REAL ANALYSIS

Second Edition (2008)

Bruckner^2·Thomson

Brian S. Thomson
Judith B. Bruckner
Andrew M. Bruckner

www.ClassicalRealAnalysis.com

This second edition of **Real Analysis** contains all the material of the first edition originally published by Prentice Hall (Pearson) in 1997, with corrections and revisions and in a new format. For further information on this title and others in the series visit our website. There are PDF files of all of our texts freely available for download as well as instructions on how to order trade paperback copies.

Original Citation: *Real Analysis*, Andrew M. Bruckner, Judith B. Bruckner, Brian S. Thomson, Prentice-Hall, 1997, xiv 713 pp. [ISBN 0-13-458886-X]

New Citation: *Real Analysis, Second Edition,*, Andrew M. Bruckner, Judith B. Bruckner, Brian S. Thomson, ClassicalRealAnalysis.com, 2008, xiv 656 pp. [ISBN 1434844129]

Cover Design and Photography: David Sprecher

Date PDF file compiled: June 2, 2008

www.classicalrealanalysis.com

CONTENTS

PREFACE

Preface to the second edition

This second edition is a corrected, revised, and reprinted version of our original textbook. We are particularly grateful to readers who have sent in suggestions for corrections. Among them we owe a huge debt to R. B. Burckel (Kansas State University). Many of his corrections and suggestions are incorporated in this new edition. Thanks too to Keith Yates (Manchester Metropolitan University) who, while working on some of the more difficult problems, found some further errors.

Original Preface to first edition

In teaching first courses in real analysis over the years, we have found increasingly that the classes form rather heterogeneous groups. It is no longer true that most of the students are first-year graduate students in mathematics, presenting more or less common backgrounds for the course. Indeed, nowadays we find diverse backgrounds and diverse objectives among students in such classes. Some students are undergraduates, others are more advanced. Many students are in other departments, such as statistics or engineering. Some students are seeking terminal master's degrees; others wish to become research mathematicians, not necessarily in analysis.

We have tried to write a book that is suitable for students with minimal backgrounds, one that does not presuppose that most students will eventually specialize in analysis.

We have pursued two goals. First, we would like all students to have an opportunity to obtain an appreciation of the tools, methods, and history of the subject and a sense of how the various topics we cover develop naturally. Our second objective is to provide those who will study analysis further with the necessary background in measure, integration, differentiation, metric space theory, and functional analysis.

To meet our first goal, we do several things. We provide a certain amount of historical perspective that may enable a reader to see why a theory was needed and sometimes, why the researchers of the time had difficulty obtaining the “right” theory. We try to motivate topics before we develop them and try to motivate the proofs of some of the important theorems that students often find difficult. We usually avoid proofs that may appear “magical” to students in favor of more revealing proofs that may be a bit longer. We describe the interplay of various subjects—measure, variation, integration, and differentiation. Finally, we indicate applications of abstract theorems such as the contraction mapping principle, the Baire category theorem, Ascoli's theorem, Hahn-Banach theorem, and the open mapping theorem, to concrete settings of various sorts.

We consider the exercise sections an important part of the book. Some of the exercises do no more than ask the reader to complete a proof given in the text, or to prove an easy result that we merely state. Others involve simple applications of the theorems. A number are more ambitious. Some of these exercises extend the theory that we developed or present some related material. Others provide examples that we believe are interesting and revealing, but may not be well known. In general, the problems at the ends of the chapters are more substantial. A few of these problems can form the basis of projects for further study. We have marked exercises that are referenced in later parts of the book with a $\diamondsuit$ to indicate this fact.

When we poll our students at the beginning of the course, we find there are a number of topics that some students have seen before, but many others have not. Examples are the rudiments of metric space theory, Lebesgue measure in $\mathbb{R}^1$, Riemann–Stieltjes integration, bounded variation and the elements of set theory (Zorn's lemma, well-ordering, and others). In Chapter 1, we sketch some of this material. These sections can be picked up as needed, rather than covered at the beginning of the course. We do suggest that the reader browse through Chapter 1 at the beginning, however, as it provides some historical perspective.

Text Organization

Many graduate textbooks are finely crafted works as intricate as a fabric. If some thread is pulled too severely, the whole structure begins to unravel. We have hoped to avoid this. It is reasonably safe to skip over many sections (within obvious limitations) and construct a course that covers your own choice of topics, with little fear that the student will be forced to cross reference back through a maze of earlier skipped sections.

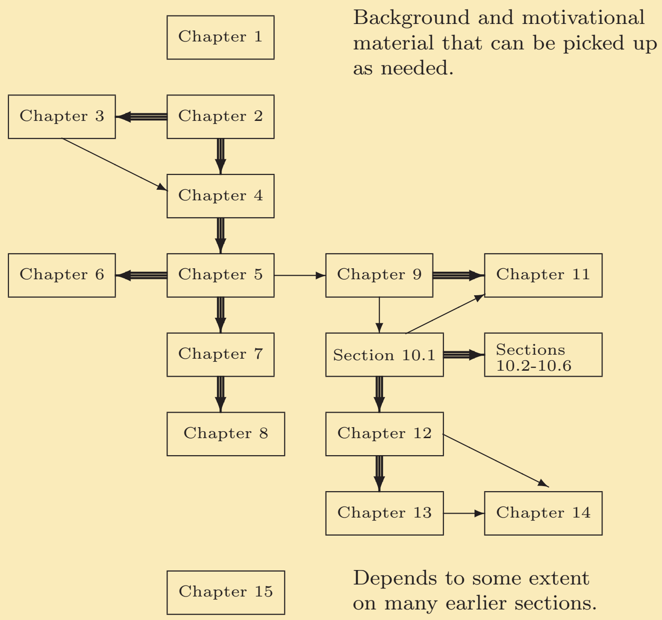
A word about the order of the chapters. The first chapter is intended as background reading. Some topics are included to help motivate ideas that reappear later in a more abstract setting. Zorn's lemma and the axiom of choice will be needed soon enough, and a classroom reference to Sections 1.3, 1.5 and 1.11 can be used.

The course can easily start with the measure theory of Chapter 2 and proceed from there. We chose to cover measure and integration before metric space theory because so many important metric spaces involve measurable or integrable functions. The rudiments of metric space theory are needed in Chapter 3, however, so we begin that chapter with a short section containing the necessary terminology.

Instructors who wish to emphasize functional analysis and reach Chapter 9 quickly can do so by omitting much of the material in the earlier chapters. One possibility is to cover Sections 2.1 to 2.6, 4.1, 4.2, and Chapter 5 and then proceed directly to Chapter 9. This will provide enough background in measure and integration to prepare the student for the later chapters.

Chapter 6 on the Fubini and Tonelli theorems is used only occasionally in the sequel (Sections 8.4 and 13.9). This is presented from the outer measure point of view because it fits better with the philosophy developed in Chapters 2 and 3. One can substitute any treatment in its place. Chapter 11 on analytic sets is not needed for the later chapters, and is presented as a subject of interest on its own merits. Chapter 13 on the L_p–spaces can be bypassed in favor of Chapter 14 or 15 except for a few points. Chapter 14 on Hilbert space could be undertaken without covering Chapters 12 and 13 since all material on the spaces ℓ_2 and L_2 is repeated as needed. Chapter 15 on Fourier series does not need the Hilbert space material in order to work, but, since it is intended as a showplace for many of the methods, it does draw on many other chapters for ideas and techniques.

The dependency chart gives a rough indication of how chapters depend on their predecessors. A strong dependency is indicated by a bold arrow, a weaker one by a fine arrow. The absence of an arrow indicates that no more than peripheral references to the earlier chapters are involved. Even when a strong dependency is indicated, the omission of certain sections near the en d of a chapter should not cause difficulties in later chapters. In addition, we have provided a number of concrete applications of abstract theorems. Many of these applications are not needed in later chapters. Thus an instructor who wishes to include material from all chapters in a year course for reasonably prepared students can do so by



1. Omitting some of the less central material such as 3.8 to 3.10, 5.10, 7.6 to 7.8, 8.4 to 8.7, 9.14 to 9.15, 10.2 to 10.6, and various material from the remaining chapters.

2. Sampling from the applications in Sections 9.8, 9.12, 9.14, 10.2 to 10.6, and 12.6.

3. Pruning sections from chapters from which no arrow emanates.

Acknowledgments

In writing this book we have benefitted from discussions with many students and colleagues. Special thanks are due to Dr. T. H. Steele who read the entire first draft of the manuscript and made many helpful suggestions. Several colleagues and many graduate students (at UCSB and SFU) worked through earlier drafts and found errors and rough spots. In particular we wish to thank Steve Agronsky, Hongjian Shi, Cristos Goodrow, Michael Saclolo, and Cliff Weil. We wish also to thank the following reviewers of the text for their helpful comments: Jack B. Brown, Auburn University; Krzysztof Ciesielski, West Virginia University; Douglas Hardin, Vanderbilt University; Hans P. Heinig, McMaster University; Morris Kalka, Tulane University; Richard J. O'Malley, University of Wisconsin–Milwaukee; Mitchell Taibelson, Washington University; Daniel C. Weiner, Boston University; and Warren R. Wogen, University of North–Carolina, Chapel Hill.

Chapter 1

BACKGROUND AND PREVIEW

In this chapter we provide a review and historical sampling of much of the background needed to embark on a study of the theory of measure, integration, and functional analysis. The setting here is the real line. In later chapters we place most of the theory in an abstract measure space or in a metric space, but the ideas all originate in the situation on the real line. The reader will have a background in elementary analysis, including such ideas as continuity, uniform continuity, convergence, uniform convergence, and sequence limits. The emphasis at this more advanced level shifts to a study of sets of real numbers and collections of sets, and this is what we shall address first in Sections 1.1 and 1.2.

Some of the basic ideas from set theory needed throughout the text are introduced in this chapter. The rudiments of cardinal and ordinal numbers appear in Sections 1.3 to 1.5. At certain points in the text we make extensive use of cardinality arguments and transfinite induction. The axiom of choice and its equivalent versions, Zermelo's theorem and Zorn's lemma, are discussed in Sections 1.3, 1.5, and 1.11. This material should be sufficient to justify these ideas, although a proper course of instruction in these concepts is recommended. We have tried to keep these considerations both minimal and intuitive. Our business is to develop the analysis without long lingering on the set-theoretic methods that are needed.

In Sections 1.7 to 1.10 we present two contrasting and competing theories of measure on the real line: the theory of Peano–Jordan content and the theory of Lebesgue measure. They serve as an introduction to the general theory that will be developed in Chapters 2 and 3. All the material here receives its full expression in the later chapters with complete proofs in the most general setting. The reader who works through the concepts and exercises in this introductory chapter should have an easier time of it when the abstract material is presented.

The notion of category plays a fundamental role in almost all aspects of analysis nowadays. In Section 1.6 the basics of this theory on the real line are presented. We shall explore this in much more detail in Chapter 10.

Borel sets and analytic sets play a key role in measure theory. These are covered briefly in Sections 1.12 and 1.13. The latter contains only a report on the origins of the theory of analytic sets. A full treatment appears in Chapter 11.

Sections 1.15 to 1.21 present the basics of integration theory on the real line. A quick review of the integral as viewed by Newton, Cauchy, Riemann, Stieltjes, and Lebesgue is a useful prelude to an approach to the modern theory of integration. We conclude with a generalized version of the Riemann integral that helps to complete the picture on the real line. We will return to these ideas in Section 5.10.

A brief study of functions of bounded variation appears in Section 1.14. This material, often omitted from an undergraduate education, is essential background for the student of general measure theory and, in any case, cannot be avoided by anyone wishing to understand the differentiation theory of real functions.

The exercises are designed to allow the student to explore the technical details of the subject and grasp new methods. The chapter can be read superficially without doing many exercises as a fast review of the background that is needed in order to appreciate the abstract theory that follows. It may also be used more intensively as a short course in the basics of analysis on the real line.

1.1 The Real Numbers

The reader is presumed to have a working knowledge of the real number system and its elementary properties. We use $\mathbb{R}$ to denote the set of real numbers. The natural numbers (positive integers) are denoted as $\mathbb{N}$, the integers (positive, negative, and zero) as $\mathbb{Z}$, and the rational numbers as $\mathbb{Q}$. The complex numbers are written as $\mathbb{C}$ and will play a role at a number of points in our investigation, even though the topic is called real analysis.

The extended real number system $\overline{\mathbb{R}}$, that is, $\mathbb{R}$ with the two infinities $+\infty$ and $-\infty$ appended, is used extensively in measure theory and analysis. One does not try to extend too many of the real operations to $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$: we shall write, though,

$$c + \infty = +\infty \quad \text{and} \quad c - \infty = -\infty$$

for any $c \in \mathbb{R}$.

Limits of sequences in $\mathbb{R}$ are defined using the metric

$$\rho(x, y) = |x - y| \quad (x, y \in \mathbb{R}).$$

This metric has the properties that one expects of a distance, properties that shall be used later in Chapter 9 to develop the concept of an abstract metric space.

1. $0 \leq \rho(x, y) < +\infty$, $(x, y \in \mathbb{R})$.

2. $\rho(x,y) = 0$ if and only if $x = y$.

3. $\rho(x,y) = \rho(y,x)$.

4. $\rho(x,y) \leq \rho(x,z) + \rho(z,y)$, $(x,y,z \in \mathbb{R})$.

We recall that sequence convergence in $\mathbb{R}$ means convergence relative to this distance. Thus $x_n \to x$ means that $\rho(x_n, x) = |x_n - x| \to 0$. A sequence $\{x_n\}$ is convergent if and only if that sequence is *Cauchy,* that is, if $\lim_{m,n\to\infty} \rho(x_m, x_n) = 0$. On the real line, sequences that are monotone and bounded are necessarily convergent. Virtually all the analysis on the real line develops from these fundamental notions.

1.1.1 Sets of real numbers

In the theory to be studied here, we require an extensive language for classifying sets of real numbers. The reader is familiar, no doubt, with most of the following concepts, which we present here to provide an easy reference and review. All these concepts will be generalized to an abstract metric space in Chapter 9.

Set notation throughout is standard. Thus union and intersection are written $A \cup B$ and $A \cap B$. Set difference is written $A \setminus B$, and so the complement of a set $A \subset \mathbb{R}$ will be written $\mathbb{R} \setminus A$. It is convenient to have a shorthand for this sometimes and we use $\widetilde{A}$ as well for this. The union and intersection of a family of sets $\mathcal{A}$ will appear as

$$\bigcup_{A \in \mathcal{A}} A \quad \text{and} \quad \bigcap_{A \in \mathcal{A}} A.$$

- A *limit point* of a set E or *point of accumulation* of a set E is any number that can be expressed as the limit of a convergent sequence of distinct points in E.
- The *closure* of a set E is the union of E together with its limit points. One writes $\overline{E}$ for the closure of E.
- An *interior point* of a set E is a point contained in an interval (a, b) that is itself entirely contained in E.
- The *interior* of a set E is the set of interior points of E. One writes E^o or perhaps $\text{int}(E)$ for the interior of E.
- An *isolated point* of a set is a member of the set that is not a limit point of the set.
- A *boundary point* of a set is a point of accumulation of the set that is not also an interior point of the set.
- A set G of real numbers is *open* if every point of G is an interior point of G.
- A set F of real numbers is *closed* if F contains all its limit points.
- A set of real numbers is *perfect* if it is nonempty, closed, and has no isolated points.
- A set of real numbers is *scattered* if it is nonempty and every nonempty subset has at least one isolated point.
- A set E of real numbers is *dense* in a set E_0 if every point in E_0 is a limit point of the set E.

- A set E of real numbers is *nowhere dense* if for every interval (a,b) there is a subinterval $(c,d) \subset (a,b)$ containing no points of E. (This is the same as asserting that E is dense in no interval.)

- A set E of real numbers is a *Cantor set* if it is nonempty, bounded, perfect, and nowhere dense.

In elementary courses one learns a variety of facts about these kinds of sets. We review some of the more important of these here, and the exercises explore further facts. All will play a role in our investigations of measure theory and integration theory on the real line.

1.1.2 Open sets and closed sets

To begin, one observes that the interval

$$(a,b) = \{x : a < x < b\}$$

is open and that the interval

$$[a,b] = \{x : a \leq x \leq b\}$$

is closed. The intervals

$$[a,b) = \{x : a \leq x < b\} \ \text{ and } \ (a,b] = \{x : a < x \leq b\}$$

are neither open, nor closed.

It is nearly universal now for mathematicians to lean toward the letter "G" to express open sets and the letter "F" to represent closed sets. The folklore is that the custom came from the French (*fermé* for closed) and the Germans (*Gebiet* for region). The following theorem describes the fundamental properties of the families of open and closed sets.

Theorem 1.1: *Let $\mathcal{G}$ denote the family of all open subsets of the real numbers and $\mathcal{F}$ the family of all closed subsets of the real numbers. Then*

1. *Each element in $\mathcal{G}$ is the complement of a unique element in $\mathcal{F}$, and vice versa.*
2. *$\mathcal{G}$ is closed under arbitrary unions and finite intersections.*
3. *$\mathcal{F}$ is closed under finite unions and arbitrary intersections.*
4. *Every set G in $\mathcal{G}$ is the union of a sequence of disjoint open intervals (called the components of G).*
5. *Given a collection $\mathcal{C} \subset \mathcal{G}$, there is a sequence $\{G_1, G_2, G_3, \dots\}$ of sets from $\mathcal{C}$ so that*

$$\bigcup_{G \in \mathcal{C}} G = \bigcup_{i=1}^{\infty} G_i.$$

Much more complicated sets than merely open sets or closed sets arise in many questions in analysis. If $\mathcal{C}$ is a class of sets, then frequently one is led to consider sets of the form

$$E = \bigcup_{i=1}^{\infty} C_i$$

for a sequence of sets $C_i \in \mathcal{C}$. We shall write $\mathcal{C}_\sigma$ for the resulting class. Similarly, we shall write $\mathcal{C}_\delta$ for the class of sets of the form

$$E = \bigcap_{i=1}^{\infty} C_i$$

for some sequence of sets $C_i \in \mathcal{C}$. The subscript σ denotes a summation (i.e., union) and δ denotes an intersection (from the German word *Durchschnitt*).

Continuing in this fashion, we can construct classes of sets of greater and greater complexity

$$\mathcal{C},\ \mathcal{C}_\delta,\ \mathcal{C}_\sigma,\ \mathcal{C}_{\delta\sigma},\ \mathcal{C}_{\sigma\delta},\ \mathcal{C}_{\delta\sigma\delta},\ \mathcal{C}_{\sigma\delta\sigma},\ldots,$$

which may play a role in the analysis of the sets $\mathcal{C}$.

These operations applied to the class $\mathcal{G}$ of open sets or the class $\mathcal{F}$ of closed sets result in sets of great importance in analysis. The class $\mathcal{G}_\delta$ and the class $\mathcal{F}_\sigma$ are just the beginning of a hierarchy of sets that form what is known as the Borel sets:

$$\mathcal{G} \subset \mathcal{G}_\delta \subset \mathcal{G}_{\delta\sigma} \subset \mathcal{G}_{\delta\sigma\delta} \subset \mathcal{G}_{\delta\sigma\delta\sigma} \ldots$$

and

$$\mathcal{F} \subset \mathcal{F}_\sigma \subset \mathcal{F}_{\sigma\delta} \subset \mathcal{F}_{\sigma\delta\sigma} \subset \mathcal{F}_{\sigma\delta\sigma\delta} \ldots.$$

A complete description of the class of Borel sets requires more apparatus than this might suggest, and we discuss these ideas in Section 1.12 along with some historical notes. Some elementary exercises now follow that will get the novice reader started in thinking along these lines.

Exercises

1:1.1 The classical *Cantor ternary set* is the subset of $[0,1]$ defined as

$$C = \left\{ x \in [0,1] : x = \sum_{n=1}^{\infty} \frac{i_n}{3^n} \text{ for } i_n = 0 \text{ or } 2 \right\}.$$

Show that C is bounded, perfect, and nowhere dense (i.e., C is a Cantor set in the terminology of this section).

1:1.2 List the intervals complementary to the Cantor ternary set in $[0, 1]$ and sum their lengths.

1:1.3 Let

$$D = \left\{ x \in [0,1] : x = \sum_{n=1}^{\infty} \frac{j_n}{3^n} \text{ for } j_n = 0 \text{ or } 1 \right\}.$$

Show $D + D = \{x + y : x, y \in D\} = [0, 1]$. From this deduce, for the Cantor ternary set C, that $C + C = [0, 2]$.

1:1.4 Criticize the following "argument" which is far too often seen:

> "If $G = (a, b)$ then $\overline{G} = [a, b]$. Similarly, if $G = \bigcup_{i=1}^{\infty}(a_i, b_i)$ is an open set, then $\overline{G} = \bigcup_{i=1}^{\infty}[a_i, b_i]$. It follows that an open set G and its closure $\overline{G}$ differ by at most a countable set."(?)

[Hint: Consider $G = (0, 1) \setminus C$ where C is the Cantor ternary set.]

1:1.5 Show that a scattered set is nowhere dense.

1:1.6 If $f : \mathbb{R} \to \mathbb{R}$ is continuous, then show that the set

$$f^{-1}(C) = \{x : f(x) = y \in C\}$$

is closed for every closed set C.

1:1.7 If f is continuous, then show that the set

$$f^{-1}(G) = \{x : f(x) = y \in G\}$$

is open for every open set G.

1:1.8◇ We define the *oscillation* of a real function f at a point x as

$$\omega_f(x) = \inf_{\delta>0} \sup \{|f(y) - f(z)| : y, z \in (x - \delta, x + \delta)\}.$$

Show that f is continuous at x if and only if $\omega_f(x) = 0$.

1:1.9 Show that the set $\{x : \omega_f(x) \geq \varepsilon\}$ is closed for each $\varepsilon \geq 0$.

1:1.10 For an arbitrary function f, show that the set of points where f is discontinuous is of type $\mathcal{F}_\sigma$.

1:1.11 For an arbitrary function f, show that the set of points where f is continuous is of type $\mathcal{G}_\delta$.

1:1.12 Prove the elementary parts (1, 2, and 3) of Theorem 1.1.

1:1.13 Prove part 4 of Theorem 1.1. Every open set G is the union of a unique sequence of disjoint open intervals, called the *components* of G.

1:1.14 Prove part 5 of Theorem 1.1 (Lindelöf's theorem). Given any collection $\mathcal{C}$ of open sets, there is a sequence $\{G_1, G_2, G_3, \dots\}$ of sets from $\mathcal{C}$ so that

$$\bigcup_{G \in \mathcal{C}} G = \bigcup_{i=1}^{\infty} G_i.$$

1:1.15 Show that every open interval may be expressed as the union of a sequence of closed intervals with rational endpoints. Thus every open interval is a $\mathcal{F}_\sigma$. (What about arbitrary open sets?)

1:1.16 What is $\mathcal{G} \cap \mathcal{F}$?

1:1.17 Show that $\mathcal{F} \subset \mathcal{G}_\delta$.

1:1.18 Show that $\mathcal{G} \subset \mathcal{F}_\sigma$.

1:1.19 Show that the complements of sets in $\mathcal{G}_\delta$ are in $\mathcal{F}_\sigma$, and conversely.

1:1.20 Find a set in $\mathcal{G}_\delta \cap \mathcal{F}_\sigma$ that is neither open nor closed.

1:1.21 Show that the set of zeros of a continuous function is a closed set. Given any closed set, show how to construct a continuous function that has precisely this set as its set of zeros.

1:1.22 A function f is upper semicontinuous at a point x if for every $\varepsilon > 0$ there is a $\delta > 0$ so that if $|x-y| < \delta$ then $f(y) > f(x) - \varepsilon$. Show that f is upper semicontinuous everywhere if and only if for every real α the set $\{x : f(x) \geq \alpha\}$ is closed.

1:1.23 Formulate a version of Exercise 1:1.22 for the notion of lower semicontinuity. [Hint: It should work in such a way that f is lower semicontinuous at a point if and only if $-f$ is upper semicontinuous there.]

1:1.24♢ Prove that, if $f_n \to f$ at every point, then

$$\{x : f(x) > \alpha\} = \bigcup_{m=1}^{\infty} \bigcup_{r=1}^{\infty} \bigcap_{n=r}^{\infty} \{x : f_n(x) \geq \alpha + 1/m\}.$$

1:1.25 Let $\{f_n\}$ be a sequence of real functions. Show that the set E of points of convergence of the sequence can be written in the form

$$E = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \bigcap_{m=N}^{\infty} \left\{x : |f_n(x) - f_m(x)| \leq \tfrac{1}{k}\right\}.$$

1:1.26 Let $\{f_n\}$ be a sequence of continuous real functions. Show that the set of points of convergence of the sequence is of type $\mathcal{F}_{\sigma\delta}$.

1:1.27 Show that every scattered set is of type $\mathcal{G}_{\delta}$.

1:1.28 Give an example of a scattered set that is not closed nor is its closure scattered.

1:1.29 Show that every set of real numbers can be written as the union of a set that is dense in itself (i.e., has no isolated points) and a scattered set.

1:1.30 Show that the union of a finite number of Cantor sets is also a Cantor set.

1.2 Compact Sets of Real Numbers

A closed, bounded set of real numbers is said to be *compact*. The concept of compactness plays a fundamental role in nearly all aspects of analysis. On the real line the notions are particularly easy to grasp and to apply. A basic theorem, often ascribed to Cantor (1845–1918), leads easily to many applications.

Theorem 1.2 (Cantor) *If $\{[a_i, b_i]\}$ is a nested sequence of closed, bounded intervals whose lengths shrink to zero, then the intersection*

$$\bigcap_{i=1}^{\infty} [a_i, b_i]$$

contains a unique point.

Here the sequence of intervals is said to be *nested* if, for each n,

$$[a_{n+1}, b_{n+1}] \subset [a_n, b_n].$$

The easy proof of this theorem can be obtained either by using the fact that monotone, bounded sequences converge (and hence a_n and b_n must converge) or by using the fact that Cauchy sequences converge (a sequence of points x_n chosen so that each $x_n \in [a_n, b_n]$ must be Cauchy). See Exercises 1:2.1 and 1:2.2.

1.2.1 Cousin covering theorem

Our next theorem is less well known. It was apparently first formulated by Pierre Cousin at the end of the nineteenth century. It asserts that a collection of intervals that contains all sufficiently small ones can be used to form a partition of any interval. The term *partition*, used often in elementary accounts of integration theory, here means a subdivision of an interval $[a,b]$ by points

$$a = x_0 < x_1 < \cdots < x_n = b$$

so that $[x_{i-1}, x_i]$ $(i = 1, 2, \ldots, n)$ are nonoverlapping subintervals of $[a,b]$ whose union is all of $[a,b]$.

