $W'(s)$ exists for some $s \in [0,1]$, then $|W'(s)| \ge \lambda$ a.s. Since $\lambda > 0$ is arbitrary, this shows that $|W'(s)| = +\infty$, a.s., which contradicts the differentiability of W at some $s \in [0,1]$. By scaling (Theorem 8.9), this shows that W is a.s. nowhere differentiable in $[0,c]$ for any $c > 0$, and therefore W is a.s. nowhere-differentiable.

Technical Aside in the Proof. I have actually proven is that there exists a null set N, such that the collection D of all ω's for which $t \mapsto W(t,\omega)$ is somewhere differentiable is inside N. The collection D need not be measurable; I do not know if it is, and do not particularly care, since we can complete the underlying probability space at no cost (Theorem 1.22). In the said completion, D *is* a null set, and we are done. □

5 The Brownian Filtration and Stopping Times

Recall the Markov property of Brownian motion (Theorem 8.9): Given any fixed $T > 0$, the "post-T" process $t \mapsto W(T+t) - W(T)$ is a Brownian motion that is independent of $\sigma\{W(u);\, 0 \le u \le T\}$. Intuitively, this states that given the value of $W(T)$, the process after time T is independent of the process before time T.

The *strong Markov property*[8.7] states that the Markov property holds for a large class of random times T that are called stopping times. We have encountered such times when studying martingales in discrete time, and their continuous-time definition is formally the same. To present it, let us start by establishing some notation. Throughout, W is a Brownian motion (with some prescribed starting point, say $W(0) = x$).

Definition 8.17 A filtration $\mathfrak{A} := \{\mathfrak{A}_t;\ t \geq 0\}$ is a collection of sub-σ-algebras of $\mathfrak{F}$ such that whenever $s \leq t$, $\mathfrak{F}_s \subseteq \mathfrak{F}_t$. If $\mathfrak{A}$ is a filtration, then a measurable function $T : \Omega \to [0,\infty]$ (infinity is allowed) is a *stopping time* (or $\mathfrak{A}$-stopping time) if for all $t \geq 0$, $\{T \leq t\}$ is $\mathfrak{A}_t$-measurable. Given a stopping time T, we can define $\mathfrak{A}_T$ by

$$\mathfrak{A}_T := \{A \in \mathfrak{F} :\ A \cap \{T \leq t\} \in \mathfrak{A}_t,\ \forall t \geq 0\}\,. \tag{8.41}$$

T is called a *simple stopping time* if there exist $0 \leq \tau_0, \tau_1, \ldots < +\infty$ such that for all $\omega \in \Omega$, $T(\omega) \in \{\tau_0, \tau_1, \ldots\}$.

An important filtration is the one supplied to us by the Brownian motion W; namely,

$$\mathfrak{F}_t^0 := \sigma\{W(u);\ 0 \leq u \leq t\}\,. \tag{8.42}$$

In light of our development of martingale theory this definition is quite natural. Here are some of the properties of $\mathfrak{F}_T^0$ when T is a stopping time.

Proposition 8.18 *If T is a finite $\mathfrak{F}^0$-stopping time, then $\mathfrak{F}_T^0$ is a σ-algebra, and T is $\mathfrak{F}_T^0$-measurable. Furthermore, if $S \leq T$ is another stopping time, then $\mathfrak{F}_S^0 \subseteq \mathfrak{F}_T^0$. If $A \subseteq \mathbb{R}$ is either open or closed, then the first hitting time $T_A := \inf\{t \geq 0 :\ W(t) \in A\}$ with $\inf \varnothing := +\infty$ is a stopping time with respect to $\bar{\mathfrak{F}}_t$, where the latter is the* P*-completion of $\mathfrak{F}_t^0$.*

Remark 8.19 Notice that $\{\bar{\mathfrak{F}}_t;\ t \geq 0\}$ is a filtration of σ-algebras.

Proof I will prove this proposition in five easy steps.

Step 1. $\mathfrak{F}_T^0$ is a Sigma-Algebra.
Since it is a monotone class, it suffices to show that $\mathfrak{F}_T^0$ is closed under

[8.7] See Kinney [Kin53], Hunt [Hun56], Dynkin and Jushkevich [DJ56], and Blumenthal [Blu57]. The phrase "strong Markov property" was coined by Dynkin and Jushkevich [DJ56].

complementation. But for each $t \geq 0$, $A^{\mathsf{C}} \cap \{T \leq t\} = \{T \leq t\} \setminus (A \cap \{T \leq t\}) \in \mathfrak{F}^0_t$; hence $\mathfrak{F}^0_T$ is a σ-algebra.

Step 2. T is $\mathfrak{F}^0_T$-Measurable.
It suffices to show that $T^{-1}([0,a]) \in \mathfrak{F}^0_T$ for all $0 \leq a < \infty$ (why?). But given any $t \geq 0$, $T^{-1}([0,a]) \cap \{T \leq t\} = \{T \leq a \wedge t\} \in \mathfrak{F}^0_{a \wedge t} \subseteq \mathfrak{F}^0_t$, which does the job.

Step 3. $\mathfrak{F}^0_S \subseteq \mathfrak{F}^0_T$.
Suppose $A \in \mathfrak{F}^0_S$, and note that for any $t \geq 0$, $A \cap \{T \leq t\} = A \cap \{S \leq t\} \cap \{T \leq t\}$. Since $A \cap \{S \leq t\}$ and $\{T \leq t\}$ are both in $\mathfrak{F}^0_t$, this shows that $A \cap \{T \leq t\} \in \mathfrak{F}^0_t$ and hence $A \in \mathfrak{F}^0_T$; i.e., $\mathfrak{F}^0_S \subseteq \mathfrak{F}^0_T$.

Step 4. T_A is a Stopping Time When A is Open.
If A is open, then $\{T_A \leq t\}$ is the event that there exists a time s before t at which $W(s) \in A$. Let C denote the collection of all ω such that $t \mapsto W(t,\omega)$ is continuous. We know that $\mathrm{P}(\mathsf{C}) = 1$, and since A is open, then $\{T_A \leq t\} \cap \mathsf{C}$ is the event that there exists a rational $s \leq t$ at which $W(s) \in A$ (why?); i.e.,

$$\begin{aligned} \{T_A \leq t\} \cap \mathsf{C} &= \bigcup_{s \leq t : s \in \mathbb{Q}} \{W(s) \in A\} \cap \mathsf{C} \\ &= \bigcup_{s \leq t : s \in \mathbb{Q}} \{W(s) \in A\} \setminus \left(\bigcup_{s \leq t : s \in \mathbb{Q}} \{W(s) \in A\} \cap \mathsf{C}^{\mathsf{C}} \right). \end{aligned} \tag{8.43}$$

This is where the completeness of $\bar{\mathfrak{F}}_t$ comes in: Since $\bar{\mathfrak{F}}_t$ is complete, all subsets of null sets are $\bar{\mathfrak{F}}_t$-measurable (and null). In particular, $\cup_{s \leq t : s \in \mathbb{Q}} \{W(s) \in A\} \cap \mathsf{C}^{\mathsf{C}} \in \bar{\mathfrak{F}}_t$. On the other hand, it is clear that $\cup_{s \leq t : s \in \mathbb{Q}} \{W(s) \in A\} \in \mathfrak{F}^0_t \subseteq \bar{\mathfrak{F}}_t$, and this implies that $\{T_A \leq t\} \cap \mathsf{C} \in \bar{\mathfrak{F}}_t$. Since $\{T_A \leq t\} \cap \mathsf{C}^{\mathsf{C}} \subseteq \mathsf{C}^{\mathsf{C}}$ is null, it is $\bar{\mathfrak{F}}_t$-measurable, thanks again to completeness. To summarize, we have $\{T_A \leq t\} = (\{T_A \leq t\} \cap \mathsf{C}) \cup (\{T_A \leq t\} \cap \mathsf{C}^{\mathsf{C}}) \in \bar{\mathfrak{F}}_t$, as desired.

Step 5. T_A is a Stopping Time When A is Closed.
For each $n = 1, 2, \ldots$, define A_n to be the set of all $x \in \mathbb{R}$ such that the Euclidean distance between x and the set A is $< \frac{1}{n}$. It is clear that A_n is open, and for the C of Step 4, $\{T_A \leq t\} \cap \mathsf{C} = \cap_n \{T_{A_n} \leq t\} \cap \mathsf{C}$. By Step 4, $\{T_A \leq t\} \cap \mathsf{C}$ is in $\bar{\mathfrak{F}}_t$, and by the completeness of $\bar{\mathfrak{F}}_t$, so is $\{T_A \leq t\}$. □

Remark 8.20 The fact that we have to complete each $\mathfrak{F}^0_t$ in order to obtain a reasonable continuous-time theory is a typical technical annoyance that does not have a discrete-time counterpart. However, appealing to the said completeness is entirely harmless since we can always complete $\mathfrak{F}_t^0$ (Theorem 1.22).

Remark 8.21 If you know what F_σ- and G_δ-sets are, convince yourself that when A is of either variety, then T_A is a stopping time. You may ask further, "What about T_A when A is measurable but is neither F_σ nor G_δ?" The answer is that T_A is a stopping time for any Borel set A, but there are no known simple proofs of this deep fact.[8.8]

With this convention in mind, let me point out that we have yet to try to prove that $W(T)$ is $\bar{\mathfrak{F}}_T$, or perhaps even $\mathfrak{F}_T^0$-measurable. To do so we need to making another round of modifications to the $\mathfrak{F}_t$'s.[8.9]

Definition 8.22 A filtration $\{\mathfrak{A}_t;\ t \geq 0\}$ is *right-continuous* if for all $t \geq 0$, $\mathfrak{A}_t = \cap_{\varepsilon > 0} \mathfrak{A}_{t+\varepsilon}$.

Definition 8.23 The *Brownian filtration* $\{\mathfrak{F}_t;\ t \geq 0\}$ is defined as the smallest right-continuous filtration that contains $\{\bar{\mathfrak{F}}_t;\ t \geq 0\}$. That is, for all $t \geq 0$, $\mathfrak{F}_t := \cap \bar{\mathfrak{F}}_s$, where the intersection is taken over all $s > t$.

Remark 8.24 Note that any $\mathfrak{F}^0$- or $\bar{\mathfrak{F}}$-stopping time is also an $\mathfrak{F}$-stopping time.

Proposition 8.25 *If T is a finite stopping time, then $W(T)$ is $\mathfrak{F}_T$-measurable where $W(T, \omega)$ is defined as $W(T(\omega), \omega)$.*

The proof of this proposition relies on a simple, though important, approximation scheme.

Lemma 8.26 *Given any finite $\mathfrak{F}$-stopping time T, one can construct a non-increasing sequence of simple stopping times $T_1 \geq T_2 \geq \cdots$ such that $\lim_n T_n(\omega) = T(\omega)$ for all $\omega \in \Omega$. In addition, $\mathfrak{F}_T = \cap_n \mathfrak{F}_{T_n}$.*

[8.8] See Hunt [Hun57].

[8.9] Actually this is not necessary since Brownian motion is a.s. continuous but it would take too long to prove this fact here.

Proof Here is a receipt for the T_n's:

$$T_n(\omega) := \sum_{k=0}^{\infty} \left(\frac{k+1}{2^n} \right) \mathbf{1}_{[k2^{-n},(k+1)2^{-n})}(T(\omega)). \tag{8.44}$$

Since every interval of the form $[k2^{-n}, (k+1)2^{-n})$ is obtained by splitting into two an interval of the form $[j2^{-n+1}, (j+1)2^{-n+1})$, we have $T_n \geq T_{n+1}$ (why?); it is also clear that $T_n \geq T$.

To check that T_n is a stopping time, note that $\{T_n \leq (k+1)2^{-n}\} = \{T \leq (k+1)2^{-n}\} \in \mathfrak{F}_{(k+1)2^{-n}}$, since T is a stopping time. Now given any $t \geq 0$, we can find k and n such that $t \in [k2^{-n}, (k+1)2^{-n})$. Therefore, $\{T_n \leq t\} = \{T_n \leq k2^{-n}\} = \{T \leq k2^{-n}\} \in \mathfrak{F}_{k2^{-n}} \subseteq \mathfrak{F}_t$, which proves that the T_n's are nonincreasing simple stopping times. Moreover, T_n converges to T since $0 \leq T_n - T \leq 2^{-n}$. It remains to prove that $\mathfrak{F}_T \supseteq \cap_n \mathfrak{F}_{T_n}$; cf. Proposition 8.18 but replace $\mathfrak{F}^0$ by $\mathfrak{F}$ everywhere.

If $A \in \cap_n \mathfrak{F}_{T_n}$, then for all $n \geq 1$ and $t \geq 0$, $A \cap \{T_n \leq t\}$ is in $\mathfrak{F}_t$. Since $\lim_n T_n = T$, then $A \cap \{T \leq t\} = \cap_{\varepsilon>0} \cap_{n=1}^{\infty} (A \cap \{T_n \leq t+\varepsilon\})$ is in $\cap_{\varepsilon>0} \mathfrak{F}_{t+\varepsilon}$, which is equal to $\mathfrak{F}_t$ thanks to the right-continuity of the latter. □

Proof of Proposition 8.25 If T is a simple $\mathfrak{F}$-stopping time, then this is not hard to prove. Indeed, suppose that T takes values in $\{\tau_0, \tau_1, \ldots\}$. In this case, given any Borel set A and any $t \geq 0$,

$$\{W(T) \in A\} \cap \{T \leq t\} = \bigcup_{\substack{n \geq 0: \\ \tau_n \leq t}} \{W(\tau_n) \in A\} \cap \{T = \tau_n\} \in \mathfrak{F}_t. \tag{8.45}$$

For a general finite stopping time T, we can find simple stopping times $T_n \downarrow T$ (Lemma 8.26) with $\mathfrak{F}_T = \cap_n \mathfrak{F}_{T_n}$. Let C denote the collection of ω's for which $t \mapsto W(t, \omega)$ is continuous and recall that $\mathrm{P}(\mathsf{C}) = 1$. Then, for any open set $A \subseteq \mathbb{R}$,

$$\begin{aligned} &\{W(T) \in A\} \cap \mathsf{C} \cap \{T \leq t\} \\ &\quad = \bigcap_{\varepsilon>0 \text{ rational}} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{W(T_n) \in A\} \cap \mathsf{C} \cap \{T_n \leq t + \varepsilon\}. \end{aligned} \tag{8.46}$$

(Why?) Since T_n is a finite simple stopping time, $\{W(T_n) \in A\} \cap \{T_n \leq t\} \in \mathfrak{F}_t$. In particular, the completeness of $\mathfrak{F}_t$ shows that the above, and hence, $\{W(T) \in A\} \cap \{T \leq t\}$ are also in $\mathfrak{F}_t$. The collection of all $A \in \mathfrak{B}(\mathbb{R})$

such that $\{W(T) \in A\} \cap \{T \le t\} \in \mathfrak{F}_t$ is a monotone class that contains all open sets. This proves that $\{W(T) \in A\} \cap \{T \le t\}$ for all $A \in \mathfrak{B}(\mathbb{R})$ (Theorem 1.27). □

6 The Strong Markov Property

We are finally in position to state and prove the strong Markov property of Brownian motion.

Theorem 8.27 (The Strong Markov Property; Kinney [Kin53]) *If T is a finite $\mathfrak{F}$-stopping time where $\mathfrak{F}$ is the Brownian filtration, then $t \mapsto W(T+t) - W(T)$ is a Brownian motion that is independent of $\mathfrak{F}_T$.*

Proof I first prove this for simple stopping times, and then approximate, using Lemma 8.26, a general stopping time with simple ones.

Step 1. Simple Stopping Times.
If T is a simple stopping time, then there exist $\tau_0 \le \tau_1 \le \dots$ such that $T \in \{\tau_0, \tau_1, \dots\}$, a.s. Now for any $A \in \mathfrak{F}_T$, and for all $B_1, \dots, B_m \in \mathfrak{B}(\mathbb{R})$,

$$\begin{aligned} &\mathrm{P}\left\{\forall i = 1, \dots, m:\ W(T+t_i) - W(t_i) \in B_i\ ,\ A\right\} \\ &= \sum_{k=0}^{\infty} \mathrm{P}\left\{\forall i = 1, \dots, m:\ W(\tau_k + t_i) - W(\tau_k) \in B_i\ ,\ T = \tau_k\ ,\ A\right\}. \end{aligned} \tag{8.47}$$

But $A \cap \{T = \tau_k\} = A \cap \{T \le \tau_k\} \cap \{T \le \tau_{k-1}\}^{\complement}$ is in $\mathfrak{F}_{\tau_k}$ since $A \in \mathfrak{F}_T$. Therefore, by the Markov property (Theorem 8.9),

$$\begin{aligned} &\mathrm{P}\left\{\forall i = 1, \dots, m:\ W(T+t_i) - W(t_i) \in B_i\ ,\ A\right\} \\ &= \sum_{k=0}^{\infty} \mathrm{P}\left\{\forall i = 1, \dots, m:\ W(\tau_k + t_i) - W(\tau_k) \in B_i\right\} \\ &\quad \times \mathrm{P}\left\{T = \tau_k\ ,\ A\right\} \\ &= \mathrm{P}\left\{\forall i = 1, \dots, m:\ W(t_i) \in B_i\right\} \mathrm{P}(A). \end{aligned} \tag{8.48}$$

This proves the theorem in the case that T is a simple stopping time. Indeed, to deduce that $t \mapsto W(t+T) - W(T)$ is a Brownian motion, simply set $A := \mathbb{R}$. The asserted independence also follows since $A \in \mathfrak{F}_T$ is arbitrary.

Step 2. The General Case.
In the general case, we approximate T by simple stopping times as in Lemma 8.26. Namely, we find $T_n \downarrow T$—all simple stopping times—such that $\cap_n \mathfrak{F}_{T_n} = \mathfrak{F}_T$. Now for any $A \in \mathfrak{F}_T$, and for all open $B_1, \ldots, B_m \subseteq \mathbb{R}$,

$$\begin{aligned}
&\mathrm{P}\left\{\forall i = 1, \ldots, m: \ W(T + t_i) - W(t_i) \in B_i \ , \ A\right\} \\
&= \lim_{n\to\infty} \mathrm{P}\left\{\forall i = 1, \ldots, m: \ W(T_n + t_i) - W(t_i) \in B_i \ , \ A\right\} \\
&= \lim_{n\to\infty} \mathrm{P}\left\{\forall i = 1, \ldots, m: \ W(t_i) \in B_i\right\} \mathrm{P}(A).
\end{aligned} \tag{8.49}$$

In the first equation we used the fact that the B's are open and W is continuous, while in the second equation we used the fact that $A \in \mathfrak{F}_{T_n}$ for all n, together with the result of Step 1 applied to T_n. This proves the theorem. □

7 The Reflection Principle

The reflection principle is a prime example of how the strong Markov property (Theorem 8.27) can be applied to make nontrivial computations for the Brownian motion.

Theorem 8.28 (The Reflection Principle; Bachelier [Bac00]) *For any nonrandom $t > 0$, $\sup_{0\le s\le t} W(s)$ has the same distribution as $|W(t)|$. Equivalently, for all $a \ge 0$ and $t \ge 0$,*

$$\mathrm{P}\left\{\sup_{0\le s\le t} W(s) \ge a\right\} = \sqrt{\frac{2}{\pi t}} \int_a^\infty e^{-z^2/(2t)}\, dz. \tag{8.50}$$

Proof I will write out a proof carefully; this translates to a picture-proof that you are encouraged to discover on your own.

Define $T_a := \inf\{s \ge 0: \ W(s) \ge a\}$ where $\inf \varnothing := +\infty$. Thanks to Proposition 8.18, T_a is an $\bar{\mathfrak{F}}$- and hence an $\mathfrak{F}$-stopping time.

Step 1. T_a is a.s. Finite.
It is not hard to see that $T_a < +\infty$, almost surely. Here is a quick proof: By scaling (Theorem 8.9), for any $t > 0$,

$$\mathrm{P}\left\{W(t) \ge \sqrt{t}\right\} = \int_1^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx := c > 0. \tag{8.51}$$

Consequently, $\limsup_{t\to\infty} t^{-1/2}W(t) \ge 1$ with probability at least c (why?). But then this and the zero-one law (Exercise 8.5) together show that with probability one, $\limsup_{t\to\infty} t^{-1/2}W(t) \ge 1$. In particular, $\limsup_t W(t) = +\infty$, a.s. Since W is continuous a.s., it must then hit a at some finite time, a.s.; i.e., $T_a < +\infty$, a.s. Once again thanks to continuity, we also have that $W(T_a) = a$, a.s.

Step 2. Reflection via The Strong Markov Property.
Note that $\{\sup_{0\le s\le t} W(s) \ge a\} = \{T_a \le t\}$, which is $\mathfrak{F}_t$-measurable. Moreover, we can write

$$\begin{aligned}
&\mathrm{P}\{T_a \le t\} \\
&= \mathrm{P}\{T_a \le t\ ,\ W(t) \ge a\} + \mathrm{P}\{T_a \le t\ ,\ W(t) < a\} \\
&= \mathrm{P}\{W(t) \ge a\} \\
&\quad + \mathrm{P}\{T_a \le t\ ,\ W(T_a + (t - T_a)) - W(T_a) < 0\} \\
&= \mathrm{P}\{W(t) \ge a\} \\
&\quad + \mathrm{E}\left(\mathrm{P}\left\{W(T_a + (t - T_a)) - W(T_a) < 0 \,\middle|\, \mathfrak{F}_{T_a}\right\}\ ;\ T_a \le t\right),
\end{aligned} \tag{8.52}$$

since T_a is $\mathfrak{F}_{T_a}$-measurable (Proposition 8.18). On the other hand, by the strong Markov property (Theorem 8.27), $\mathrm{P}\{W(T_a + (t - T_a)) - W(T_a) < 0 \,|\, \mathfrak{F}_{T_a}\}$ is the probability that a Brownian motion independent of $\mathfrak{F}_{T_a}$ is below zero at time $t - T_a$, conditional on the value of T_a. The stated independence together with symmetry (Theorem 8.9) show that the said probability is a.s. equal to $\mathrm{P}\{W(T_a+(t-T_a))-W(T_a) > 0 \,|\, \mathfrak{F}_{T_a}\}$ (why a.s.?).[8.10] Therefore, we make this change and backtrack in the preceding display to get the following:

$$\begin{aligned}
&\mathrm{P}\{T_a \le t\} \\
&= \mathrm{P}\{W(t) \ge a\} \\
&\quad + \mathrm{E}\left(\mathrm{P}\left\{W(T_a + (t - T_a)) - W(T_a) > 0 \,\middle|\, \mathfrak{F}_{T_a}\right\}\ ;\ T_a \le t\right) \\
&= \mathrm{P}\{W(t) \ge a\} + \mathrm{P}\{T_a \le t\ ,\ W(T_a + (t - T_a)) - W(T_a) > 0\} \\
&= \mathrm{P}\{W(t) \ge a\} + \mathrm{P}\{T_a \le t\ ,\ W(t) > a\} = 2\mathrm{P}\{W(t) \ge a\},
\end{aligned} \tag{8.53}$$

as desired. Let me also point out that the last equality used the evident fact that $\mathrm{P}\{W(t) = a\} = 0$. By symmetry (Theorem 8.9), $2\mathrm{P}\{W(t) \ge a\} = \mathrm{P}\{W(t) \ge a\} + \mathrm{P}\{-W(t) \ge a\} = \mathrm{P}\{|W(t)| \ge a\}$, as desired. □

[8.10] In other words, we have reflected the post-T_a process to get another Brownian motion, whence the term "reflection principle."

The reflection principle has two curious consequences. The first is that while we expect Brownian motion to reach a level a at some finite time, this time has infinite expectation. That is,

Corollary 8.29 *If* $T_a := \inf\{s \geq 0 : \; W(s) = a\}$*, then for any* $a \neq 0$*,* $T_a < +\infty$*, a.s. but* $\mathrm{E}\{T_a\} = +\infty$*.*

Proof We have already seen that T_a is a.s. finite; let us prove that it has infinite expectation. Without loss of generality, we can assume that $a > 0$ (why?). Notice that thanks to Theorem 8.28 we have

$$\begin{aligned}\mathrm{P}\left\{T_a \geq t\right\} &= \mathrm{P}\left\{\sup_{0\leq s\leq t} W(s) \leq a\right\} = \mathrm{P}\left\{|W(t)| \leq a\right\} \\ &= \int_{-a}^{a} \frac{e^{-x^2/(2t)}}{\sqrt{2\pi t}}\, dx = \int_{-a/\sqrt{t}}^{a/\sqrt{t}} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dy.\end{aligned} \tag{8.54}$$

This shows that $\lim_{t\to\infty} \sqrt{t}\mathrm{P}\{T_a \geq t\} = a\sqrt{2/\pi}$; therefore, $\sum_{n=1}^{\infty} \mathrm{P}\{T_a \geq n\} = +\infty$, and Lemma 5.9 finishes the proof. □

The second surprising consequence of reflection principle is that Brownian motion started at zero crosses zero infinitely-many times immediately after starting out!

Corollary 8.30 *With probability one, we can find random times* $\sigma_n, \sigma_n' \downarrow 0$*, such that* $W(\sigma_n) > 0$ *and* $W(\sigma_n') < 0$*. In particular, given any* $\varepsilon > 0$*, with probability one, there are infinitely-many zeros of* W *in the time interval* $[0, \varepsilon]$*.*

Proof Thanks to the reflection principle (Theorem 8.28) given any $\varepsilon > 0$, $\mathrm{P}\{\sup_{s\leq\varepsilon} W(t) \leq 0\} = \mathrm{P}\{|W(\varepsilon)| \leq 0\}$, which is zero since $\mathrm{P}\{W(\varepsilon) = 0\} = 0$. Combining countably-many of the null sets leads to the statement that outside one null set, $\sup_{0\leq s\leq\varepsilon} W(s) > 0$ for all rational $\varepsilon > 0$. But $\varepsilon \mapsto \sup_{0\leq s\leq\varepsilon} W(s)$ is nondecreasing; therefore, we have shown that with probability one, $\sup_{0\leq s\leq\varepsilon} W(s) > 0$ for all $\varepsilon > 0$, and the null set does not depend on ε. Since $W(0) = 0$, there a.s. must exist a random sequence $\sigma_n \downarrow 0$ along which W is strictly positive. We can also apply this very statement to the Brownian motion $-W$, and obtain $\sigma_n' \downarrow 0$ along which $-W$ is strictly positive. This completes our proof. □

8 Exercises

Exercise 8.1 Prove the following: If X and Y are respectively $\mathbb{R}^n$- and $\mathbb{R}^m$-valued random variables, then X and Y are independent if and only if for all $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$,

$$\mathrm{E}\left\{e^{ia\cdot X+ib\cdot Y}\right\} = \mathrm{E}\left\{e^{iaX}\right\} \cdot \mathrm{E}\left\{e^{ib\cdot Y}\right\}. \tag{8.55}$$

(Hint: To prove that the display implies independence, you may appeal to the multidimensional inversion formula for characteristic functions.)

Exercise 8.2 Suppose for every $n = 1, 2, \ldots,$ $G^n := (G_1^n, \ldots, G_k^n)$ is an $\mathbb{R}^k$-valued centered normal random variable. Suppose further that $Q_{i,j} := \lim_{n\to\infty} \mathrm{E}\{G_i^n G_j^n\}$ exists and is finite. Then prove that Q is a symmetric nonnegative-definite matrix, and that G^n converges weakly to a centered normal random variable $G := (G_1, \ldots, G_k)$ whose covariance matrix is Q.

Exercise 8.3 Let (Z_1, Z_2, Z_3) be three independent random variables; Z_1 and Z_2 having a standard normal distribution, and $Z_3 = \pm 1$ with probability $\frac{1}{2}$ each. Define $X_1 := Z_3|Z_1|$ and $X_2 := Z_3|Z_2|$, and prove that X_1 and X_2 are each standard normal although (X_1, X_2) is not a Gaussian random variable.

Exercise 8.4 If W denotes a Brownian motion, then prove the following LIL's: With probability one,

$$\limsup_{t\to\infty} \frac{W(t)}{\sqrt{2t\ln\ln t}} = \limsup_{t\to 0} \frac{W(t)}{\sqrt{2t\ln\ln(\frac{1}{t})}} = 1. \tag{8.56}$$

(Hint: First prove that $\limsup_n W(n) \div \sqrt{2n\ln\ln n} = 1$, where n is an integer; cf. the law of the iterated logarithm for random walks (Theorem 7.48). Then devise a maximal inequality and use it to prove that not much happens in-between n and $n+1$. For the "$t \to 0$" case, use time-inversion.)

Exercise 8.5 If W denotes a Brownian motion, define the *tail σ-algebra* $\mathfrak{T}$ as follows: First, for any $t \geq 0$, define $\mathfrak{T}_t$ to be the P-completion of $\sigma\{W(u);\ u \geq t\}$. Then, define $\mathfrak{T} := \cap_{t\geq 0}\mathfrak{T}_t$.

1. Prove that $\mathfrak{T}$ is trivial; i.e., for any $A \in \mathfrak{T}$, $\mathrm{P}(A) \in \{0, 1\}$.

2. Let $\mathfrak{F}_t^0 := \sigma\{W(u); u \leq t\}$, define $\bar{\mathfrak{F}}_t$ to be the P-completion of $\mathfrak{F}_t^0$, and let $\mathfrak{F}_t$ be the right-continuous extension; i.e., $\mathfrak{F}_t := \cap \bar{\mathfrak{F}}_s$, where the intersection is taken over all rational $s > t$. Prove *Blumenthal's zero-one law*: $\mathfrak{F}_0$ is trivial; i.e., for all $A \in \mathfrak{F}_0$, $\mathrm{P}(A) \in \{0, 1\}$.

(Hint: Theorem 5.15.)

Exercise 8.6 In this exercise, we refine Theorem 8.15 by showing that with probability one, $\lim_{n\to\infty} W_n(t) = W(t)$ uniformly for all $t \in [0, 1]$.

1. Check that for for any fixed $n \geq 1$ and $t \in [0, 1]$, $\{W_m(t) - W_{2^n}(t);\ m \geq 2^n\}$ is a martingale.

2. Conclude that $m \mapsto \sup_{0\leq t\leq 1} |W_m(t) - W_{2^n}(t)|^2$ is a submartingale as m varies over $2^n, \ldots, 2^{n+1}$.

3. Prove that a.s., $W(t) = \lim_n W_n(t)$ uniformly over all $t \in [0, 1]$.

(Hint: In the last part you can use (8.32) in conjunction with Doob's inequality (Theorem 7.43).)

Exercise 8.7 Our proof of Theorem 8.16 can be refined to produce a stronger statement. Indeed, suppose $\alpha > \frac{1}{2}$ is fixed, and then prove that with probability one,

$$\lim_{t\to s} \frac{|W(s) - W(t)|}{|s - t|^\alpha} = +\infty, \qquad \forall s \in [0, 1]. \tag{8.57}$$

In other words, prove that there exists a null set outside which the above holds simultaneously for all $s \in [0, 1]$.
(Hint: First check that the very proof of Theorem 8.16 shows that this holds for $\alpha > \frac{3}{4}$. To improve this, subdivide every $D(j; n)$ into l subintervals, where l can be chosen as large as we please. The existing proof of Theorem 8.16 uses $l = 4$.)

Exercise 8.8 Given a fixed $s > 0$, consider the stopping time $\tau_s := \inf\{u \geq s :\ W(u) = 0\}$.

1. By first conditioning on $\mathfrak{F}_s$ and then appealing to the reflection principle (Theorem 8.28), prove that for all $x \geq s$,

$$\begin{aligned} \mathrm{P}\left\{\tau_s \leq x\right\} &= \int_{-\infty}^{\infty} \frac{e^{-z^2/(2s)}}{\sqrt{2\pi s}} \mathrm{P}\left\{|W(x-s)| \geq |z|\right\} dz \\ &= \int_{-\infty}^{\infty} \frac{e^{-z^2/(2s)}}{\sqrt{2\pi s}} \mathrm{P}\left\{|W(1)| \geq \frac{|z|}{\sqrt{x-s}}\right\} dz. \end{aligned} \tag{8.58}$$

2. Conclude that the distribution of $s^{-1}\tau_s$ is the same as that of $1+(g/G)^2$, where g and G are independent standard normal random variables. In particular, prove that the density function of τ_s is

$$f(x) = \begin{cases} \left[2\pi x\sqrt{s(x-s)}\right]^{-1}, & \text{if } x > s, \\ 0, & \text{otherwise.} \end{cases} \tag{8.59}$$

3. Characterize all $\beta \in \mathbb{R}$ such that $\mathrm{E}\{\tau_s^\beta\} < +\infty$.

Exercise 8.9 Given $a \neq 0$, define $T_a := \inf\{s \geq 0 : \ W(s) = a\}$.

1. Find the density function, as well as the characteristic function, of T_a. This gives rise to a so-called *stable distribution with index* $\frac{1}{2}$.

2. Show that the stochastic process $\{T_a; \ a \geq 0\}$ has i.i.d. increments.

(Hint: For the first part, study Corollary 8.29.)

Exercise 8.10 Let $T_0 := 0$, and successively define $T_{k+1} := \inf\{s > T_k : |W(s) - W(T_k)| = 1\}$ for $k = 1, 2, \ldots$.

1. Prove that the T_j's are stopping times.

2. Prove that the vectors $(W(T_{k+1}) - W(T_k), T_{k+1} - T_k)$ $(k = 0, 1, \ldots)$ are i.i.d.

3. Conclude that the process $k \mapsto W(T_k)$ is an embedding of a simple random walk inside Brownian motion. [In fact, Skorohod [Sko61] has shown that every mean-zero finite-variance random walk can be embedded inside Brownian motion.]

Chapter 9

A Taste of Stochastic Integration

As these notes started with measure theory and integration, it is only appropriate that they end with stochastic integration. Although it is one of the highlights of the theory of continuous-time stochastic processes, its analysis, and more generally the analysis of continuous-time processes, has inherent technical difficulties that are insurmountable in the amount of time that is left to us. Therefore, I will conclude these lectures with a very incomplete, somewhat nonrigorous, but hopefully coherent introduction to aspects of stochastic integration. You can learn much more about this topic by reading more specialized texts.

1 The Indefinite Itô Integral

Rather than present a general theory of stochastic integration, I will discuss a special case that is: (i) Broad enough to be applicable for our needs. (ii) Concrete enough so as to make the main ideas clear. The definitive treatment is Dellacherie and Meyer [DM82].

If $H := \{H(s);\, s \geq 0\}$ is a "nice" stochastic process, I more or less follow Itô [Itô44],[9.1] and construct the integral $\int H\, dW = \int_0^\infty H(s)\, W(ds)$ despite the fact that W is nowhere differentiable a.s. (Theorem 8.16). Now let us

[9.1] See also Bru and Yor [BY02] to learn about the recently-rediscovered work of W. Doeblin on stochastic integrals.

go ahead and officially *redefine* what we mean by a stochastic process in continuous-time.[9.2]

Definition 9.1 A *stochastic process* (or *process* for brevity) $X := \{X(t);\, t \geq 0\}$ is a product-measurable function $X : [0, \infty) \times \Omega \to \mathbb{R}$. We often write $X(t)$ in place of $X(t, \omega)$; this is similar to what we did in discrete-time.