Theorem 1.3 (Cousin) *Let $\mathcal{C}$ be a collection of closed subintervals of $[a,b]$ with the property that for every $x \in [a,b]$ there is a $\delta > 0$ so that $\mathcal{C}$ contains all intervals $[c,d] \subset [a,b]$ that contain x and have length smaller than δ. Then there are points*

$$a = x_0 < x_1 < \cdots < x_n = b$$

from $[a,b]$ so that each interval $[x_{i-1}, x_i] \in \mathcal{C}$ for all $i = 1, 2, \ldots, n$.

A proof is sketched in Exercises 1:2.3. Note that it can be made to follow from the Cantor theorem. We introduce some language that is useful in applying this theorem. Let us say that a collection of closed intervals $\mathcal{C}$ is *full* if it has the property of the theorem that it contains all sufficiently small intervals at any point x. Let us say that $\mathcal{C}$ is *additive* if whenever $[c,d]$ and $[d,e]$ are in $\mathcal{C}$ it follows that $[c,e] \in \mathcal{C}$. Then Cousin's theorem implies that any collection $\mathcal{C}$ of closed intervals that is both additive and full must contain *all* intervals.

1.2.2 Heine-Borel and Bolzano-Weierstrass theorems

Our remaining theorems are all consequences of the Cantor theorem or the Cousin theorem. The most economical approach to proving each is apparently provided by the Cousin theorem.

In each case, define a collection $\mathcal{C}$ of closed intervals, check that it is full and additive, and conclude that $\mathcal{C}$ contains all intervals. The exercises give the necessary hints on how to start as well as explain the terminology.

Theorem 1.4 (Heine–Borel) *Every open covering of a closed and bounded set of real numbers has a finite subcover.*

Theorem 1.5: *Every collection of closed, bounded sets of real numbers that has the finite intersection property, has a nonempty intersection.*

Theorem 1.6 (Bolzano–Weierstrass) *A bounded, infinite set of real numbers has a limit point.*

By a *compactness argument* in the study of sets and functions on $\mathbb{R}$, we understand any application of one of the theorems of this section. Often one can recognize a compactness argument most clearly in the process of reducing open covers to finite subcovers (Heine–Borel) or passing from a sequence to a convergent subsequence (Bolzano–Weierstrass). The reader is encouraged to try for a variety of proofs of the exercises that ask for a compactness argument. Hints are given that allow an application of Cousin's theorem. But one should develop the other techniques too, especially since in more general settings (metric spaces, topological spaces) a version of Cousin's theorem may not be available, and a version of the Heine–Borel theorem or the Bolzano–Weierstrass theorem may be.

Exercises

1:2.1 If $\{[a_i, b_i]\}$ is a nested sequence of closed, bounded intervals whose lengths shrink to zero, then the intersection $\bigcap_{i=1}^{\infty}[a_i, b_i]$ contains a unique point. Prove this by showing that both $\lim a_i$ and $\lim b_i$

exist and are equal.

1:2.2 If $\{[a_i, b_i]\}$ is a nested sequence of closed, bounded intervals whose lengths shrink to zero, then the intersection $\bigcap_{i=1}^{\infty}[a_i, b_i]$ contains a unique point. Prove this by selecting a point x_i in each $[a_i, b_i]$ and showing that $\{x_i\}$ is Cauchy.

1:2.3 Prove Theorem 1.3. [Hint: If there is no partition of $[a, b]$, then either there is no partition of $[a, \frac{1}{2}(a + b)]$ or else there is no partition of $[\frac{1}{2}(a + b), b]$. Construct a nested sequence of intervals and obtain a contradiction.]

1:2.4 Prove Theorem 1.3. [Hint: Consider the set S of all points $z \in (a, b]$ for which there is a partition of $[a, t]$ whenever $t < z$. Write $z_0 = \sup S$. Then $z_0 \in S$ (why?), $z_0 > a$ (why?), and $z_0 < b$ is impossible (why?). Hence $z_0 = b$ and the theorem is proved.]

1:2.5 Prove the Heine–Borel theorem: Let $\mathcal{S}$ be a collection of open sets covering a closed set E. Then, for every interval $[a, b]$, there is a finite subset of $\mathcal{S}$ that covers $E \cap [a, b]$. [Hint: Let $\mathcal{C}$ be the collection of closed subintervals I of $[a, b]$ for which there is a finite subset of $\mathcal{S}$ that covers $E \cap I$.]

1:2.6 Prove Theorem 1.5 directly from the Heine–Borel theorem. Here a family of sets has the *finite intersection property* if every finite subfamily has a nonempty intersection. [Hint: Take complements of the closed sets.]

1:2.7 Prove the Bolzano–Weierstrass theorem: If a set S has no limit points, then $S \cap [a, b]$ is finite for every interval $[a, b]$. [Hint: If x is not a limit point of S, then $S \cap [c, d]$ is finite for small intervals containing x.]

1:2.8 Show that if a function $f : \mathbb{R} \to \mathbb{R}$ is continuous, then it is uniformly continuous on every closed bounded interval. [Hint: Let $\varepsilon > 0$ and let $\mathcal{C}$ denote the set of intervals I such that, for some $\delta > 0$, $x, y \in I$ and $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$. Try also for other compactness arguments than Cousin's theorem.]

1:2.9 If f is continuous it is bounded on every closed bounded interval. [Hint: Let $\mathcal{C}$ denote the set of intervals I such that, for some $M > 0$ and all $x \in I$, $|f(x)| \leq M$.]

1:2.10 Prove the intermediate-value property: If f is continuous and never vanishes, then it is either always positive or always negative. [Hint: Let $\mathcal{C}$ denote the set of intervals $[a, b]$ such that $f(b)f(a) > 0$.]

1:2.11 If $f : \mathbb{R} \to \mathbb{R}$ is continuous and $K \subset \mathbb{R}$ is compact, show that $f(K)$ is compact. Is $f^{-1}(K)$ also necessarily compact?

1:2.12 [Dini] Suppose that $f_n : \mathbb{R} \to \mathbb{R}$ is continuous for each $n = 1, 2, 3, \ldots$, and $f_1(x) \geq f_2(x) \geq f_3(x) \geq \ldots$ and $\lim_{n\to\infty} f_n(x) = 0$ at each point. Prove that the convergence is uniform on every compact interval. [Hint: Consider all intervals $[a, b]$ such that there is a p so that, for all $n \geq p$ and all $x \in [a, b]$, $f_n(x) < \varepsilon$.]

1.3 Countable Sets

The *cardinality* of a finite set is merely the number of elements that the set possesses. For infinite sets a similar notion was made available by the fundamental work of Cantor in the 1870s. We can say that a finite set S has cardinality n if the elements of S can be placed in a one-one correspondence with the elements of the set $\{1, 2, 3, 4, \ldots, n\}$.

Similarly, we say an infinite set S has cardinality $\aleph_0$ if the elements of S can be placed in a one-one correspondence with the elements of the set $\mathbb{N}$ of natural numbers. More simply put, this says that the elements of S can be listed:

$$S = \{s_1, s_2, s_3, \ldots\}.$$

A set is *countable* (some authors say it is “at most countable”) if it has finite cardinality or cardinality $\aleph_0$. A set is *uncountable* if it is infinite but does not have cardinality $\aleph_0$. The choice of the first letter in the Hebrew alphabet (aleph, $\aleph$) to represent the transfinite cardinal numbers was made quite carefully by Cantor himself, and the notation is standard today.

To illustrate that these notions are not trivial, Cantor showed that any interval of real numbers is uncountable. Thus the points of an interval cannot be written in a list. The easiest and clearest proof is based on the fact that a nested sequence of intervals shrinks to a point. Cantor based his proof on a diagonal argument.

Theorem 1.7 (Cantor) *No interval $[a, b]$ is countable.*

Proof. Suppose not. Then the elements of $[a, b]$ can be arranged into a sequence $c_1, c_2, c_3, \ldots$. Select an interval $[a_1, b_1] \subset [a, b]$ so that $c_1 \notin [a_1, b_1]$ and so that $b_1 - a_1 < 1/2$. Continuing inductively, we find a nested sequence of intervals $\{[a_i, b_i]\}$ with lengths $b_i - a_i < 2^{-i} \to 0$ and with $c_i \notin [a_i, b_i]$ for each i.

By Theorem 1.2, there is a unique point $c \in [a, b]$ common to each of the intervals. This point cannot be equal to any c_i and this is a contradiction, since the sequence $c_1, c_2, c_3, \ldots$ was to contain every point of the interval $[a, b]$. ■

A comment must be made here about the method of proof. It is undoubtedly true that there is an interval $[a_1, b_1]$ with the properties that we require. It is also true that there is an interval $[a_2, b_2]$ with the properties that we require. But is it legitimate to make an infinite number of selections? One way to justify this is to make explicit in the rules of mathematics that we can make such infinite selections. This is provided by the axiom of choice that can be invoked when needed.

1.3.1 The axiom of choice

1.8 (Axiom of Choice) *Let $\mathcal{C}$ be any collection of nonempty sets. Then there is a function f defined on $\mathcal{C}$ so that $f(E) \in E$ for each $E \in \mathcal{C}$.*

The function f is called a choice function. That such a function exists is the same for us as the claim that an element can be chosen from each of the (perhaps) infinitely many sets. The original wording (translated from the German) of E. Zermelo from 1904 is instructive:

> For every subset M', imagine a corresponding element m_1', which is itself a member of M' and may be called the "distinguished" *[ausgezeichnete]* element of M'.

We can invoke this axiom in order to justify the proof we have just given. Alternatively, we can puzzle over whether, in this specific instance, we can obtain our proof without using this principle. Here is how to avoid using the axiom of choice in this particular instance, replacing it with an ordinary inductive argument. Suppose that I_1, I_2, I_3, ... is a list of all the closed intervals with rational endpoints. (See Exercise 1:3.7.) Then in our proof we announce a recipe for the choice of $[a_i, b_i]$ at each stage. At the kth step in the proof we simply find the first interval I_p in the sequence I_1, I_2, I_3, ... that has the three properties that

1. $I_p \subset [a_{k-1}, b_{k-1}]$,

2. $c_k \notin I_p$, and

3. the length of I_p is less than 2^{-k}.

Then we set $[a_k, b_k] = I_p$. Since, at each stage, only a finite number of intervals need be considered in order to arrive at our interval I_p, we need much less than the full force of the axiom of choice to make the determination for us.

In most aspects of real analysis the use of the axiom of choice is unavoidable and is undertaken without apology (or perhaps even without explicit mention). Later, in Section 1.10, when we construct a nonmeasurable set we shall have to invoke the axiom of choice; there we shall mention the fact quite clearly and comment on what is known about the situation if the axiom of choice were not to be allowed. In many other parts of this work we shall follow the usual custom of real analysts and apply the axiom when needed without much concern as to whether it can be avoided or not. This attitude has taken some time to develop. The early French analysts Baire, Borel, and Lebesgue relied on the axiom implicitly in their early works and then, after Zermelo gave a formal enunciation, reacted negatively. For most of his life Lebesgue remained deeply opposed, on philosophical grounds, to its use.[1]

Further material on the axiom of choice appears in Section 1.11. This axiom is known to be independent of the rest of the axioms of set theory known as ZF (Zermelo–Fraenkel set theory, without the axiom of choice). Kurt Gödel (1906–1978) showed that the axiom of choice is consistent with the remaining axioms provided one assumes that the remaining axioms are consistent themselves. (This is something that cannot be proved, only assumed.)

[1] For an interesting historical essay on the subject, see G. H. Moore, "Lebesgue's measure problem and Zermelo's axiom of choice: the mathematical effect of a philosophical dispute," *Ann. N. Y. Acad. Sci.*, 412 (1983), pp. 129–154.

Exercises

1:3.1 Show Theorem 1.7 using a diagonal argument (or find a proof in a standard text).

1:3.2 Prove that every subset of a countable set is countable.

1:3.3 Let S be countable and let S^k ($k \in \mathbb{N}$) denote the set of all sequences of length k formed of elements of S. Show that S^k is countable.

1:3.4 Prove that a union of a sequence of countable sets is countable.

1:3.5 Let S be countable. Show that the set of all sequences of finite length formed of elements of S is countable.

1:3.6 Show that the set of rational numbers is countable.

1:3.7$\diamondsuit$ Show that the set of intervals with rational numbers as endpoints is countable.

1:3.8 Show that the set of algebraic numbers is countable.

1:3.9 Show that every subset of a countable $\mathcal{G}_\delta$ set is again a countable $\mathcal{G}_\delta$ set.

1:3.10 Show that scattered sets are countable. [Hint: Consider all intervals (a, b) with rational endpoints such that $S \cap (a, b)$ is countable.]

1:3.11 Show that every Cantor set is uncountable.

1:3.12 Prove that every infinite set contains a subset that is infinite and countable. [Hint: Use the axiom of choice.]

1:3.13 (Cantor–Bendixson) Show that every closed set C of real numbers can be written as the union of a perfect set and a countable set. Moreover, there is only one decomposition of C into two disjoint sets, one perfect and the other countable.

1:3.14 Show that the set of discontinuities of a monotone, nondecreasing function f is (at most) countable. [Hint: Use the fact that the right-hand and left-hand limits $f(x+0)$ and $f(x-0)$ must both exist. Consider the sets

$$\{x : f(x+0) - f(x-0) < 1/n\}.$$

1:3.15 Let C be any countable set. Show that there is a monotone function f such that C is precisely the set of discontinuities of f. [Hint: Write $C = c_1, c_2, c_3, \ldots$ and construct $f(x) = \sum_{c_i<x} 2^{-i}$.]

1:3.16 Show that the family of all finite subsets of a countable set is countable.

1:3.17 Let $E \subset \mathbb{R}$ and let A consist of the right-isolated points of E (that is, $x \in A$ if $x \in E$ and there exists some $y > x$ so that $(x, y) \cap E = \emptyset$). Show that A is countable.

1:3.18$\diamondsuit$ Let $\mathcal{S}$ be a collection of nondegenerate closed intervals covering a set $E \subset \mathbb{R}$. Prove that there is a countable subset of $\mathcal{S}$ that also covers E. Show by example that there need not be a finite subset of $\mathcal{S}$ that covers E. [Hint: You may wish to use Exercise 1:3.17.]

1.4 Uncountable Cardinals

Every set can be assigned a *cardinal number* that denotes its size. So far we have listed just the cardinal numbers

$$0, 1, 2, 3, 4, \ldots, \aleph_0, \tag{1}$$

and we recall that the set of real numbers must have a cardinality different from these since it is infinite and is uncountable.

To handle cardinality questions for arbitrary sets, we require the following definitions and facts that can be developed from the axioms of set theory. If the elements of two sets A and B can be placed into a one-one correspondence, then we say that A and B are *equivalent* and we write $A \sim B$. For any two sets A and B, only three possibilities can arise:

1. A is equivalent to some subset of B and, in turn, B is equivalent to some subset of A.

2. A is equivalent to some subset of B, but B is equivalent to no subset of A.

3. B is equivalent to some subset of A, but A is equivalent to no subset of B.

The other possibility that might be imagined (that A is equivalent to no subset of B and B is equivalent to no subset of A) can be proved not to occur. In the first of these three cases, it can be proved that $A \sim B$ (Bernstein's theorem). These facts allow us to assign to every set A a symbol called *the cardinal number of* A. Then, if a is the cardinal number of A and if b is the cardinal number of B, cases 1, 2, and 3 can be described by the relations

1. $a = b$.

2. $a < b$.

3. $a > b$.

This orders the cardinal numbers and allows us to extend the list (1) above. We write $\aleph_1$ for the next cardinal in the list,

$$0 < 1 < 2 < 3 < 4 < \cdots < \aleph_0 < \aleph_1,$$

and we write c for the cardinality of the set $\mathbb{R}$. That the cardinals can be, in fact, written in such a list and that there is a "next" cardinal is one of the most important features of this subject. (This is called a well-order and is discussed in the next section.)

Cantor presumed that $c = \aleph_1$ but, despite great effort, was unable to prove it. It has since been established that this cannot be determined within the axioms of set theory and that those axioms are consistent if it is assumed and also consistent if it is negated. (More precisely, *if* the axioms of set theory are consistent, then they remain consistent if $c = \aleph_1$ is added or if $c > \aleph_1$ is added.) The assumption that $c = \aleph_1$ is called the *continuum hypothesis* (abbreviated CH) and is often assumed in order to construct exotic examples. But in all such cases one needs to announce clearly that the construction has invoked the continuum hypothesis.

Here are some of the rudiments of cardinal arithmetic, adequate for all the analysis that we shall pursue.

1. Let a and b be cardinal numbers for disjoint sets A and B. Then $a + b$ denotes the cardinality of the set $A \cup B$.

2. Let a and b be cardinal numbers for sets A and B. Then $a \cdot b$ denotes the cardinality of the Cartesian product set $A \times B$.

3. Let a_i $(i \in I)$ be cardinal numbers for mutually disjoint sets A_i $(i \in I)$. Then $\sum_{i\in I} a_i$ denotes the cardinality of the set $\bigcup_{i\in I} A_i$.

4. Let b be the cardinal number for a set B; then 2^b denotes the cardinality of the set of all subsets of B.

5. Finally, let a and b be cardinal numbers for sets A and B. Then a^b denotes the cardinality of the set of all functions mapping B into A.

For finite sets A and B, it is easy to count explicitly the sets in (iv) and (v). There are 2^b distinct subsets of B and there are a^b distinct functions mapping B into A. Note that with $A = \{0, 1\}$, so that $a = 2$, these two meanings in (iv) and (v) give the same cardinal in general.

(That is, the set of all subsets of B is equivalent to the set of all mappings from $B \to \{0,1\}$. See Exercise 1:4.5.)

This suggests a notation that we shall use throughout. By A^B we mean the set of functions mapping B into A. Hence by 2^B we mean the set of all subsets of B (sometimes called the power set of B).

One might wish to know the following theorems:

Theorem 1.9: *For every cardinal number a, $2^a > a$.*

Theorem 1.10: $\aleph_0 \cdot \aleph_0 = \aleph_0$.

Theorem 1.11: $c + \aleph_0 = c$ *and* $c + c = c$.

Theorem 1.12: $c \cdot c = c$.

Theorem 1.13: $2^{\aleph_0} = c$.

In particular, the continuum hypothesis can then be written as

$$\text{CH: } 2^{\aleph_0} = \aleph_1$$

which is its most familiar form.

Exercises

1:4.1 Prove that $(0,1) \sim \mathbb{R}$.

1:4.2 (Bernstein's theorem) If $A \sim B_1 \subset B$ and $B \sim A_1 \subset A$, then $A \sim B$. (Not at all an easy theorem.)

1:4.3 Prove that any open interval is equivalent to any closed interval without invoking Bernstein's theorem.

1:4.4 Show that every Cantor set has cardinality c.

1:4.5 Show that the set of all subsets of B is equivalent to the set of all mappings from $B \to \{0, 1\}$. [Hint: Consider χ_A for any $A \subset B$.]

1:4.6 Show that the class of functions continuous on the interval $[0, 1]$ has cardinality c. [Hint: If two continuous functions agree on each rational in $[0, 1]$, then they are identical.]

1:4.7$\diamondsuit$ Show that the family of all closed subsets of $\mathbb{R}$ has cardinality c.

1.5 Transfinite Ordinals

The set $\mathbb{N}$ of natural numbers is the simplest, nontrivial example of what we shall call a well-ordered set. The usual order (that is, $m < n$) on the natural numbers has the following properties.

1. For any $n \in \mathbb{N}$, it is not true that $n < n$.

2. For any distinct $n, m \in \mathbb{N}$, either $m < n$ or $n < m$.

3. For any n, m, $p \in \mathbb{N}$, if $n < m$ and $m < p$, then $n < p$.

4. Every nonempty subset $S \subset \mathbb{N}$ has a first element (i.e., there is an element $n_0 \in S$ so that $n_0 < s$ for every other element s of S).

It is precisely this set of properties that allows mathematical induction. Let P be a set of integers with the following properties:

1. $1 \in P$.

2. For all $n \in \mathbb{N}$, $m \in P$ for each $m < n$ implies that $n \in P$.

Then $P = \mathbb{N}$. Indeed, if P is not $\mathbb{N}$, then $P' = \mathbb{N} \setminus P$ is nonempty and so has a first element n_0. That element cannot be 1. All predecessors of n_0 are in P, which, by property (ii), implies that $n_0 \in P$, which is not possible.

Mathematical induction can be carried out on any set that has these four properties, and so we are not confined to induction on integers. We say that a set X is *linearly ordered* and that "$<$" is a *strict linear order* on X if properties (i), (ii), and (iii) hold for this set and this relation. We say that X is *well-ordered* if all four properties (i)–(iv) hold. If X is well-ordered and x_0 is in X, then the set of all elements that precede x_0 is called an *initial segment* of X.

The following two facts are fundamental. The first can be proved from the axiom of choice and is, in fact, equivalent to the axiom of choice. The second essentially defines the countable ordinals.

1.14 (Well-ordering principle) *Every set can be well-ordered. That is, for any nonempty set X there is a relation $<$ that is a strict linear order on X making it a well-ordered set.*

1.15 (Countable ordinals) *There exists an uncountable, well-ordered set X with an order relation $<$ so that*

1. *X has a last element denoted Ω.*

2. *For every $x_0 \in X$ with $x_0 \neq \Omega$ the initial segment*

$$\{x \in X : x < x_0\}$$

is countable.

3. *There is an element $\omega \in X$ such that*

$$\{x \in X : x < \omega\} = \{0, 1, 2, 3, \dots\}$$

and $<$ has its usual meaning in the set of nonnegative integers.

Thus the set $\{0, 1, 2, 3, \dots\}$ of nonnegative integers is an initial segment of X. We can think of X as looking like a long list starting with 0 and continuing just until uncountably many elements have been listed:

$$0 < 1 < 2 < \cdots < \omega < \omega + 1 < \omega + 2 < \cdots < \omega^2 < \omega^2 + 1 < \cdots < \Omega.$$

We call all the elements of X *ordinals.* Each element prior to ω is called a *finite ordinal.* Each element from then, but prior to the last one Ω, is called a *countable ordinal.* The element Ω is called *the first uncountable ordinal.*

We can identify an element x with the initial segment consisting of the elements that precede it. Thus each element of X can be thought of as a subset of X, and we see that each element (other than the last element Ω) is finite or countable considered as a set. The first infinite ordinal is ω and the first uncountable ordinal is Ω. The cardinality of Ω (i.e., the cardinality of $X \setminus \{\Omega\}$ or, the same thing, the cardinality of X) is $\aleph_1$. Unless we assume the continuum hypothesis, we do not know if this is c.

One can develop a bit of intuition about this situation by making the following observation. Any finite collection of finite ordinals ξ_1, ξ_2, $\dots \xi_n$ will stay away from ω in the sense that there is a finite ordinal ξ so that, for each i,

$$\xi_i < \xi < \omega.$$

The reason for this is that a finite union of finite sets is again finite. Similarly any countable collection of countable ordinals ξ_1, ξ_2, ... will stay away from Ω in the sense that there is a countable ordinal ξ so that, for each i,

$$\xi_i < \xi < \Omega.$$

The reason for this is that a countable union of countable sets is again countable. This observation is most useful.

If we do assume the continuum hypothesis (CH), then the real numbers (or any set of cardinality $2^{\aleph_0}$) can be well-ordered as described above. If we do not wish to assume CH, we can still perform a transfinite induction. In this case the version of Theorem 1.15 that we shall use is the following:

Lemma 1.16: *Any set X of cardinality $2^{\aleph_0}$ can be well-ordered in such a way that for each $x \in X$ the set of all predecessors of x has cardinality strictly less than $2^{\aleph_0}$.*

Every element, except the last, of a well-ordered set has an immediate successor defined as the first element of the set of all later elements; for any $x \in X$, if x is not the last element then the immediate successor of x can be written as $x + 1$. Note, however, that elements need not have immediate predecessors. Any element (ω and Ω in Theorem 1.15 are examples) that does not have an immediate predecessor is called a *limit ordinal.* We shall later define ordinals as *even* and *odd* in a way that extends the usual meaning. The first element 0 and every limit ordinal is thought of as even, a successor of an even is odd, and a successor of an odd is even. In this way every ordinal is designated as either odd or even.

This is admittedly a very sketchy introduction to the ordinals, but adequate for our purposes. The serious reader will take a course in transfinite arithmetic or consult textbooks that take the time to develop this subject from first principles.

1.5.1 A transfinite covering argument

As an illustration of the method of transfinite induction, let us prove a simple covering property of intervals using the ideas. We show that from a certain family of subintervals $[x, y) \subset [a, b)$ a disjoint subcover can be selected. The argument is, perhaps, the most transparent and intuitive use of a transfinite sequence.

Lemma 1.17: *Let $\mathcal{C}$ be a family of subintervals of $[a, b)$ such that for every $a \leq x < b$ there exists y, $x < y < b$ so that $[x, y) \in \mathcal{C}$. Then there is a countable disjoint subfamily $\mathcal{E} \subset \mathcal{C}$ so that*

$$\bigcup_{[x,y)\in\mathcal{E}} [x, y) = [a, b).$$

Proof. Set $x_0 = a$. By the hypotheses, we can choose an interval $[x_0, x_1) \in \mathcal{C}$ and then an interval $[x_1, x_2) \in \mathcal{C}$ and, once again, $[x_2, x_3) \in \mathcal{C}$, and so on. If $x_n \to b$, then take $\mathcal{E} = \{[x_{i-1}, x_i)\}$ and we are done. Otherwise, $x_n \to c$ with $c < b$. Then we can carry on with $[c, y_1)$, $[y_1, y_2)$, and so on, until we eventually reach b.

Well not quite! The idea seems sound, but a proper expression of this requires a transfinite sequence and transfinite induction. Set $x_0 = a$ and choose $x_1 < b$ so that $[x_0, x_1) \in \mathcal{C}$. Suppose that for each ordinal α we have chosen $x_\beta < b$ in such a way that $[x_\beta, x_{\beta+1}) \in \mathcal{C}$ for every β for which $\beta + 1 < \alpha$. Then we can choose x_α as follows: (i) If α is a limit ordinal, take $x_\alpha = \sup_{\beta<\alpha} x_\beta$. (ii) If α is not a limit ordinal, let α_0 be the immediate predecessor of α and suppose that $x_{\alpha_0} < b$. Take $x_\alpha < b$ so that $[x_{\alpha_0}, x_\alpha) \in \mathcal{C}$. The process stops if $x_{\alpha_0} = b$.

Inside each interval $[x_{\alpha-1}, x_\alpha)$ we can choose distinct rationals. Hence this process must stop in a countable number of steps. The family $\mathcal{E} = \{[x_{\alpha-1}, x_\alpha)\}$ is a countable disjoint subfamily of $\mathcal{C}$ so that $\bigcup_{[x,y)\in\mathcal{E}}[x, y) = [a, b)$. ■

Exercises

1:5.1 Prove the assertion 1.17 without using transfinite induction.

[Hint: Say that a point $z > a$ can be reached if there is a countable disjoint subfamily $\mathcal{E} \subset \mathcal{C}$ so that $\bigcup_{[x,y)\in\mathcal{E}}[x, y) \supset [a, z)$. Take the sup of all points that can be reached.]

1:5.2 Define a "natural" order on $\mathbb{N} \times \mathbb{N}$ and determine if it is a well-ordering.

1:5.3 Let A and B be linearly ordered sets. A natural order (the lexicographic order) on $A \times B$ is defined as $(a, b) \preceq (c, d)$ if $a \preceq c$ or if $a = c$ and $b \preceq d$. Show that this is a linear order. If A, B are well-ordered, then is this a well-ordering of $A \times B$? Describe the initial segments of $A \times B$.

1:5.4 A limit ordinal is an ordinal with no immediate predecessor. Show that ω and Ω are limit ordinals.

1.6 Category

Recall that a set E of real numbers is *nowhere dense* if for every open interval (a, b) there is a subinterval $(c, d) \subset (a, b)$ that contains no points of E. That is, it is nowhere dense if it is dense in no interval. Loosely, a nowhere dense set is shot full of holes.

A set is *first category* if it can be expressed as a union of a sequence of nowhere dense sets. Any set not of the first category is said to be of the *second category*. Nowhere dense sets are, in a certain sense, *very* small. Thus first category sets are, in the same sense, merely small. Second category sets are then not small. The complement of a first category set must apparently be quite large; such sets are said to be *residual.* Here, this notion of smallness should be taken as merely providing an intuitive guide to how these concepts can be interpreted.

1.6.1 The Baire category theorem on the real line

A fundamental theorem of René Baire (1874–1932) proved in 1899 asserts that every interval is second category. (It was proved too by W. F. Osgood two years earlier, but credit is almost always assigned to Baire.) Note that the proof here is nearly identical with the proof of the fact that intervals are uncountable; indeed, this theorem contains Theorem 1.7.

Theorem 1.18 (Baire) *No interval $[a, b]$ is first category.*

Proof. Suppose not. Then $[a, b]$ can be written as the union of a sequence of sets $C_1, C_2, C_3, \dots$ each of which is nowhere dense. Select an interval $[a_1, b_1] \subset [a, b]$ so that $C_1 \cap [a_1, b_1] = \emptyset$ and so that $b_1 - a_1 < 1/2$. Continuing inductively, we find a nested sequence of intervals $\{[a_i, b_i]\}$ with lengths $b_i - a_i < 2^{-i} \to 0$ and with $C_i \cap [a_i, b_i] = \emptyset$ for each i.

By Theorem 1.2, there is a unique point $c \in [a, b]$ common to each of the intervals. This point cannot belong to any C_i and this is a contradiction, since every point of the interval $[a, b]$ was to belong to some member of the sequence $C_1, C_2, C_3, \dots$. ■

A *category argument* is one that appeals to Baire's theorem. One can prove the existence of sets or points (or even functions) by these means. It has become one of the standard tools of the analyst and plays a fundamental role in many investigations.

1.6.2 An illustration of a category argument

We illustrate with an application showing that an important class of functions has certain continuity properties. A function f is said to be *in the first class of Baire* or *Baire 1* if it can be written as the pointwise limit of a sequence of continuous functions. A Baire 1 function need not be continuous. Does a Baire 1 function have any points of continuity? The existence of such points is obtained by a category argument.

Theorem 1.19 (Baire) *Every Baire 1 function is continuous except at the points of a set of the first category.*

Proof. Recall that we use $\omega_f(x)$ to denote the oscillation of the function f at a point x (see Exercise 1:1.8). The proof follows from the fact that for each $\varepsilon > 0$ the set of points

$$F(\varepsilon) = \{x : \omega_f(x) \geq \varepsilon\}$$

is nowhere dense. [This is because the set of points of discontinuity of f can be written as $\bigcup_{n=1}^{\infty} F(\frac{1}{n}$
Let I be any interval; let us search for a subinterval $J \subset I$ that misses $F(\varepsilon)$. The proof is complete once we find J.

Let f be the pointwise limit of a sequence of continuous functions $\{f_i\}$ and write

$$E_n = \bigcap_{i=n}^{\infty} \bigcap_{j=n}^{\infty} \{x \in I : |f_i(x) - f_j(x)| \leq \varepsilon/2\}.$$

Each set E_n is closed (since the f_i are continuous), and the sequence of sets E_n expands to cover all of I (since $\{f_i\}$ converges everywhere). By Baire's theorem (Theorem 1.18), there must be an interval $J \subset I$ and a set E_n dense in J. (Otherwise, we have just expressed I as the union of a sequence of nowhere dense sets, which is impossible.) But the sets here are closed, so this means merely that E_n *contains* the interval J. For this n (which is now fixed) we have

$$|f_i(x) - f_j(x)| \leq \varepsilon/2$$

for all i, $j \geq n$ and for all $x \in J$. In this inequality set $j = n$, and let $i \to \infty$ to obtain

$$|f(x) - f_n(x)| \leq \varepsilon/2.$$

Now we see that J misses the set $F(\varepsilon)$. Our last inequality shows that f is close to the continuous function f_n on J, too close to allow the oscillation of f at any point in J to be greater than ε. Thus there is no point in J that is also in $F(\varepsilon)$. ■

Theorem 1.19 very nearly characterizes Baire 1 functions. One needs to state it in a more general form, but one that can be proved by the same method. A function f is Baire 1 if and only if f has a point of continuity relative to any perfect set.

Exercises

1:6.1 Prove Theorem 1.18 using induction in place of the axiom of choice. (We used this axiom here without comment.) [Hint: See the discussion in Section 1.3.]

1:6.2 Show that every subset of a set of first category is first category.

1:6.3 Show that every finite set is nowhere dense, and show that every countable set is first category.

1:6.4 Show that every union of a sequence of sets of first category is first category.

1:6.5 Show that every intersection of a sequence of residual sets is residual.

1:6.6 Show that the complement of a set of second category may be either first or second category.

1:6.7 Prove that, if $\overline{E}$ is first category, then E is nowhere dense.

1:6.8 Show that a set of type $\mathcal{G}_\delta$ that is dense (briefly, "a dense $\mathcal{G}_\delta$") is residual.

1:6.9 Let $S \subset \mathbb{R}$. Call a point $x \in \mathbb{R}$ first category relative to S if there is some interval (a, b) containing x so that $(a, b) \cap S$ is first category. Show that the set

$$\{x \in S : \ x \text{ is first category relative to } S\}$$

is first category.

1:6.10 The rationals $\mathbb{Q}$ form a set of type $\mathcal{F}_\sigma$. Are they of type $\mathcal{G}_\delta$?

1:6.11 Does there exist a function continuous at every rational and discontinuous at every irrational? Does there exist a function continuous at every irrational and discontinuous at every rational? [Hint: Use Exercises 1:1.10 and 1:1.11.]

1:6.12 Let $f_n : [0, 1] \to \mathbb{R}$ be a sequence of continuous functions converging pointwise to a function f. Prove that, if the convergence is uniform, then there is a finite number M so that $|f_n(x)| < M$ for all n and all $x \in [0, 1]$. Even if the convergence is not uniform, show that there must be a subinterval $[a, b] \subset [0, 1]$ and a finite number M so that $|f_n(x)| < M$ for all n and all $x \in [a, b]$.

1:6.13 Theorem 1.19 as stated does not characterize Baire 1 functions. Show that a function is discontinuous except at the points of a first category set if and only if it is continuous at a dense set of points.

1:6.14 (Fort's theorem) If f is discontinuous at the points of a dense set, show that the set of points x, where $f'(x)$ exists, is of the first category.

1:6.15 If f is Baire 1, show that every set of the form $\{x : f(x) > \alpha\}$ is of type $\mathcal{F}_\sigma$ and every set of the form $\{x : f(x) \geq \alpha\}$ is of type $\mathcal{G}_\delta$. (The converse is also true.) [Hint: Use Exercise 1:1.24.]

1.7 Outer Measure and Outer Content

By the 1880s it was recognized that integration theory was intimately linked to the notion of measuring the "length" of subsets of $\mathbb{R}$ or the "area" of subsets of $\mathbb{R}^2$. Peano (1858–1932), Jordan (1838–1922), Cantor (1845–1918), Borel (1871–1956) and Lebesgue (1875–1941) are the main contributors to this development, but many authors addressed these problems.

At the end of the century there were two main competing notions that allowed the concept of length to be applied to all sets of real numbers. The Peano–Cantor–Jordan treatment defines a notion of *outer content* in terms of approximations that employ finite sequences of intervals. The Borel–Lebesgue method defines a notion of *outer measure* in terms of approximations that employ infinite sequences of intervals. The two methods are closely related, and it is, perhaps, best to study them together. The outer measure concept now dominates analysis and has left the outer content idea as a historical curiosity. Nonetheless, by seeing the two together and appreciating the difficulties that the early mathematicians had in coming to the correct ideas about measure, we can more easily learn this theory.

For any interval I we shall write $|I|$ for its length. Thus $|[a,b]| = |(a,b)| = b - a$ and $|(-\infty, a)| = |(b, +\infty)| = +\infty$. We include the empty set as an open interval and consider it to have zero length.

Definition 1.20: Let E be an arbitrary set of real numbers. We write

$$c^*(E) = \inf \left\{ \sum_{i=1}^{n} |I_i| : E \subset \bigcup_{i=1}^{n} I_i \right\}$$

and

$$\lambda^*(E) = \inf \left\{ \sum_{i=1}^{\infty} |I_i| : E \subset \bigcup_{i=1}^{\infty} I_i \right\},$$

where in the two cases $\{I_i\}$ is a finite (infinite) sequence of open intervals covering E.

We refer to the set function c^* as the outer content (or Peano-Jordan content) and λ^* as (Lebesgue) outer measure. Note that c^* is not of much interest for unbounded sets since it must assign the value $+\infty$ to each. Each of these set functions assigns a value (thought of as a "length") to each subset $E \subset \mathbb{R}$.

The following properties are essential and can readily be proved directly from the definitions. All the properties claimed for the Lebesgue outer measure in this chapter will be fully justified in Chapters 2 and 3.

Theorem 1.21: *The outer content and the outer measure have the following properties:*

1. $c^*(\emptyset) = \lambda^*(\emptyset) = 0$.
2. *For every interval* I, $c^*(I) = \lambda^*(I) = |I|$.
3. *For every set* E, $c^*(E) \geq \lambda^*(E)$.
4. *For every compact set* K, $c^*(K) = \lambda^*(K)$.
5. *For a finite sequence of sets* $\{E_i\}$, $c^*(\bigcup_{i=1}^n E_i) \leq \sum_{i=1}^n c^*(E_i)$.
6. *For any sequence of sets* $\{E_i\}$, $\lambda^*(\bigcup_{i=1}^\infty E_i) \leq \sum_{i=1}^\infty \lambda^*(E_i)$.
7. *Both* c^* *and* λ^* *are translation invariant.*
8. *For any set* E, $c^*(E) = c^*(\overline{E})$.

This last property, $c^*(E) = c^*(\overline{E})$, would nowadays be considered a flaw in the definition of a generalized length function. For a long time, though, it was felt that this property was essential: if a set $A \subset B$ is dense in B, then "surely" the two sets should be assigned the same length.

Exercises

1:7.1 Show that, for every interval I, $c^*(I) = \lambda^*(I) = |I|$.

1:7.2 Show that, for every set E, $c^*(E) \geq \lambda^*(E)$, and give an example to show that the inequality can occur.

1:7.3 Show that, for every compact set K, $c^*(K) = \lambda^*(K)$.

1:7.4 Show that, for any set E, $c^*(E) = c^*(\overline{E})$.

1:7.5$\diamondsuit$ Show that, for every finite sequence of sets $\{E_i\}$,

$$c^*\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} c^*(E_i).$$

1:7.6$\diamondsuit$ Show that, for every infinite sequence of sets $\{E_i\}$,

$$\lambda^*\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \lambda^*(E_i).$$

1:7.7 Show that both c^* and λ^* are translation invariant.

1:7.8$\diamondsuit$ Let G be an open set with components $\{(a_i, b_i)\}$. Show that

$$\lambda^*(G) = \sum_{i=1}^{\infty}(b_i - a_i),$$

but that $c^*(G)$ may be strictly larger.

1:7.9$\diamondsuit$ Let G be an open subset of an interval $[a, b]$ and write $K = [a, b] \setminus G$. Show that

$$c^*(K) = \lambda^*(K) = b - a - \lambda^*(G)$$

but that $c^*(K) = b - a - c^*(G)$ may be false.

1.8 Small Sets

In many studies of analysis there is a natural class of sets whose members are "small" or "negligible" for some purposes. We have already encountered the classes of countable sets, nowhere dense sets, and first category sets that can, with some justice, be considered small. In addition, the class of sets of zero outer content and the class of sets of zero outer measure also play the role of small sets in many investigations. Each of these classes enters into certain problems in that if a set is small in one of these senses it may be neglected in the analysis.

After some thought, one expects that in order to apply the term "small" to the members of some class of sets $\mathcal{S}$ one would require that finite (or perhaps countable) unions of small sets be small, that subsets of small sets be small, and that no interval be allowed to be small. More formally, the properties of $\mathcal{S}$ that seem to be desirable are as follows:

1. The union of a finite [countable] collection of sets in $\mathcal{S}$ is itself in $\mathcal{S}$.

2. Any subset of a set in $\mathcal{S}$ is itself in $\mathcal{S}$.

3. No interval (a, b) belongs to $\mathcal{S}$.

We say that $\mathcal{S}$ is an *ideal* of sets if properties (i) and (ii) hold. If the stronger version of (i) holds (with countable unions), then we say that $\mathcal{S}$ is a σ*-ideal* of sets. We have, by now, a number of different ideals of sets that can be viewed as composed of small sets. Let us summarize.

Theorem 1.22:

1. *The nowhere dense sets form an ideal.*
2. *The first category sets form a σ-ideal.*
3. *The finite sets form an ideal.*
4. *The countable sets form a σ-ideal.*
5. *The sets of outer content zero form an ideal.*
6. *The sets of outer measure zero form a σ-ideal.*

There are some obvious connections and some surprising contrasts. Certainly, finite sets are nowhere dense and of outer content zero. Countable sets are first category and of outer measure zero. The other relations are not so easy or so immediate. Let us first compare perfect, nowhere dense sets and sets of outer content zero.

1.8.1 Cantor sets

In the early days of the study of the Riemann integral (before the 1870s) it was recognized that sets of zero outer content played an important role as the sets that could be neglected in arguments. Nowhere dense sets at first appeared to be equally negligible, and there was some confusion as to the distinction. It is easy to check that a set of zero outer content must be nowhere dense; lacking any easy examples to the contrary, one might assume, as did a number of mathematicians, that the converse is also true. The following construction then comes as a bit of a surprise and shook the intuition of many nineteenth-century mathematicians. This shows that Cantor sets (nonempty, bounded, perfect, nowhere dense sets) can have relatively large measure (or content, since the two notions agree for compact sets) even though they appear to be small in some other sense. Constructions of this sort were given by H. J. Smith (1826–1883), du Bois-Reymond (1831–1889) and others.

Theorem 1.23: *Let $0 \leq \alpha < 1$. Then there is a Cantor set $C \subset [0,1]$ whose outer content (measure) is exactly α.*

Proof. Let $\alpha_1, \alpha_2, \dots$ be a sequence of positive numbers with

$$\sum_{k=1}^{\infty} \alpha_k = 1 - \alpha.$$

Let I_1 be an open subinterval of $I_0 = [0,1]$, with $|I_1| = \alpha_1$ chosen in such a way that the set $A_1 = I_0 \setminus I_1$ consists of two closed intervals, each of length less than $1/2$. At the second stage we shall remove from A_1 two further intervals, one from inside each of the two closed intervals, leaving $A_2 = I_0 \setminus (I_1 \cup I_2 \cup I_3)$ consisting of four intervals. We define the procedure inductively. After the nth stage, we have selected

$$1 + 2 + 2^2 + \cdots + 2^{n-1} = 2^n - 1$$

nonoverlapping open intervals $I_1, \dots, I_{2^n-1}$ with

$$\sum_{k=1}^{2^n-1} |I_k| = \sum_{i=1}^{n} \alpha_i,$$

and the set

$$A_n = I_0 \setminus \bigcup_{k=1}^{2^n-1} I_k$$

consists of 2^n closed intervals, each of length less than $1/n$, and $\lambda^*(A_n) = 1 - \sum_{i=1}^{n} \alpha_i$. (Note that the lengths of the closed intervals go to zero as n goes to infinity.)

Now let $C = \bigcap_{n=1}^{\infty} A_n$ and $B = I_0 \setminus C$. Then C is closed, B is open, and $B = \bigcup_{k=1}^{\infty} I_k$, with the intervals I_k pairwise disjoint. We see, by Exercise 1:7.8, that

$$\lambda^*(B) = \sum_{k=1}^{\infty} |I_k| = \sum_{k=1}^{\infty} \alpha_k = 1 - \alpha$$

and hence, by Exercise 1:7.9, that

$$\lambda^*(C) = 1 - \lambda^*(B) = \alpha.$$

Thus C is a nowhere dense closed subset of I_0 of measure α, and B is a dense open subset of I_0 of measure $1 - \alpha$. ■

1.8.2 Expressing the real line as the union of two "small" sets

Theorem 1.23 shows the contrast between sets of zero content and nowhere dense sets. As a result, we should not be surprised that there is a similar contrast between sets of outer measure zero and sets of the first category. The next theorem expresses this in a remarkable way. Every set of reals can be expressed as the union of two "small" sets (small in different ways). Be sure to notice that we are using outer *measure*, not outer content, in the theorem.

Theorem 1.24: *Every set of real numbers can be written as the disjoint union of a set of outer measure zero and a set of the first category.*

Proof. Let $\{q_i\}$ be a listing of all the rational numbers. Denote by I_{ij} that open interval centered at q_i and with length 2^{-i-j}. Write $G_j = \bigcup_{i=1}^{\infty} I_{ij}$ and $B = \bigcap_{j=1}^{\infty} G_j$. Each G_j is a dense open set, and so B is residual and hence its complement $\mathbb{R} \setminus B$ is first category. But it is easy to check that B has measure zero. Thus every set $A \subset \mathbb{R}$ can be written as

$$A = (A \cap B) \cup (A \setminus B)$$

which is, evidently, the union of a set of outer measure zero and a set of the first category. ■

Exercises

1:8.1 Show that every set of outer content zero is nowhere dense, but there exist dense sets of outer measure zero.

1:8.2 Show that every set of outer measure zero that is also of type $\mathcal{F}_\sigma$ is first category.

1:8.3 Show that no interval can be written as the union of a set of outer content zero and a set of the first category.

1:8.4 Show that a set E of real numbers has outer measure zero if and only if there is a sequence of intervals $\{I_k\}$ such that each point of E belongs to infinitely many of the intervals and $\sum_{k=1}^{\infty} |I_k| < +\infty$.

1:8.5 Let B and C be the sets referenced in the proof of Theorem 1.23.

(a) Prove that B is dense and open in $[0, 1]$, so C is nowhere dense and closed.

(b) Prove that C is perfect.

(c) Let $\{q_i\}$ be a listing of all the rational numbers. Denote by I_{ij} that open interval centered at q_i and with length 2^{-i-j}. Write $G_j = \bigcup_{i=1}^{\infty} I_{ij}$ and $B = \bigcap_{j=1}^{\infty} G_j$. Show that $\lambda^*(B) \leq \lambda^*(G_j) \leq 2^{-j}$ for each j, and deduce that $\lambda^*(B) = 0$.

(d) Prove Theorem 1.24 by using the fact that, in every interval $[a, b]$ and for every $\varepsilon > 0$, there is a Cantor set $C \subset [a, b]$ with measure exceeding $b - a - \varepsilon$.

1:8.6 Let $\mathcal{Z}$ be the class of all sets of real numbers that are expressible as countable unions of sets of outer content zero.

(a) Show that $\mathcal{Z}$ is a σ-ideal.

(b) Show that $\mathcal{Z}$ is precisely the σ-ideal of subsets of sets that are outer measure zero and $\mathcal{F}_\sigma$.

(c) Show that $\mathcal{Z}$ is not the σ-ideal of sets that are outer measure zero.

[Hint: Let C be a Cantor set whose intersection with each open interval is either empty or of positive outer measure. Choose a countable subset $D \subset C$, dense in C, and a $\mathcal{G}_\delta$ set $E \supset D$ of outer measure zero. Then $E \cap C$ is also outer measure zero but cannot be in $\mathcal{Z}$. (Use a Baire category argument.)]

1.9 Measurable Sets of Real Numbers

The outer measure and outer content have many desirable properties, but lack one that would seem to be an essential ingredient of a theory of lengths. They are not additive. If E_1 and E_2 are disjoint sets, then one expects the length of the union $E_1 \cup E_2$ to be the sum of the two lengths. In general, we have only that

$$c^*(E_1 \cup E_2) \leq c^*(E_1) + c^*(E_2)$$

and
$$\lambda^*(E_1 \cup E_2) \leq \lambda^*(E_1) + \lambda^*(E_2).$$
It is, however, not difficult to see that if E_1 and E_2 are not too "intertangled," then equality would hold. One seeks a class of sets on which the outer content or the outer measure is additive.

The key to creating these classes rests on a notion used by the Greeks in their investigations into area of plane figures. They considered that the area had been successfully found only if it had been computed by successive approximations from outside and by successive approximations from inside and that the two methods gave the same answer. Here our outer measure and outer content are obtained from outside approximations. Evidently, we should introduce an *inside* approximation, hence an inner measure and an inner content, and look for the class of sets on which the outer and inner estimates agree. In the case of content, this theory is due to Peano and Jordan. In the case of measure, the corresponding definition was used by Lebesgue.

Definition 1.25: Let E be a bounded set contained in an interval $[a, b]$. We write
$$c_*(E) = b - a - c^*([a, b] \setminus E)$$
and refer to $c_*(E)$ as the *inner content* of E and the set function c_* as the *inner content*.

Definition 1.26: Let E be a bounded set contained in an interval $[a, b]$. We write
$$\lambda_*(E) = b - a - \lambda^*([a, b] \setminus E)$$
and refer to $\lambda_*(E)$ as the *inner measure* of E and the set function λ_* as the *inner measure*.

It is left as an exercise to show that, in these two definitions, the particular interval $[a, b]$

that is chosen to contain the set E need not be specified. Measurability for bounded sets is defined as agreement of the inner and outer estimates.

Definition 1.27: A bounded set E is said to be *Peano–Jordan measurable* if $c_*(E) = c^*(E)$. A bounded set E is said to be *Lebesgue measurable* if $\lambda_*(E) = \lambda^*(E)$. An unbounded set E is measurable (in either sense) if $E \cap [a,b]$ is measurable in the same sense for each interval $[a,b]$. The class of Peano–Jordan measurable sets shall be denoted as $\mathcal{PJ}$. The class of Lebesgue measurable sets shall be denoted as $\mathcal{L}$.

When the inner and outer estimates agree, it makes sense to drop the subscripts and superscripts. Thus on the sets where $c_* = c^*$ we write $c = c_* = c^*$ and refer to c as the *content* or perhaps Peano–Jordan content. Similarly, on the Lebesgue measurable sets we write $\lambda = \lambda_* = \lambda^*$ and refer to λ as *Lebesgue measure.*

The families of sets so formed have strong properties, and the set functions c and λ defined on those families will have our desired additive properties. To have some language to express these facts, we shall use the following:

Definition 1.28: Let X be any set, and let $\mathcal{A}$ be a nonempty class of subsets of X. We say $\mathcal{A}$ is an *algebra of sets* if it satisfies the following conditions:

1. $\emptyset \in \mathcal{A}$.

2. If $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

3. If $A \in \mathcal{A}$, then $X \setminus A \in \mathcal{A}$.

It is easy to verify that an algebra of sets is closed also under differences, finite unions, and finite intersections. For any set X, the class 2^X of all subsets of X is obviously an algebra. So is the class $\mathcal{A} = \{\emptyset, X\}$. An algebra that is also closed under *countable* unions is said to be a *σ–algebra.* Many of the classes of sets that arise in measure theory are algebras or σ–algebras.

Definition 1.29: Let $\mathcal{A}$ be an algebra of sets and let ν be an extended real-valued function defined on $\mathcal{A}$. If ν satisfies the following conditions, we say that ν is an *additive set function.*

1. $\nu(\emptyset) = 0$.

2. If $A \in \mathcal{A}$, $B \in \mathcal{A}$, and $A \cap B = \emptyset$, then $\nu(A \cup B) = \nu(A) + \nu(B)$.

A nonnegative additive set function is often called a *finitely additive measure.* Note that, for an additive set function ν and every finite disjoint sequence $\{E_1, E_2, \dots E_n\}$ of sets from $\mathcal{M}$,

$$\nu\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n \nu(E_i).$$

In general, we shall prefer a countable version of this definition. We say that ν is a *countably additive set function* if, for every infinite disjoint sequence $\{E_1, E_2, \dots\}$ of sets from $\mathcal{M}$ whose union $\bigcup_{i=1}^\infty E_i$ is also in $\mathcal{M}$,

$$\nu\left(\bigcup_{i=1}^\infty E_i\right) = \sum_{i=1}^\infty \nu(E_i).$$

Using this language, we can now describe the classical measure theory developed in the nineteenth century by Peano, Jordan, and others and by Lebesgue at the beginning of the twentieth century. Peano–Jordan content is a finitely additive set function on an algebra of sets;

Lebesgue measure is a countably additive set function on a σ–algebra of sets. The theorems that now follow describe this formally. The first is not difficult. The second will be proved in full as part of our more general development in Chapter 2. It is worth attempting a proof of these two theorems now in order to appreciate the technical problems that arise in the subject.

Theorem 1.30: *Let $\mathcal{PJ}[a,b]$ denote the family of all Peano–Jordan measurable subsets of an interval $[a,b]$. Then the class $\mathcal{PJ}[a,b]$ forms an algebra of subsets of $[a,b]$, and $c = c_* = c^*$ is a finitely additive set function on that algebra.*

Theorem 1.31: *The class $\mathcal{L}$ forms a σ–algebra of subsets of $\mathbb{R}$, and $\lambda = \lambda_* = \lambda^*$ is a countably additive set function on that σ–algebra.*

Theorem 1.30 is largely a historical curiosity. Theorem 1.31 is one of the fundamental results of elementary measure theory. Chapter 2 contains a complete proof of this in a more general setting.