Remark 9.2 Check that the Brownian motion of the previous section is indeed a stochastic process.

Now if H is nicely-behaved, then it stands to reason that we should define $\int H\, dW$ as $\lim_{n\to\infty} \mathcal{I}_n(H)$, where "$\lim_{n\to\infty}$" implies an as-yet-unspecified form of a limit, and[9.3]

$$\mathcal{I}_n(H) := \sum_{k=0}^{\infty} H\left(\frac{k}{2^n}\right) \times \left[W\left(\frac{k+1}{2^n}\right) - W\left(\frac{k}{2^n}\right)\right]. \tag{9.1}$$

It is abundantly clear that $\mathcal{I}_n(H)$ is a well-defined random variable if, for instance, H has compact support; i.e., $H(s) = 0$ for all s sufficiently large. The following performs some of the requisite book-keeping about $n \mapsto \mathcal{I}_n(H)$.

Lemma 9.3 *Suppose there exists a (random or nonrandom) $T > 0$ such that with probability one, $H(s) = 0$ for all $s \geq T$. Then, $\mathcal{I}_n(H)$ is a.s. a finite sum, and*

$$\begin{aligned}\mathcal{I}_{n+1}(H) - \mathcal{I}_n(H) = \sum_{j=0}^{\infty} &\left[H\left(\frac{2j+1}{2^{n+1}}\right) - H\left(\frac{j}{2^n}\right)\right] \\ &\times \left[W\left(\frac{j+1}{2^n}\right) - W\left(\frac{2j+1}{2^{n+1}}\right)\right].\end{aligned} \tag{9.2}$$

[9.2] This is a nonstandard definition, but reduces technical difficulties without endangering the essence of the theory. One can often show that our notion of a stochastic process is a consequence of much weaker technical assumptions on X; cf. Doob [Doo53, Chapter II] under the general heading of "separability."

[9.3] Notice the left-hand-rule approximation is being used here. This is the hallmark of Itô's theory of stochastic integration. In contrast to Riemann integration, in Itô's theory, the left-end-rule cannot be replaced by other rules (such as the midpoint- or the right-hand-rule) without changing the resulting stochastic integral.

Proof (Optional) The fact that the sum is finite is obvious. I will derive the stated identity for $\mathcal{I}_{n+1}(H) - \mathcal{I}_n(H)$. Throughout I will write $H_{k,n}$ in place of $H(k2^{-n})$, and $\Delta_{j,n}^{k,m} := W(j2^{-n}) - W(k2^{-m})$. Although this makes the reader's task a bit more difficult, that of the typesetter is greatly simplified.

Consider the expression $\mathcal{I}_{n+1}(H) = \sum_{k=0}^{\infty} H_{k,n+1}\Delta_{k+1,n+1}^{k,n+1}$, and split the sum according to whether $k = 2j$ or $k = 2j + 1$:

$$\begin{aligned}\mathcal{I}_{n+1}(H) &= \sum_{j=0}^{\infty} H_{j,n}\Delta_{2j+1,n+1}^{j,n} + \sum_{j=0}^{\infty} H_{2j+1,n+1}\Delta_{j+1,n}^{2j+1,n+1} \\ &= \sum_{j=0}^{\infty} H_{j,n}\Delta_{2j+1,n+1}^{j,n} + \sum_{j=0}^{\infty} H_{j,n}\Delta_{j+1,n}^{2j+1,n+1} \\ &\quad - \sum_{j=0}^{\infty} \left(H_{2j+1,n+1} - H_{j,n}\right)\Delta_{j+1,n}^{2j+1,n+1} \qquad (9.3) \\ &= \sum_{j=0}^{\infty} H_{j,n}\left(\Delta_{2j+1,n+1}^{j,n} + \Delta_{j+1,n}^{2j+1,n+1}\right) \\ &\quad - \sum_{j=0}^{\infty} \left(H_{2j+1,n+1} - H_{j,n}\right)\Delta_{j+1,n}^{2j+1,n+1}.\end{aligned}$$

Because $\Delta_{2j+1,n+1}^{j,n} + \Delta_{j+1,n}^{2j+1,n+1} = \Delta_{j+1,n}^{j,n}$, the first term is equal to $\mathcal{I}_n(H)$, whence the lemma. □

Definition 9.4 A stochastic process $H := \{H(s); t \geq 0\}$ is said to be *adapted* to the Brownian filtration $\mathfrak{F}$ if for each $s \geq 0$, $H(s)$ is $\mathfrak{F}_s$-measurable. It is a *compact-support process* if there exists a nonrandom $T \geq 0$ such that with probability one, $H(s) = 0$ for all $s \geq T$. Finally, given $p \geq 1$, H is said to be *Dini-continuous in* $L^p(\mathrm{P})$ if $H(s) \in L^p(\mathrm{P})$ for all $s \geq 0$, and $\int_0^1 \psi_p(r)r^{-1}\,dr < +\infty$, where $\psi_p(r) := \sup_{s,t:\,|s-t|\leq r} \|H(s) - H(t)\|_p$ denotes the *modulus of continuity* of H in $L^p(\mathrm{P})$.

Example 9.5

(a) Note that whenever H is compact-support, continuous, and a.s.-bounded by a nonrandom quantity, then it is a.s. uniformly continuous in $L^p(\mathrm{P})$ for any $p \geq 1$. In particular, $\psi_p(t) \to 0$ as $t \to 0$. The extra assumption of Dini-continuity in $L^p(\mathrm{P})$ states that in fact ψ_p has to converge to zero at some minimum rate. Here are some examples:

(b) Suppose H is (a.s.) differentiable with a derivative that satisfies $K := \sup_t \|H'(t)\|_p < +\infty$.[9.4] By the fundamental theorem of calculus, if $t \geq s \geq 0$, then $\|H(s)-H(s)\|_p \leq \int_s^t \|H'(r)\|_p \, dr \leq K|t-s|$. Therefore, $\psi_p(r) \leq Kr$, and H is easily seen to be Dini-continuous in $L^p(\mathrm{P})$.

(c) Another class of important examples is found by considering processes H of the form $H(s) := f(W(s))$, where f is a nonrandom differentiable function with $L := \sup_x |f'(x)| < +\infty$. In such a case, $|H(s)-H(t)| \leq L|W(s)-W(t)|$, and we have $\psi_p(r) \leq Lc_p\sqrt{r}$, where $c_p = \|Z\|_p$ where Z is a standard normal variable (why?). This yields the Dini-continuity of H in $L^p(\mathrm{P})$ for any $p \geq 1$.

(d) More generally still, suppose we have $H(s) := f(W(s), s)$, where $f(x,t)$ is nonrandom, differentiable in each variable, and satisfies: (i) There exists a nonrandom $T \geq 0$ such that $f(x,s) = 0$ for all $s \geq T$; and (ii) $M := \sup_{x,t} |\partial_x f(x,t)| + \sup_{x,t} |\partial_t f(x,t)| < +\infty$.[9.5] Then, $|H(s) - H(t)| \leq |f(W(s),s) - f(W(t),s)| + |f(W(t),s) - f(W(t),t)|$. Applying the fundamental theorem of calculus, we arrive at the following (why?):

$$|H(s)-H(t)| \leq M\left(|W(s)-W(t)| + |t-s|\right). \tag{9.4}$$

By Minkowski's inequality (Theorem 2.25), for any $p \geq 1$, $\|H(s) - H(s)\|_p \leq M(\|W(s)-W(t)\|_p + |t-s|) = M(c_p|t-s|^{1/2} + |t-s|)$, where $c_p = \|Z\|_p$ (why?). In particular, whenever $r \in [0,1]$, we have $\psi_p(r) \leq M(c_p+1)\sqrt{r}$, from which the Dini-continuity of H follows in any $L^p(\mathrm{P})$ $(p \geq 1)$.

Remark 9.6 (Cauchy Summability test) Dini-continuity in $L^p(\mathrm{P})$ is equivalent to the summability of $\psi_p(2^{-n})$. Indeed, we can write $\int_0^1 \psi_p(t)t^{-1}\, dt =$

[9.4] Since $(s,\omega) \mapsto H(s,\omega)$ is product-measurable, $\int |H'(r)|^p \, dr$ is a random variable, and hence $\|H'(t)\|_p$ are well-defined (check this!).

[9.5] As is customary, $\partial_z g(x,y,z,w,\cdots)$ denotes the derivative of g with respect to the variable z.

$\sum_{n=0}^{\infty}\int_{2^{-n-1}}^{2^{-n}} t^{-1}\psi_p(t)\,dt$. Because ψ_p is nondecreasing, for the t in this integral, $\psi_p(2^{-n-1}) \le \psi_p(t) \le \psi_p(2^{-n})$, and $2^n \le t^{-1} \le 2^{n+1}$. Therefore,

$$\frac{1}{2}\sum_{n=1}^{\infty}\psi_p\left(2^{-n}\right) = \frac{1}{2}\sum_{n=0}^{\infty}\psi_p\left(2^{-n-1}\right) \le \int_0^1 \frac{\psi_p(t)}{t}\,dt \le \sum_{n=0}^{\infty}\psi_p\left(2^{-n}\right). \tag{9.5}$$

In particular, H is Dini continuous in $L^p(\mathrm{P})$ if and only if $\sum_n \psi_p(2^{-n}) < +\infty$. This method is generally ascribed to A. L. Cauchy.

We can now define $\int H\,dW$ for adapted compact-support processes that are Dini-continuous in $L^2(\mathrm{P})$. I will then show how one can improve the assumptions on H.

Theorem 9.7 (Itô [Itô44]) *Suppose H is an adapted compact-support stochastic process that is Dini-continuous in $L^2(\mathrm{P})$. Then the stochastic integral $\int H\,dW := \lim_{n\to\infty}\mathcal{I}_n(H)$ exists in $L^2(\mathrm{P})$, and $\int H\,dW$ has mean zero and variance*

$$\mathrm{E}\left\{\left(\int H\,dW\right)^2\right\} = \mathrm{E}\left\{\int_0^{\infty} H^2(s)\,ds\right\}. \tag{9.6}$$

Finally, if $a, b \in \mathbb{R}$, and V is another adapted compact-support stochastic process that is Dini-continuous in $L^2(\mathrm{P})$, then with probability one,

$$\int (aH + bV)\,dW = a\int H\,dW + b\int V\,dW. \tag{9.7}$$

Definition 9.8 Equation (9.6) is called the *Itô isometry.*

Proof We employ Lemma 9.3, square both sides of the equation therein, and take expectations, and obtain

$$\begin{aligned}\|\mathcal{I}_{n+1}(H) - \mathcal{I}_n(H)\|_2^2 = &\sum_{0\le j\le 2^nT-1}\left\|H\left(\frac{2j+1}{2^{n+1}}\right) - H\left(\frac{j}{2^n}\right)\right\|_2^2 \\ &\times\left\|W\left(\frac{j+1}{2^n}\right) - W\left(\frac{2j+1}{2^{n+1}}\right)\right\|_2^2.\end{aligned} \tag{9.8}$$

To obtain this, we only need the facts that: (i) For $t \ge s$, $W(t) - W(s)$ is independent of $\mathfrak{F}_s$ (Theorem 8.27); and (ii) $H(u)$ is adapted to $\mathfrak{F}_s$ for $u \le s$. But $\|W(s) - W(t)\|_2^2 = t - s$. Hence, in the preceding display, the expectation involving Brownian motion is equal to $2^{-n} - 2^{-n-1} = 2^{-n-1}$, whereas the first expectation there is no more than $\psi_2^2(2^{-n-1})$. Consequently, $\|\mathcal{I}_{n+1}(H) - \mathcal{I}_n(H)\|_2 \le \sqrt{T}\psi_2(2^{-n-1})$. In particular, we can use the monotonicity of ψ_2 to see that for any nonrandom $N, M \ge 1$,

$$\begin{aligned} \|\mathcal{I}_{N+M}(H) - \mathcal{I}_N(H)\|_2 &\le \sum_{n=N+1}^{N+M-1} \|\mathcal{I}_{n+1}(H) - \mathcal{I}_n(H)\|_2 \\ &\le \sqrt{T} \sum_{n=N+1}^{N+M-1} \psi_2\left(2^{-n-1}\right). \end{aligned} \tag{9.9}$$

Thanks to this and Dini continuity (Remark 9.6) in $L^2(\mathrm{P})$, the above goes to zero as $N, M \to \infty$; i.e., $n \mapsto \mathcal{I}_n(H)$ is a Cauchy sequence in $L^2(\mathrm{P})$, which proves the assertion about $L^2(\mathrm{P})$-convergence.

To compute $\mathrm{E}\{\int H\, dW\}$, I merely note that $\mathrm{E}\{\mathcal{I}_n(H)\} = 0$ (why?); this is a consequence of the fact that for $t \ge s$, $W(t) - W(s)$ has mean zero, and is independent of $\mathfrak{F}_s$, whereas $H(u)$ is $\mathfrak{F}_s$-measurable for each $u \le s$. Similarly, we can prove (9.6):

$$\begin{aligned} \mathrm{E}\left\{\left(\int H\, dW\right)^2\right\} &= \lim_{n\to\infty} \|\mathcal{I}_n(H)\|_2^2 \\ &= \lim_{n\to\infty} \sum_{k=0}^{\infty} \mathrm{E}\left\{H^2\left(k2^{-n}\right)\right\} 2^{-n} \\ &= \mathrm{E}\left\{\int_0^\infty H^2(s)\, ds\right\}, \end{aligned} \tag{9.10}$$

where the many exchanges of limits and integrals are all justified by the compact-support assumption on H, together with the continuity of the function $t \mapsto \|H(t)\|_2$ (check this!).

Finally, I need to verify (9.7); but this follows from the linearity of $H \mapsto \mathcal{I}_n(H)$ and the existence of $L^2(\mathrm{P})$-limits. □

We now drop many of the technical assumptions in Theorem 9.7.

Theorem 9.9 (Itô [Itô44]) *Suppose that H is an adapted stochastic process, and $\mathrm{E}\{\int_0^\infty H^2(s)\, ds\} < +\infty$. Then one can define a stochastic integral $\int H\, dW$ that has mean zero and variance $\mathrm{E}\{\int_0^\infty H^2(s)\, ds\}$. Moreover, if V*

is another such integrand-process, then for all $a, b \in \mathbb{R}$

$$\begin{aligned}\int (aH + bV)\, dW &= a \int H\, dW + b \int V\, dW, \quad \textit{a.s.} \\ \mathrm{E}\left\{ \int H\, dW \cdot \int V\, dW \right\} &= \mathrm{E}\left\{ \int_0^\infty H(s)V(s)\, ds \right\}.\end{aligned} \tag{9.11}$$

Throughout, let m denote the Lebesgue measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, and let $L^2(m \times \mathrm{P})$ denote the corresponding product L^2-space. In particular, note that

$$\mathrm{E}\left\{ \int_0^\infty H^2(s)\, ds \right\} = \iint_{\Omega \times [0,\infty)} H^2\, dm\, d\mathrm{P} = \|H\|^2_{L^2(m\times\mathrm{P})}, \tag{9.12}$$

and $\mathrm{E}\{\int H(s)V(s)\, ds\}$ is the $L^2(m \times \mathrm{P})$-inner product between H and V.

The following technical result is the main step in constructing the general stochastic integral.

Proposition 9.10 *Given any stochastic process $H \in L^2(m \times \mathrm{P})$ we can find stochastic processes $H_1, H_2, \cdots$, all compact-support and Dini-continuous in $L^2(\mathrm{P})$, such that* $\lim_n H_n = H$ *in $L^2(m \times \mathrm{P})$.*

Theorem 9.9 follows from immediately from this.

Proof of Theorem 9.9 Thanks to Proposition 9.10 we can find adapted stochastic processes H_n that are compact-support Dini-continuous in $L^2(\mathrm{P})$, and converge to H in $L^2(m \times \mathrm{P})$. Thanks to the Itô isometry (equation 9.6), $\int H_n\, dW$ is a Cauchy sequence in $L^2(\mathrm{P})$, since H_n is a Cauchy sequence in $L^2(m \times \mathrm{P})$. Consequently, $\int H\, dW := \lim_n \int H_n\, dW$ exists in $L^2(\mathrm{P})$. The properties of $\int H\, dW$ follow readily from those of $\int H_n\, dW$, and the $L^2(\mathrm{P})$-convergence that we proved earlier. □

I will conclude this section by proving the one remaining proposition.

Proof of Proposition 9.10 (Optional) I will proceed in three steps, each of which reduces the problem to a more restrictive class of processes H.

Step 1. Reduction to the Compact-Support Case.
Define $H_n(t) := H(t)\mathbf{1}_{[0,n]}(t)$, and note that H_n is an adapted compact-support stochastic process. Moreover,

$$\lim_{n\to\infty} \|H - H_n\|^2_{L^2(m\times\mathrm{P})} = \lim_{n\to\infty} \mathrm{E}\left\{ \int_n^\infty H^2(s)\, ds \right\} = 0. \tag{9.13}$$

In other words, for the remainder of the proof, we can and will assume without loss of generality that H is also compact-support.