Exercises

1:9.1 Let E be a bounded set contained in an interval $[a,b] \subset [a_1,b_1]$. Show that

$$c_*(E) = b - a - c^*([a,b] \setminus E) = b_1 - a_1 - c^*([a_1,b_1] \setminus E).$$

This shows that the definition of the inner content does not depend on the containing interval.

1:9.2 Let E be a bounded set contained in an interval $[a,b] \subset [a_1,b_1]$. Show that

$$\lambda_*(E) = b - a - \lambda^*([a,b] \setminus E) = b_1 - a_1 - \lambda^*([a_1,b_1] \setminus E).$$

This shows that the definition of the inner measure does not depend on the containing interval.

1:9.3 Verify that an algebra of sets is closed also under differences, finite unions, and finite intersections.

1:9.4 Show that each of the following classes of subsets of a set X is an algebra:

(a) The class $\{\emptyset, X\}$.

(b) The class of all subsets of X.

(c) The class of subsets E of X such that either E or $X \setminus E$ is finite.

(d) The class of subsets of X that have outer content zero or whose complement has outer content zero (here $X \subset \mathbb{R}$).

1:9.5 Show that each of the following classes of subsets of a set X is a σ–algebra:

(a) The class of all subsets of X.

(b) The class of all subsets of X that are countable or have a countable complement.

(c) The class of subsets of X that have outer measure zero or whose complement has outer measure zero (here $X \subset \mathbb{R}$).

1:9.6 Let $\mathcal{A}_i$ be an algebra of subsets of a set X for each $i \in I$. Show that $\bigcap_{i\in I} \mathcal{A}_i$ is also an algebra.

1:9.7 Let $\mathcal{A}_i$ be a σ–algebra of subsets of a set X for each $i \in I$. Show that $\bigcap_{i\in I} \mathcal{A}_i$ is also a σ–algebra.

1:9.8$\diamondsuit$ Let $\mathcal{S}$ be a collection of subsets of a set X. Show that there is a smallest σ–algebra containing $\mathcal{S}$. (We call this the σ–algebra *generated by* $\mathcal{S}$.) [Hint: Consider the family of all σ–algebras that contain $\mathcal{S}$ (are there any?) and use Exercise 1:9.7.]

1:9.9 Show that every interval (closed, open, or half-closed) is both Peano–Jordan measurable and Lebesgue measurable.

1:9.10 Show that every set of outer content zero is Peano–Jordan measurable.

1:9.11 Show that every set of outer measure zero is Lebesgue measurable.

1:9.12♢ Suppose that a set E is Peano–Jordan measurable or Lebesgue measurable. Show that every translate $E + r = \{x + r : x \in E\}$ is also measurable in the same sense and has the same measure.

1:9.13♢ Show that the class of Peano–Jordan measurable sets and the class of Lebesgue measurable sets must both have cardinality 2^c. [Hint: Consider the subsets of a Cantor set of measure zero.]

1:9.14 Show that every Peano–Jordan measurable set is also Lebesgue measurable, but not conversely.

1:9.15 Theorems 1.30 and 1.31 might be misrepresented by saying that "c is merely finitely additive while λ is countably additive." Explain why it is that c is also countably additive.

1:9.16♢ Let E be a bounded subset of $\mathbb{R}$. Show that

$$\lambda_*(E) = \sup\{\lambda^*(F) : F \subset E, F \text{ closed}\}.$$

1:9.17 Prove that if $E_1 \subset E_2$ then $\lambda^*(E_1) \leq \lambda^*(E_2)$ and $\lambda_*(E_1) \leq \lambda_*(E_2)$.

1:9.18 Prove that both outer measure λ^* and inner measure λ_* are translation invariant functions defined on the class of all subsets of $\mathbb{R}$.

1:9.19 Show that $\lambda_*(E) \leq \lambda^*(E)$ for all $E \subset \mathbb{R}$.

1:9.20 Show that every σ–algebra of sets has either finitely many elements or uncountably many elements.

1.10 Nonmeasurable Sets

The measurability concept allows us to restrict the set functions c^* and λ^* to certain algebras of sets on which they are well behaved, in particular on which they are additive. Have we excluded any sets from consideration by this device? Are there sets that are so badly misbehaved with respect to the measurability definition that we cannot use them?

It is easy enough to characterize the class of Peano–Jordan measurable sets. Then we easily see which sets are not measurable and we see how to construct nonmeasurable sets. We address this first. The situation for Lebesgue measure is considerably more subtle and requires entirely different arguments.

Theorem 1.32: *A bounded set E of real numbers is Peano–Jordan measurable if and only if its set of boundary points has outer content zero.*

Proof. We may suppose that $\overline{E} \subset (a,b)$. Let $E_1 = \operatorname{int}(E)$, $E_2 = \overline{E} \setminus E_1$, and $E_3 = (a,b) \setminus \overline{E}$. Suppose that $c^*(E_2) = 0$; we show that E is Peano–Jordan measurable. Let $\varepsilon > 0$. Choose a finite collection of disjoint open subintervals $\{I_i\}$ of (a,b) covering E_2 so that $\sum |I_i| < \varepsilon$. Let us consider the intervals complementary to $\{\overline{I_i}\}$ in (a,b). These are of two types, the ones interior to E_1 and the ones interior to E_3. We call the former $\{J_i\}$ and the latter $\{K_i\}$. Note that $\{I_i\}$, $\{J_i\}$ together cover E and $\{I_i\}$, $\{K_i\}$ together cover $(a,b) \setminus E$.

We have

$$b - a = \sum |I_i| + \sum |J_i| + \sum |K_i|.$$

Hence

$$\begin{aligned} b - a &= \left(\sum |I_i| + \sum |J_i|\right) + \left(\sum |I_i| + \sum |K_i|\right) - \sum |I_i| \\ &\geq c^*(E) + c^*((a,b) \setminus E) - \varepsilon. \end{aligned}$$

Since ε is arbitrary, we can deduce that

$$c^*(E) + c^*((a,b) \setminus E) \leq b - a.$$

But the inequality

$$c^*(E) + c^*([a,b] \setminus E) \geq b - a$$

is true and

$$c^*([a,b] \setminus E) = c^*((a,b) \setminus E).$$

Thus $c^*(E) + c^*((a,b) \setminus E) = b - a$, and this establishes the measurability of the set E.

Conversely, suppose that we have this equality. Take a partition $\{I_i\}$ of $[a,b]$ using open intervals in such a way that

$$\sum \{|I_i| : I_i \cap E \neq \emptyset\} \leq c^*(E) + \varepsilon$$

and

$$\sum \{|I_i| : I_i \cap ([a,b] \setminus E) \neq \emptyset\} \leq c^*([a,b] \setminus E) + \varepsilon.$$

(We can do this by refining two partitions that handle each inequality separately.) Note that intervals that are used in both of these sums must contain a boundary point of E. Thus, because $b - a = \sum |I_i|$ and $c^*(E) + c^*([a,b] \setminus E) = b - a$, we can argue that

$$c^*(\overline{E} \setminus \operatorname{int}(E)) \leq \sum \{|I_i| : I_i \text{ contains a boundary point of } E\} \leq 2\varepsilon.$$

Since ε is arbitrary, $c^*(\overline{E} \setminus \operatorname{int}(E)) = 0$ as required. ■

In particular, note that it is an easy matter now to exhibit sets that are not Peano–Jordan measurable. The set of rational numbers in any interval must be nonmeasurable since every point is a boundary point. For a more interesting example, any Cantor set C will be Peano–Jordan measurable if and only if $c^*(C) = 0$ (see Exercise 1:10.1). We have seen in Theorem 1.23 how to construct Cantor sets in $[0,1]$ of positive outer content.

1.10.1 Existence of sets of real numbers not Lebesgue measurable

We turn now to a search for Lebesgue nonmeasurable sets. We can characterize Lebesgue measurable sets in a variety of ways. None of these, however, does anything to help to see whether there might exist sets that are nonmeasurable. The first proof that nonmeasurable sets must exist is due to G. Vitali (1875–1932). He showed that there cannot possibly exist a set function defined for all subsets of real numbers that is translation invariant, is countably additive, and extends the usual notion of length.

Theorem 1.33: *There exist subsets of $\mathbb{R}$ that are not Lebesgue measurable.*

Proof. Let $I = [-\frac{1}{2}, \frac{1}{2}]$. For $x, y \in I$, write $x \sim y$ if $x - y \in \mathbb{Q}$. For all $x \in I$, let

$$K(x) = \{y \in I : x - y \in \mathbb{Q}\} = \{x + r \in I : r \in \mathbb{Q}\}.$$

We show that $\sim$ is an equivalence relation. It is clear that $x \sim x$ for all $x \in I$ and that if $x \sim y$ then $y \sim x$. To show transitivity of $\sim$, suppose that $x, y, z \in I$ and $x - y = r_1$ and $y - z = r_2$ for $r_1, r_2 \in \mathbb{Q}$. Then $x - z = (x - y) + (y - z) = r_1 + r_2$, so $x \sim z$. Thus the set of all equivalence classes $K(x)$ forms a partition of I: $\bigcup_{x \in I} K(x) = I$, and if $K(x) \neq K(y)$, then $K(x) \cap K(y) = \emptyset$.

Let A be a set containing exactly one member of each equivalence class. (The existence of such a set A follows from the axiom of choice.) We show that A is nonmeasurable. Let $0 = r_0, r_1, r_2,$ be an enumeration of $\mathbb{Q} \cap [-1, 1]$, and define

$$A_k = \{x + r_k : x \in A\}$$

so that A_k is obtained from A by the translation $x \to x + r_k$.

Then

$$[-\tfrac{1}{2}, \tfrac{1}{2}] \subset \bigcup_{k=0}^{\infty} A_k \subset [-\tfrac{3}{2}, \tfrac{3}{2}]. \tag{2}$$

To verify the first inclusion, let $x \in [-\frac{1}{2}, \frac{1}{2}]$ and let x_0 be the representative of $K(x)$ in A. We have $\{x_0\} = A \cap K(x)$. Then $x - x_0 \in \mathbb{Q} \cap [-1, 1]$, so there exists k such that $x - x_0 = r_k$. Thus $x \in A_k$. The second inclusion is immediate: the set A_k is the translation of $A \subset [-\frac{1}{2}, \frac{1}{2}]$ by the rational number $r_k \in [-1, 1]$.

Suppose now that A is measurable. It follows (Exercise 1:9.12) that each of the translated sets A_k is also measurable and that $\lambda(A_k) = \lambda(A)$ for every k. But the sets $\{A_i\}$ are pairwise disjoint. If $z \in A_i \cap A_j$ for $i \neq j$, then $x_i = z - r_i$ and $x_j = z - r_j$ are in different equivalence classes. This is impossible, since $x_i - x_j \in \mathbb{Q}$. It now follows from (2) and the countable additivity of λ on $\mathcal{L}$ that

$$1 = \lambda([-\tfrac{1}{2}, \tfrac{1}{2}]) \leq \lambda(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \lambda(A_k) \leq \lambda([-\tfrac{3}{2}, \tfrac{3}{2}]) = 3. \tag{3}$$

Let $\alpha = \lambda(A) = \lambda(A_k)$. From (3), we infer that

$$1 \leq \alpha + \alpha + \cdots \leq 3. \tag{4}$$

But it is clear that no number α can satisfy both inequalities in (4). The first inequality implies that $\alpha > 0$, but the second implies that $\alpha = 0$. Thus A is nonmeasurable.

A variant of our argument (using Exercise 1:22.11) shows that $\lambda_*(A) = 0$ while $\lambda^*(A) > 0$. This, again, reveals why it is that A is nonmeasurable. ■

Many of the ideas that appear in this section, including the exercises, will reappear, in abstract settings as well as in concrete settings, in later chapters.

The proof has invoked the axiom of choice in order to construct the nonmeasurable set. One might ask whether it is possible to give a more constructive proof, one that does not use this principle. This question belongs to the subject of logic rather than analysis, and the logicians have answered it. In 1964, R. M. Solovay showed that, in Zermelo–Fraenkel set theory with a weaker assumption than the axiom of choice, it is consistent that all sets are Lebesgue measurable. On the other hand, the existence of nonmeasurable sets does not imply the axiom of choice. Thus it is no accident that our proof had to rely on the axiom of choice: it would have to appeal to some further logical principle in any case.

Exercises

1:10.1 Show that a Cantor set is Peano–Jordan measurable if and only if it has outer content zero.

1:10.2 Show that every set of positive outer measure contains a nonmeasurable set.

1:10.3 Show that there exist disjoint sets $\{E_k\}$ so that

$$\lambda^*\left(\bigcup_{k=1}^{\infty} E_k\right) < \sum_{k=1}^{\infty} \lambda^*\left(E_k\right).$$

1:10.4 Show that there exists a decreasing sequence of sets $E_1 \supset E_2 \supset E_3 \dots$ so that each $\lambda^*(E_k) < +\infty$ and

$$\lambda^*\left(\bigcap_{k=1}^{\infty} E_k\right) < \lim_{k\to\infty} \lambda^*\left(E_k\right).$$

1.11 Zorn's Lemma

In our brief survey we have already seen several points where an appeal to the axiom of choice was needed. This fundamental logical principle can be formulated in a variety of equivalent ways, each of use in certain situations.

The form we shall discuss now is called Zorn's lemma after Max Zorn (1906–1994). To express this, we need some terms from the language of partially ordered sets. A partially ordered set is a relaxation of a linearly ordered set as defined in Section 1.5. A relation $a \preceq b$, defined for certain pairs in a set S, is said to be a *partial order* on S, and $(S, \preceq)$ is said to be a *partially ordered set* if

1. For all $a \in S$, $a \preceq a$.
2. If $a \preceq b$ and $b \preceq a$, then $a = b$.
3. If $a \preceq b$ and $b \preceq c$, then $a \preceq c$.

The word "partial" indicates that not all pairs of elements need be comparable, only that the three properties here hold. A *maximal element* in a partially ordered set is an element $m \in S$ with nothing further in the order; that is, if $m \preceq a$ is true, then $a = m$.

The existence of maximal elements in partially ordered sets is of great importance. Zorn's lemma provides a criterion that can be checked in order to claim the existence of maximal elements. A *chain* in a partially ordered set is any subset that is itself linearly ordered. An upper bound of a chain is simply an element beyond every element in the chain. The language is suggestive, and pictures should help keep the concepts in mind.

Lemma 1.34 (Zorn) *If every chain in a partially ordered set has an upper bound, then the set has a maximal element.*

This assertion is, in fact, equivalent to the axiom of choice. We shall prove one direction just as an indication of how Zorn's lemma can be used in practice.

Let $\{A_i : i \in I\}$ be a collection of sets, each nonempty. We wish to show the existence of a choice function, that is, a function f with domain I such that $f(i) \in A_i$ for each $i \in I$. For any single given element $i_1 \in I$, we are assured that A_{i_1} is nonempty and hence we can choose some element $f(i_1) \in A_{i_1}$. We could do the same for any finite collection $\{i_1, i_2, \ldots, i_n\}$, but without appealing to some logical principle we cannot do this for *all* elements of I.

Zorn's lemma offers a technique. Define $\mathcal{F}$ as the family of all functions f such that

1. The domain of f is contained in I.

2. $f(i) \in A_i$ for each i in the domain of f.

We already know that there are some functions in $\mathcal{F}$. The choice function we want is presumably there too: it is any element of $\mathcal{F}$ with domain I.

Use $\operatorname{dom} f$ to denote the domain of a function f. Define a partial order on $\mathcal{F}$ by writing $f \preceq g$ to mean that $\operatorname{dom} f \subset \operatorname{dom} g$ and g is an extension of f. A maximal element of $\mathcal{F}$ must be our choice function. For, if f is maximal and yet the domain of f is not all of I, we can choose $i_0 \in I \setminus \operatorname{dom} f$ and some $x_{i_0} \in A_{i_0}$. Define g on $\operatorname{dom} f \cup \{i_0\}$ so that $g(i_0) = x_{i_0}$. Then g is an extension of f, and this contradicts the fact that f is to be maximal.

How do we prove the existence of a maximal element? Zorn's lemma allows us merely to verify that every chain has an upper bound. If $\mathcal{C} \subset \mathcal{F}$ is a chain, then there is a function h defined on $\bigcup_{g\in\mathcal{C}} \operatorname{dom} g$ so that h is an extension of each $g \in \mathcal{C}$. Simply take $h(i) = g(i)$ for any $g \in \mathcal{C}$ for which $i \in \operatorname{dom} g$. The fact that $\mathcal{C}$ is linearly ordered shows that this definition is unambiguous.

This completes the proof that Zorn's lemma implies the axiom of choice. All applications of Zorn's lemma will look something like this. The cleverness that may be needed is to interpret the problem at hand as a maximal problem in an appropriate partially ordered set.

Exercises

1:11.1 Let 2^X denote the set of all subsets of a nonempty set X. Show that the relation $A \subset B$ is a partial order on 2^X. Is it ever a linear order?

1:11.2 Let $\mathcal{F}$ denote the family of all functions $f : X \to Y$. Write $f \preceq g$ if the domain of g includes the domain of f and g is an extension of f. Show in detail that $(\mathcal{F}, \preceq)$ is a partially ordered set in which every chain has an upper bound.

1:11.3◇ Prove that there is a Hamel basis for the real numbers; that is, there exists a set $H \subset \mathbb{R}$ that is linearly independent over the rationals and that spans $\mathbb{R}$. (A set H is linearly independent over the rationals if given distinct elements $h_1, h_2, \ldots h_n \in H$ and any $r_1, r_2, \ldots r_n \in \mathbb{Q}$ with $\sum_{i=1}^n r_i h_i = 0$ then necessarily

$$r_1 = r_2 = \cdots = r_n = 0.$$

A set H spans $\mathbb{R}$ if for any $x \in \mathbb{R}$ there exist

$$h_1, h_2, \ldots h_n \in H \text{ and } r_1, r_2, \ldots r_n \in \mathbb{Q}$$

so that $\sum_{i=1}^n r_i h_i = x$.) [Hint: Find a maximal linearly independent set.]

1:11.4 Prove the axiom of choice assuming the well-ordering principle (that every set can be well-ordered). [Hint: Given $\{A_i : i \in I\}$ a collection of sets, each nonempty, well order the set $\bigcup_{i \in I} A_i$. Consider $c(A_i)$ as the first element in the set A_i in the order.]

1:11.5 Show that the following statement is equivalent to the axiom of choice: If $\mathcal{C}$ is a family of disjoint, nonempty subsets of a set X, then there is a set C that has exactly one element in common with each set in $\mathcal{C}$.

1.12 Borel Sets of Real Numbers

We have already defined several classes of sets that form the start of what is known as the Borel sets:

$$\mathcal{G} \subset \mathcal{G}_\delta \subset \mathcal{G}_{\delta\sigma} \subset \mathcal{G}_{\delta\sigma\delta} \subset \mathcal{G}_{\delta\sigma\delta\sigma} \dots$$

and

$$\mathcal{F} \subset \mathcal{F}_\sigma \subset \mathcal{F}_{\sigma\delta} \subset \mathcal{F}_{\sigma\delta\sigma} \subset \mathcal{F}_{\sigma\delta\sigma\delta} \dots .$$

Now, with transfinite ordinals available to us, we can continue this construction. The reason the transfinite ordinals are needed is that this process, which evidently can continue following a sequence of operations, does not terminate using an ordinary sequence.

The notation used above, while useful at the start of the process, will not serve us for long. Recall that the first ordinal 0 and every limit ordinal is thought of as even, the successor of an even ordinal is odd, and a successor of an odd ordinal is even.

We define the classes $\mathcal{F}_\alpha$ and $\mathcal{G}_\alpha$ for every ordinal $\alpha < \Omega$. We start by writing $\mathcal{F}_0 = \mathcal{F}$ and $\mathcal{G}_0 = \mathcal{G}$, $\mathcal{F}_1 = \mathcal{F}_\sigma$ and $\mathcal{G}_1 = \mathcal{G}_\delta$, $\mathcal{F}_2 = \mathcal{F}_{\sigma\delta}$ and $\mathcal{G}_2 = \mathcal{G}_{\delta\sigma}$. The classes $\mathcal{F}_\alpha$ and $\mathcal{G}_\alpha$ for every ordinal α are defined by taking countable intersections or countable unions of sets from the corresponding classes $\mathcal{F}_\beta$ and $\mathcal{G}_\beta$ for ordinals $\beta < \alpha$. If α is odd, then take $\mathcal{F}_\alpha$ as the class formed from countable unions of members from any classes $\mathcal{F}_\beta$ for $\beta < \alpha$. If α is even, then take $\mathcal{F}_\alpha$ as the class formed from countable intersections of members from any classes $\mathcal{F}_\beta$ for $\beta < \alpha$.

Similarly, if α is odd, then take $\mathcal{G}_\alpha$ as the class formed from countable intersections of members from any classes $\mathcal{G}_\beta$ for $\beta < \alpha$. If α is even, then take $\mathcal{G}_\alpha$ as the class formed from countable unions of members from any classes $\mathcal{G}_\beta$ for $\beta < \alpha$.

This process continues through all the countable ordinals by transfinite induction. For $\alpha = \Omega$, we find that the formation of countable intersections (to form $\mathcal{F}_\Omega$) or countable unions (to form $\mathcal{G}_\Omega$) does not create new sets (see Exercise 1:12.5). The collection of all sets formed by this process is called the Borel sets.

We list without proof some properties of the Borel sets on the line to give the flavor of the theory.

1.35: *The complement of a set of type $\mathcal{F}_\alpha$ is a set of type $\mathcal{G}_\alpha$, and the complement of a set of type $\mathcal{G}_\alpha$ is a set of type $\mathcal{F}_\alpha$.*

1.36: *The union and intersection of a* finite *number of sets of type $\mathcal{F}_\alpha$ ($\mathcal{G}_\alpha$) is of the same type.*

1.37: *Let $\alpha < \Omega$ be odd. Then the union of a* countable *number of sets of type $\mathcal{F}_\alpha$ is of the same type, and the intersection of a* countable *number of sets of type $\mathcal{G}_\alpha$ is of the same type.*

1.38: *Every set of type $\mathcal{F}_\alpha$ is of type $\mathcal{G}_{\alpha+1}$. Every set of type $\mathcal{G}_\alpha$ is of type $\mathcal{F}_{\alpha+1}$.*

1.39: *The Borel sets form the smallest σ–algebra of sets that contains the closed sets (the open sets).*

Thus one says that the Borel sets are generated by the closed sets (or by the open sets). (Exercise 1:9.8 shows that there must exist, independent of this theorem, a "smallest" σ–algebra containing any given collection of sets.) It is this form that we take as a definition in Chapter 3 for the Borel sets in a metric space.

Exercises

1:12.1 Show that the Borel sets form the smallest family of subsets of $\mathbb{R}$ that (i) contains the closed sets, (ii) is closed under countable unions, and (iii) is closed under countable intersections.

1:12.2 Show that the Borel sets form the smallest family of subsets of $\mathbb{R}$ that (i) contains the closed sets, (ii) is closed under countable *disjoint* unions, and (iii) is closed under countable intersections.

1:12.3 Show that the collection of all Borel sets has cardinality c.

1:12.4 Show that there must exist Lebesgue measurable sets that are not Borel sets. [Hint: Use Exercise 1:9.13.]

1:12.5 Show that the formation of countable intersections (to form $\mathcal{F}_\Omega$) or countable unions (to form $\mathcal{G}_\Omega$) does not create new sets. [Hint: All members of any sequence of sets from these classes must belong to one of the classes.]

1.13 Analytic Sets of Real Numbers

The Borel sets *clearly* form the largest class of respectable sets. This class is closed under all the reasonable operations that one might perform in analysis. Or so it seems.

In an important paper in 1905, Lebesgue made the observation that the projections of Borel sets in $\mathbb{R}^2$ onto the line are again Borel sets. The statement seems so reasonable and expected that he gave no detailed proof, assuming it to follow by methods he just sketched. The reader may know that the projection of a compact set in $\mathbb{R}^2$ is a compact set in $\mathbb{R}$ (any continuous image of a compact set is compact), and so any set that is a countable union of compact sets must project to a Borel set. It seems likely that one could prove that projections of other Borel sets must also be Borel by some obvious argument.

Lebesgue's assertion went unchallenged for ten years until the error was spotted by a young student in Moscow. Suslin, a student of Lusin, not only found the error, but reported to his professor that he was able to characterize the sets that could be expressed as projections of Borel sets and that he could produce an example of a projection of a Borel set that was not itself a Borel set.

Suslin calls a set $E \subset \mathbb{R}$ *analytic* if it can be expressed in the form

$$E = \bigcup_{(n_1,n_2,n_3,\dots)} \bigcap_{k=1}^{\infty} I_{n_1,n_2,n_3,\dots,n_k}$$

where each $I_{n_1,n_2,n_3,\dots,n_k}$ is a nonempty, closed interval for each

$$(n_1, n_2, n_3, \dots, n_k) \in \mathbb{N}^k$$

and each $k \in \mathbb{N}$, and where the union is taken over all possible sequences $(n_1, n_2, n_3, \dots)$ of natural numbers. Note that while the family of sets under consideration,

$$\{I_{n_1,n_2,n_3,\dots,n_k} : (n_1, n_2, n_3, \dots, n_k) \in \mathbb{N}^k\},$$

is countable the union involves uncountably many sets. Accordingly, this operation is substantially more complicated than the operations that preserve Borel sets. We shall call this the *Suslin operation*, although some authors, following Suslin himself, call it operation A.

In a short space of time Suslin, with the evident assistance of Lusin, established the basic properties of analytic sets and laid the groundwork for a vast amount of mathematics that has proved to be of importance for analysts, topologists, and logicians. We shall study this in some detail in Chapter 11. Here let us merely announce some of his discoveries. He obtained each of the following facts about analytic sets:

- All Borel sets are analytic.
- There is an analytic set that is not Borel.
- A set is Borel if and only if it and its complement are both analytic.
- Every analytic set in $\mathbb{R}$ is the projection of some $\mathcal{G}_\delta$ set in $\mathbb{R}^2$.
- Every uncountable analytic set has cardinality c.
- The projections of analytic sets are again analytic.

Thus in his short career (he died in 1919) Suslin established the fundamental properties of analytic sets, properties that exhibit the role that they must play. Lusin and his Polish colleague Sierpiński carried on the study in subsequent years, and by the end of the 1930s the study was quite complete and extensive. Let us mention two of their results that are important from the perspective of measure theory.

- All analytic sets are Lebesgue measurable.
- The Suslin operation applied to a family of Lebesgue measurable sets produces again a Lebesgue measurable set.

The study of analytic sets was well developed and well known in certain circles (mostly in Poland), but it did not receive a great deal of general attention until two main developments. In the 1950s a number of important problems in analysis were solved by employing the techniques associated with the study of analytic sets. In another direction it was discovered that most of the theory played an essential role in the study of descriptive set theory; since then all the methods and results of Suslin, Lusin, Sierpiński, and others have been absorbed by the logicians in their development of this subject.

We shall return to these ideas in Chapter 11 where we will explore the methods used to prove the statements listed here.

1.14 Bounded Variation

The following two problems attracted some attention in the latter years of the nineteenth century.

1.40: *What is the smallest linear space containing the monotonic functions?*

1.41: *For what class of functions f does the graph*

$$\{(x, y) : y = f(x)\}$$

have finite length?

Du Bois-Reymond, for one, attempted to solve Problem 1.40. He noted that, for a function f that is the integral of its derivative, one could write

$$f(x) = f(a) + \int_a^x [f'(t)]^+ \, dt - \int_a^x [f'(t)]^- \, dt,$$

where we are using the useful notation

$$[a]^+ = \max\{a, 0\} \quad \text{and} \quad [a]^- = \max\{-a, 0\}.$$

Clearly, this expresses f as a difference of monotone functions. This led him to a more difficult problem, which he was unable to resolve: Which functions are indefinite integrals of their derivatives? Unfortunately, this leads to a problem that will not resolve the original problem in any case.

Camille Jordan (1838–1922) solved both problems by introducing the class of functions of bounded variation. The functions of bounded variation play a central role in many investigations, notably in studies of rectifiability (as Problem 1.41 would suggest) and fundamental questions involving integrals and derivatives. They also lead to natural generalizations in the abstract study of measure and integration. For that reason, the student should be aware of the basic facts and methods that are developed in the exercises.

Let f be a real-valued function defined on a compact interval $[a, b]$. As in Section 1.2.1, let $\mathcal{P}$ be a partition of $[a, b]$, i.e., choose points

$$a = x_0 < x_1 < \cdots < x_n = b$$

and then

$$\mathcal{P} = \{[x_{i-1}, x_i] : i = 1, 2, \ldots, n\}$$

is a collection of nonoverlapping subintervals of $[a, b]$ whose union is all of $[a, b]$. Let

$$V(f, \mathcal{P}) = \sum_{j=1}^{n} |f(x_j) - f(x_{j-1})|.$$

The *variation* of f on $[a,b]$ is defined as

$$V(f;[a,b]) = \sup\{V(f,\mathcal{P}) : \mathcal{P} \text{ is a partition of } [a,b]\,\}.$$

When $V(f;[a,b])$ is finite, we say that f is of *bounded variation* on $[a,b]$. We then write f is BV on $[a,b]$, or f is BV when the interval is understood. (The variant VB is also in common usage because of the French *variation bornée*.)

The function $T(x) = V(f;[a,x])$ measures the variation on the interval $[a,x]$ and evidently is an increasing function. This is called the *total variation* of f. It is this that allows the solution of Problem 1.40, for one shows that

$$f(x) = T(x) - (T(x) - f(x))$$

expresses f as a difference of monotone functions (Exercise 1:14.10).

For the problems on arc length, we need the following definitions. Let f and g be real functions on an interval $[a,b]$. A curve C in the plane is considered to be the pair of parametric equations

$$x = f(t),\ y = g(t) \qquad (a \le t \le b).$$

The graph of the curve C is the set of points

$$\{(x,y) : x = f(t),\, y = g(t)\ (a \le t \le b)\}.$$

The length $\ell(C)$ of the curve C is defined as

$$\sup \sum_{j=1}^{n} \sqrt{(f(x_j) - f(x_{j-1}))^2 + (g(x_j) - g(x_{j-1}))^2},$$

where, as above, the supremum is taken over all partitions of $[a,b]$. The curve is said to be *rectifiable* if this is finite. Such a curve is rectifiable precisely when both functions f and g have bounded variation (Exercise 1:14.14). The graph of a function f is rectifiable precisely when f has bounded variation (Exercise 1:14.16).

Exercises

1:14.1 Show that a monotonic function on $[a,b]$ is BV.

1:14.2 Show that a continuous function with a finite number of local maxima and minima on $[a,b]$ is BV.

1:14.3 Show that a continuously differentiable function on $[a,b]$ is BV.

1:14.4 Show that a function that satisfies a Lipschitz condition on $[a,b]$ is BV.

[A function f is said to satisfy a *Lipschitz condition* if, for some constant M, $|f(x) - f(y)| \leq M|y - x|$. These conditions were introduced by Rudolf Lipschitz (1832–1903) in an 1876 study of differential equations.]

1:14.5 Estimate the variation of the function $f(x) = x \sin x^{-1}$, $f(0) = 0$, on the interval $[0,1]$.

1:14.6 Estimate the variation of the function $f(x) = x^2 \sin x^{-1}$, $f(0) = 0$, on the interval $[0,1]$.

1:14.7 Prove that, if f is BV on $[a,b]$, then f is bounded on $[a,b]$.

1:14.8 Show that the class of functions of bounded variation on $[a,b]$ is closed under addition, subtraction, and multiplication. If f and g are BV, and g is bounded away from zero, then f/g is BV.

1:14.9♢ Show that if f is BV on $[a,b]$ and $a \leq c \leq b$, then

$$V(f;[a,b]) = V(f;[a,c]) + V(f;[c,b]).$$

1:14.10♢ Show that a function f is BV on $[a,b]$ if and only if there exist functions f_1 and f_2 that are nondecreasing on $[a,b]$, and $f(x) = f_1(x) - f_2(x)$ for all $x \in [a,b]$. [Hint: Let $V(x) = V(f;[a,x])$. Verify that $V - f$ is nondecreasing on $[a,b]$ and use $f = V - (V - f)$.]

1:14.11 Show that the set of discontinuities of a function of bounded variation is (at most) countable. [Hint: See Exercise 1:3.14.]

1:14.12 Show that if f is BV on $[a,b]$, with variation $V(x) = V(f;[a,x])$, then

$$\{x : f \text{ is right continuous at } x\} = \{x : V \text{ is right continuous at } x\}.$$

1:14.13 Let $\{f_n\}$ be a sequence of functions, each BV on $[a,b]$ with variation less than or equal to some number M. If $f_n \to f$ pointwise on $[a,b]$, show that f is BV on $[a,b]$ with variation no greater than M.

1:14.14 Show that the graph of a curve C in the plane, given by the pair of parametric equations

$$x = f(t),\ y = g(t) \qquad (a \le t \le b)$$

is rectifiable if and only if both f and g have bounded variation on $[a,b]$. [Hint: $|x|,\ |y| \le \sqrt{x^2+y^2} \le |x|+|y|$.]

1:14.15 Show that the length of a curve C in the plane, given by the pair of parametric equations $x = f(t)$, $y = g(t)$ $(a \le t \le b)$, is the integral

$$\int_a^b \sqrt{[f'(t)]^2 + [g'(t)]^2}\, dt$$

if f and g are continuously differentiable.

1:14.16 Show that the graph of a function f is rectifiable if and only if f has bounded variation on $[a,b]$.

1:14.17$\diamondsuit$ Let $f : [a,b] \to \mathbb{R}$. We say that f is *absolutely continuous* if for each $\varepsilon > 0$ there exists $\delta > 0$ such that, if $\{[a_n,b_n]\}$ is any finite or countable collection of nonoverlapping closed intervals in $[a,b]$ with $\sum_{k=1}^{\infty}(b_k - a_k) < \delta$, then

$$\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| < \varepsilon.$$

This concept plays a significant role in the integration theory of real functions. Show that an absolutely continuous function is both continuous and of bounded variation.

1:14.18 Give a natural definition for a complex-valued function on a real interval $[a,b]$ to have bounded variation. Prove that a complex-valued function has bounded variation if and only if its real and imaginary parts have bounded variation.

1.15 Newton's Integral

We embark now on a tour of classical integration theory leading up to the Lebesgue integral. The reader will be familiar to various degrees with much of this material, since it appears in a variety of undergraduate courses. Here we need to clarify many different themes that come together in an advanced course in measure and integration.

The simplest starting point is the integral as conceived by Newton. For him the integral is just an inversion of the derivative. In the same spirit (but not in the same technical way that he would have done it) we shall make the following definition.

Definition 1.42: A real-valued function f defined on an interval $[a,b]$ is said to be Newton integrable on $[a,b]$ if there exists an antiderivative of f, that is, a function F on $[a,b]$ with $F'(x) = f(x)$ everywhere there. Then we write

$$(N)\int_a^b f(x)\,dx = F(b) - F(a).$$

The mean-value theorem shows that the value is well defined and does not depend on the particular primitive function F chosen to evaluate the integral. This integral must be considered *descriptive* in the sense that the property of integrability and the value of the integral are determined by the existence of some object for which no construction or recipe is available. If, perchance, such a function F can be found, then the value of the integral is determined, but otherwise there is no hope, a priori, of finding the integral or even of knowing whether it exists.

One might wish to call this the *calculus integral* since, in spite of the many texts that teach constructive definitions for integrals, most freshman calculus students hardly ever view an integral as anything more than a determination of an antiderivative.

At this point let us remark that this integral is handling functions that are not handled by other methods. The integrals of Cauchy and of Riemann, discussed next, require a fair bit of continuity in the function and do not tolerate much unboundedness. But derivatives can be unbounded and derivatives can be badly discontinuous. We know that a derivative is Baire 1 and that Baire 1 functions are continuous except at the points of a first category set; this first category set can, however, have positive measure, and this will interfere with integrability in the senses of Cauchy or Riemann. Thus, while this integral may seem quite simple and unassuming, it is involved in a process that is more mysterious than might appear at first glance. Attempts to understand this integral will take us on a long journey.

Exercises

1:15.1 Show that the mean-value theorem can be used to justify the definition of the Newton integral.

1:15.2 Show that a derivative f' of a continuous function f is Baire 1 and has the intermediate-value property. [Hint: Consider $f_n(x) = n^{-1}(f(x+n^{-1}) - f(x))$. The intermediate-value property can be deduced from the mean-value theorem.]

1:15.3 Show that a derivative on a finite interval can be unbounded.

1:15.4 Which of the elementary properties of the Riemann integral hold for the Newton integral? For example, can we write

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx?$$

1.16 Cauchy's Integral

A first course in calculus will include a proper definition of the integral that dates back to the middle of the nineteenth century and is generally attributed to Bernhard Riemann (1826–1866). Actually, Augustin Cauchy (1789–1857) had conceived of such an integral a bit earlier, but Cauchy limited his study to continuous functions. Here is Cauchy's definition, stated in modern language but essentially as he would have given it in 1823 in his lessons at the École Polytechnique.

Let f be continuous on $[a,b]$ and consider a partition P of this interval:

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

Form the sum

$$S(f,P) = \sum_{i=1}^{n} f(x_{i-1})(x_i - x_{i-1}).$$

Let $\|P\| = \max_{1\le i\le n}(x_i - x_{i-1})$ and define

$$\int_a^b f\,(x)\;dx = \lim_{\|P\|\to 0} S(f,P).$$

Cauchy showed that this limit exists.

Prior to Cauchy, such a definition of integral might not have been possible. The modern notion of "continuity" was not available (it was advanced by Cauchy in 1821), and even the proper definition of "function" was in dispute. Cauchy also established a form of the fundamental theorem of calculus.

Theorem 1.43: *Let f be continuous on $[a, b]$, and let*

$$F(x) = \int_a^x f(t)\, dt \quad (a \leq x \leq b).$$

Then F is differentiable on $[a, b]$, and $F'(x) = f(x)$ for all $x \in [a, b]$.

Theorem 1.44: *Let F be continuously differentiable on $[a, b]$. Then*

$$F(b) - F(a) = \int_a^b F'(x)\, dx.$$

Thus, for continuous functions, Cauchy offers an integral that is constructive and agrees with the Newton integral. There are, however, unbounded derivatives, and so the Newton integral remains more general than Cauchy's version.

1.16.1 Cauchy's extension of the integral to unbounded functions

To handle unbounded functions, Cauchy introduces the following idea, one that survives to this day in elementary calculus courses, usually under the unfortunate term "improper integral." Let us introduce it in a more formal manner, one that leads to a better understanding of the structure.

Let f be a real function on an interval $[a, b]$. A point $x_0 \in [a, b]$ is a point of unboundedness of f if f is unbounded in every open interval containing x_0. Let S_f denote the set of points of unboundedness. If S_f is a finite set and f is continuous at every point of $[a,b]\backslash S_f$, there is some hope of obtaining an integral of f. Certainly, we know the value of $\int_c^d f(t)\,dt$ for every interval $[c,d]$ disjoint from S_f. It is a matter of extending these values. Cauchy's idea is to obtain, for any $c,d \in S_f$ with $(c,d)\cap S_f=\emptyset$,

$$\int_c^d f(t)\,dt = \lim_{\varepsilon_1\searrow 0,\,\varepsilon_2\searrow 0} \int_{c+\varepsilon_1}^{d-\varepsilon_2} f(t)\,dt$$

Then, in a finite number of steps, one can extend the integral to $[a,b]$, providing only that each limit as above exists. A function is *Cauchy integrable* on an interval $[a,b]$ provided that S_f is finite, f is continuous at each point of $[a,b]$ excepting the points in S_f and all the limits above exist.

One important feature of this integral is its *nonabsolute* character. A function f may be integrable in Cauchy's sense on an interval $[a,b]$ and yet the absolute value $|f|$ may not be. An easy example is the function $f(x)=F'(x)$ on $[0,1]$, where $F(x)=x^2\sin x^{-2}$. Here $S_f=\{0\}$ and f is continuous away from 0. Obviously, f is Cauchy integrable on $[0,1]$, and yet $|f|$ is not. Somehow the "cancelations" that take place for integrating f do not occur for $|f|$, since

$$\lim_{\varepsilon\searrow 0}\int_\varepsilon^1 |f(t)|\,dt=+\infty.$$

This can be considered as the analog in integration theory of the fact that $\sum_{i=1}^\infty (-1)^i/i$ exists and yet $\sum_{i=1}^\infty 1/i=+\infty$.

Finally, we mention Cauchy's method for handling unbounded intervals. The procedure above for determining the integral of a continuous function on a bounded interval $[a,b]$ does not immediately extend to the unbounded intervals $(-\infty,a]$, $[a,+\infty)$, or $(-\infty,+\infty)$. Cauchy handled these in a now familiar way. He defines

$$\int_{-\infty}^{+\infty} f(x)\,dx = \lim_{s,t\to+\infty} \int_{-s}^{t} f(x)\,dx.$$

Note that this integral, too, is a nonabsolute integral.

Exercises

1:16.1 Let S_f denote the set of points of unboundedness of a function f. Show that S_f is closed.

1:16.2 Cauchy also considered symmetric limits of the form

$$\lim_{t\to 0+} \left(\int_{a}^{b-t} f(x)\,dx + \int_{b+t}^{c} f(x)\,dx \right)$$

as "principal-value" limits. Give an example to show that these can exist when the ordinary Cauchy integral does not.

1:16.3 Cauchy also considered symmetric limits for unbounded intervals

$$\lim_{t\to+\infty} \int_{-t}^{t} f(x)\,dx.$$

as "principal value" limits. Give an example to show that this can exist when the ordinary Cauchy integral does not.

1:16.4 Let $f(x) = x^2 \sin x^{-2}$, $f(0) = 0$ and show that f' is an unbounded derivative on $[0,1]$ integrable by both Cauchy and Newton's methods to the same value. Show that $|f|$ is not integrable by either method.

1.17 Riemann's Integral

Riemann extended Cauchy's concept of integral to include some bounded functions that are discontinuous. All the definitions one finds in standard calculus texts are equivalent to his. Using exactly the language we have given for one of the results of Cauchy from the preceding section, we can give a definition of Riemann's integral. Note that it merely turns a theorem (for continuous functions) into a definition of the meaning of the integral for discontinuous functions. This shift represents a quite modern point of view, one that Cauchy and his contemporaries would never have made.

Definition 1.45: Let f be a real-valued function defined on $[a,b]$, and consider a partition P of this interval

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

supplied with associated points $\xi_i \in [x_{i-1}, x_i]$. Form the sum

$$S(f,P) = \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1})$$

and let

$$\|P\| = \max_{1\le i\le n} (x_i - x_{i-1}).$$

Then we define

$$\int_a^b f(x)\,dx = \lim_{\|P\|\to 0} S(f,P)$$

and call f *Riemann integrable* if this limit exists.

1.17.1 Necessary and sufficient conditions for Riemann integrability

The structure of Riemann integrable functions is quite easy to grasp. They are bounded (this is evident from the definition) and they are "mostly" continuous. This was established by Riemann himself. His analysis of the continuity properties of integrable functions lacked only an appropriate language in which to express it. With Lebesgue measure at our disposal, the characterization is immediate and compelling. It reveals too just why the Riemann integral must be considered so limited in application.

Theorem 1.46 (Riemann-Lebesgue) *A necessary and sufficient condition for a function f to be Riemann integrable on an interval $[a,b]$ is that f is bounded and that its set of points of discontinuity in $[a,b]$ forms a set of Lebesgue measure zero.*

Perhaps we should give a version of this theorem that would be more accessible to the mathematicians of the nineteenth century, who would have known Peano–Jordan content but not Lebesgue measure. The set of points of discontinuity has an easy structure: it is the countable union $\bigcup_{n=1}^{\infty} F_n$ of the sequence of closed sets

$$F_n = \{x : \omega_f(x) \geq 1/n\},$$

where the oscillation of the function is greater than the positive value $1/n$. [Exercise 1:1.8 defines $\omega_f(x)$.] That the set of points of continuity of f has measure zero is seen to be equivalent to each of the sets F_n having content zero. Thus the theorem could have been expressed in this, rather more clumsy, way. Note that, so expressed, one may miss the obvious fact that it is only the nature of the set of discontinuity points itself that plays a role, not some other geometric property of the function. In particular, this serves as a good illustration of the merits of the Lebesgue measure over the Peano–Jordan content.

Exercises

1:17.1 Show that a Riemann integrable function must be bounded.

1:17.2$\diamondsuit$ (Riemann) Let f be a real-valued function defined on $[a, b]$, and consider a partition P of this interval:

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

Form the sum

$$O(f, P) = \sum_{i=1}^{n} \omega(f, [x_{i-1}, x_i])(x_i - x_{i-1}),$$

where

$$\omega(f, I) = \sup\{|f(x) - f(y)| : x, y \in I\}$$

is called the oscillation of f on the interval I. Show that in order for f to be Riemann integrable on $[a, b]$ it is necessary and sufficient that

$$\lim_{\|P\| \to 0} O(f, P) = 0.$$

1:17.3 Relate Exercise 1:17.2 to the problem of finding the Peano–Jordan content (Lebesgue measure) of the closed set of points where the oscillation $\omega_f(x)$ of f is greater or equal to some positive number c.

1:17.4 Relate Exercise 1:17.2 to the problem of finding the Lebesgue measure of the set of points where f is continuous (i.e., where the oscillation ω_f of f is zero).

1:17.5 Riemann's integral does not handle unbounded functions. Define a Cauchy–Riemann integral using Cauchy's extension method to handle unbounded functions.

1:17.6 Let S_f denote the set of points of unboundedness of a function f in an interval $[a, b]$. Suppose that S_f has content zero (i.e., measure zero since it is closed) and that f is Riemann integrable in every interval $[c,d] \subset [a,b]$ disjoint from S_f. Define $f_{st}(x) = f(x)$ if $-s \leq f(x) \leq t$, $f_{st}(x) = t$ if $f(x) > t$ and $f_{st}(x) = -s$ if $-s > f(x)$. Define

$$\int_a^b f(x)\,dx = \lim_{s,t\to+\infty} \int_a^b f_{st}(x)\,dx$$

if this exists. Show that $\int_a^b f(x)\,dx$ does exist under these assumptions. This is the way de la Vallée Poussin proposed to handle unbounded functions. Show that this method is different from the Cauchy–Riemann integral by showing that this integral is an absolutely convergent integral.

1:17.7 Prove that a function f on an interval $[a,b]$ is Riemann integrable if f has a finite limit at every point.

1:17.8 Prove that a bounded function on an interval $[a,b]$ is Riemann integrable if and only if f has a finite right-hand limit at every point except only a set of measure zero. [Hint: The set of points at which f is discontinuous and yet has a finite right-hand limit is countable.]

1.18 Volterra's Example

By the end of the nineteenth century, many limitations to Riemann's approach were apparent. All these flaws related to the fact that the class of Riemann integrable functions is too small for many purposes.

The most obvious problem is that a Riemann integrable function must be bounded. Much attention was given to the problem of integrating unbounded functions by the analysts of the that era and less to the fact that, even for bounded functions, the integrability criteria were too strict. This fact was put into startling clarity by an example of Volterra. He produces an everywhere differentiable function F such that F' is bounded but not Riemann integrable. Thus the fundamental theorem of calculus fails for this function, and the formula

$$\int_a^b F'(x)\,dx = F(b) - F(a)$$

is invalid.

Here are some of the details of a construction due to Casper Goffman (1913–2006). For a version closer to Volterra's actual construction, see Exercise 5:5.5. Note that we have only to construct a derivative F' that is discontinuous on a set of positive measure (or a closed set of positive content). For this we take a Cantor set of positive measure (Theorem 1.23). It was the existence of such sets that provided the key to Volterra's construction.

Let $C \subset [0,1]$ be a Cantor set of measure $1/2$ and let $\{I_n\}$ denote the sequence of open intervals complementary to C in $(0,1)$. Then $\sum_{i=1}^{\infty} |I_i| = 1/2$. Choose a closed subinterval $J_n \subset I_n$ centered in I_n such that $|J_n| = |I_n|^2$. Define a function f on $[0,1]$ with values $0 \leq f(x) \leq 1$ such that f is continuous on each interval J_n and is 1 at the centers of each interval J_n and vanishes outside of every J_n. It is straightforward to check that f cannot be Riemann integrable on $[0,1]$. Indeed, since the intervals $\{I_n\}$ are dense and have total length $1/2$, and the oscillation of f is 1 on each I_n, this function violates Riemann's criterion (Exercise 1:17.2).

That f is a derivative follows immediately from advanced considerations (it is bounded and everywhere approximately continuous and hence the derivative of its Lebesgue integral). This can also be seen without any technical apparatus. We can construct a continuous primitive function F for f on each interval J_n. To define a primitive F on all of $[0,1]$, we write

$$F(x) = \sum_{n=1}^{\infty} \int_{J_n \cap [0,x]} f(t)\,dt.$$

Let $I \subset [0,1]$ be an interval that meets the Cantor set C, and let n be any integer so that $I \cap$

$J_n \neq \emptyset$. Let $\ell_n = |I_n|$. Since $\ell_n \leq \frac{1}{2}$, it follows that

$$|I \cap I_n| \geq \tfrac{1}{2}(\ell_n - \ell_n^2) \geq \tfrac{1}{4}\ell_n.$$

Then

$$|I \cap J_n| \leq |J_n| = \ell_n^2 \leq 16|I \cap I_n|^2.$$

If N is the set of integers n for which $I \cap J_n \neq \emptyset$, then

$$\sum_{n\in N} |I \cap J_n| \leq \sum_{n \in N} 16|I \cap I_n|^2 \leq 16|I|^2.$$

From this we can check that $F'(x) = f(x) = 0$ for each $x \in C$. For $x \in [0,1] \setminus C$, it is obvious that $F'(x) = f(x)$. Thus f is a derivative and bounded (between 0 and 1).

Other flaws that reveal the narrowness of the Riemann integral emerge by comparison with later theories. One would like useful theorems that assert a series of functions can be integrated term by term. More precisely, if $\{f_n\}$ is a sequence of integrable functions on $[a, b]$, and $f(x) = \sum_{n=1}^{\infty} f_n(x)$, then f is integrable, and

$$\int_a^b f\,(x)\;dx = \sum_{n=1}^{\infty} \int_a^b f_n\,(x)\;dx.$$

Riemann's integral does not do very well in this connection since the limit function f can be badly discontinuous even if the functions f_n are themselves each continuous. Many authors in the first half of the nineteenth century routinely assumed the permissibility of term-by-term integration. It was not until 1841 that the notion of uniform convergence appeared, and its role in theorems about term-by-term integration, continuity of the sum, and the like, followed soon thereafter. By the end of the century there was felt a strong need to go beyond uniform convergence in theorems of this kind.

Yet another type of limitation is that Riemann's integral is defined only over intervals. For many purposes, one needs to be able to deal with the integral over a set E that need not be an interval. The Riemann integral can, in fact, be defined over Peano–Jordan measurable sets, but we have seen that this class of sets is rather limited and does not embrace many sets (Cantor sets of positive measure for example) that arise in applications. One often needs a larger class of sets over which an integral makes sense.

We shall deal in this text with a notion of integral, essentially due to Henri Lebesgue, that does much better. The class of integrable functions is sufficiently large to remove, or at least reduce, the limitations we discussed, and it allows natural generalizations to functions defined on spaces much more general than the real line.

Exercises

1:18.1 Check the details of the construction of the function F whose derivative is bounded and not Riemann integrable.

1:18.2 Construct a sequence of continuous functions converging pointwise to a function that is not Riemann integrable.

1:18.3 Define

$$\int_E f(x)\,dx = \int_a^b \chi_E(x) f(x)\,dx$$

when $E \subset [a,b]$ and f is continuous on $[a,b]$. For what sets E is this generally possible?

1.19 Riemann–Stieltjes Integral

T. J. Stieltjes (1856–1894) introduced a generalization of the Riemann integral that would seem entirely natural. He introduced a weight function g into the definition and considered limits of sums of the form

$$\sum_{i=1}^{n} f(\xi_i)\,(g(x_i) - g(x_{i-1}))$$

where, as usual, $x_0, x_1, \ldots, x_n$ is a partition of an interval and each $\xi_i \in [x_{i-1}, x_i]$. Although it was introduced for the specific purpose of representing functions in a problem in continued fractions, it should have been clear that this object (the Riemann–Stieltjes integral) had some independent merit. Stieltjes himself died before the appearance of his paper, and the idea attracted almost no attention for the next 15 years. Then F. Riesz showed that this integral gave a precise characterization of the general continuous linear functions on the space of continuous function on an interval. (See Section 12.8.) Since then it has become a mainstream tool of analysis. It also played a fundamental role in the development [notably by J. Radon (1887-1956) and M. Fréchet (1878-1973)] of the abstract theory of measure and integration. For these reasons the student should know at least the rudiments of the theory as presented here.