Step 2. Reduction to the L^2-Bounded Case.
I first extend the definition of H to all $\mathbb{R}$ by assigning $H(t) = 0$ if $t < 0$. Next, define for all $n \geq 1$,

$$H_n(t) := n \int_{t-(1/n)}^{t} H(s)\, ds, \qquad \forall t \geq 0. \tag{9.14}$$

You should check that H is an adapted stochastic process. Moreover, for all $t \geq T+1$, $H_n(t) = 0$, so that H_n is also compact-support. Next, I claim that H_n is bounded in $L^2(\mathrm{P})$; i.e., $\sup_t \|H_n(t)\|_2 < +\infty$. Indeed, by the Cauchy–Bunyakovsky–Schwarz inequality (Corollary 2.26), and by the Fubini–Tonelli theorem (Theorem 3.6),

$$\begin{aligned} \sup_{t\geq 0} \|H_n(t)\|_2^2 &\leq n \sup_{t\geq 0} \int_{t-(1/n)}^{t} \|H(s)\|_2^2\, ds \\ &\leq n \int_0^\infty \|H(s)\|_2^2\, ds = n\|H\|_{L^2(m\times\mathrm{P})}^2. \end{aligned} \tag{9.15}$$

(Why?) It remains to prove that H_n converges in $L^2(m \times \mathrm{P})$ to H.

Since $\int_0^\infty H^2(s)\, ds < +\infty$ a.s., then thanks to the Lebesgue differentiation theorem (Theorem 7.49), with probability one, $H_n(t) \to H(t)$ for almost every $t \geq 0$. Therefore, by Fubini–Tonelli (Theorem 3.6), $\lim_n H_n = H$, $(m\times \mathrm{P})$-almost surely (why?). According to the dominated convergence theorem (Theorem 2.22), to finish this step we need to only prove that $\sup_n |H_n| \in L^2(m \times \mathrm{P})$. In fact, I will prove (9.17) below which is slightly stronger still.

Now note that $\sup_n |H_n| \leq \mathcal{M}H$, where the latter is the "maximal function,"

$$\mathcal{M}H(t) := \sup_{n\geq 1} \left(n \int_{t-(1/n)}^{t} |H(s)|\, ds \right), \qquad \forall t \geq 0. \tag{9.16}$$

For each ω, $\mathcal{M}H(t + n^{-1})$ is our good-old Hardy–Littlewood maximal function of H, and you should check that $\mathcal{M}H$ is an adapted stochastic process. In addition, by applying Corollary 7.51 with $p = 2$ we obtain, $\int_0^\infty |\mathcal{M}H(s)|^2\, ds \leq 64 \int_0^\infty H^2(s)\, ds$. This is useful only if the right-hand side is finite. But since $H(t) = 0$ for all $t \geq T$ and $\sup_t \|H(t)\|_2 < \infty$, the right-hand side of the preceding inequality is finite for almost-all ω. In particular, we can appeal to Fubini–Tonelli (Theorem 3.6) to take expectations, and then square-roots to deduce that

$$\|\mathcal{M}H\|_{L^2(m\times\mathrm{P})} \le 8\|H\|_{L^2(m\times\mathrm{P})}, \tag{9.17}$$

which is finite. This is the desired inequality, and reduces the problem to H's that are bounded in $L^2(\mathrm{P})$ and compact-support.

Step 3. The Conclusion.
Finally, if H is bounded in $L^2(\mathrm{P})$ and compact-support, then we define H_n by (9.14) and note that H_n is differentiable, and $H_n'(t) = n\{H(t) - H(t - n^{-1})\}$. Therefore, $\sup_t \|H_n'(t)\|_2 \le 2n \sup_t \|H(t)\|_2 < +\infty$, and part (b) of Example 9.5 proves the asserted Dini-continuity of H_n. On the other hand, the argument developed in Step 2 proves that $H_n \to H$ in $L^2(m \times \mathrm{P})$, and this concludes the proof. □

2 Continuous Martingales in $L^2(\mathrm{P})$

The theories of continuous-time martingale and stochastic integration are intimately connected. Thus, before proceeding further, we take a side-step, and have a quick look at martingale-theory in continuous-time. To avoid unnecessary abstraction, $\mathfrak{F}$ will denote the Brownian filtration throughout.[9.6]

Definition 9.11 A process $M := \{M(t);\, t \ge 0\}$ is a (continuous-time) *martingale* if:

1. For all $t \ge 0$, $M(t) \in L^1(\mathrm{P})$.

2. If $t \ge s \ge 0$, then $\mathrm{E}\{M(t)\,|\,\mathfrak{F}(s)\} = M(s)$, a.s.

M is said to be a *continuous L^2-martingale* if $t \mapsto M(t)$ is almost-surely continuous, and $M(t) \in L^2(\mathrm{P})$ for all $t \ge 0$.

Essentially all of the theory of martingales in discrete-time has continuous-time translations for continuous $L^2(\mathrm{P})$-martingales. Here is a first sampler.

Theorem 9.12 (Optional Stopping) *If M is a continuous L^2-martingale and $S \le T$ are bounded $\mathfrak{F}$-stopping times, then $\mathrm{E}\{M(T)\,|\,\mathfrak{F}_S\} = M(S)$, a.s.*

[9.6] This section's proofs are optional reading.

Proof (Optional) Throughout, choose and fix some nonrandom $K > 0$ such that almost-surely, $T \leq K$.

This is a consequence of Theorem 7.33 if S and T are simple stopping times. In general, let $S_n \downarrow S$ and $T_n \downarrow T$ be the simple stopping times of Lemma 8.26, and note that the condition $S \leq T$ imposes $S_m \leq T_n$ for all $n \leq m$ (why?). We have already seen that for all $n \geq m$,

$$\mathrm{E}\left\{ M(T_n) \,\middle|\, \mathfrak{F}_{S_m} \right\} = M(S_m), \qquad \text{a.s.} \tag{9.18}$$

Moreover, this very argument implies that $M(T_1), M(T_2), \ldots$ is a (discrete-time) martingale in its own filtration. Since $T_n \leq T + 2^{-n} \leq K + 2^{-n}$, by Exercise 7.9,

$$\mathrm{E}\left\{ \sup_{n\geq 1} M^2(T_n) \right\} \leq 4 \sup_{n\geq 1} \mathrm{E}\left\{ M^2(T_n) \right\} \leq 4\mathrm{E}\left\{ M^2\left(K + \frac{1}{2} \right) \right\}, \tag{9.19}$$

which is finite. Since M is continuous, a.s., $M(T_n) \to M(T)$, a.s.. Therefore, by the dominated convergence theorem (Theorem 2.22), we also have $M(T_n) \to M(T)$ in $L^2(\mathrm{P})$. This and conditional Fatou's lemma (Theorem 7.6) together imply that

$$\begin{aligned} &\lim_{n\to\infty} \left\| \mathrm{E}\left\{ M(T_n) \,\middle|\, \mathfrak{F}_{S_m} \right\} - \mathrm{E}\left\{ M(T) \,\middle|\, \mathfrak{F}_{S_m} \right\} \right\|_2 \\ &\quad \leq \lim_{n\to\infty} \left\| M(T) - M(T_n) \right\|_2 = 0. \end{aligned} \tag{9.20}$$

Therefore, by (9.18), $M(S_m) = \mathrm{E}\{M(T) \,|\, \mathfrak{F}_{S_m}\}$, a.s. To finish the proof, we simply let $m \to \infty$, and appeal to the time-reversed martingale convergence theorem (in discrete time; Theorem 7.47). □

The following is a related result whose proof is relegated to the exercises.

Theorem 9.13 (Doob's Inequalities) *If M is a continuous L^2-martingale, then for all $\lambda, t > 0$,*

$$\mathrm{P}\left\{ \sup_{0\leq s\leq t} |M(s)| \geq \lambda \right\} \leq \frac{1}{\lambda} \mathrm{E}\left\{ |M(t)| \,;\; \sup_{0\leq s\leq t} |M(s)| \geq \lambda \right\}. \tag{9.21}$$

In particular, for all $p > 1$,

$$\mathrm{E}\left\{ \sup_{0\leq s\leq t} |M(s)|^p \right\} \leq \left(\frac{p}{p-1} \right)^p \mathrm{E}\left\{ |M(t)|^p \right\}. \tag{9.22}$$

3 The Definite Itô Integral

It is a good time to also mention the definite Itô integral, which is simply defined as $\int_0^t H\,dW := \int H\mathbf{1}_{[0,t)}\,dW$ for all adapted processes H such that $\mathrm{E}\{\int_0^t H^2(s)\,ds\} < +\infty$ for all $t \geq 0$. This defines a collection of random variables $\int_0^t H\,dW$—one for each $t \geq 0$. The following is of paramount importance, since it says something about the properties of the random function $t \mapsto \int_0^t H\,dW$.

Theorem 9.14 *If H is an adapted process such that $\mathrm{E}\{\int_0^t H^2(s)\,ds\} < +\infty$, then we can construct the process $t \mapsto \int_0^t H\,dW$ such that it is a continuous L^2-martingale.*

Proof (Optional) According to Theorem 9.9, $\int_0^t H\,dW$ exists, so we can proceed by verifying the assertions of the theorem. This will be done in three steps.

Step 1. Reduction to H that is Dini-Continuous in $L^2(\mathrm{P})$.
Suppose we have proved the theorem for all processes H that are adapted and Dini-continuous in $L^2(\mathrm{P})$. In this first step we prove that this implies the remaining assertions of the theorem.

Let H be an adapted process such that for all $t \geq 0$, $\mathrm{E}\{\int_0^t H^2(s)\,ds\} < +\infty$. We can find adapted processes H_n that are Dini-continuous in $L^2(\mathrm{P})$ and

$$\lim_{n\to\infty} \mathrm{E}\left\{\int_0^t \left[H_n(s) - H(s)\right]^2\,ds\right\} = 0. \tag{9.23}$$

Indeed, we can apply Proposition 9.10 to $H\mathbf{1}_{[0,t]}$, and use the recipe of the said proposition for H_n. Then apply the proposition to $H_n\mathbf{1}_{[0,t]}$. This shows that in fact H_n can even be chosen independently of t as well.

By the Itô isometry (equation 9.6),

$$\begin{aligned}\lim_{n\to\infty} \mathrm{E}&\left\{\left(\int_0^t H\,dW - \int_0^t H_n\,dW\right)^2\right\}\\ &= \lim_{n\to\infty} \mathrm{E}\left\{\int_0^t \Big(H(s) - H_n(s)\Big)^2\,ds\right\} = 0.\end{aligned} \tag{9.24}$$

But $\int_0^t H_n\,dW - \int_0^t H_{n+k}\,dW = \int_0^t (H_n - H_{n+k})\,dW$, a.s., and is a continuous L^2-martingale. Therefore, by Doob's maximal inequality (Theorem 9.13), for any nonrandom but fixed $T > 0$,

$$\lim_{n\to\infty} \mathrm{E}\left\{ \sup_{0\le t\le T} \left(\int_0^t H_n\, dW - \int_0^t H_{n+k}\, dW \right)^2 \right\} = 0. \tag{9.25}$$

In particular, for each $T > 0$, there exists a stochastic process $X := \{X(t); t \geq 0\}$ and a (random) subsequence $n' \to \infty$ such that with probability one, $\lim_{n'\to\infty} \sup_{0\le t\le T} |\int_0^t H_{n'}\, dW - X(t)| = 0$. Moreover, the same uniform convergence holds in $L^2(\mathrm{P})$, and along the original subsequence $n \to \infty$. Consequently, this and (9.24) together show that X is a particular construction of $t \mapsto \int_0^t H\, dW$ that is a.s.-continuous and adapted. In other words, X is (obviously) adapted and a.s.-continuous, but it also satisfies

$$\mathrm{P}\left\{ X(t) = \int_0^t H\, dW \right\} = 1, \quad \forall t \geq 0. \tag{9.26}$$

(Why?) Finally, $X(t) \in L^2(\mathrm{P})$ for all $t \geq 0$, so it remains to prove that X is a martingale. But remember that we are assuming that $t \mapsto \int_0^t H_n\, dW$ is a martingale.

By the conditional Jensen inequality (7.6), and by $L^2(\mathrm{P})$-convergence,

$$\begin{aligned} &\left\| \mathrm{E}\{X(t+s) \,|\, \mathfrak{F}_s\} - \mathrm{E}\left\{ \int_0^{t+s} H_n\, dW \,\middle|\, \mathfrak{F}_s \right\} \right\|_2 \\ &\quad = \left\| \mathrm{E}\left\{ X(t+s) - \int_0^{t+s} H_n\, dW \,\middle|\, \mathfrak{F}_s \right\} \right\|_2 \\ &\quad \leq \left\| X(t+s) - \int_0^{t+s} H_n\, dW \right\|_2, \end{aligned} \tag{9.27}$$

which goes to zero as $n \to \infty$. On the other hand, since $t \mapsto \int_0^t H_n\, dW$ is a martingale, this also shows that $\int_0^t H_n\, dW \to \mathrm{E}\{X(t+s) \,|\, \mathfrak{F}_s\}$ in $L^2(\mathrm{P})$. But we have already seen that $\int_0^t H_n\, dW \to X(t)$ in $L^2(\mathrm{P})$. Therefore, with probability one, $\mathrm{E}\{X(t+s) \,|\, \mathfrak{F}_s\} = X(s)$; i.e., X is martingale as claimed.

Step 2. A Continuous Martingale in the Dini-Continuous Case. Now we suppose that H is in addition Dini-continuous in $L^2(\mathrm{P})$, and prove the theorem in this special case. Together with Step 1, this completes the proof. The argument is based on a trick. Define,

$$\begin{aligned}\mathcal{J}_n(H)(t) := \sum_{0\le k<2^n t-1} & H\left(\frac{k}{2^n}\right)\times\left[W\left(\frac{k+1}{2^n}\right)-W\left(\frac{k}{2^n}\right)\right]\\ &+H\left(\frac{\lfloor 2^n t-1\rfloor}{2^n}\right)\times\left[W(t)-W\left(\frac{\lfloor 2^n t-1\rfloor}{2^n}\right)\right].\end{aligned}\tag{9.28}$$

This is a minor variant of $\mathcal{I}_n(H\mathbf{1}_{[0,t]})$. Indeed, you should check that

$$\begin{aligned}&\mathcal{J}_n(H)(t)-\mathcal{I}_n\left(H\mathbf{1}_{[0,t]}\right)\\ &\quad=H\left(\frac{\lfloor 2^n t-1\rfloor}{2^n}\right)\times\left[W(t)-W\left(\frac{\lfloor 2^n t-1\rfloor+1}{2^n}\right)\right],\end{aligned}\tag{9.29}$$

whose $L^2(\mathrm{P})$-norm goes to zero as $n\to\infty$. But $\mathcal{J}_n(H)$ is also a stochastic process that is (a) adapted, and (b) continuous in t. In fact, it is also a martingale. Here is why: Suppose $t \ge s \ge 0$. Then there exist integers $0\le k\le K\le 2^n s-1$ such that $s\in D(k;n) := [k2^{-n},(k+1)2^{-n})$ and $t\in D(K;n)$. Then,

$$\begin{aligned}&\mathcal{J}_n(H)(t)-\mathcal{J}_n(H)(s)\\ &\quad=\sum_{k\le j<K} H\left(\frac{j}{2^n}\right)\times\left[W\left(\frac{j+1}{2^n}\right)-W\left(\frac{j}{2^n}\right)\right]\\ &\quad\quad+H\left(\frac{K}{2^n}\right)\times\left[W(t)-W\left(\frac{K}{2^n}\right)\right]\\ &\quad\quad-H\left(\frac{k}{2^n}\right)\times\left[W(s)-W\left(\frac{k}{2^n}\right)\right],\end{aligned}\tag{9.30}$$

where $\sum_{k\le j<k}(\cdots) := 0$ (in the case that $k=K$). Since W has independent increments, $\mathrm{E}\{[\cdots]\,|\,\mathfrak{F}_s\}=0$, a.s. where $[\cdots]$ is any of the terms of the preceding display in the square-brackets. The adaptedness of H and Corollary 8.10 together show that $\mathrm{E}\{\mathcal{J}_n(H)(t)-\mathcal{J}_n(H)(s)\,|\,\mathfrak{F}_s\}=0$, a.s., which then proves the martingale property.

Step 3. The Conclusion.

To finish the proof, suppose H is an adapted process that is Dini-continuous in $L^2(\mathrm{P})$. A calculation similar to that of Lemma 9.3 reveals that for any nonrandom $T>0$,

$$\lim_{n\to\infty}\sup_{0\le t\le T}\mathrm{E}\left\{\Big(\mathcal{J}_{n+1}(H)(t)-\mathcal{J}_n(H)(t)\Big)^2\right\}=0.\tag{9.31}$$

Therefore, by Doob's maximal inequality (Theorem 9.13),

$$\lim_{n\to\infty} \mathrm{E}\left\{ \sup_{0\le t\le T} \Big(\mathcal{J}_{n+1}(H)(t) - \mathcal{J}_n(H)(t) \Big)^2 \right\} = 0. \tag{9.32}$$

This implies that a subsequence of $\mathcal{J}_n(H)$ converges a.s. and uniformly for all $t \in [0,T]$ to some process X. Since $\mathcal{J}_n(H)$ is a continuous process, X is necessarily continuous a.s. Furthermore, the argument applied in Step 1 shows that here too X is a martingale. Finally, we have already seen that for any fixed $t \ge 0$, $\mathcal{J}_n(H)(t) - \mathcal{I}_n(H\mathbf{1}_{[0,t]}) \to 0$ in $L^2(\mathrm{P})$. Since $\mathcal{I}_n(H\mathbf{1}_{[0,t]}) \to \int_0^t H\,dW$ in $L^2(\mathrm{P})$, this shows that $\mathrm{P}\{X(t) = \int_0^t H\,dW\} = 1$, which proves the result. □

4 Quadratic Variation

I now elaborate a little on quadratic variation; cf. Theorem 8.9. Quadratic variation is a central theme in continuous-time martingale theory, but it requires too much time to study properly. Therefore, we will only develop the portions for which we have immediate use.

Throughout, we define the *second-order* analogue of $\mathcal{I}_n$ (9.1);

$$\mathcal{Q}_n(H) := \sum_{k=0}^{\infty} H\left(\frac{k}{2^n}\right) \times \left[W\left(\frac{k+1}{2^n}\right) - W\left(\frac{k}{2^n}\right) \right]^2. \tag{9.33}$$

Theorem 9.15 *Suppose H is an adapted compact-support process that is uniformly continuous in $L^2(\mathrm{P})$; i.e., $\lim_{r\to 0}\psi_2(r) = 0$. Then, $\lim_{n\to\infty} \mathcal{Q}_n(H) = \int_0^\infty H(s)\,ds$ in $L^2(\mathrm{P})$.*

Proof To simplify the notation, I write for all integers $k \ge 0$ and $n \ge 1$,

$$H_{j,n} := H\left(\frac{j}{2^n}\right), \quad d_{k,n} := W((k+1)2^{-n}) - W(k2^{-n}). \tag{9.34}$$

Recall next that we can find a nonrandom $T > 0$ such that for all $s \ge T$, $H(s) = 0$, a.s. Throughout, we keep such a T fixed.

Step 1. Approximating the Lebesgue Integral.
I begin by proving that the Riemann–integral approximation of the Lebesgue

integral $\int_0^\infty H(s)\, ds$ converges in $L^2(\mathrm{P})$. Namely, note that

$$\begin{aligned} &\left| \sum_{k=0}^{\infty} H_{k,n} 2^{-n} - \int_0^\infty H(s)\, ds \right| \\ &\qquad \leq \sum_{0 \leq k \leq 2^n T - 1} \int_{k2^{-n}}^{(k+1)2^{-n}} \left| H_{k,n} - H(s) \right| \, ds, \end{aligned} \tag{9.35}$$

since if we remove the absolute values, the preceding becomes an identity. In particular, apply Minkowski's inequality (Theorem 2.25) to deduce that

$$\begin{aligned} &\left\| \sum_{k=0}^{\infty} H_{k,n} 2^{-n} - \int_0^\infty H(s)\, ds \right\|_2 \\ &\qquad \leq \sum_{0 \leq k \leq 2^n T - 1} 2^{-n} \psi_2\left(2^{-n}\right) \leq T \psi_2\left(2^{-n}\right). \end{aligned} \tag{9.36}$$

As $n \to \infty$, the above converges to zero, and Step 1 follows.

Step 2. Completing the Proof.
Note that $H(k2^{-n})$ is independent of $d_{k,n}$, and the latter has mean zero and variance 2^{-n}. Therefore, $\mathcal{Q}_n(t) - \sum_{k=0}^\infty H_{k,n} 2^{-n} = \sum_{k=0}^\infty H_{k,n} \left[d_{k,n}^2 - 2^{-n} \right].$ Next, we square this and take expectations.

$$\begin{aligned} &\left\| \mathcal{Q}_n(t) - \sum_{k=0}^{\infty} H_{k,n} 2^{-n} \right\|_2^2 \\ &\quad = \sum_{0 \leq k \leq 2^n T - 1} \mathrm{E}\left\{ H_{k,n}^2 \right\} \mathrm{E}\left\{ \left[d_{k,n}^2 - 2^{-n} \right]^2 \right\} \\ &\qquad + 2 \sum\sum_{0 \leq j < k \leq 2^n T - 1} \mathrm{E}\left\{ H_{k,n} H_{j,n} \left[d_{k,n}^2 - 2^{-n} \right] \left[d_{j,n}^2 - 2^{-n} \right] \right\} \\ &\quad = \sum_{0 \leq k \leq 2^n T - 1} \mathrm{E}\left\{ H_{k,n}^2 \right\} \mathrm{E}\left\{ \left[d_{k,n}^2 - 2^{-n} \right]^2 \right\} \\ &\qquad + 2 \sum\sum_{0 \leq j < k \leq 2^n T - 1} \mathrm{E}\left\{ H_{k,n} H_{j,n} \left[d_{j,n}^2 - 2^{-n} \right] \right\} \times \mathrm{E}\left\{ d_{k,n}^2 - 2^{-n} \right\} \\ &\quad = \sum_{0 \leq k \leq 2^n T - 1} \mathrm{E}\left\{ H_{k,n}^2 \right\} \mathrm{E}\left\{ \left[d_{k,n}^2 - 2^{-n} \right]^2 \right\}. \end{aligned} \tag{9.37}$$

But $d_{k,n}$ is normal with mean zero and variance 2^{-n}; so $2^{-n/2} d_{k,n}$ is standard normal, and so, $[d_{k,n}^2 - 2^{-n}]$ has the same distribution as $2^{-n}[Z^2 - 1]$ where Z is standard normal. But $\mathrm{E}\{[Z^2-1]^2\} = \mathrm{E}\{Z^4\}-1 = 2$, thanks to footnote 8.5. Therefore, $\|\mathcal{Q}_n(t) - \sum_k H_{k,n}2^{-n}\|_2^2 = 2\cdot 4^{-n}\sum_{0\le k\le 2^nT-1}\|H_{k,n}\|_2^2$. But Dini-continuity insures that $t\mapsto \|H(t)\|_2$ is continuous, and hence bounded on $[0,T]$ by some constant K_T. Thus, $\|\mathcal{Q}_n(t)-\sum_k H_{k,n}2^{-n}\|_2^2 \le TK_T^2 2^{-n+1}\to 0$. This and Step 1 together prove the result. □

5 Itô's Formula and Two Applications

The chain rule of calculus states that given two continuously-differentiable functions f and g, $(f\circ g)' = f'(g)g'$. In its integrated form, this is integration-by-parts, and states that for all $t\ge s\ge 0$ (say),

$$f(g(t)) - f(g(s)) = \int_s^t f'(g(u))g'(u)\,du. \tag{9.38}$$

To cite a typical example, let me mention that when we use $f(x)=x^2$, this yields $g^2(t)-g^2(0) = \int_0^t g\,dg$, where $dg(s)$ is formally the same thing as $g'(s)\,ds$. Itô's formula tells us what happens if we replace g by the nowhere-differentiable function W. A consequence of this is that $W^2(t)-W^2(0) = \int_0^t W\,dW + \frac{1}{2}t$. Consequently, the stochastic integration-by-parts formula has an extra $\frac{1}{2}t$ factor.

Theorem 9.16 (Itô's Formula; [Itô44]) *If $f:\mathbb{R}\to\mathbb{R}$ has two continuous derivatives, then for all $t\ge s\ge 0$, the following holds a.s.:*

$$f(W(t)) - f(W(s)) = \int_s^t f'(W(r))\,W(dr) + \frac{1}{2}\int_s^t f''(W(r))\,dr, \tag{9.39}$$

provided that there exist constants $A,B>0$ such that $|f'(x)|\le Ae^{B|x|}$.

Remark 9.17

1. In other words, the nowhere-differentiability of W forces us to replace the right-hand side of (9.38) with a stochastic integral plus a second-derivative term.

2. Itô's formula continues to hold even if we assume only that f'' exists almost-everywhere, and that $\int_s^t (f'(W(r))^2\,dr < +\infty$, a.s. Of course, then we have to make sense of the stochastic integral, etc. Rather than prove such refinements here, I urge you to have a look at [DM82] for a definitive account.

3. The strange-looking condition on f' ensures that it does not grow too fast. This condition can be removed nearly altogether but this development rests on introducing a new family of processes that are called local martingales. To see how this exponential-growth condition comes up in the context of Theorem 9.9, note that for $\int_0^t f'(W(s))\,W(ds)$ to be well defined, we need $\mathcal{E} := \mathrm{E}\{\int_0^t [f'(W(s))]^2\,ds\}$ be finite for all $t \geq 0$. But then the said condition on f' implies:

$$\begin{aligned}\mathcal{E} &= \int_{-\infty}^{\infty}\int_0^t \Big[f'(x)\Big]^2 \frac{e^{-x^2/(2s)}}{\sqrt{2\pi s}}\,ds\,dx \\ &\leq A\int_{-\infty}^{\infty}\int_0^t e^{2B|x|}\frac{e^{-x^2/(2t)}}{\sqrt{2\pi s}}\,ds\,dx < +\infty.\end{aligned} \tag{9.40}$$

Proof in the Case that f''' is Bounded and Continuous Without loss of generality, $s := 0$ (why?).

The proof of Itô's formula starts out in the same manner as that of (9.38). Namely, by telescoping the sum, we first write,

$$\begin{aligned}&f\left(W\left(2^{-n}\left\lfloor 2^n t - 1\right\rfloor\right)\right) - f(0) \\ &\quad= \sum_{0\leq k\leq 2^n t-1}\left[f\left(W\left(\frac{k+1}{2^n}\right)\right) - f\left(W\left(\frac{k}{2^n}\right)\right)\right].\end{aligned} \tag{9.41}$$

To this we apply Taylor's expansion with remainder, and write

$$\begin{aligned}&f\left(W\left(2^{-n}\left\lfloor 2^n t - 1\right\rfloor\right)\right) - f(0) \\ &\quad= \sum_{0\leq k\leq 2^n t-1} f'\left(W\left(\frac{k}{2^n}\right)\right) d_{k,n} \\ &\qquad+ \frac{1}{2}\sum_{0\leq k\leq 2^n t-1} f''\left(W\left(\frac{k}{2^n}\right)\right) d_{k,n}^2 + \sum_{0\leq k\leq 2^n t-1} R_{k,n} d_{k,n}^3,\end{aligned} \tag{9.42}$$

where $d_{k,n} := W((k+1)2^{-n}) - W(k2^{-n})$, and $|R_{k,n}| \leq \sup_x |f'''(x)| := M < \infty$, uniformly for all k, n. In the last identity, the first term converges in $L^2(\mathrm{P})$ to $\int_0^t f'(W(s))\, W(ds)$ by Theorem 9.7; see also Example 9.5. The second term, on the other hand, converges in $L^2(\mathrm{P})$ to $\frac{1}{2}\int_0^t f''(W(s))\, ds$; cf. Theorem 9.15. In addition, $f(W(2^{-n}\lfloor 2^n t - 1\rfloor)) \to f(W(t))$ a.s. and in $L^2(\mathrm{P})$ by continuity, and thanks to the dominated convergence theorem (Theorem 2.22). It, therefore, suffices to prove that as $n \to \infty$, $\sum_{0 \leq k \leq 2^n t - 1} R_{k,n} d_{k,n}^3 \xrightarrow{L^1(\mathrm{P})} 0$. On the other hand, as $n \to \infty$,

$$\begin{aligned} &\mathrm{E}\left\{ \sum_{0 \leq k \leq 2^n t - 1} |R_{k,n}| \times |d_{k,n}|^3 \right\} \\ &\quad \leq M \sum_{0 \leq k \leq 2^n t - 1} \mathrm{E}\left\{ |d_{k,n}|^3 \right\} \leq M 2^n t \cdot 2^{-3n/2} \mathrm{E}\{|Z|^3\} \longrightarrow 0, \end{aligned} \tag{9.43}$$

where Z is standard normal. This completes the proof. □

Itô's formula is particularly useful because it identifies various martingales. This in turn leads to explicit calculations. The following is a brief sampler; it is proved by applying the Itô formula with $f(x) = x$, and $f(x) = x^2$, respectively.

Corollary 9.18 *W and $t \mapsto W^2(t) - t$ are continuous L^2-martingales. In addition, $W^2(t) - t = \frac{1}{2}\int_0^t W\, dW$, a.s.*

Next is an interesting refinement; it is proved by similar arguments involving Taylor series expansions that were used to derive Theorem 9.16.

Theorem 9.19 *Suppose W is Brownian motion started at a given point $W(0) = x_0 \in \mathbb{R}$. If $f = f(x,t)$ is twice continuously-differentiable in x and continuously-differentiable in t, and if for all $t \geq 0$, $\mathrm{E}\{\int_0^t |\partial_x f(W(s), s)|^2\, ds\} < +\infty$, then with probability one,*

$$\begin{aligned} &f\left(W(t), t\right) - f(x_0, 0) \\ &\quad = \int_0^t \partial_x f\left(W(s), s\right) W(ds) \\ &\quad\quad + \int_0^t \left[\frac{1}{2}\partial_{x,x}^2 f\left(W(s), s\right) + \partial_t f\left(W(s), s\right) \right] ds, \end{aligned} \tag{9.44}$$

where $\partial_{x,x}^2 f(x,t) := \partial_x \partial_x f(x,t)$. This theorem is valid even if f is complex-valued.

Of course, I owe you the definition of $\int H\,dW$ when H is complex-valued, but this is easy: Whenever possible, $\int H\,dW := \int \mathrm{Re}(H)\,dW + i\int \mathrm{Im}(H)\,dW$.

Rather than use up the remainder of our time to prove this theorem, I close these notes by applying Theorem 9.19 to make two fascinating computations.[9.7]

5.1 A First Look at Exit Distributions

If W denotes the Brownian motion started at some fixed point $x_0 \in (-1,1)$, one might wish to know when it leaves a given interval $(-1,1)$ (say). The following remarkable formula of Paul Lévy answers a generalization of this question.[9.8]

Theorem 9.20 (Lévy [Lév51]) *Choose and fix $a, b > 0$, and define $T_{-b,a} := \inf\{s > 0 : W(s) = a \text{ or } -b\}$, where $\inf\varnothing := +\infty$. If $W(0) = x_0 \in (-b,a)$ if also fixed, then the characteristic function of $T_{-b,a}$ is given by the following: For all real numbers $\lambda \neq 0$,*

$$\mathrm{E}\left\{e^{i\lambda T_{-b,a}}\right\} = \frac{e^{(1+i)x_0\sqrt{\lambda}}}{e^{(1+i)a\sqrt{\lambda}} + e^{-(1+i)b\sqrt{\lambda}}} + \frac{e^{-(1+i)x_0\sqrt{\lambda}}}{e^{(1+i)b\sqrt{\lambda}} + e^{-(1+i)a\sqrt{\lambda}}}. \tag{9.45}$$

Proof Let us apply Itô's formula (Theorem 9.19) with $f(x,t) := \psi(x)e^{i\lambda t}$, where $\lambda \neq 0$ is fixed, and the function ψ satisfies the following (complex) *eigenvalue problem*:

$$\frac{1}{2}\psi''(x) = i\lambda\psi(x) \qquad \psi(a) = \psi(-b) = 1. \tag{9.46}$$

You can directly check that the solution is

$$\psi(x) := \frac{e^{(1+i)x\sqrt{\lambda}}}{e^{(1+i)a\sqrt{\lambda}} + e^{-(1+i)b\sqrt{\lambda}}} + \frac{e^{-(1+i)x\sqrt{\lambda}}}{e^{(1+i)b\sqrt{\lambda}} + e^{-(1+i)a\sqrt{\lambda}}}. \tag{9.47}$$

[9.7] Although this may be a natural way to end these notes, you should be made aware that this chapter's treatment of the subject is far from complete. Perhaps its greatest omission is connections between Itô's formula and William Feller's characterization of one-dimensional diffusions—one of the crowning achievements of its day; cf. Feller [Fel55b, Fel55a, Fel56]. For a pedagogic account that includes some of the most recent progress in this area see Bass [Bas98] and Revuz and Yor [RY99].

[9.8] For this and much more, see Knight [Kni81, Chapter 4].

Clearly, $|\partial_x f(x,t)|$ is bounded in (x,t) so that $\mathrm{E}\{\int_0^t |\partial_x f(W(s),s)|^2\, ds\} < +\infty$. Moreover, the eigenvalue problem for ψ implies that $\frac{1}{2}\partial_{x,x} f(x,t) + \partial_t f(x,t) = 0$. Therefore, Theorem 9.19 tells us that $f(W(t),t) - f(x_0,0)$ is a mean-zero (complex) martingale. By the optional stopping theorem (Theorem 9.12),

$$\mathrm{E}\left\{f\left(W\left(T_{-b,a} \wedge t\right), T_{-b,a} \wedge t\right)\right\} = f\left(x_0, 0\right), \tag{9.48}$$

which equals $\psi(x_0)$. Thanks to the dominated convergence theorem (Theorem 2.22), and by the a.s.-continuity of W, we can let $t \to \infty$ to deduce that

$$\mathrm{E}\left\{f\left(W\left(T_{-b,a}\right), T_{-b,a}\right)\right\} = \psi(x_0). \tag{9.49}$$

But $f(W(T_{-b,a})) = \psi(W(T_{-b,a}))e^{i\lambda T_{-b,a}} = e^{i\lambda T_{-b,a}}$, since with probability one, $W(T_{-b,a}) \in \{-b, a\}$, and $\psi(a) = \psi(-b) = 1$ (check the details!). This proves the theorem. □

5.2 A Second Look at Exit Distributions

Let us have a second look at Theorem 9.20 in the simplest setting where $x_0 := 0$ and $a = b = 1$. In this case, (9.45) somewhat simplifies to the following elegant form: For all $\lambda \in \mathbb{R} \setminus \{0\}$,

$$\mathrm{E}\left\{e^{i\lambda T_{-1,1}}\right\} = \frac{2}{e^{\sqrt{2i\lambda}} + e^{-\sqrt{2i\lambda}}} = \left[\cosh\left(\sqrt{2i\lambda}\right)\right]^{-1}. \tag{9.50}$$

(I have used the elementary fact that $(1+i) = \sqrt{2i}$.) In principle, the uniqueness theorem for characteristic functions (Theorem 6.20) tells us that the preceding formula determines the distribution of $T := T_{-1,1}$. However, in reality it is not always so easy to extract the right piece of information from (9.50). For instance, if it were not for (9.50), then we could not even easily prove that $[\cosh(\sqrt{2i\lambda})]^{-1}$ is a characteristic function of a probability measure. Or for that matter, can you see from (9.50) that T has finite moments of all orders?