Definition 1.47: Let f, g be real-valued functions defined on $[a, b]$, and consider a partition P of this interval

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b,$$

supplied with associated points $\xi_i \in [x_{i-1}, x_i]$. Form the sum

$$S(f, dg, P) = \sum_{i=1}^{n} f(\xi_i)\,(g(x_i) - g(x_{i-1}))$$

and let

$$\|P\| = \max_{1 \le i \le n} (x_i - x_{i-1}).$$

Then we define

$$\int_a^b f(x)\,dg(x) = \lim_{\|P\| \to 0} S(f, dg, P)$$

and call f *Riemann–Stieltjes integrable with respect to* g if this limit exists.

Clearly, the case $g(x) = x$ is just the Riemann integral. For g continuously differentiable, the integral reduces to a Riemann integral of the form

$$\int_a^b f(x)\,dg(x) = \int_a^b f(x)g'(x)\,dx.$$

If g is of a very simple form, then the integral can be computed by hand. Suppose that g is a step function; that is, for some partition P of this interval,

$$a = c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k = b,$$

the function g is constant on each interval (c_{i-1}, c_i). Let j_i be the jumps of g at c_i; that is $j_0 = g(c_0+) - g(c_0)$, $j_k = g(c_k) - g(c_k-)$, and $j_i = g(c_i+) - g(c_i-)$ for $1 \leq i \leq k-1$. Then one easily checks for a continuous function f that

$$\int_a^b f(x)\,dg(x) = \sum_{i=1}^{k} f(c_i) j_i.$$

The most natural applications of this integral occur for f continuous and g of bounded variation. In this case the integral exists and there is a useful estimate for its magnitude. We state this as a theorem; it is assigned as an exercise in Section 12.8 where it is needed. We leave the rest of the theoretical development of the integral to the exercises.

Theorem 1.48: *If f is continuous and g has bounded variation on an interval $[a, b]$, then f is Riemann–Stieltjes integrable with respect to g and*

$$\left|\int_a^b f(x)\,dg(x)\right| \leq \left(\max_{x\in[a,b]} |f(x)|\right) V(g; [a, b]).$$

The exercises can be used to sense the structure of the theory that emerges without working through the details. We do not require this theory in the sequel; but, as there are many applications of the Riemann–Stieltjes integral in analysis, the reader should emerge with some familiarity with the ideas, if not a full technical appreciation of how the proofs go. The study of $\int_a^b f(x)\,dg(x)$ is easiest if f is continuous and g monotonic (or of bounded variation). The details are harder if one wants more generality.

Exercises

1:19.1 What is $\int_a^b f(x)\,dg(x)$ if f is constant? If g is constant?

1:19.2 Writing

$$I(f,g) = \int_a^b f(x)\,dg(x)$$

establish the linearity of $f \to I(f,g)$ and $g \to I(f,g)$; that is, show that $I(f_1+f_2,g) = I(f_1,g) + I(f_2,g)$, $I(cf,g) = I(f,cg) = cI(f,g)$, and $I(f,g_1+g_2) = I(f,g_1) + I(f,g_2)$.

1:19.3 Give an example to show that both $\int_a^b f(x)\,dg(x)$ and $\int_b^c f(x)\,dg(x)$ may exist and yet $\int_a^c f(x)\,dg(x)$ may not.

1:19.4 Show that

$$\int_a^c f(x)\,dg(x) = \int_a^b f(x)\,dg(x) + \int_b^c f(x)\,dg(x)$$

under appropriate assumptions.

1:19.5 Suppose that g is continuously differentiable and f is continuous. Prove that

$$\int_a^b f(x)\,dg(x) = \int_a^b f(x)g'(x)\,dx.$$

[Hint: Write $f(\xi_i)(g(x_i) - g(x_{i-1}))$ as $f(\xi_i)g'(\eta_i)(x_i - x_{i-1})$, where ξ_i, $\eta_i \in [x_{i-1}, x_i]$ using the mean-value theorem.]

1:19.6 Let g be a step function, constant on each interval (c_{i-1}, c_i) of the partition

$$a = c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k = b.$$

Then, for a continuous function f,

$$\int_a^b f(x)\,dg(x) = \sum_{i=1}^{k} f(c_i)j_i,$$

where j_i are the jumps of g at c_i; that is, $j_0 = g(c_0+) - g(c_0)$, $j_k = g(c_k) - g(c_k-)$, and $j_i = g(c_i+) - g(c_i-)$ for $1 \le i \le k-1$.

1:19.7 Show that if $\int_a^b f(x)\,dg(x)$ exists then f and g have no common point of discontinuity.

1:19.8 (Integration by parts) Establish the formula

$$\int_a^b f(x)\,dg(x) + \int_a^b g(x)\,df(x) = f(b)g(b) - f(a)g(a)$$

under appropriate assumptions on f and g.

1:19.9 (Mean-value theorem) Show that

$$\int_a^b f(x)\,dg(x) = f(\xi)(g(b) - g(a))$$

for some $\xi \in [a,b]$ under appropriate assumptions on f and g.

1:19.10 Suppose that f_1, f_2 are continuous and g is of bounded variation on $[a,b]$, and define

$$h(x) = \int_a^x f_1(t)\,dg(t)$$

for $a \le x \le b$. Show that

$$\int_a^b f_2(t)\,dh(t) = \int_a^b f_1(t)f_2(t)\,dg(t).$$

1:19.11 Let g, g_1, g_2, ... be BV functions on $[a,b]$ such that $g(a) = g_1(a) = \cdots = 0$. Suppose that the variation of $g - g_n$ on $[a,b]$ tends to zero as $n \to \infty$. Show that

$$\lim_{n\to\infty} \int_a^b f(x)\, dg_n(x) = \int_a^b f(x)\, dg(x)$$

for every continuous f. [Hint: Use Theorem 1.48.]

1.20 Lebesgue's Integral

The mainstream of modern integration theory is based on the notion of integral due to Lebesgue. A formal development of the integral must wait until Chapter 5, where it is done in full generality. Here we give some insight into what is involved.

Suppose that you have several coins in your pocket to count: 4 dimes, 2 nickels, and 3 pennies. There are two natural ways to count the total value of the coins.

Computation 1. Count the coins in the order in which they appear as you pull them from your pocket, for example,

$$10 + 10 + 5 + 10 + 1 + 5 + 10 + 1 + 1 = 53.$$

Computation 2. Group the coins by value, and compute

$$(10)(4) + (5)(2) + (1)(3) = 53.$$

The first computation corresponds to Riemann integration, while the second computation is closely related to the methods of Lebesgue integration. Let's look at this in more detail. Figure 1.1 is the graph of a function that models our counting problem using the order from computation 1.

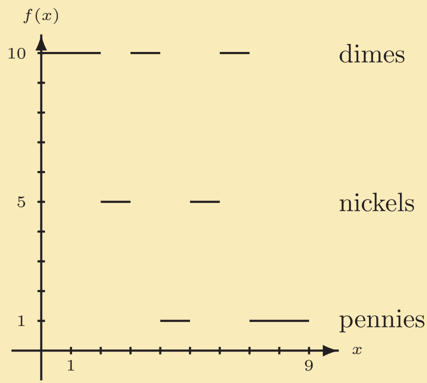


Figure 1.1. A function that models our counting problem.

One can check easily that $\int_0^9 f(x)\, dx = 53$, the integral being Riemann's. Because of the simple nature of this function, one sees that one needs no finer partition than the partition obtained by dividing $[0, 9]$ into 9 congruent intervals. This partition gives the sum corresponding to the first method.

To consider the second method of counting, we use the notation of measure theory. If I is an interval, we write, as usual, $\lambda(I)$ for the length of I. If E is a finite union of pairwise-disjoint intervals, $E = I_1 \cup \cdots \cup I_n$, then the measure of E is given by the sum

$$\lambda(E) = \lambda(I_1) + \cdots + \lambda(I_n).$$

Now let

$$E_1 = \{x : f(x) = 1\},$$

$$E_5 = \{x : f(x) = 5\},$$

and

$$E_{10} = \{x : f(x) = 10\}.$$

Then $\lambda(E_1) = 3$, $\lambda(E_5) = 2$, and $\lambda(E_{10}) = 4$. In computation 2 we formed the sum

$$(1)\lambda(E_1) + (5)\lambda(E_5) + (10)\lambda(E_{10}).$$

Note that the numbers 1, 5, and 10 represent the values of the function f, and $\lambda(E_i)$ indicates "how often" the value i is taken on.

We have belabored this simple example because it contains the seed of the Lebesgue integral. Let us try to imitate this example for an arbitrary bounded function f defined on $[a, b]$. Suppose that $m \leq f(x) < M$ for all $x \in [a, b]$. Instead of partitioning the interval $[a, b]$, we partition the interval $[m, M]$:

$$m = y_0 < y_1 < \cdots < y_n = M.$$

For $k = 1, \ldots, n$, let

$$E_k = \{x : y_{k-1} \leq f(x) < y_k\}.$$

Thus the partition of the range induces a partition of the interval $[a, b]$:

$$[a, b] = E_1 \cup E_2 \cup \cdots \cup E_n$$

where the sets $\{E_k\}$ are clearly pairwise disjoint. We can form the sums

$$\sum y_k\,\lambda(E_k) \quad \text{and} \quad \sum y_{k-1}\,\lambda(E_k)$$

in the expectation that these can be used to approximate our integral, the first from above and the second from below. We hope two things: that such approximating sums approach a limit as the norm of the partition approaches zero and that the two limits are the same. If each of the sets E_k happens to be always a finite union of intervals (e.g., if f is a polynomial), then the upper and lower sums do have the same limit. This is just another way of describing a well-known development of the Riemann integral via upper and lower sums.

But the sets E_k may be much more complicated than this. For example, each E_k might contain no interval. Thus one needs to know in advance the measure of quite arbitrary sets. This attempt at an integral will break down unless we restrict things in such a way that the sets that arise are Lebesgue measurable. This means we must restrict our attention to classes of functions for which all such sets are measurable, the measurable functions (Chapter 4).

After we understand the basic ideas of measures (Chapter 2) and measurable functions (Chapter 4), we will be ready to develop the integral. The idea of considering sums of the form

$$\sum y_k\,\lambda(E_k) \text{ and } \sum y_{k-1}\,\lambda(E_k)$$

taken over a partition of the interval

$$[a,b] = E_1 \cup E_2 \cup \cdots \cup E_n$$

did not originate with Lebesgue; Peano had used it earlier. But the idea of partitioning the range in order to induce this partition seems to be Lebesgue's contribution, and it points out very clearly the class of functions that should be considered; that is, functions f for which the associated sets

$$E = \{x : \alpha \leq f(x) < \beta\}$$

are Lebesgue measurable.

The preceding paragraphs represent an outline of how one could arrive at the Lebesgue integral. Our development will be more general; it will include a theory of integration that applies to functions defined on general "measure spaces." The fascinating evolution of the theory of integration is delineated in Hawkins book on this subject.[2] A reading of this book allows one to admire the genius of some leading mathematicians of the time. It also allows one to sympathize with their misconceptions and the frustration these misconceptions must have caused.

1.21 The Generalized Riemann Integral

The main motivation that Lebesgue gave for generalizing the Riemann integral was Volterra's example of a bounded derivative that is not Riemann integrable. Lebesgue was able to prove that his integral would handle all bounded derivatives. His integral is, however, by its very nature an absolute integral. That is, in order for $\int_a^b f(x)\,dx$ to exist, it must be true that

$$\int_a^b |f(x)|\,dx$$

also exists. The problem of inverting derivatives cannot be solved by an absolute integral, as we know from the elementary example F' with

$$F(x) = x^2 \sin x^{-2}.$$

[2] T. Hawkins, *Lebesgue's Theory of Integration*, Chelsea Publishing Co., (1979).

Thus we are still left with a curious situation. Despite a century of the best work on the subject, the integration theories of Cauchy, Riemann, and Lebesgue do not include the original Newton integral. There are derivatives (necessarily unbounded) that are not integrable in any of these three senses. In general, how can one invert a derivative then?

To answer this, we can take a completely naive approach and start with the definition of the derivative itself. If $F' = f$ everywhere, then, at each point ξ and for every $\varepsilon > 0$, there is a $\delta > 0$ so that

$$|F(x'') - F(x') - f(\xi)(x'' - x')| < \varepsilon(x'' - x') \tag{5}$$

for $x' \leq \xi \leq x''$ and $0 < x'' - x' < \delta$.

We shall attempt to recover $F(b) - F(a)$ as a limit of Riemann sums for f, even though this is a misguided attempt, since we know that the Riemann integral must fail in general to accomplish this. Even so, let us see where the attempt takes us.

Let

$$a = x_0 < x_1 < x_2 \ldots x_n = b$$

be a partition of $[a, b]$, and let $\xi_i \in [x_{i-1}, x_i]$. Then

$$F(b) - F(a) = \sum_{i=1}^{n} (F(x_{i-1}) - F(x_i)) = \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) + R$$

where

$$R = \sum_{i=1}^{n} (F(x_i) - F(x_{i-1}) - f(\xi_i)(x_i - x_{i-1}))\,.$$

Thus $F(b) - F(a)$ has been given as a Riemann sum for f plus some error term R. But it appears now that, if the partition is finer than the number δ so that (5) may be used, we have

$$\begin{aligned} |R| &\leq \sum_{i=1}^{n} \Big| F(x_i) - F(x_{i-1}) - f(\xi_i)(x_i - x_{i-1}) \Big| \\ &< \sum_{i=1}^{n} \varepsilon(x_i - x_{i-1}) = \varepsilon(b-a). \end{aligned}$$

Evidently, then, *if there are no mistakes here* we have just proved that f is Riemann integrable and that

$$\int_a^b f(t)\,dt = F(b) - F(a).$$

This is false of course. Even the Lebesgue integral does not invert all derivatives, and the Riemann integral cannot invert even all bounded derivatives. The error is that the choice of δ depends on the point ξ considered and so is not a constant. But, instead of abandoning the argument, one can change the definition of the Riemann integral to allow a variable δ. The definition then changes to look like this.

Definition 1.49: A function f is *generalized Riemann integrable* on $[a,b]$ with value I if for every ε there is a positive function δ on $[a,b]$ so that

$$\left| \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) - I \right| < \varepsilon$$

whenever $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ is a partition of $[a,b]$ with $\xi_i \in [x_{i-1}, x_i]$ and $0 < x_i - x_{i-1} < \delta(\xi_i)$.

To justify the definition requires knowing that such partitions actually exist for any such gauge δ; this is supplied by the Cousin theorem (Theorem 1.3).

This defines a Riemann-type integral that includes the Lebesgue integral and the Newton integral. It is equivalent to the integrals invented by A. Denjoy (1884–1974) and O. Perron in 1912. The generalized Riemann integral was discovered in the 1950s, independently, by Ralph Henstock (1923–2007) and Jaroslav Kurzweil, and these ideas have led to a number of other integration theories that exploit the geometry of the underlying space in the same way that this integral exploits the geometry of derivatives on the real line.

In Section 5.10 we shall present a property of the Lebesgue integral that shows how it is included in a generalized Riemann integral. We do not develop this theme any further as these ideas should be considered, for the moment anyway, as rather specialized. A development of these ideas can be found in the recent monographs of Pfeffer[3] or Gordon.[4] The main tool of modern analysis is the standard theory of measure and integration developed in subsequent chapters, and we confine our interests in integration theory to its exposition.

Exercises

1:21.1 Develop the elementary properties of the generalized Riemann integral directly from its definition (e.g., the integral of a sum $f+g$, the integral formula $\int_a^b + \int_b^c = \int_a^c$, etc.).

1:21.2 Show directly from the definition that the function f defined as $f(x) = 0$ for x rational and $f(x) = 1$ for x irrational is not Riemann integrable, but is generalized Riemann integrable on any

[3]W. F. Pfeffer, *The Riemann Approach to Integration: Local Geometric Theory.* Cambridge (1993).

[4]R. A. Gordon, *The Integrals of Lebesgue, Denjoy, Perron and Henstock.* Grad. Studies in Math, Vol. 4, Amer. Math. Soc. (1994).

interval, and that $\int_0^1 f(x)\,dx = 1$.

1:21.3 Show that the generalized Riemann integral is closed under the extension procedure of Cauchy from Section 1.16.

1.22 Additional Problems for Chapter 1

1:22.1 For an arbitrary function $F:\mathbb{R}\to\mathbb{R}$, prove that the set

$$\{x : F \text{ assumes a strict local maximum or minimum at } x\}$$

is countable. [Hint: Consider

$$A_n = \left\{x : F(t) < F(x)\ \forall\ t \neq x \text{ in } \left(x-\frac{1}{n}, x+\frac{1}{n}\right)\right\}.$$

1:22.2 For an arbitrary function $F:\mathbb{R}\to\mathbb{R}$, prove that the set

$$\left\{x : \limsup_{t\to x} F(t) > \limsup_{t\to x+} F(t)\right\}$$

is countable.

1:22.3 For an arbitrary function $F:\mathbb{R}\to\mathbb{R}$, prove that the set

$$\left\{x : F(x) \notin \left[\liminf_{t\to x} F(t), \limsup_{t\to x} F(t)\right]\right\}$$

is countable.

1:22.4 For an arbitrary function $F:\mathbb{R}\to\mathbb{R}$, prove that the set

$$\left\{x : F \text{ is discontinuous at } x \text{ and } \lim_{t\to x} F(t) \text{ exists}\right\}$$

is countable.

1:22.5 Show that the set of irrationals in $[0,1]$ has inner measure 1 and the set of rationals in $[0,1]$ has outer measure 0.

1:22.6 Prove (or find somewhere a proof) that the following three logical principles are equivalent:

(a) The axiom of choice,

(b) The well-ordering principle [Zermelo's theorem].

(c) Zorn's lemma.

1:22.7$\diamondsuit$ An uncountable set S of real numbers is said to be *totally imperfect* if it contains no nonempty perfect set. A set S of real numbers is said to be a *Bernstein set* if neither S nor $\mathbb{R} \setminus S$ contains a nonempty perfect set. Prove the existence of such sets assuming the continuum hypothesis and using Statement 1.15. (Incidentally, no Borel set can be totally imperfect.) [Hint: Let $\mathcal{C}$ be the collection of all perfect sets. This has cardinality c (see Exercise 1:4.7). Under CH we can well order $\mathcal{C}$ as in Statement 1.15, say indexing as $\{P_\alpha\}$, so that each element has only countably many predecessors. Construct S by picking two distinct points x_α, y_α from each P_α in such a way that at each stage we pick new points. (You will have to justify this by a cardinality argument.) Put the x_α in S.]

1:22.8$\diamondsuit$ Show the existence of Bernstein sets without assuming CH.

[Hint: Use Lemma 1.16, and basically the same proof as Exercise 1:22.7, but with a little more attention to the cardinality arguments.]

1:22.9$\diamondsuit$ Assuming CH, show that there is an uncountable set U of real numbers (called a *Lusin set*) such that every dense open set contains all but countably many points from U. [Hint: Let $\{G_\alpha\}$ be a well ordering of the open dense sets so that every element has only countably many predecessors. Choose distinct points x_α from $\bigcap_{\beta \leq \alpha} G_\beta$. Then U consists of all the points x_α. (The steps have to be justified. Remember that a countable intersection of dense open sets is residual and therefore uncountable.)]

1:22.10 Recall (Exercise 1:7.5) that the outer content c^* is *finitely subadditive*; that is, if $\{E_k\}$ is a sequence of subsets of an interval $[a,b]$, then

$$c^*\left(\bigcup_{k=1}^{n} E_k\right) \leq \sum_{k=1}^{n} c^*(E_k).$$

Show that c_* is *finitely superadditive*; that is, if $\{E_k\}$ is a *disjoint* sequence of subsets of $\mathbb{R}$,

$$c_*\left(\bigcup_{k=1}^{n} E_k\right) \geq \sum_{k=1}^{n} c_*(E_k).$$

1:22.11 Recall (Exercise 1:7.6) that the outer measure λ^* is *countably subadditive*; that is, if $\{E_k\}$ is a sequence of subsets of $\mathbb{R}$, then

$$\lambda^*\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} \lambda^*\left(E_k\right).$$

Similarly, show that λ_* is *countably superadditive*; that is, if $\{E_k\}$ is a *disjoint* sequence of subsets of an interval $[a,b]$, then

$$\lambda_*\left(\bigcup_{k=1}^{\infty} E_k\right) \geq \sum_{k=1}^{\infty} \lambda_*(E_k).$$

[Hint: Use Exercise 1:9.16.]

1:22.12 Let $\{c_k\}$ be complex numbers with $\sum_{k=1}^{\infty} |c_k| < +\infty$ and write $f(z) = \sum_{k=1}^{\infty} c_k z^k$ for $|z| \leq 1$. Show that f is BV on each radius of the circle $|z| = 1$.

1:22.13$\diamondsuit$ Let C and B be the sets referenced in the proof of Theorem 1.23. Define a function f in the following way. On I_1, let $f = 1/2$; on I_2, $f = 1/4$; on I_3, $f = 3/4$. Proceed inductively. On the $2^{n-1} - 1$ open intervals appearing at the nth stage, define f to satisfy the following conditions:

(*i*) f is constant on each of these intervals.

(*ii*) f takes the values

$$\frac{1}{2^n}, \frac{3}{2^n}, \ldots, \frac{2^n-1}{2^n}$$

on these intervals.

(*iii*) If x and y are members of different nth-stage intervals with $x < y$, then $f(x) < f(y)$.

This description defines f on B. Extend f to all of $[0,1]$ by defining $f(0) = 0$ and, for $x \neq 0$, $f(x) = \sup\{f(t) : t \in B, t < x\}$.

(a) Show that $f(B)$ is dense in I_0.

(b) Show that f is nondecreasing on I_0.

(c) Infer from (a) and (b) that f is continuous on I_0.

(d) Show that $f(C) = I_0$, and thus C has the same cardinality as I_0.

As an example, Figure 1.2 corresponds to the case in which, every time an interval I_k is selected, it is the middle third of the closed component of A_n from which it is chosen. In this case, the set C is called the *Cantor set* (or Cantor ternary set) and f is called the *Cantor function*. The set and function are named for the German mathematician Georg Cantor (1845–1918). Observe that f "does all its rising" on the set C, which here has measure zero. More precisely, $\lambda(f(B)) = 0$, $\lambda(f(C)) = 1$. This example will be important in several places in Chapters 4 and 5.

1:22.14 Using some of the ideas in the construction of the Cantor function (Exercise 1:22.13), obtain a continuous function that is not of bounded variation on any subinterval of $[0,1]$.

1:22.15 Using some of the ideas in the construction of the Cantor function (Exercise 1:22.13), obtain a continuous function that is of bounded variation on $[0,1]$, but is not monotone on any subinterval of $[0,1]$.

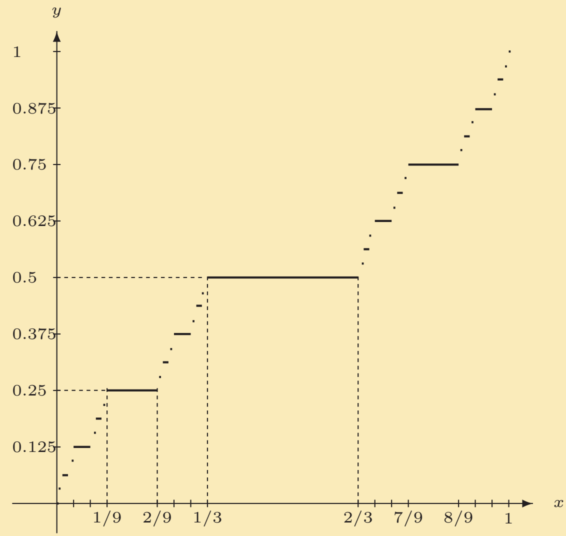


Figure 1.2. The Cantor function.

1:22.16 Show that the Cantor function is not absolutely continuous (Exercise 1:14.17).

Chapter 2

MEASURE SPACES

With the help of the Riemann version of the integral, calculus students can study such notions as the length of a curve, the area of a region in the plane, the volume of a region in space, and mass distributions on the line, in the plane, or in space. But there are serious limitations and many awkward difficulties associated with Riemann's methods. These length, area, and volume notions, as well as many others, are better studied within the framework of measure theory.

In this framework, one has a set X, a class $\mathcal{M}$ of subsets of X, and a measure μ defined on $\mathcal{M}$. The class $\mathcal{M}$ satisfies certain natural conditions (See Sections 2.2 and 2.3), and μ satisfies conditions one would expect of such notions as length, area, volume, or mass.

Our objective in this chapter is to provide the reader with a working knowledge of basic measure theory. In Section 2.1, we provide an outline of Lebesgue measure on the line via the notions of *inner measure* and *outer measure*. Then, in Sections 2.2 and 2.3, we begin our development of abstract measure theory by extracting features of Lebesgue measure that one would want for any notion of measure.

This abstract approach has the advantage of being quite general and therefore of being applicable to a variety of phenomena. But it does not tell us *how* to obtain a measure with which to model a given phenomenon. Here we take our cue from the development in Section 2.1. We find that a measure can always be obtained from an *outer measure* (Section 2.8).

We also find that when we have a primitive notion of our phenomenon, for example, length of an interval, area of a square, volume of a cube, or mass in a square or cube, this primitive notion determines an outer measure in a natural way. The outer measure, in turn, defines a measure that extends this primitive notion to a large class of sets $\mathcal{M}$ that is suitable for a coherent theory.

Many measures possess special properties that make them particularly useful. Lebesgue measure has most of these. For example, the Lebesgue outer measure of any set E can be obtained as the Lebesgue measure of a larger set $H \supset E$ that is measurable. Every subset of a set of Lebesgue measure zero is measurable and has, again, Lebesgue measure zero. In Sections 2.10 to 2.13 we develop such properties abstractly. Finally, Section 2.11 addresses the problem of nonmeasurable sets in a very general setting.

2.1 One-Dimensional Lebesgue Measure

We begin our study of measures with a heuristic development of Lebesgue measure in $\mathbb{R}$ that will provide a concrete example that we can recall when we develop the abstract theory. This is independent of the sketch given in the first chapter. Our development will be heuristic for two reasons. First, a development including all details would obscure the major steps we wish to highlight. Some of these details are covered by the exercises. Second, our development of the abstract theory in the remainder of the chapter, which does not depend on Lebesgue measure in any way, will verify the correctness of our claims. Thus Lebesgue measure serves as our motivating example to guide the development of the theory and our illustrative example to show the theory in application.

We begin with the primitive notion of the length of an interval. We then extend this notion in a natural way first to open sets, then to closed sets. Finally, by the method of inner and outer measures, this is extended to a large class of "measurable" sets.

1. The measure of open intervals. We define

$$\lambda(I) = b - a,$$

where I denotes the open interval (a, b). This is the beginning of a process that can, with some adjustments, be applied to a variety of situations.