The following contains a different representation of the distribution of T that answers the question about the moments of the stopping time T.[9.9]

[9.9] See Ciesielski and Taylor [CT62] for an interesting multidimensional analogue.

Theorem 9.21 (Chung [Chu47]) *For any $t > 0$,*

$$\mathrm{P}\left\{T > t\right\} = \frac{4}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}e^{-\frac{1}{8}n^2\pi^2 t}. \tag{9.51}$$

Consequently, $\mathrm{P}\{T > t\} = \frac{4}{\pi}\exp(-\frac{1}{8}\pi^2 t) \times (1 + \Theta(t))$*, where* $\Theta(t) \to 0$ *as* $t \to \infty$*.*

In particular, when t is large, $\mathrm{P}\{T > t\} \leq 2\exp(-\frac{1}{8}\pi^2 t)$. In lieu of Lemma 5.9, for any $p \geq 1$,

$$\mathrm{E}\left\{T^p\right\} = p\int_0^{\infty} t^{p-1}\mathrm{P}\{T > t\}\, dt < +\infty, \tag{9.52}$$

which shows that T has moments of all orders, as asserted earlier. In fact, we can even carry out the computation further to produce a formula for the pth moment of T:

$$\mathrm{E}\left\{T^p\right\} = \frac{\Gamma(p+1)2^{3p+2}}{\pi^{1+2p}}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^{1+2p}}, \qquad \forall p \geq 1, \tag{9.53}$$

where $\Gamma(p) := \int_0^{\infty} s^{p-1}e^{-s}\, ds$ denotes the Gamma function (check!).

Theorem 9.21 implies also the following unusual formula:

Corollary 9.22 (Chung [Chu47]) *We have*

$$\mathrm{P}\left\{\sup_{0\leq s\leq 1}|W(s)| \leq x\right\} = \frac{4}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\exp\left(-\frac{n^2\pi^2}{8x^2}\right). \tag{9.54}$$

In particular, $\mathrm{P}\{\sup_{0\leq s\leq 1}|W(s)| \leq x\} = \frac{4}{\pi}\exp(-\frac{1}{8}\pi^2 x^{-2}) \times (1 + \Theta(x^{-2}))$*, where* $\Theta(x^{-2}) \to 0$ *as* $x \to 0$*.*

I will prove only Theorem 9.21. While it is possible to write a very quick proof, I will sketch an argument that I believe shows the motivation behind the proof. [It is a sketch only because I will not develop some of the details of the PDE and Fourier-analysis arguments.]

Proof of Theorem 9.21 (Sketch) This proof is presented in three quick steps.

Step 1. Itô's Formula and a Sturm–Liouville Problem. Itô's formula (Theorem 9.19) tells us that modulo technical conditions, $f(W(t),t) - f(0,0)$ is a mean-zero martingale provided that f satisfies the partial differential equation, $\frac{1}{2}\partial^2_{x,x} f + \partial_t f = 0$. It is known that f must have the form $f(x,t) = \psi(x)e^{\lambda t}$; this is called *separation of variables.* Rather than prove this fact, I will merely be guided by it, and seek functions of the type $f(x,t) := \psi(x)e^{\lambda t}$ that satisfy the said PDE. This is an easy task, for in terms of ψ the PDE is: $\psi''+2\lambda\psi = 0$ (check!). Any ψ that satisfies this ODE yields a function f for which we then have (modulo technical integrability), $\mathrm{E}\{f(W(T\wedge t), T\wedge t\} = f(0,0)$. In the last part I used the optional stopping theorem (Theorem 9.12). Equivalently,

$$\mathrm{E}\left\{f\left(W(T),T\right); T \le t\right\} + \mathrm{E}\left\{f\left(W(t),t\right); T > t\right\} = f(0,0). \tag{9.55}$$

Of course, $W(T) \in \{-1,1\}$. Therefore, if we add to the ODE the conditions that $\psi(\pm 1) = 0$, then we obtain

$$\mathrm{E}\left\{f\left(W(t),t\right); T > t\right\} = f(0,0). \tag{9.56}$$

Equivalently, suppose ψ solves the *Sturm–Liouville problem*: $\psi'' = -2\lambda\psi$ and $\psi(\pm 1) = 0$. Then,

$$\mathrm{E}\left\{\psi\left(W(t)\right); T > t\right\} = e^{-\lambda t}\psi(0). \tag{9.57}$$

The typical solution to the said Sturm–Liouville problem is $\psi(x) = \cos(\frac{1}{2}n\pi x)$ where $n = 1, 2, \ldots$ (check!). This function solves $\psi'' = -2\lambda\psi$, $\psi(\pm 1) = 0$ with $\lambda := \frac{1}{8}n^2\pi^2$. Since $\psi(0) = 1$, we have

$$\mathrm{E}\left\{\cos\left(\frac{n\pi W(t)}{2}\right); T > t\right\} = e^{-\frac{1}{8}n^2\pi^2 t}. \tag{9.58}$$

Step 2. Fourier Series. Let $L^2(-2,2)$ denote the collection of all measurable functions $g : [-2,2] \to \mathbb{R}$ such that $\int_{-2}^{2} g^2(x)\,dx < +\infty$. Then, Theorem B.1 and a little fidgeting with the variables shows us that $L^2(-2,2)$ has the following complete orthonormal basis: $\frac{1}{2}$, $\frac{1}{\sqrt{2}}\sin(\frac{1}{2}n\pi x)$, $\frac{1}{\sqrt{2}}\cos(\frac{1}{2}m\pi x)$ $(n, m = 1, 2, \ldots)$. In particular, any $\phi \in L^2(-2,2)$ has the representation,

$$\phi(x) = \frac{A_0}{2} + \frac{1}{\sqrt{2}}\sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{2}\right) + \frac{1}{\sqrt{2}}\sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{2}\right). \tag{9.59}$$

In the preceding, the convergence holds in $L^2(-2,2)$, $A_0 := \int_{-2}^2 \phi(x)\,dx$, and for all $n = 1,2,3,\ldots$, $A_n := \frac{1}{\sqrt{2}}\int_{-2}^2 \phi(x)\cos(\frac{1}{2}n\pi x)\,dx$, whereas $B_n := \frac{1}{\sqrt{2}}\int_{-2}^2 \phi(x)\sin(\frac{1}{2}n\pi x)\,dx$.

Step 3. Putting it Together.

Apply Step 2 to the function $\phi(x) := \mathbf{1}_{(-1,1)}(x)$ to obtain $A_0 = 1$, $A_n = 2^{3/2}(n\pi)^{-1}(-1)^{n+1}$, and $B_n = 0$ ($n = 1,2,\ldots$). Thus,

$$\mathbf{1}_{(-1,1)}(x) = \frac{1}{2} + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\cos\left(\frac{n\pi x}{2}\right). \tag{9.60}$$

Plug in $x := W(t,\omega)$, multiply by $\mathbf{1}_{\{T(\omega)>t\}}$, and integrate $[\mathrm{P}(d\omega)]$ to obtain

$$\begin{aligned}&\mathrm{P}\left\{W(t) \in (-1,1)\ ,\ T > t\right\}\\ &\quad = \frac{1}{2}\mathrm{P}\{T>t\} + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\mathrm{E}\left\{\cos\left(\frac{n\pi W(t)}{2}\right)\ ;\ T>t\right\}.\end{aligned} \tag{9.61}$$

[A word of caution: In a fully-rigorous treatment, we need to justify this exchange of $L^2(\mathrm{P})$-limit and expectation.] Two quick observations are in order: The left-hand side equals $\mathrm{P}\{T > t\}$; this follows from the fact that $\{T > t\} = \{\sup_{s\le 1}|W(s)| < 1\}$. The second observation is that the right-hand side is computable via (9.58). After a little algebra, this completes our proof. □

The following is a corollary of the proof. It turns out to be the starting-point of some of the many deep connections between Markov processes and the boundary-theory of second-order differential equations.

Corollary 9.23 *Suppose $\phi \in L^2(-2,2)$ has the Fourier series representation (9.59), and $\phi(\pm 1) = 0$; i.e., for all $n \ge 1$, $B_n = 0$. Then for any $t \ge 0$,*

$$\begin{aligned}&\mathrm{E}\left\{\phi\left(W(t)\right)\ ;\ T>t\right\}\\ &\quad = \frac{2A_0}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}e^{-\frac{1}{8}n^2\pi^2 t} + \frac{1}{\sqrt{2}}\sum_{n=1}^{\infty}A_n e^{-\frac{1}{8}n^2\pi^2 t}.\end{aligned} \tag{9.62}$$

6 Exercises

Exercise 9.1 In this exercise, we construct a Dini-continuous process in $L^p(\mathrm{P})$ that is not a.s. continuous.

1. If $t > s > 0$, then prove that
$$\mathrm{P}\left\{W(s)W(t) < 0\right\} = \int_0^\infty \frac{e^{-y^2/(2s)}}{\sqrt{2\pi s}} \mathrm{P}\left\{W(t-s) > y\right\} dy. \tag{9.63}$$
2. Prove that $\mathrm{P}\{W(t-s) > y\} \le e^{-y^2/(2(t-s))}$.
3. Conclude that $\mathrm{P}\{W(s)W(t) < 0\} \le \frac{1}{2}\sqrt{\frac{t-s}{t}}$.
4. Use this to prove that $H(s) := \mathbf{1}_{(0,\infty)}(W(e^{s\wedge 1}))$ is Dini-continuous in $L^2(\mathrm{P})$, but H is not a.s. continuous.

Exercise 9.2 In this exercise, you are asked to construct a rather general abstract integral that is due to Young [You70].[9.10]

A function $f : [0,1] \to \mathbb{R}$ is said to be *Hölder continuous* of order $\alpha > 0$ if there exists a finite constant K such that for all $s, t \in [0,1]$ $|f(s) - f(t)| \le K|t-s|^\alpha$. Let C^α denote the collection of all such functions. When $\alpha = 0$, we define C^0 to be the collection of all continuous real functions on $[0,1]$.

1. Prove that when $\alpha > 1$, C^α contains only constants, whereas C^1 includes but is not limited to all continuously-differentiable functions.
2. If $0 < \alpha < 1$, then prove that C^α is a complete normed linear space that is normed by
$$\|f\|_{C^\alpha} := \sup_{\substack{s,t\in[0,1]\\ s\neq t}} \frac{|f(s) - f(t)|}{|s-t|^\alpha}. \tag{9.64}$$
3. Given two functions f and g, define for all $n \ge 1$,
$$\int_0^1 f\,\delta_n g := \sum_{k=0}^{2^n-1} f\left(\frac{k}{2^n}\right) \times \left[g\left(\frac{k+1}{2^n}\right) - g\left(\frac{k}{2^n}\right)\right]. \tag{9.65}$$

   Suppose for some $\alpha, \beta \le 1$, $f \in C^\alpha$ and $g \in C^\beta$. Prove that whenever $\alpha + \beta > 1$, then $\int_0^1 f\,\delta g := \lim_n \int_0^1 f\,\delta_n g$ exists. Note that when we let $g(x) := x$ we recover the Riemann integral of f; i.e., that $\int_0^1 f\,\delta g = \int_0^1 f(x)\,dx$.

[9.10] See also McShane [McS69].

4. Prove that $\int_0^1 g\,\delta f$ is also well-defined, and verify the following:

$$\int_0^1 f\,\delta g = f(1)g(1) - f(0)g(0) - \int_0^1 g\,\delta f. \tag{9.66}$$

The integral $\int f\,\delta g$ is called the *Young integral.*
(Hint: Lemma 9.3.)

Exercise 9.3 In this exercise you are asked to prove Theorem 9.13 and its variants in steps. We say that M is a *submartingale* if it is defined as a martingale, except whenever $t \geq s \geq 0$, we have $\mathrm{E}\{M(t) \mid \mathfrak{F}_s\} \geq M(s)$, a.s. M is a *supermartingale* if $-M$ is a submartingale. A process M is said to be a *continuous* L^2*-submartingale* (respectively, continuous L^2-supermartingale) if it is a submartingale (respectively supermartingale), $t \mapsto M(t)$ is a.s.-continuous, and for all $t \geq 0$, $M(t) \in L^2(\mathrm{P})$.

1. If Y is in $L^2(\mathrm{P})$, then prove that $M(t) := \mathrm{E}\{Y \mid \mathfrak{F}_t\}$ is a martingale. (This is the *Doob martingale* in continuous-time.)

2. If M is a martingale and ψ is convex, then $\psi(M)$ is a submartingale provided that $\psi(M(t)) \in L^1(\mathrm{P})$ for each $t \geq 0$.

3. If M is a submartingale and ψ is a nondecreasing convex function, and if $\psi(M(t)) \in L^1(\mathrm{P})$ for all $t \geq 0$, then $\psi(M)$ is a submartingale.

4. Prove that the first inequality in Theorem 9.13 holds if $|M|$ is replaced by any a.s.-continuous submartingale

(Hint: In the last part, you will need to prove that $\sup_{0\leq s\leq t} M(s)$ is measurable!)

Exercise 9.4 (Gambler's Ruin Formula) If W denotes a Brownian motion, then for any $a \in \mathbb{R}$, define $T_a := \inf\{s \geq 0 : \ W(s) = a\}$ where $\inf \varnothing := \infty$. Recall that T_a is an $\mathfrak{F}$-stopping time (Proposition 8.18). If $a, b > 0$, then carefully prove that $\mathrm{P}\{T_a < T_{-b}\} = b \div (a+b)$. Finally, compute $\mathrm{E}\{T_a\}$.
(Hint: Use Corollaries 8.10, 9.18, and Theorem 9.12.)

Exercise 9.5 Prove Corollary 9.22.
(Hint: Use Theorem 9.21 and the scaling property of Brownian motion; cf. Theorem 8.9.)

Part IV

Appendices

Appendix A

Hilbert Spaces

Throughout, let $\mathbb{H}$ be a set, and recall that it is a (real) Hilbert space if it is linear and if there exists an inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{H}\times\mathbb{H}$ such that $f\mapsto\langle f,f\rangle =: \|f\|^2$ norms $\mathbb{H}$. We recall that inner product means that for all $f,g,h\in\mathbb{H}$ and all $\alpha,\beta\in\mathbb{R}$, we have

$$\langle\alpha f+\beta g,h\rangle=\langle h,\alpha f+\beta g\rangle=\alpha\langle f,h\rangle+\beta\langle g,h\rangle. \tag{1.1}$$

Hilbert spaces come naturally equipped with a notion of angles: If $\langle f,g\rangle=0$, then f and g are *orthogonal.*

Definition A.1 Given any subspace $\mathbb{S}$ of $\mathbb{H}$, we let $\mathbb{S}^\perp$ denote the collection of all elements of $\mathbb{H}$ that are orthogonal to all of the elements of $\mathbb{S}$. That is, $\mathbb{S}^\perp:=\{f\in\mathbb{H}:\ \langle f,g\rangle=0,\ \forall g\in\mathbb{S}\}$.

It is easy to see that $\mathbb{S}^\perp$ is itself a subspace of $\mathbb{H}$, and that $\mathbb{S}\cap\mathbb{S}^\perp=\{0\}$. We now show that in fact $\mathbb{S}$ and $\mathbb{S}^\perp$ have a sort of complementary property.

Theorem A.2 (Orthogonal Decomposition) *If $\mathbb{S}$ is a subspace of a complete Hilbert space $\mathbb{H}$, then $\mathbb{H}=\mathbb{S}+\mathbb{S}^\perp:=\{f+g:\ f\in\mathbb{S}\ ,\ g\in\mathbb{S}^\perp\}$.*

In order to prove this, we need a lemma.

Lemma A.3 *If $\mathbb{X}$ is a closed and convex subset of a complete Hilbert space $\mathbb{H}$, then there exists a unique $f\in\mathbb{X}$ such that $\|f\|=\inf_{g\in\mathbb{X}}\|g\|$.*

Proof Existence is easy to prove: There exists $f_n \in \mathbb{X}$ such that $\lim_n \|f_n\| = \inf_{h\in\mathbb{X}} \|h\|$. By completeness and closedness, there exists $f \in \mathbb{X}$ such that $f_n \to f \in \mathbb{X}$, and since the norm is continuous functional, $\|f\| = \lim_n \|f_n\| = \inf_{h\in\mathbb{X}} \|h\|$.

For the uniqueness portion, suppose there were two norm-minimizing functions $f, g \in \mathbb{X}$. Note that $\|f-g\|^2 = 2\inf_{h\in\mathbb{X}} \|h\|^2 - 2\langle f, g\rangle$ and $\|f+g\|^2 = 2\inf_{h\in\mathbb{X}} \|h\|^2 + 2\langle f, g\rangle$. Multiply both identities by $\frac{1}{4}$ and add to obtain:

$$\frac{1}{4}\|f-g\|^2 = \inf_{h\in\mathbb{X}} \|h\|^2 - \left\|\frac{f+g}{2}\right\|^2 \leq 0, \tag{1.2}$$

since $\frac{1}{2}(f+g) \in \mathbb{X}$ by convexity. This yields the desired uniqueness. □

Proof of Theorem A.2 We are about to define two operators $\mathcal{P}$ and $\mathcal{P}^\perp$ that are in fact projection operators onto $\mathbb{S}$ and $\mathbb{S}^\perp$, respectively. Our definitions are motivated by well-known facts in linear algebra.