2. The measure of open sets. Define

$$\lambda(G) = \sum \lambda(I_k),$$

where G is an open set and $\{I_k\}$ is the sequence of component intervals of G. If one of the components is unbounded, we let $\lambda(G) = \infty$. [If $G \neq \emptyset$, then G can be expressed as a finite or countably infinite disjoint union of open intervals: $G = \bigcup I_k$. If $G = \emptyset$, the empty set, define $\lambda(G) = 0$.] This definition is a natural one; it conforms to our intuitive requirement that "the whole is equal to the sum of the parts."

3. The measure of bounded closed sets. Define

$$\lambda(E) = b - a - \lambda((a, b) \setminus E),$$

where E is a bounded closed set and $[a,b]$ is the smallest closed interval containing E. Since $[a,b] = E \cup ([a,b] \setminus E)$, our intuition would demand that

$$\lambda(E) + \lambda((a,b) \setminus E) = b - a$$

and this becomes our definition.

So far, we have a notion of measure for arbitrary open sets and for bounded closed sets. We shall presently use these notions to extend the measure to a larger class of sets—the measurable sets. Let us pause first to look at an intuitive example.

Example 2.1: Let $0 \leq \alpha < 1$. There is a nowhere dense closed set $C \subset [0,1]$ that is of measure α. (For the full details of the construction see Section 1.8.) Its complement $B = [0,1] \setminus C$ is a dense open subset of $[0,1]$ of measure $1 - \alpha$. In particular, if $\alpha > 0$, C has positive measure. In any case, C is a nonempty nowhere dense perfect subset of $[0,1]$ and therefore has cardinality of the continuum. (See Exercise 1:22.13.)

While the construction of the set C is relatively simple, the existence of such sets was not known until late in the nineteenth century. Prior to that, mathematicians recognized that a nowhere dense set could have limit points, even limit points of limit points, but could not conceive of a nowhere dense set as possibly having positive measure. Since dense sets were perceived as large and nowhere dense sets as small, this example, with $\alpha > 0$, would have begun the process of clarifying the ideas that would lead to a coherent development of measure theory.

We shall now use our definitions of measure for bounded open sets and bounded closed sets to obtain a large class $\mathcal{L}$ of Lebesgue measurable sets to which the measure λ can be extended. To each set $E \in \mathcal{L}$, we shall assign a nonnegative number $\lambda(E)$ called the Lebesgue measure of

E. Our intuition demands that a certain "monotonicity" condition be satisfied for measurable sets: if E_1 and E_2 are measurable and $E_1 \subset E_2$, then

$$\lambda(E_1) \leq \lambda(E_2).$$

In particular, if G is any open set containing a set E, we would want $\lambda(E) \leq \lambda(G)$, so $\lambda(G)$ provides an upper bound for $\lambda(E)$, if E is to be measurable.

2.1.1 Lebesgue outer measure

We can now define the *outer measure* of an arbitrary set E by choosing the open set G "economically."

Definition 2.2: Let E be an arbitrary subset of $\mathbb{R}$. Let

$$\lambda^*(E) = \inf\{\lambda(G) : \ E \subset G, G \text{ open}\}.$$

Then $\lambda^*(E)$ is called the Lebesgue *outer measure* of E.

We point out, for later reference, that the outer measure can also be obtained by approximating from outside with sequences of open intervals (Exercise 2:1.10):

$$\lambda^*(E) = \inf\left\{\sum_{k=1}^{\infty} \lambda(I_k) : E \subset \bigcup_{k=1}^{\infty} I_k, \text{ each } I_k \text{ an open interval}\right\}.$$

Now $\lambda^*(E)$ may seem like a good candidate for $\lambda(E)$. It meets the monotonicity requirement and it is well defined for all bounded subsets of $\mathbb{R}$. It is also true, but by no means obvious, that $\lambda^*(E) = \lambda(E)$ when E is open or closed. (See Exercise 2:1.4.) But λ^* lacks an essential property: we cannot conclude for a pair of disjoint sets E_1, E_2 that

$$\lambda^*(E_1 \cup E_2) = \lambda^*(E_1) + \lambda^*(E_2).$$

The whole need not equal the sum of its parts.

2.1.2 Lebesgue inner measure

Here is how Lebesgue remedied this flaw. So far we have used only part of what is available to us—*outside* approximation of E by *open* sets. Now we use *inside* approximation by *closed* sets.

Definition 2.3: Let E be an arbitrary subset of $\mathbb{R}$. Let

$$\lambda_*(E) = \sup\{\lambda(F) : F \subset E,\ F \text{ compact}\}.$$

Then $\lambda_*(E)$ is called the Lebesgue *inner measure* of E.

Since E need not contain any intervals, there is no inner approximation by intervals, analogous to the approximation of the outer measure by intervals. We have, however, the following formula for a bounded set E.

2.4: Let $[a, b]$ be the smallest interval containing a bounded set E. Then

$$\lambda_*(E) = b - a - \lambda^*([a, b] \setminus E).$$

This shows the important fact that the inner measure is definable directly in terms of the outer measure. In particular, it suggests already that a theory based on the outer measure alone may be feasible. We illustrate these definitions with an example.

Example 2.5: Let $I_0 = [0, 1]$, and let $\mathbb{Q}$ denote the rational numbers in I_0. Let $\varepsilon > 0$ and let $\{q_k\}$ be an enumeration of $\mathbb{Q}$. For each positive integer n, let I_n be an open interval such that $q_n \in I_n$ and $\lambda(I_n) < \varepsilon/2^n$. Then $\mathbb{Q} \subseteq \bigcup I_n$ and $\sum \lambda(I_n) < \varepsilon$. Thus $\lambda^*(\mathbb{Q}) = 0$. The set $P = I_0 \setminus \bigcup I_k$ is closed, and $P \subset I_0 \setminus \mathbb{Q}$. We see, using the assertion 2.4 and Exercise 2:1.12, that $\lambda(P) > 1 - \varepsilon$. It follows that

$$1 - \varepsilon < \lambda_*(P) \leq \lambda_*(I_0 \setminus \mathbb{Q}),$$

so that $\lambda_*(I_0 \setminus \mathbb{Q}) = 1$. Thus the set of irrationals in I_0 has inner measure 1, and the set of rationals has outer measure 0.

2.1.3 Lebesgue measurable sets

Inner measure λ_* has the same flaw as outer measure λ^*. The key to obtaining a large class of measurable sets lies in the observation that we would like outside approximation to give the same result as inside approximation.

Definition 2.6: Let E be a bounded subset of $\mathbb{R}$, and let $\lambda^*(E)$ and $\lambda_*(E)$ denote the outer and inner measures of E. If

$$\lambda^*(E) = \lambda_*(E),$$

we say that E is *Lebesgue measurable* with *Lebesgue measure* $\lambda(E) = \lambda^*(E)$. If E is unbounded, we say that E is measurable if $E \cap I$ is measurable for every interval I and again write $\lambda(E) = \lambda^*(E)$.

One can verify that the class $\mathcal{L}$ of Lebesgue measurable sets is closed under countable unions and under set difference. If $\{E_k\}$ is a sequence of measurable sets, so is $\bigcup E_k$, and the difference of two measurable sets is measurable. In addition, Lebesgue measure λ is *countably additive on the class* $\mathcal{L}$: if $\{E_k\}$ is a sequence of pairwise disjoint sets from $\mathcal{L}$, then

$$\lambda(\bigcup E_k) = \sum \lambda(E_k).$$

We shall not prove these statements at this time. They will emerge as consequences of the theory developed in Section 2.10. Observe for later reference that λ^* *is* countably additive on $\mathcal{L}$, since $\lambda^* = \lambda$ on $\mathcal{L}$. Thus we can view λ as the restriction of λ^*, which is defined for *all* subsets of $\mathbb{R}$, to $\mathcal{L}$, the class of Lebesgue measurable sets.

Not all subsets of $\mathbb{R}$ can be measurable. In Section 1.10 we have given the details of the proof of this fact. But we shall discover that all sets that arise in practice are measurable.

Many of the ideas that appear in this section, including the exercises, will reappear, in abstract settings as well as in concrete settings, throughout the remainder of this chapter.

Exercises

2:1.1 In the definition of $\lambda(G)$ for G a bounded open set, how do we know that the sum $\sum \lambda(I_k)$ is finite?

2:1.2 Prove that both the outer measure and inner measure are monotone: If $E_1 \subset E_2$, then $\lambda^*(E_1) \leq \lambda^*(E_2)$ and $\lambda_*(E_1) \leq \lambda_*(E_2)$.

2:1.3 Prove that the outer measure λ^* and inner measure λ_* are translation-invariant functions defined on the class of all subsets of $\mathbb{R}$.

2:1.4 Prove that $\lambda^*(E) = \lambda_*(E) = \lambda(E)$ when E is open or closed and bounded. (Thus the definition of measure for open sets and for compact sets in terms of λ^* and λ_* is consistent with the definition given at the beginning of the section.) [Hint: If E is an open set with component intervals

$\{(a_i, b_i)\}$, then show how $\lambda_*(E)$ can be approximated by the measure of a compact set of the form

$$\bigcup_{i=1}^{N} \left[a_i + \varepsilon 2^{-i}, b_i - \varepsilon 2^{-i}\right]$$

for large N and small $\varepsilon > 0$.]

2:1.5 Let $[a, b]$ be the smallest interval containing a bounded set E. Prove that

$$\lambda_*(E) = b - a - \lambda^*([a, b] \setminus E).$$

[Hint: Split the equality into two inequalities and prove each directly from the definition.]

2:1.6 For all $E \subset \mathbb{R}$, show that $\lambda_*(E) \leq \lambda^*(E)$. [Hint: If $F \subset E \subset G$ with F compact and G open, we know already that $\lambda(F) \leq \lambda(G)$. Take first the infimum over G and then the supremum over F.]

2:1.7 Show that if $\lambda^*(E) = 0$ then E and all its subsets are measurable.

2:1.8 Show that there exist 2^c Lebesgue measurable sets (where c is, as usual, the cardinality of the real numbers).

2:1.9 Show that if $\{G_k\}$ is a sequence of open subsets of $\mathbb{R}$ then

$$\lambda\left(\bigcup_{k=1}^{\infty} G_k\right) \leq \sum_{k=1}^{\infty} \lambda(G_k).$$

[Hint: If $(a, b) \subset \bigcup_{k=1}^{\infty} G_k$, show that $b - a \leq \sum_{k=1}^{\infty} \lambda(G_k)$ by considering that

$$[a + \varepsilon, b - \varepsilon] \subset \bigcup_{k=1}^{N} G_k$$

for small ε and sufficiently large N.]

2:1.10 Using Exercise 2:1.9, show that

$$\lambda^*(E) = \inf\left\{\sum_{k=1}^{\infty} \lambda(I_k) : E \subset \bigcup_{k=1}^{\infty} I_k, \text{ each } I_k \text{ an open interval}\right\}.$$

2:1.11 Show that if $\{F_k\}$ is a sequence of compact disjoint subsets of $\mathbb{R}$ then

$$\lambda\left(\bigcup_{k=1}^{n} F_k\right) \geq \sum_{k=1}^{n} \lambda(F_k).$$

[Hint: If F_1 and F_2 are disjoint compact sets, then there are disjoint open sets $G_1 \supset F_1$ and $G_2 \supset F_2$.]

2:1.12 Show that λ^* is *countably subadditive*: if $\{E_k\}$ is a sequence of subsets of $\mathbb{R}$, then

$$\lambda^*\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} \lambda^*(E_k).$$

[Hint: Choose open sets $G_k \supset E_k$ so that $\lambda^*(E_k) + \varepsilon 2^{-k} \geq \lambda(G_k)$ and use Exercise 2:1.9.]

2:1.13 Similarly to Exercise 2:1.12, show that λ_* is *countably superadditive*: if $\{E_k\}$ is a *disjoint* sequence of subsets of $\mathbb{R}$,

$$\lambda_*\left(\bigcup_{k=1}^{\infty} E_k\right) \geq \sum_{k=1}^{\infty} \lambda_*(E_k).$$

[Hint: Choose compact sets $F_k \subset E_k$ so that $\lambda_*(E_k) - \varepsilon 2^{-k} \geq \lambda(F_k)$ and use Exercise 2:1.11.]

2:1.14$\diamondsuit$ We recall that a set is of type $\mathcal{F}_\sigma$ if it can be expressed as a countable union of closed sets, and it is of type $\mathcal{G}_\delta$ if it can be expressed as a countable intersection of open sets. (See the discussion of these ideas in Sections 1.1 and 1.12.)

(a) Prove that every closed set $F \subset \mathbb{R}$ is of type $\mathcal{G}_\delta$ and every open set $G \subset \mathbb{R}$ is of type $\mathcal{F}_\sigma$.

(b) Prove that for every set $E \subset \mathbb{R}$ there exists a set K of type $\mathcal{F}_\sigma$ and a set H of type $\mathcal{G}_\delta$ such that $K \subset E \subset H$ and

$$\lambda(K) = \lambda_*(E) \leq \lambda^*(E) = \lambda(H).$$

The set K is called a *measurable kernel* of E, while the set H is called a *measurable cover* for E.

(c) Prove that if $E \in \mathcal{L}$ there exist K, H as above such that

$$\lambda(K) = \lambda(E) = \lambda(H).$$

[The point of this exercise is to show that one can approximate measurable sets by relatively simple sets on the inside and on the outside. By use of the Baire category theorem (see Section 1.6), one can show that the roles played by sets of type $\mathcal{F}_\sigma$ and of type $\mathcal{G}_\delta$ cannot be exchanged in parts (b) and (c).]

(d) Show that "$\mathcal{F}_\sigma$" cannot be improved to "closed" and "$\mathcal{G}_\delta$" cannot be improved to "open" in parts (b) and (c).

2:1.15 Give an example of a nonmeasurable set E for which $\lambda_*(E) = \lambda^*(E) = \infty$. [Hint: Use Theorem 1.33.]

2.2 Additive Set Functions

We begin now our study of structures suggested by Lebesgue measure. The class of sets that are Lebesgue measurable has certain natural properties: it is closed under the formation of unions, intersections, and set differences. This leads to our first abstract definition.

Definition 2.7: Let X be any set, and let $\mathcal{A}$ be a nonempty family of subsets of X. We say $\mathcal{A}$ is an *algebra of sets* if it satisfies the following conditions:

1. $\emptyset \in \mathcal{A}$.

2. If $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

3. If $A \in \mathcal{A}$, then $X \setminus A \in \mathcal{A}$.

It is easy to verify that an algebra of sets is closed also under differences, finite unions, and finite intersections. (See Exercise 2:2.1.) For any set X, the family 2^X of all subsets of X is obviously an algebra. So is the family $\mathcal{A} = \{\emptyset, X\}$. We have noted that the family $\mathcal{L}$ of Lebesgue measurable sets is an algebra. Here is another example, to which we shall return later.

Example 2.8: Let $X = (0, 1]$. Let $\mathcal{A}$ consist of $\emptyset$ and all finite unions of half-open intervals $(a, b]$ contained in X. Then $\mathcal{A}$ is an algebra of sets.

Our next notion, that of *additive set function*, might be viewed as the forerunner of the notion of measure. If we wish to model phenomena such as area, volume, or mass, we would like our model to conform to physical laws, reflect our intuition, and make precise concepts, such as "the whole is the sum of its parts." We can do this as follows.

Definition 2.9: Let $\mathcal{A}$ be an algebra of sets and let ν be an extended real-valued function defined on $\mathcal{A}$. If ν satisfies the following conditions, we say ν is an *additive set function.*

1. $\nu(\emptyset) = 0$.

2. If A, $B \in \mathcal{A}$ and $A \cap B = \emptyset$, then $\nu(A \cup B) = \nu(A) + \nu(B)$.

Note that such a function is allowed to take on infinite values, but cannot take on both $-\infty$ and ∞ as values. (See Exercise 2:2.8.) A nonnegative additive set function is often called a *finitely additive measure*.

Example 2.10: Let $X = (0, 1]$ and $\mathcal{A}$ be as in Example 2.8. Let f be an arbitrary function on $[0, 1]$. Define $\nu_f((a, b]) = f(b) - f(a)$, and extend ν_f to be additive on $\mathcal{A}$. Then ν_f is an additive set function. (See Exercise 2:2.14.)

2.2.1 Example: Distributions of mass

Example 2.10 plays an important role in the general theory, both for applications and to illustrate many ideas. Note that if f is nondecreasing, then the set function ν_f is nonnegative and can model many concepts. If $f(x) = x$ for all $x \in X$, then $\nu_f(A) = \lambda(A)$ for all $A \in \mathcal{A}$. Here, ν_f models a uniform distribution of mass—the amount of mass in an interval is proportional to the length of the interval. Another nondecreasing function would give rise to a different mass distribution. For example, if $f(x) = x^2$, $\nu_f((0, \frac{1}{2}]) = \frac{1}{4}$, while $\nu_f((\frac{1}{2}, 1]) = \frac{3}{4}$; in this case the mass is not uniformly distributed. As yet another example, let

$$f(x) = \begin{cases} 0, & 0 \le x < x_0 < 1; \\ 1, & x_0 \le x \le 1. \end{cases}$$

Then f has a jump discontinuity at x_0, and

$$\nu_f(A) = \begin{cases} 0, & \text{if } x_0 \notin A; \\ 1, & \text{if } x_0 \in A. \end{cases}$$

We would like to say that x_0 is a "point mass" and that the set function assigns the value 1 to the singleton set $\{x_0\}$, but $\{x_0\} \notin \mathcal{A}$. Since point masses arise naturally as models in nature, this algebra $\mathcal{A}$ is not fully adequate to discuss finite mass distributions on $(0,1]$. This flaw will disappear when we consider *measures* on *σ-algebras* in Section 2.3. In that setting, $\{x_0\}$ will be a member of the σ-algebra and will have unit mass. These ideas are the forerunner of Lebesgue–Stieltjes measures, which we study in Section 3.5.

In Example 2.10 we can take f *nonincreasing* and we can model "negative mass." This is analogous to the situation in elementary calculus where one often interprets an integral $\int_a^b g(x)\,dx$ in terms of negative area when the integrand is negative on the interval.

One can combine positive and negative mass. If f has a decomposition into a difference of monotonic functions

$$f = f_1 - f_2 \text{ with } f_1 \text{ and } f_2 \text{ nondecreasing on } X, \tag{1}$$

then it is easy to check that ν_f has a similar decomposition:

$$\nu_f = \nu_{f_1} - \nu_{f_2}.$$

Unless f is monotonic on X, there will be intervals of positive mass and intervals of negative mass. Functions f that admit the representation (1) are those that are of *bounded variation*. (We have reviewed some properties of such functions in Section 1.14. Note particularly Exercise 1:14.10.) It appears then that we can model a mass distribution ν_f on $[a,b]$ that involves both positive and negative mass as a difference of two nonnegative mass distributions. This is so if, in Example 2.10, f has bounded variation; is it true for an arbitrary function f?

2.2.2 Positive and negative variations

This leads us to variational questions for additive set functions that parallel the ideas and methods employed in the study of functions of bounded variation.

Definition 2.11: Let X be any set, let $\mathcal{A}$ be an algebra of subsets of X and let ν be additive on $\mathcal{A}$. For $E \in \mathcal{A}$, we define the *positive variation of ν on E* by

$$\overline{V}(\nu, E) = \sup \{\nu(A) : A \in \mathcal{A}, A \subset E\}.$$

Similarly, we define the *negative variation of ν on E* by

$$\underline{V}\,(\nu, E) = \inf \{\nu(A) : A \in \mathcal{A}, A \subset E\}.$$

Finally, we define the (total) *variation of ν on E* by

$$V(\nu, E) = \overline{V}\,(\nu, E) - \underline{V}\,(\nu, E).$$

Note that the positive variation is indeed positive or nonnegative since $\overline{V}\,(\nu, E) \geq \nu(\emptyset) = 0$. Similarly the negative variation is negative or nonpositive since $\underline{V}\,(\nu, E) \leq \nu(\emptyset) = 0$. The total variation, defined as the difference of the two expressions, is well-defined even if one or both of the postive and negative variations is infinite. [Some authors, thinking of these notions as sups and infs, call them *upper variation* and *lower variation* instead.]

Exercise 2:2.16 displays the total variation $V(\nu, E)$ in an equivalent form

$$V(\nu, E) = \sup \sum_{k=1}^{n} |\nu(A_k)|\,,$$

where the supremum is taken over all finite collections of pairwise disjoint subsets A_k of E, with each A_k in $\mathcal{A}$. For that reason some authors call the total variation the *absolute variation*. Note that it is reminiscent of the usual definition of variation for a real-valued function.

Theorem 2.12: *If ν is additive on an algebra $\mathcal{A}$ of subsets of X, then all the variations are additive set functions on $\mathcal{A}$.*

Proof. We show that the positive variation is additive on $\mathcal{A}$, the other proofs being similar. That $\overline{V}(\nu,\emptyset)=0$ is clear. To verify condition 2 of Definition 2.9, let A and B be disjoint members of $\mathcal{A}$. Assume first that

$$\overline{V}(\nu, A\cup B)<\infty.$$

Let $\varepsilon>0$. There exist A' and B' in $\mathcal{A}$ such that $A'\subset A$, $B'\subset B$, $\nu(A')>\overline{V}(\nu,A)-\varepsilon/2$, and $\nu(B')>\overline{V}(\nu,B)-\varepsilon/2$. Thus

$$\begin{aligned}\overline{V}(\nu, A\cup B)\geq \nu(A'\cup B') &= \nu(A')+\nu(B') \\ &> \overline{V}(\nu,A)+\overline{V}(\nu,B)-\varepsilon.\end{aligned} \tag{2}$$

In the other direction, there exists a set $C\in\mathcal{A}$ such that $C\subset A\cup B$ and $\nu(C)>\overline{V}(\nu,A\cup B)-\varepsilon$. Thus

$$\begin{aligned}\overline{V}(\nu, A\cup B)-\varepsilon<\nu(C) &= \nu(A\cap C)+\nu(B\cap C) \\ &\leq \overline{V}(\nu,A)+\overline{V}(\nu,B).\end{aligned} \tag{3}$$

Since ε is arbitrary, it follows from (2) and (3) that

$$\overline{V}(\nu, A\cup B)=\overline{V}(\nu,A)+\overline{V}(\nu,B).$$

It remains to consider the case $\overline{V}(\nu,A\cup B)=\infty$. Here one can easily verify that either $\overline{V}(\nu,A)=\infty$ or $\overline{V}(\nu,B)=\infty$, and the conclusion follows. ■

2.2.3 Jordan decomposition theorem

Theorem 2.13 provides an abstract version in the setting of additive set functions of the Jordan decomposition theorem for functions of bounded variation (Exercise 1:14.10). It indicates how, in many cases, a mass distribution can be decomposed into the difference of two nonnegative mass distributions. Here we shall show that

$$\nu(A) = \overline{V}\,(\nu, A) + \underline{V}\,(\nu, A)$$

or, equivalently,

$$\nu(A) = \overline{V}\,(\nu, A) - [-\underline{V}\,(\nu, A)].$$

Since $\underline{V}(\nu, A)$ is *nonpositive*, this latter identity expressed the decomposition as a difference of two nonnegative additive set functions.

Theorem 2.13 (Jordan decomposition) *Let ν be an additive set function on an algebra $\mathcal{A}$ of subsets of X, and suppose that ν has finite total variation. Then, for all $A \in \mathcal{A}$,*

$$\nu(A) = \overline{V}\,(\nu, A) + \underline{V}\,(\nu, A)\,. \tag{4}$$

Proof. Let A, $E \in \mathcal{A}$ and $E \subset A$. Since

$$\nu(E) = \nu(A) - \nu(A \setminus E),$$

we have

$$\nu(A) - \overline{V}(\nu, A) \leq \nu(E) \leq \nu(A) - \underline{V}(\nu, A). \tag{5}$$

Expression (5) is valid for all $E \in \mathcal{A}$, $E \subset A$. Noting the definition of $\overline{V}(\nu, A)$, we see from the second inequality that

$$\overline{V}(\nu, A) \leq \nu(A) - \underline{V}(\nu, A). \tag{6}$$

Similarly, from the first inequality, we infer that

$$\underline{V}(\nu, A) \geq \nu(A) - \overline{V}(\nu, A). \tag{7}$$

Comparing (6) and (7), we obtain our desired conclusion, (4). ■

Exercises

2:2.1 Show that an algebra of sets is closed under differences, finite unions, and finite intersections.

2:2.2 Let X be a nonempty set. Show that 2^X (the family of all subsets of X) and $\{\emptyset, X\}$ are both algebras of sets, in fact the largest and the smallest of the algebras of subsets of X.

2:2.3$\diamondsuit$ Let $\mathcal{S}$ be any family of subsets of a nonempty set X. The smallest algebra containing $\mathcal{S}$ is called *the algebra generated by* $\mathcal{S}$. Show that this exists. [Hint: This can be described as the intersection of all algebras containing $\mathcal{S}$. Make sure to check that there are such algebras and that the intersection of a collection of algebras is again an algebra.]

2:2.4$\diamondsuit$ Let $\mathcal{S}$ be a family of subsets of a nonempty set X such that (i) $\emptyset$, $X \in \mathcal{S}$ and (ii) if A, $B \in \mathcal{S}$ then both $A \cap B$ and $A \cup B$ are in $\mathcal{S}$. Show that the algebra generated by $\mathcal{S}$ is the family of all sets of the form $\bigcup_{i=1}^n A_i \setminus B_i$ for A_i, $B_i \in \mathcal{S}$ with $B_i \subset A_i$.

2:2.5$\diamondsuit$ Let X be an arbitrary nonempty set, and let $\mathcal{A}$ be the family of all subsets $A \subset X$ such that either A or $X \setminus A$ is finite. Show that $\mathcal{A}$ is the algebra generated by the singleton sets $\mathcal{S} = \{\{x\} : x \in X\}$.

2:2.6$\diamondsuit$ Let X be an arbitrary nonempty set, and let $\mathcal{A}$ be the algebra generated by a collection $\mathcal{S}$ of subsets of X. Let A be an arbitrary element of $\mathcal{A}$. Show that there is a finite family $\mathcal{S}_0 \subset \mathcal{S}$ so that A belongs to the algebra generated by $\mathcal{S}_0$. [Hint: Consider the union of all the algebras generated by finite subfamilies of $\mathcal{S}$.]

2:2.7 Show that Example 2.8 provides an algebra of sets.

2:2.8$\diamondsuit$ Show how it follows from Definition 2.9 that an additive set function ν cannot take on both $-\infty$ and ∞ as values. [Hint: If $\nu(A) = -\nu(B) = +\infty$, then find disjoint subsets A', B' with $\nu(A') = +\infty$ and $\nu(B') = -\infty$. Consider what this means for $\nu(A' \cup B')$.]