Given any $f \in \mathbb{H}$, the set $f+\mathbb{S} := \{f+s : s \in \mathbb{S}\}$ is closed and convex. In particular, $f+\mathbb{S}$ has a unique element $\mathcal{P}^\perp(f)$ of minimal norm (Lemma A.3). We also define $\mathcal{P}(f) := f - \mathcal{P}^\perp(f)$.

Whenever $f \in \mathbb{H}$, and because $\mathcal{P}^\perp(f) \in f + \mathbb{S}$, it follows that $\mathcal{P}(f) \in \mathbb{S}$. Since $\mathcal{P}(f) + \mathcal{P}^\perp(f) = f$, it suffices to show that for all $f \in \mathbb{H}$, $\mathcal{P}^\perp(f) \in \mathbb{S}^\perp$. But by the definition of $\mathcal{P}^\perp$, for all $g \in \mathbb{S}$, $\|\mathcal{P}^\perp(f)\| \leq \|f-g\|$. Instead of g write $\alpha g + \mathcal{P}(f)$ where $\|g\| = 1$ (this is in $\mathbb{S}$), where $\alpha \in \mathbb{R}$, and deduce that for all $g \in \mathbb{S}$ with $\|g\| = 1$ and all $\alpha \in \mathbb{R}$,

$$\begin{aligned}\|\mathcal{P}^\perp(f)\|^2 &\leq \|\mathcal{P}^\perp(f) - \alpha g\|^2 \\ &= \|\mathcal{P}^\perp(f)\|^2 - 2\alpha \left\langle \mathcal{P}^\perp(f), g\right\rangle + \alpha^2.\end{aligned} \tag{1.3}$$

Let $\alpha := \langle \mathcal{P}^\perp(f), g\rangle$ to deduce that for all $g \in \mathbb{S}$, $\langle \mathcal{P}^\perp(f), g\rangle = 0$, which is the desired result. □

Theorem A.4 *To every bounded linear functional $\mathcal{L}$ on a complete Hilbert space $\mathbb{H}$, there corresponds a unique $\pi \in \mathbb{H}$ such that for all $f \in \mathbb{H}$, $\mathcal{L}(f) = \langle f, \pi\rangle$.*

Proof I f $\mathcal{L}(f) = 0$ for all $f \in \mathbb{H}$, then define $\pi \equiv 0$ and we are done. If not, then $\mathbb{S} := \{f \in \mathbb{H} : \mathcal{L}(f) = 0\}$ is a closed subspace of $\mathbb{H}$ that does not span all of $\mathbb{H}$, i.e, there exists $g \in \mathbb{S}^\perp$ with $\|g\| = 1$ and $\mathcal{L}(g) > 0$; this follows from the decomposition theorem for $\mathbb{H}$ (Theorem A.2). We will show that $\pi := g\mathcal{L}(g)$ is the function that we seek, all the time remembering that $\mathcal{L}(g) \in \mathbb{R}$. But this is elementary. For any $f \in \mathbb{H}$, consider the function $h := \mathcal{L}(g)f - \mathcal{L}(f)g$, and note that $h \in \mathbb{S}$, since $\mathcal{L}(h) = 0$ by design. This means that $\langle \pi, h\rangle = 0$, but $\langle \pi, h\rangle = \mathcal{L}(g)\langle \pi, f\rangle - \mathcal{L}(f)\langle \pi, g\rangle = \mathcal{L}(g)\langle \pi, f\rangle - \mathcal{L}(g)\mathcal{L}(f)$. Since $\mathcal{L}(g) > 0$, we have $\mathcal{L}(f) = \langle f, \pi\rangle$ for all $f \in \mathbb{H}$. It remains to prove uniqueness, but this too is easy for if there were two of these functions, say π_1 and π_2, then for all $f \in H$, $\langle f, \pi_1 - \pi_2\rangle = 0$. In particular, let $f := \pi_1 - \pi_2$ to see that $\pi_1 = \pi_2$. □

1 Exercises

Exercise 1.1 Recall from the proof of Theorem A.2 the operators $\mathcal{P}$ and $\mathcal{P}^\perp$, and show that $\mathcal{P}$ and $\mathcal{P}^\perp$ are linear operators with $\mathcal{P} : \mathbb{H} \to \mathbb{S}$ and $\mathcal{P}^\perp : \mathbb{H} \to \mathbb{S}^\perp$. Furthermore, prove that they are projection operators, i.e., that whenever $f \in \mathbb{S}$ then $\mathcal{P}(f) = f$, and whenever $f \in \mathbb{S}^\perp$, then $\mathcal{P}^\perp(f) = f$.

Appendix B

Fourier Series

Throughout this section, we let $\mathbb{T} := [-\pi, \pi]$ denote the torus of length 2π, and consider some elementary facts about the trigonometric Fourier series on $\mathbb{T}$ that are based on the following:

$$\phi_n(x) := \frac{e^{inx}}{\sqrt{2\pi}}, \qquad \forall x \in \mathbb{T},\ n = 0, \pm 1, \pm 2, \dots . \tag{2.1}$$

Let $L^2(\mathbb{T})$ denote the Hilbert space of all measurable functions $f : \mathbb{T} \to \mathbb{C}$ such that $\|f\|_{\mathbb{T}}^2 := \int_{\mathbb{T}} |f(x)|^2 \, dx < +\infty$. As usual, $L^2(\mathbb{T})$ is equipped with the (semi-)norm $\|f\|_{\mathbb{T}}$ and inner-product $\langle f, g \rangle := \int_{\mathbb{T}} f(x)\overline{g(x)} \, dx$. Our goal is to prove the following theorem.

Theorem B.1 *The collection $\{\phi_n\}_{n \in \mathbb{Z}}$ is a complete orthonormal system in $L^2(\mathbb{T})$. Consequently, given any $f \in L^2(\mathbb{T})$, $f = \sum_n \langle f, \phi_n \rangle \phi_n$, where the convergence takes place in $L^2(\mathbb{T})$. Furthermore, $\|f\|_{\mathbb{T}}^2 = \sum_n |\langle f, \phi_n \rangle|^2$.*

The proof is not difficult, but requires some preliminary developments.

Definition B.2 A *trigonometric polynomial* is a finite linear combination of the ϕ_n's. An *approximation to the identity* is a sequence of integrable functions $\psi_0, \psi_1, \dots : \mathbb{T} \to \mathbb{R}_+$ such that

1. (i) $\int_{\mathbb{T}} \psi_n(x) \, dx = 1$ for all n.
2. (ii) For any $\varepsilon \in (0, \pi]$, however small, $\lim_{n \to \infty} \int_{-\varepsilon}^{\varepsilon} \psi_n(x) \, dx = 1$.

Note that (a) all the ψ_n's are nonnegative; and (b) the preceding display shows that all of the area under ψ_n is concentrated near the origin when n is large. In other words, as $n \to \infty$, ψ_n looks more and more like a point mass.

For $n = 0, 1, 2, \ldots$ consider

$$\kappa_n(x) := \frac{(1+\cos(x))^n}{\alpha_n}, \qquad \forall x \in \mathbb{T}, \tag{2.2}$$

where $\alpha_n := \int_{\mathbb{T}} (1+\cos(x))^n \, dx$.

Lemma B.3 $\kappa_0, \kappa_1, \ldots$ *is an approximation to the identity.*

Proof I need to only prove part (ii) of Definition B.2. First, note that for any fixed $\varepsilon \in (0, \pi]$,

$$\int_\varepsilon^\pi (1+\cos(x))^n \, dx \le 2\pi(1+\cos(\varepsilon))^n. \tag{2.3}$$

By symmetry, this estimates the integral away from the origin. To estimate the integral near the origin, we use a method of P.-S. Laplace and write $\int_0^\varepsilon (1+\cos(x))^n \, dx = \int_0^\varepsilon e^{ng(x)} \, dx$, where $g(x) := \ln(1+\cos(x))$. Apply Taylor's theorem with remainder to deduce that for any $x \in [0, \varepsilon]$, there exists $\zeta \in [0, x]$ such that

$$\begin{aligned} g(x) &= g(0) + g'(0)x + \frac{1}{2}g''(\zeta)x^2 \\ &= \ln(2) - \left(\frac{x}{1+\cos(\zeta)}\right)^2 \ge \ln(2) - \left(\frac{x}{1+\cos(\varepsilon)}\right)^2. \end{aligned} \tag{2.4}$$

Thus,

$$\begin{aligned} \int_0^\varepsilon (1+\cos(x))^n \, dx &\ge 2^n \int_0^\varepsilon \exp\left(-\frac{nx^2}{(1+\cos(\varepsilon))^2}\right) dx \\ &= \frac{2^n}{\sqrt{n}} \int_0^{\sqrt{n}\,\varepsilon} \exp\left(-\frac{z^2}{(1+\cos(\varepsilon))^2}\right) dz \\ &\ge A_\varepsilon \frac{2^n}{\sqrt{n}}, \qquad \forall n \ge 1, \end{aligned} \tag{2.5}$$

where $A_\varepsilon := \int_0^\varepsilon \exp\{-z^2(1+\cos(\varepsilon))^{-2}\} \, dz$. In particular, $\alpha_n \ge 2A_\varepsilon n^{-1/2} 2^n$, which is many orders of magnitude larger than $\int_{\varepsilon \le |x| \le \pi} (1+\cos(x))^n \, dx$, thanks to (2.3). This proves the lemma. □

Proposition B.4 *If $f \in L^2(\mathbb{T})$ and $\varepsilon > 0$, then there is a trigonometric polynomial T such that $\|T - f\|_{\mathbb{T}} \leq \varepsilon$.*

In other words, trigonometric polynomials are dense in $L^2(\mathbb{T})$.

Proof Since continuous functions (endowed with uniform topology) are dense in $L^2(\mathbb{T})$ it suffices to prove that trigonometric polynomials are dense in the space of all continuous function on $\mathbb{T}$ (why?). We first of all observe that the functions κ_n are trigonometric polynomials. Indeed, by the binomial theorem,

$$\begin{aligned}\kappa_n(x) &= \frac{1}{\alpha_n}\sum_{k=0}^{n}\binom{n}{k}(\cos(x))^k = \frac{1}{\alpha_n}\sum_{k=0}^{n}\binom{n}{k}\left(\frac{e^{ix}+e^{-ix}}{2}\right)^k \\ &= \frac{1}{\alpha_n}\sum_{k=0}^{n}\binom{n}{k}2^{-k}\sum_{l=0}^{k}\binom{k}{l}e^{ix(2l-k)},\end{aligned} \tag{2.6}$$

which is clearly a linear combination of $\phi_{-n}(x), \ldots, \phi_n(x)$. Having established this, note that the convolution of κ_n and f is also a trigonometric polynomial, where the said convolution is the function

$$\kappa_n \star f(x) := \int_{\mathbb{T}} f(y)\kappa_n(x-y)\,dy. \tag{2.7}$$

[This is a trigonometric polynomial since we can write $\kappa_n(x) = \sum_{j=-n}^{n} c_j\phi_j(x)$, from which we get $\kappa_n \star f(x) = \sum_{j=-n}^{n} q_j\phi_j(x)$, where $q_j := c_j \int_{\mathbb{T}} f(y)e^{-iny}\,dy$.] Now fix $\varepsilon \in (0, \pi)$ and consider

$$\begin{aligned}&|\kappa_n \star f(x) - f(x)| \\ &\quad = \left|\int_{-\varepsilon}^{\varepsilon} [f(y) - f(x)]\,\kappa_n(y-x)\,dy\right| \\ &\quad \leq \int_{\substack{y\in\mathbb{T}:\\ |y-x|\leq\varepsilon}} |f(x)-f(y)|\kappa_n(y-x)\,dy \\ &\qquad + 2\sup_{w\in\mathbb{T}}|f(w)| \cdot \int_{\substack{y\in\mathbb{T}:\\ |y-x|>\varepsilon}} \kappa_n(y-x)\,dy \\ &\quad \leq \sup_{\substack{y,u\in\mathbb{T}:\\ |y-u|\leq\varepsilon}} |f(y)-f(u)| + 2\sup_{w\in\mathbb{T}}|f(w)| \cdot \int_{\varepsilon\leq|z|\leq\pi} \kappa_n(z)\,dz.\end{aligned} \tag{2.8}$$

[The first line follows from $\int_{\mathbb{T}} \kappa_n(x)\,dx = 1$, the second from $\kappa_n(a) \geq 0$, and the third from both of the said properties of κ_n.] The third property of being an approximation to the identity shows that as $n \to \infty$ the last term vanishes, and we obtain the following: For all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{T}} |\kappa_n \star f(x) - f(x)| \leq \sup_{\substack{y,u\in\mathbb{T}:\\ |y-u|\leq\varepsilon}} |f(y) - f(u)|. \tag{2.9}$$

Let $\varepsilon \to 0$ to see that the left-hand side is zero, whence the proposition follows. □

We are ready to prove Theorem B.1.

Proof of Theorem B.1 It is easy to see that $\{\phi_n\}_{n\in\mathbb{Z}}$ is an orthonormal sequence in $L^2(\mathbb{T})$; i.e., that $\langle \phi_n, \phi_m\rangle$ equals one if $n = m$ and zero otherwise. To prove its completeness, suppose that $f \in L^2(\mathbb{T})$ is orthogonal to all ϕ_n's; i.e., $\langle f, \phi_n\rangle = 0$ for all $n \in \mathbb{Z}$. If ε and T are as in the preceding proposition, then: (i) $\langle f, T\rangle = 0$; and (ii) $\|f - T\|_{\mathbb{T}}^2 \leq 2\pi \sup_w |f(w) - T(w)| \leq 2\pi\varepsilon$. But $\|f - T\|_{\mathbb{T}}^2 = \|f\|_{\mathbb{T}}^2 + \|T\|_{\mathbb{T}}^2 - 2\langle f, T\rangle \geq \|f\|_{\mathbb{T}}^2$ by (ii). Since ε is arbitrary, $\|f\|_{\mathbb{T}} = 0$ from which we deduce that $f = 0$, almost everywhere. This proves completeness. The remainder is easy to prove but requires the material from Section A of Chapter 4.

Let $\mathcal{P}_n$ and $\mathcal{P}_n^\perp$ respectively denote the projections onto $\mathbb{S}_n :=$ the linear span of $\{\phi_j; |j| \leq n\}$ and $\mathbb{S}_n^\perp$. If $f \in L^2(\mathbb{T})$, then $\mathcal{P}_n f$ is the a.e. unique function $g \in \mathbb{S}_n$ that minimizes $\|f - g\|_{\mathbb{T}}$. We can write $g = \sum_{|j|\leq n} c_j\phi_j$ and expand the said L^2-norm to obtain the following optimization problem: Minimize over all $\{c_j\}$,

$$\left\| f - \sum_{|j|\leq n} c_j\phi_j \right\|_{\mathbb{T}}^2 = \|f\|_{\mathbb{T}}^2 + \sum_{|j|\leq n} c_j^2 - 2\sum_{|j|\leq n} c_j\langle f, \phi_j\rangle. \tag{2.10}$$

This is a calculus exercise and yields the optimal value of $c_j := \langle f, \phi_j\rangle$. Thus, $\mathcal{P}_n f = \sum_{|j|\leq n}\langle f, \phi_j\rangle\phi_j$, $\mathcal{P}_n^\perp f = f - \mathcal{P}_n f$, $\|\mathcal{P}_n f\|_{\mathbb{T}}^2 = \sum_{|j|\leq n} |\langle f, \phi_j\rangle|^2$, and $\|\mathcal{P}_n^\perp f\|_{\mathbb{T}}^2 = \|f\|_{\mathbb{T}}^2 - \sum_{|j|\leq n} |\langle f, \phi_j\rangle|^2$. The last inequality in turn yields *Bessel's inequality*:

$$\sum_{j=-\infty}^{\infty} |\langle f, \phi_j\rangle|^2 \leq \|f\|_{\mathbb{T}}^2. \tag{2.11}$$

Our goal is to show that this is sharp. If not, then $\| \liminf_{n\to\infty} \mathcal{P}_n^{\perp} f\|_{\mathbb{T}} > 0$ (Fatou's lemma, Theorem 2.20). In particular, $g := \liminf_n \mathcal{P}_n^{\perp} f$ is not zero almost everywhere. But note that $g \in \mathcal{P}_n^{\perp}$ for all n. Thus, the preceding proposition shows that $g = 0$ almost everywhere, which is the desired contradiction. [To see this fix $\varepsilon > 0$ and find a trigonometric polynomial $T \in \mathbb{S}_n$ for some large n such that $\|g - T\|_{\mathbb{T}} \leq \varepsilon$. Now expand $\varepsilon^2 \geq \|g - T\|_{\mathbb{T}}^2 = \|g\|_{\mathbb{T}}^2 + \|T\|_{\mathbb{T}}^2 - 2\langle g, T\rangle = \|g\|_{\mathbb{T}}^2 + \|T\|_{\mathbb{T}}^2 \geq \|g\|_{\mathbb{T}}^2$.] In fact, this argument shows that any subsequential limit of $\mathcal{P}_n^{\perp} f$ must be zero almost everywhere, and hence $\mathcal{P}_n^{\perp} f \to 0$ in $L^2(\mathbb{T})$, from which we get $f = \lim_n \mathcal{P}_n f = \sum_j \langle f, \phi_j\rangle \phi_j$ (in $L^2(\mathbb{T})$) as desired. □

Bibliography

[Ada74] W. J. Adams, *The Life and Times of the Central Limit Theorem*, Kaedmon Publishing Co., New York, 1974.

[Arc26] R. C. Archibald, *A rare pamphlet of de Moivre and some of his discoveries*, Isis **VIII** (1926), no. 4, 671–683.

[AS91] N. Alon and J. H. Spencer, *The Probabilistic Method*, first ed., Wiley, New York, 1991, With an appendix with P. Erdős.

[Bac00] L. Bachelier, *Théorie de la spéculation*, Ann. Sci. École Norm. Sup. **17** (1900), 21–86, [See also the 1995 reprint. Sceaux: Gauthier-Villars].

[Ban31] S. Banach, *über die baire'sche kategorie gewisser funkionenmengen*, Studia. Math. **t. III** (1931), 174–179.

[Bas98] R. F. Bass, *Diffusions and Elliptic Operators*, Springer-Verlag, New York, 1998.

[Bay63] T. Bayes, *An essay towards solving a problem in the doctrine of chances*, Phil. Trans. of the Royal Soc. **53** (1763), 370–418.

[BC01] I. Berkés and E. Csáki, *A universal result in almost sure central limit theory*, Stochastic Process. Appl. **94** (2001), no. 1, 105–134.

[BCH98] I. Berkés, E. Csáki, and L. Horváth, *Almost sure central limit theorems under minimal conditions*, Statist. Probab. Lett. **37** (1998), no. 1, 67–76.

[Ber13] J. Bernoulli, *Ars conjectandi (the art of conjecture)*, Basel, 1713.

[Ber13] S. N. Bernstein, *Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités*, Comm. Soc. Math. Kharkow **13** (1912/1913), 1–2.

[Ber96] J. Bertoin, *Lévy Processes*, Cambridge University Press, Cambridge, 1996.

[Ber98] I. Berkés, *Results and problems related to the pointwise central limit theorem*, Asymptotic Methods in Probability and Statistics (Ottawa, ON, 1997), North-Holland, Amsterdam, 1998, pp. 59–96.

[Blu57] R. M. Blumenthal, *An extended Markov property*, Trans. Amer. Math. Soc. **85** (1957), 52–72.

[Bor09] É. Borel, *Les probabilités dénombrables et leurs applications arithmetique*, Rend. Circ. Mat. Palermo **27** (1909), 247–271.

[BR96] M. Baxter and A. Rennie, *Financial Calculus: An Introduction to Derivative Pricing*, Cambridge University Press, Cambridge, 1996, Second (1998) reprint.

[Bro28] R. Brown, *A brief Account of Microscopical Observations made in the months of June, July, and August, 1827, on the Particles contained in the Pollen of Plants; and on the general Existence of active Molecules in Organic and Inorganic Bodies*, Philosophical Magazine N. S. **4** (1828), 161–173.

[Bro88] G. A. Brosamler, *An almost everywhere central limit theorem*, Math. Proc. Cambridge Philos. Soc. **104** (1988), no. 3, 561–574.

[BS73] F. Black and M. Scholes, *Pricing of options and corporate liabilities*, J. Political Econ. **81** (1973), 637–654.

[BY02] B. Bru and M. Yor, *Comments on the life and mathematical legacy of Wolfgang Doeblin*, Finance Stoch. **6** (2002), no. 1, 3–47.

[Can33] F. Cantelli, *Sulla probabilita come limita della frequenza*, Rend. Accad. Lincei **26** (1933), no. 1, 3.

[Car48] C. Carathéodory, *Vorlesungen über reelle Funktionen*, Chelsea Publishing Company, New York, 1948.

[Che46] P. L. Chebyshev, *Démonstration élementaire d'une proposition générale de la théorie des probabilités*, Crelle J. Math. **33** (1846), no. 2, 259–267.

[Che67] P. L. Chebyshev, *Des valeurs moyennes*, J. Math. Pures Appl. **12** (1867), no. 2, 177–184.

[Chu47] K. L. Chung, *On the maximum partial sum of independent random variables*, Proc. Nat. Acad. Sci. U. S. A. **33** (1947), 132–136.

[Cra36] H. Cramér, *Über eine eigenschaft der normalen verteilungsfunktion*, Math. Z. **41** (1936), 405–415.

[CT62] Z. Ciesielski and S. J. Taylor, *First passage times and sojourn times for Brownian motion in space and the exact Hausdorff measure of the sample path*, Trans. Amer. Math. Soc. **103** (1962), 434–450.

[dA83] A. de Acosta, *A new proof of the Hartman-Wintner law of the iterated logarithm*, Ann. Probab. **11** (1983), no. 2, 270–276.

[DJ56] E. Dynkin and A. Jushkevich, *Strong Markov processes*, Teor. Veroyatnost. i Primenen. **1** (1956), 149–155.

[dM18] A. de Moivre, *The Doctrine of Chances; or a Method of Calculating the Probabilities of Events in Play*, first ed., W. Pearson, London, 1718.

[dM33] A. de Moivre, *Approximatio ad Summam terminorum Binomii* $(a + b)^n$ *in Serium expansi*, Privately printed (1733), For a facsimile see Archibald [Arc26].

[dM38] A. de Moivre, *The Doctrine of Chances; or a Method of Calculating the Probabilities of Events in Play*, second ed., H. Woodfall, London, 1738.

[dM56] A. de Moivre, *The Doctrine of Chances; or a Method of Calculating the Probabilities of Events in Play*, third ed., A. Millar, London, 1756, Reprinted in 1967 by the Chelsea Publishing Co., New York.

[DM82] C. Dellacherie and P.-A. Meyer, *Probabilities and Potential. B*, North-Holland Publishing Co., Amsterdam, 1982, Theory of martingales, Translated from the French by J. P. Wilson.

[Doo40] J. L. Doob, *Regularity properties of certain families of chance variables*, Trans. Amer. Math. Soc. **47** (1940), 455–486.

[Doo49] J. L. Doob, *Application of the theory of martingales*, Le Calcul des Probabilités et ses Applications, Centre National de la Recherche Scientifique, Paris, 1949, pp. 23–27.

[Doo53] J. L. Doob, *Stochastic Processes*, John Wiley & Sons Inc., New York, 1953.

[Doo71] J. L. Doob, *What is a martingale?*, Amer. Math. Monthly **78** (1971), 451–463.

[Doo89] J. L. Doob, *Commentary on probability*, A Century of Mathematics in America, Part II, Amer. Math. Soc., Providence, RI, 1989, pp. 353–354.

[Ein05] A. Einstein, *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Ann. Phys. **322** (1905), 549–560.

[Erd48] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc. **53** (1948), 292–294.

[ES35] P. Erdős and G. Szekeres, *A combinatorial problem in geometry*, Composito. Math. **2** (1935), 463–470.

[Ete81] N. Etemadi, *An elementary proof of the strong law of large numbers*, Z. Wahr. verw Geb. **55** (1981), 119–122.

[Fat06] P. J. L. Fatou, *Séries trigonométriques et séries de taylor*, Acta Math. **69** (1906), 372–433.

[Fel55a] W. Feller, *On differential operators and boundary conditions*, Comm. Pure Appl. Math. **8** (1955), 203–216.

[Fel55b] W. Feller, *On second order differential operators*, Ann. of Math. (2) **61** (1955), 90–105.

[Fel56] W. Feller, *On generalized Sturm-Liouville operators*, Proceedings of the conference on differential equations (dedicated to A. Weinstein), University of Maryland Book Store, College Park, Md., 1956, pp. 251–270.

[Fel66] W. Feller, *An Introduction to Probability Theory and Its Applications. Vol. II*, John Wiley & Sons Inc., New York, 1966.

[Fis87] A. Fisher, *Convex-invariant means and a pathwise central limit theorem*, Adv. in Math. **63** (1987), no. 3, 213–246.

[Fré30] M. R. Fréchet, *Sur la convergence en probabilité*, Metron **8** (1930), 1–48.

[GK68] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley Publishing Co., Reading, Mass., 1968, [Translated from the original Russian, annotated, and revised by Kai Lai Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition].

[Gli33] V. Glivenko, *Sulla determinazione empirica delle leggi di probabilita*, Giornale d. Istituto Italiano Attuari **4** (1933), 92.

[Gli36] V. I. Glivenko, *Sul teorema limite della teoria funjioni caratteristische*, Giornale d. Instituto Italiano attuari **7** (1936), 160–167.

[Gne69] B. V. Gnedenko, *On Hilbert's sixth problem*, Hilbert's Problems (Russian), Izdat. "Nauka", Moscow, 1969, pp. 116–120.

[Hau27] F. Hausdorff, *Mengenlehre*, Walter De Gruyter & Co., Berlin, 1927.

[HK79] J. M. Harrison and D. Kreps, *Martingales and arbitrage in multiperiod securities markets*, J. Econ. Theory **20** (1979), 381–408.

[HL30] G. H. Hardy and J. E. Littlewood, *A maximal theorem with function-theoretic applications*, Acta Math. **54** (1930), 81–166.

[HP81] J. M. Harrison and S. R. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stoch. Proc. Their Appl. **11** (1981), 215–260.

[Hun56] G. A. Hunt, *Some theorems concerning Brownian motion*, Trans. Amer. Math. Soc. **81** (1956), 294–319.

[Hun57] G. A. Hunt, *Markoff processes and potentials. I, II*, Illinois J. Math. **1** (1957), 44–93, 316–369.

[Hun66] G. A. Hunt, *Martingales et processus de Markov*, Dunod, Paris, 1966.

[HW41] P. Hartman and A. Wintner, *On the law of the iterated logarithm*, Amer. J. Math. **63** (1941), 169–176.

[Isa65] R. Isaac, *A proof of the martingale convergence theorem*, Proc. Amer. Math. Soc. **16** (1965), 842–844.

[Itô44] K. Itô, *Stochastic integral*, Proc. Imp. Acad. Tokyo **20** (1944), 519–524.

[Khi24] A. Khinchine, *Ein satz der wahrscheinlichkeitsrechung*, Fund. Math. **6** (1924), 9–10.

[Khi29] A. Ya. Khintchine, *Sue la loi des grands nombres*, C. R. Acad. Sci. Paris **188** (1929), 477–479.

[Kin53] J. R. Kinney, *Continuity properties of sample functions of Markov processes*, Trans. Amer. Math. Soc. **74** (1953), 280–302.

[KK25] A. Khintchine and A. Kolmogorov, *On the convergence of series*, Rec. Math. Soc. Moscow **32** (1925), 668–677.

[Kni81] F. B. Knight, *Essentials of Brownian Motion and Diffusion*, American Mathematical Society, Providence, R.I., 1981.

[Kol30] A. Kolmogorov, *Sur la loi forte des grandes nombres*, C. R. Acad. Sci. Paris **191** (1930), 910–911.

[Kol33] A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin, 1933.

[Kol50] A. N. Kolmogorov, *Foundations of Probability*, Chelsea Publishig Company, New York, 1950, [Translation edited by Nathan Morrison].

[Kri63] K. Krickeberg, *Wahrscheinlichkeitstheorie*, Teubner, Stuttgart, 1963.

[Kri65] K. Krickeberg, *Probability Theory*, Addison–Wesley, Reading, Massachusetts, 1965.

[Lap05] P.-S. Laplace, *Traité de Mécanique Céleste*, vol. 4, 1805, [Reprinted by the Chelsea Publishing Co. (1967). Translated by N. Bowditch.].

[Lap10] P.-S. Laplace, *Mémoire sur les approximations des formules qui sont fonctions de trés grands nombres, et sur leur application aux probabilités*, Mémoires de la classe des sciences mathématiques et physiques de l'institut impéiral de France (1810), Année, See also pages 347–412 (1811).

[Lap12] P.-S. Laplace, *Théorie Analytique des Probabilités, Vol. I and II*, 1812, Reprinted in *Oeuvres complétes de Laplace*, Volume VII (1886), Paris: Gauthier–Villars.

[Leb10] H. Lebesgue, *Sur l'intégration des fonctions discontinues*, Ann. Ecole Norm. Sup. **27** (1910), no. 3, 361–450.

[Lev06] B. Levi, *Sopra l'integrazione delle serie*, Rend. Instituto Lombardino di Sci. e Lett. **39** (1906), no. 2, 775–780.

[Lév25] P. Lévy, *Calcul des Probabilités*, Gauthier–Villars, Paris, 1925.

[Lév37] P. Lévy, *Théorie de l'Addition des Variables Aleatoires*, Gauthier–Villars, Paris, 1937.

[Lév51] P. Lévy, *La mesure de Hausdorff de la courbe du mouvement brownien à n dimensions*, C. R. Acad. Sci. Paris **233** (1951), 600–602.

[Lin22] J. W. Lindeberg, *Eine neue herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung*, Math. Z. **15** (1922), 211–225.

[LP90] M. T. Lacey and W. Philipp, *A note on the almost sure central limit theorem*, Statist. Probab. Lett. **9** (1990), no. 3, 201–205.

[LPX02] W. Li, Y. Peres, and Y. Xiao, *Small ball probabilities of Brownian motion on thin sets and their applications*, Preprint (2002).

[Mat99] L. Mattner, *Product measurability, parameter integrals, and a fubini–tonelli counterexample*, Enseign. Math. (2) **45** (1999), no. 3-4, 271–279.

[Maz31] S. Mazurkiewicz, *Sue les fonctions non dérivales*, Studia. Math. **t. III** (1931), 92–94.

[McS69] E. J. McShane, *A Riemann-type integral that includes Lebesgue-Stieltjes, Bochner and stochastic integrals*, Memoirs of the American Mathematical Society, No. 88, American Mathematical Society, Providence, R.I., 1969.

[Mum00] D. Mumford, *The dawning of the age of stochasticity*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. (2000), no. Special Issue, 107–125, Mathematics towards the third millennium (Rome, 1999).

[Per03] J. B. Perrin, *Les Atoms*, 1903, [See also *Atoms*. The reprint of the second (1923) English edition. Connecticut: Ox Bow Press. Translated by D. Ll. Hammick].

[Pla10] M. Plancherel, *Contribution à l'étude de la répresentation d'une fonction arbitraire par des intégrales définies*, Rend. Circ. Mat. Palermo **30** (1910), 289–335.

[Pla33] M. Plancherel, *Sur les formules de réciprocité du type de Fourier*, J. London Math. Soc. **8** (1933), 220–226.

[Poi12] H. Poincaré, *Calcul dés Probabilités*, Gauthier-Villars, Paris, 1912.

[PWZ33] R. E. A. C. Paley, N. Wiener, and A. Zygmund, *Notes on random functions*, Math. Z. **37** (1933), 647–668.

[PZ32] R. E. A. C. Paley and A. Zygmund, *A note on analytic functions in the unit circle*, Proc. Camb. Phil. Soc. **28** (1932), 366–372.

[Ram30] F. P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), no. 2, 264–286.

[RY99] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, third ed., Springer-Verlag, Berlin, 1999.

[Sat99] K.-I. Sato, *Lévy processes and infinitely divisible distributions*, Cambridge University Press, Cambridge, 1999, [Translated from the 1990 Japanese original, Revised by the author].

[Sch88] P. Schatte, *On strong versions of the central limit theorem*, Math. Nachr. **137** (1988), 249–256.

[Sha48] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423, 623–656.

[Sie20] W. Sierpinski, *Sur les rapport entre l'existence des integrales* $\int_0^1 f(x,y)\,dx,\ \int_0^1 f(x,y)\,dy\ \ et\ \int_0^1 dx \int_0^1 f(x,y)\,dy$, Fundamenta Mathimaticae **1** (1920), 142–147.

[Sko33] Th. Skolem, *Ein kombinatorischer satz mit anwendung auf ein logisches entscheidungsproblem*, Fund. Math. **20** (1933), 254–261.

[Sko61] A. V. Skorohod, *Issledovaniya po teorii sluchainykh protsessov*, Izdat. Kiev. Univ., Kiev, 1961, [*Studies in the Theory of Random Processes.* Translated from Russian by Scripta Technica, Inc., Addison-Wesley Publishing Co., Inc., Reading, Mass. (1965). See also the second edition (1985), Dover, New York.].

[Sol70] R. M. Solovay, *A model of set-theory in which every set of reals is Lebesgue measurable*, Ann. Math. **92** (1970), 1–56.

[Ste30] H. Steinhaus, *Sur la probabilité de la convergence de serié*, Studia Math. **2** (1930), 21–39.

[Sti30] J. Stirling, *Methodus differentialis*, Whiston & White, London, 1730.

[SW49] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, Ill., 1949.

[Tan83] K. Tandori, *The Life and Works of Lipót Fejér*, Functions, series, operators, Vol. I, II (Budapest, 1980), Colloq. Math. Soc. János Bolyai, vol. 35, North-Holland, Amsterdam, 1983, pp. 77–85.

[Tuk77] J. W. Tukey, *Exploratory Data Analysis*, Addison–Wesley, Reading, Mass., 1977.

[Tur34] A. M. Turing, *On the Gaussian error function*, Unpublished Fellowship Dissertation, King's College Library, Cambridge, 1934.

[vN40] J. von Neumann, *On rings of operators, III*, Ann. Math. **41** (1940), 94–161.

[vS18] M. von Smoluchowski, Die Naturwissenschaften **6** (1918), 253–263.

[Wie23] N. Wiener, *Differential space*, J. Math. Phys. **2** (1923), 131–174.

[Wil91] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, 1991.

[You70] L. C. Young, *Some new stochastic integrals and Stieltjes integrals. I. Analogues of Hardy-Littlewood classes*, Advances in Probability and Related Topics, Vol. 2, Dekker, New York, 1970, pp. 161–240.

[Zab95] S. L. Zabell, *Alan Turing and the central limit theorem*, Amer. Math. Monthly **102** (1995), no. 6, 483–494.

Index