2:2.9 Suppose that ν is an additive set function on an algebra $\mathcal{A}$. Let E_1 and E_2 be members of $\mathcal{A}$ with $E_1 \subset E_2$ and $\nu(E_2)$ finite. Show that

$$\nu(E_2 \setminus E_1) = \nu(E_2) - \nu(E_1).$$

2:2.10 Let μ be a finitely additive measure and suppose that A, B and C are sets in the domain of μ with $\mu(A)$ finite. Show that

$$|\mu(A \cap B) - \mu(A \cap C)| \leq \mu(B \triangle C)$$

where $B \triangle C = (B \setminus C) \cup (C \setminus B)$ is called the *symmetric difference* of B and C.

2:2.11$\diamondsuit$ Suppose that ν is additive on an algebra $\mathcal{A}$. If $B \subset A$ with $A, B \in \mathcal{A}$ and $\nu(B) = +\infty$, then $\nu(A) = +\infty$.

2:2.12 Use Exercise 2:2.9 to show that the condition $\nu(\emptyset) = 0$ in Definition 2.9 is superfluous unless ν is identically infinite.

2:2.13 Let X be any infinite set, and let $\mathcal{A} = 2^X$. For $A \subset X$, let

$$\nu(A) = \begin{cases} 0, & \text{if } A \text{ is finite;} \\ \infty, & \text{if } A \text{ is infinite.} \end{cases}$$

Show that ν is additive. Let

$$\mathcal{B} = \{A \subset X : A \text{ is finite or } X \setminus A \text{ is finite}\},$$

let $B \in \mathcal{B}$, and let

$$\tau(B) = \begin{cases} 0, & \text{if } B \text{ is finite;} \\ \infty, & \text{if } X \setminus B \text{ is finite.} \end{cases}$$

Show that $\mathcal{B}$ is an algebra and τ is additive.

2:2.14$\diamondsuit$ Show that, in Example 2.10, ν_f is additive on $\mathcal{A}$ and ν_f is nonnegative if and only if f is nondecreasing. [Hint: This involves verifying that, for $A \in \mathcal{A}$, $\nu_f(A)$ does not depend on the choice of intervals whose union is A.]

2:2.15 Complete the proof of Theorem 2.12 by showing that the negative and total variations are additive on $\mathcal{A}$.

2:2.16 Establish the formula

$$V(\nu, E) = \sup \sum_{k=1}^{n} |\nu(A_k)|,$$

where the supremum is taken over all finite collections of pairwise disjoint subsets A_k of E, with each A_k in $\mathcal{A}$.

2:2.17 Suppose that ν is additive on $\mathcal{A}$ and is bounded above. Prove that $\overline{V}(\nu, A)$ is finite for all $A \in \mathcal{A}$. Similarly, if ν is bounded from below, $\underline{V}(\nu, A)$ is finite for all $A \in \mathcal{A}$.

2:2.18 Use Exercise 2:2.17 to obtain the Jordan decomposition for additive set functions that are bounded either above or below.

2:2.19 Show that to every finitely additive set function of finite total variation on the algebra of Example 2.8 corresponds a function f of bounded variation, such that $\nu(\,(a, b]\,) = f(b) - f(a)$ for every $(a, b] \in \mathcal{A}$.

2:2.20 We have already seen that if f is BV on $[0, 1]$ then Example 2.10 models a finite mass distribution that may have negative, as well as positive, mass. What happens if f is not of bounded variation? Is there necessarily a decomposition into a difference of nonnegative additive set functions then?

2.3 Measures and Signed Measures

Additive set functions defined on algebras have limitations as models for mass distributions or areas. These limitations are in some way similar to limitations of the Riemann integral. The

Riemann integral fails to integrate enough functions. Similarly, an algebra of sets may not include all the sets that one expects to be able to handle. In Example 2.10, for example, one can discuss the mass of an interval or a finite union of intervals, but one cannot define mass for more general sets.

We have mentioned several times that to obtain a coherent theory of measure the class of measurable sets should be "large." What do we mean by that statement? Roughly, we should require that the class of sets to be considered measurable encompass all the sets that one reasonably expects to encounter while applying the normal operations of analysis. The situation on the real line with Lebesgue measure will illustrate.

In a study of a continuous function $f : \mathbb{R} \to \mathbb{R}$ we could expect to investigate sets of the form $\{x : f(x) \geq c\}$ or $\{x : f(x) > c\}$. The first of these is closed and the second open if f is continuous. We would hope that these sets are measurable, as indeed they are for Lebesgue measure. In Chapter 3 we shall make the measurability of closed and open sets a key requirement in our study of general measures on metric spaces.

Again, if f is the limit of a convergent sequence of continuous functions (a common enough operation in analysis), what can we expect for the set

$$\{x : f(x) > c\}?$$

We can rewrite this as

$$\{x : f(x) > c\} = \bigcup_{m=1}^{\infty} \bigcup_{r=1}^{\infty} \bigcap_{n=r}^{\infty} \{x : f_n(x) \geq c + 1/m\}$$

(using Exercise 1:1.24). It follows that the set that we are interested in is measurable provided that the class of measurable sets is closed under the operations of taking countable unions and countable intersections. An algebra of sets need only be closed under the operations of taking finite unions and finite intersections.

2.3.1 σ–algebras of sets

This, and other considerations, leads us to Definition 2.14. We shall see that with this definition we can develop a coherent theory of measure and integration.

Definition 2.14: Let X be a set, and let $\mathcal{M}$ be a family of subsets of X. We say that $\mathcal{M}$ is a *σ–algebra of sets* if $\mathcal{M}$ is an algebra of sets and $\mathcal{M}$ is closed under countable unions; that is, if $\{A_k\} \subset \mathcal{M}$, then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{M}$.

2.3.2 Signed measures

It is now natural to replace the notion of additive set function with *countably* additive set function or *signed measure*.

Definition 2.15: Let $\mathcal{M}$ be a σ-algebra of subsets of a set X, and let μ be an extended real-valued function on $\mathcal{M}$. We say that μ is a *signed measure* if $\mu(\emptyset) = 0$, and whenever $\{A_k\}$ is a sequence of pairwise disjoint elements of $\mathcal{M}$, then $\sum_{n=1}^{\infty} \mu(A_n)$ is defined as an extended real number with

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{8}$$

If $\mu(A) \geq 0$ for all $A \in \mathcal{M}$, we say that μ is a *measure*. In this case we call the triple $(X, \mathcal{M}, \mu)$ a *measure space*. The members of $\mathcal{M}$ are called *measurable sets*.

We mention that the term *countably additive set function* μ indicates that μ satisfies (8). We shall also use the term *σ-additive set function.*

Example 2.16: Let $X = \mathbb{N}$ (the set of natural numbers) and $\mathcal{M} = 2^{\mathbb{N}}$, the family of all subsets of $\mathbb{N}$. It is clear that $\mathcal{M}$ is a σ-algebra of sets. For $A \in \mathcal{M}$, let

$$\mu_1(A) = \textstyle\sum_{n\in A} 1/2^n, \qquad \mu_2(A) = \textstyle\sum_{n\in A} 1/n,$$

$$\mu_3(A) = \textstyle\sum_{n\in A} (-1)^n/2^n, \qquad \mu_4(A) = \textstyle\sum_{n\in A} (-1)^n/n.$$

One verifies easily that μ_1 and μ_2 are measures, with $\mu_1(X) = 1$ and $\mu_2(X) = \infty$. The set function μ_3 is a signed measure. Since the series $\sum_{n=1}^{\infty} (-1)^n/n$ is conditionally convergent, $\mu_4(A)$ is not defined for all subsets of X, and μ_4 is not a signed measure.

An inspection of the example μ_3 reveals that it is the difference of two measures,

$$\mu_3(A) = \sum_{n\,\in\,A,\ n\text{ even}} 1/2^n - \sum_{n\,\in\,A,\ n\text{ odd}} 1/2^n,$$

just as we have seen that every additive set function is the difference of two nonnegative additive set functions. In Section 2.5 we will show that this is always the case for signed measures; thus we will be able to reduce the study of signed measures to the study of measures. Signed measures will again return to a position of importance in Chapter 5. At the moment, our focus will be on measures.

2.3.3 Computations with signed measures

We shall require immediately some skill in handling measures. Often we are faced with a set expressed as a countable union of measurable sets. If the sets are disjoint, then the measure of the union can be obtained as a sum. What do we do if the sets are not pairwise disjoint? Our first theorem shows how to unscramble these sets in a useful way. (We leave the straightforward proof of Theorem 2.17 as Exercise 2:3.11. Recall that we use $\mathbb{N}$ to denote the set of natural numbers.)

Theorem 2.17: *Let $\{A_n\}$ be a sequence of subsets of a set X, and let $A = \bigcup_{n=1}^{\infty} A_n$. Let $B_1 = A_1$ and, for all $n \in \mathbb{N}$, $n \geq 2$, let*

$$B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1}).$$

Then $A = \bigcup_{n=1}^{\infty} B_n$, the sets B_n are pairwise disjoint and $B_n \subset A_n$ for all $n \in \mathbb{N}$. If the sets A_n are members of an algebra $\mathcal{M}$, then $B_n \in \mathcal{M}$ for all $n \in \mathbb{N}$.

We next show that measures are *monotonic* and *countably subadditive*.

Theorem 2.18: *Let $(X, \mathcal{M}, \mu)$ be a measure space.*

1. *If $A, B \in \mathcal{M}$ with $B \subset A$, then $\mu(B) \leq \mu(A)$. If, in addition, $\mu(B) < \infty$, then $\mu(A \setminus B) = \mu(A) - \mu(B)$.*
2. *If $\{A_k\}_{k=1}^{\infty} \subset \mathcal{M}$, then $\mu(\bigcup_{k=1}^{\infty} A_k) \leq \sum_{k=1}^{\infty} \mu(A_k)$.*

Proof. Part (i) follows from the representation

$$A = B \cup (A \setminus B).$$

To verify part (ii), let $\{A_k\} \in \mathcal{M}$, and let $A = \bigcup_{k=1}^{\infty} A_k$. Let $\{B_k\}$ be the sequence of sets appearing in Theorem 2.17. Since $\mathcal{M}$ is an algebra of sets, $B_k \in \mathcal{M}$ for all $k \in \mathbb{N}$. It follows that $A = \bigcup_{k=1}^{\infty} B_k$ and that the sets B_k are pairwise disjoint. Since μ is a measure, $\mu(A) = \sum_{k=1}^{\infty} \mu(B_k)$. But for each $k \in \mathbb{N}$, $\mu(B_k) \leq \mu(A_k)$, by part (i). Thus $\mu(A) \leq \sum_{k=1}^{\infty} \mu(A_k)$. ■

We end

2.3.4 The σ-algebra generated by a family of sets

Note that any family $\mathcal{S}$ of subsets of a nonempty set X is contained in the σ-algebra 2^X of all subsets of X. It is also contained in a smallest σ-algebra.

Definition 2.19: The smallest σ-algebra containing a family of sets $\mathcal{S}$ is called *the σ-algebra generated by* $\mathcal{S}$.

This can be described as the intersection of all σ-algebras containing $\mathcal{S}$. Indeed, to prove that a smallest σ-algebra containing a given family of sets $\mathcal{S}$ exists, one simply checks that the intersection of all σ-algebras containing $\mathcal{S}$ is itself a σ-algebra.

The σ-algebra generated by the open (or closed) subsets of $\mathbb{R}$ is called the class of *Borel sets*. It contains all sets of type $\mathcal{F}_\sigma$ or of type $\mathcal{G}_\delta$, but it also contains many other sets. The σ-algebra generated by the algebra $\mathcal{A}$ of Example 2.10 also consists of the Borel sets.

Exercises

2:3.1 Let X be a nonempty set. Show that 2^X (the family of all subsets of X) and $\{\emptyset, X\}$ are both σ–algebras of sets, in fact the largest and the smallest of the σ–algebras of subsets of X.

2:3.2 Let $\mathcal{S}$ be any family of subsets of a nonempty set X. The smallest σ–algebra containing $\mathcal{S}$ is called *the σ–algebra generated by* $\mathcal{S}$. Show that this exists. [Hint: This is described in the last paragraph of this section. Compare with Exercise 2:2.3.]

2:3.3 Let $\mathcal{S}$ be a family of subsets of a nonempty set X such that (i) $\emptyset$, $X \in \mathcal{S}$ and (ii) if A, $B \in \mathcal{S}$, then both $A \cap B$ and $A \cup B$ are in $\mathcal{S}$. Show that the σ–algebra generated by $\mathcal{S}$ is, in general, *not* the family of all sets of the form $\bigcup_{i=1}^{\infty} A_i \setminus B_i$ for A_i, $B_i \in \mathcal{S}$ with $B_i \subset A_i$. This contrasts with what one might have expected in view of Exercise 2:2.4. [Hint: Take $\mathcal{S}$ as the collection of intervals $[0, n^{-1}]$ along with $\emptyset$.]

2:3.4 Let X be an arbitrary nonempty set, and let $\mathcal{A}$ be the family of all subsets $A \subset X$ such that either A or $X \setminus A$ is countable. Show that $\mathcal{A}$ is the σ–algebra generated by the singleton sets $\mathcal{S} = \{\{x\} : x \in X\}$.

2:3.5 Let X be an arbitrary nonempty set, and let $\mathcal{A}$ be the σ–algebra generated by a collection $\mathcal{S}$ of subsets of X. Let A be an arbitrary element of $\mathcal{A}$. Show that there is a countable family $\mathcal{S}_0 \subset \mathcal{S}$ so that A belongs to the σ–algebra generated by $\mathcal{S}_0$. [Hint: Compare with Exercise 2:2.6.]

2:3.6 Let $\mathcal{A}$ be an algebra of subsets of a set X. If $\mathcal{A}$ is finite, prove that $\mathcal{A}$ is in fact a σ–algebra. How many elements can $\mathcal{A}$ have?

2:3.7 Describe the domain of the set function μ_4 defined in Example 2.16.

2:3.8 Show that a σ-algebra of sets is closed under countable intersections.

2:3.9$\diamondsuit$ Let X be any set, and let $\mu(A)$ be the number of elements in A if A is finite and ∞ if A is infinite. Show that μ is a measure. (Commonly, μ is called the *counting measure* on X.)

2:3.10$\diamondsuit$ Let μ be a signed measure on a σ-algebra. Show that the associated variations are countably additive. Thus, by Theorem 2.13, each signed measure of finite total variation is a difference of two measures. (See Theorem 2.23 for an improvement of this statement.)

2:3.11 Prove Theorem 2.17.

2:3.12 Let ν be a signed measure on a σ-algebra. If $E_0 \subset E_1 \subset E_2 \dots$ are members of the σ–algebra, then the limit $\lim_{n\to\infty} E_n$ of the sequence is defined to be $\bigcup_{n=0}^{\infty} E_n$. Prove that

$$\nu(\lim_{n\to\infty} E_n) = \lim_{n\to\infty} \nu(E_n).$$

[The method of Theorem 2.21 can be used, but try to prove without looking ahead. The same remark applies to the next exercise.]

2:3.13◇ Let ν be a signed measure on a σ-algebra. If $E_0 \supset E_1 \supset E_2 \dots$ are members of the σ–algebra, then the limit $\lim_{n\to\infty} E_n$ of the sequence is defined to be $\bigcap_{n=0}^{\infty} E_n$. Prove that if $\nu(E_0)$ is finite then

$$\nu(\lim_{n\to\infty} E_n) = \lim_{n\to\infty} \nu(E_n).$$

2.4 Limit Theorems

The countable additivity of a signed measure allows a number of limit theorems not possible for the general additive set function. To formulate some of these theorems, we need a bit of set-theoretic terminology.

2.4.1 Limsup and liminf of a sequence of sets

First, recall that if A is a subset of a set X then the *characteristic function* of A is defined by

$$\chi_A(x) = \begin{cases} 1, & \text{if } x \in A; \\ 0, & \text{if } x \in X \setminus A. \end{cases}$$

Suppose, now, that we are given a sequence $\{A_n\}$ of subsets of X. Then there exist sets B_1 and B_2 with

$$\chi_{B_1} = \limsup \chi_{A_n}$$

and

$$\chi_{B_2} = \liminf \chi_{A_n}.$$

The set B_1 consists of those $x \in X$ that belong to infinitely many of the sets A_n, while the set B_2 consists of those $x \in X$ that belong to all but a finite number of the sets A_n. We call these sets the $\limsup A_n$ and $\liminf A_n$, respectively. Our formal definition has the advantage of involving only set-theoretic notions.

Definition 2.20: Let $\{A_n\}$ be a sequence of subsets of a set X. We define

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

and

$$\liminf A_n = \bigcup_{m=1}^{\infty} \left(\bigcap_{n=m}^{\infty} A_n \right).$$

If

$$\limsup A_n = \liminf A_n = A,$$

we say that *the sequence* $\{A_n\}$ *converges to* A and we write

$$A = \lim A_n.$$

2.4.2 Monotone limits in a measure space

Observe that monotone sequences, either expanding or contracting, converge to their union and intersection, respectively. Furthermore, if all the sets A_n belong to a σ-algebra $\mathcal{M}$, then

$\limsup A_n \in \mathcal{M}$ and $\liminf A_n \in \mathcal{M}$.

For monotone sequences of measurable sets, limit theorems are intuitively clear.

Theorem 2.21: *Let $(X, \mathcal{M}, \mu)$ be a measure space, and let $\{A_n\}$ be a sequence of measurable sets.*

1. *If $A_1 \subset A_2 \subset \ldots$, then $\lim \mu(A_n) = \mu(\lim A_n)$.*
2. *If $A_1 \supset A_2 \supset \ldots$ and $\mu(A_m) < \infty$ for some $m \in I\!N$, then $\lim \mu(A_n) = \mu(\lim A_n)$.*

Proof. Let $A_0 = \emptyset$. Then

$$\lim_n A_n = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} (A_n \setminus A_{n-1}).$$

Since the last union is a disjoint union, we can infer that

$$\begin{aligned}\mu(\lim_n A_n) &= \sum_{n=1}^{\infty} \mu(A_n \setminus A_{n-1}) = \lim_k \sum_{n=1}^{k} \mu(A_n \setminus A_{n-1}) \\ &= \lim_k \mu\left(\bigcup_{n=1}^{k} (A_n \setminus A_{n-1})\right) = \lim_k \mu(A_k).\end{aligned}$$

This proves part (i). For part (ii), choose m so that $\mu(A_m) < \infty$. A similar argument shows that

$$\mu(A_m \setminus \lim_n A_n) = \lim_n \left(\mu(A_m) - \mu(A_n)\right).$$

Because these are finite, assertion (ii) follows. ■

2.4.3 Liminfs and limsups in a measure space

Theorem 2.22: *Let $(X, \mathcal{M}, \mu)$ be a measure space, and let $\{A_n\}$ be a sequence of sets from $\mathcal{M}$. Then*

1. $\mu(\liminf A_n) \leq \liminf \mu(A_n)$*;*
2. *if* $\mu(\bigcup_{n=1}^{\infty} A_n) < \infty$*, then* $\mu(\limsup A_n) \geq \limsup \mu(A_n)$*;*
3. *if* $\{A_n\}$ *converges and* $\mu(\bigcup_{n=1}^{\infty} A_n) < \infty$*, then*
$$\mu(\lim A_n) = \lim \mu(A_n).$$

Proof. We prove (i), the remaining parts following readily. For $m \in \mathbb{N}$, let $B_m = \bigcap_{n=m}^{\infty} A_n$. Since $B_m \subset A_m$, $\mu(B_m) \leq \mu(A_m)$. It follows that

$$\liminf \mu(B_m) \leq \liminf \mu(A_m). \tag{9}$$

The sequence$\{B_m\}$ is expanding, so $\lim_m B_m = \bigcup_{m=1}^{\infty} B_m$. Using Theorem 2.21, we then obtain

$$\mu(\lim_m B_m) = \lim_m \mu(B_m).$$

Thus

$$\begin{aligned} \mu(\liminf A_n) \ &= \ \mu\left(\bigcup_{m=1}^{\infty} B_m\right) = \mu(\lim_m B_m) = \lim_m \mu(B_m) \\ &= \ \liminf \mu(B_m) \leq \liminf \mu(A_m), \end{aligned}$$

the last inequality being (9). ∎

Exercises

2:4.1 Verify that in Definition 2.20

$$\limsup_{n\to\infty} A_n = \{x : x \in A_n \text{ for infinitely many } n\}$$

and

$$\liminf_{n\to\infty} A_n = \{x : x \in A_n \text{ for all but finitely many } n\}\,.$$

2:4.2 Supply all the details needed to prove part (ii) of Theorem 2.21.

2:4.3 For any $A \subset \mathbb{N}$, let

$$\nu(A) = \begin{cases} \sum_{n\in A} 2^{-n}, & \text{if } A \text{ is finite;} \\ \infty, & \text{if } A \text{ is infinite.} \end{cases}$$

(a) Show that ν is an additive set function, but not a measure on $2^{\mathbb{N}}$.

(b) Show that ν does not have the limit property expressed in part (i) of Theorem 2.21 for measures.

2:4.4 Verify parts (ii) and (iii) of Theorem 2.22.

2:4.5 Show that the finiteness assumptions in parts 2 and 3 of Theorem 2.22 cannot be dropped.

2:4.6 State and prove an analog for Theorem 2.21 for signed measures.

2:4.7◊ Verify the following criterion for an additive set function to be a signed measure: If ν is additive on a σ-algebra $\mathcal{M}$, and $\lim_n \nu(A_n) = \nu(\lim_n A_n)$ for every expanding sequence $\{A_n\}$ of sets from $\mathcal{M}$, then ν is a signed measure on $\mathcal{M}$.

2:4.8◊ (Borel–Cantelli lemma) Let $(X, \mathcal{M}, \mu)$ be a measure space, and let $\{A_n\}$ be a sequence of sets with $\sum_{n=1}^{\infty} \mu(A_n) < \infty$. Then

$$\mu(\limsup A_n) = 0.$$

2:4.9 Let C be a Cantor set in $[0,1]$ of measure α ($0 \leq \alpha < 1$) (see Example 2.1). Does there exist a sequence $\{J_k\}$ of intervals with $\sum_{k=1}^{\infty} \lambda(J_k) < \infty$ such that every point of the set C lies in infinitely many of the intervals J_k?

2:4.10$\diamondsuit$ Let $\mathcal{A}$ be the algebra of Example 2.10, let

$$f(x) = \begin{cases} 0, & \text{if } 0 \leq x < x_0 < 1; \\ 1, & \text{if } x_0 \leq x \leq 1. \end{cases},$$

and let ν_f be as in that example. We shall see later that ν_f can be extended to a measure μ_f defined on the σ-algebra $\mathcal{B}$ of Borel sets in $(0,1]$. Assume this, for the moment. Show that $\mu_f(\{x_0\}) = 1$; thus $\{x_0\}$ represents a point mass.

2.5 The Jordan and Hahn Decomposition Theorems

Let us return to the Jordan decomposition theorem, but applied now to signed measures. Certainly, since a signed measure is also an additive set function, we see that any signed measure with finite variation can be expressed as the difference of two nonnegative additive set functions. We expect the latter to be measures, but this does not yet follow. In the setting of signed measures there is also a technical simplification that comes about. An additive set function may be itself finite and yet have both of its variations infinite. For this reason, in the proof of Theorem 2.13, we needed to assume that both variations were finite; otherwise, the proof collapsed. For signed measures this does not occur.

2.5.1 Jordan Decomposition

Thus we have the correct version of the decomposition for signed measures, with better hypotheses and a stronger conclusion.

Theorem 2.23 (Jordan decomposition) *Let ν be a signed measure on a σ–algebra $\mathcal{A}$ of subsets of X. Then, for all $A \in \mathcal{A}$,*

$$\nu(A) = \overline{V}\,(\nu, A) + \underline{V}\,(\nu, A)$$

and the set functions $\overline{V}\,(\nu, \cdot)$ and $-\underline{V}\,(\nu, \cdot)$ are measures, at least one of which must be finite.

Proof. This follows by the same methods used in the proof of Theorem 2.13, provided we establish some facts. We can prove (see Exercise 2:3.10) that if ν is σ–additive on $\mathcal{A}$ then so too are both variations. We prove also that if ν is finite then both variations are finite. Thus, with these two facts, the theorem (for finite-valued signed measures) follows directly from Theorem 2.13.

If ν is not finite, then we shall show that precisely one of the two variations is infinite. In fact, if $\nu(E)$ takes the value $+\infty$, then $\overline{V}\,(\nu, E) = +\infty$ and $-\underline{V}\,(\nu, \cdot)$ is everywhere finite. With this information the proof of Theorem 2.13 can be repeated to obtain the decomposition.

Evidently then, the theorem can be obtained from the following assertion which we will now prove.

2.24: *Let ν be a signed measure on a σ–algebra $\mathcal{A}$ of subsets of X. If $E \in \mathcal{A}$ and $\overline{V}\,(\nu, E) = +\infty$, then $\nu(E) = +\infty$. If $E \in \mathcal{A}$ and $\underline{V}\,(\nu, E) = -\infty$, then $\nu(E) = -\infty$.*

It is sufficient to prove the first statement. Suppose that $\overline{V}\,(\nu, E) = +\infty$. Because of Exercise 2:2.11, we may obtain that $\nu(E) = +\infty$ by finding a subset $A \subset E$ with $\nu(A) = +\infty$. There must exist a set $E_1 \subset E$ such that

$$\nu(E_1) > 1.$$

As $\overline{V}(\nu, \cdot)$ is additive and $\overline{V}(\nu, E) = +\infty$, it follows that either $\overline{V}(\nu, E_1) = \infty$ or else $\overline{V}(\nu, E \setminus E_1) = \infty$. Choose A_1 to be either E_1 or $E \setminus E_1$, according to which of these two is infinite, so that $\overline{V}(\nu, A_1) = +\infty$.

Inductively choose $E_n \subset A_{n-1}$ so that

$$\nu(E_n) > n$$

and choose A_n to be either E_n or $A_{n-1} \setminus E_n$ according to which of these two is infinite so that $\overline{V}(\nu, A_n) = +\infty$.

There are two case to consider:

1. For an infinite number of n, $A_n = A_{n-1} \setminus E_n$.

2. For all sufficiently large n (say for $n \geq n_0$), $A_n = E_n$.

In the first of these cases we obtain a sequence of disjoint sets $\{E_{n_k}\}$ so that we can use the σ–additivity of ν to obtain

$$\nu\left(\bigcup_{k=1}^{\infty} E_{n_k}\right) = \sum_{k=1}^{\infty} \nu\left(E_{n_k}\right) \geq \sum_{k=1}^{\infty} n_k = +\infty.$$

This would give us a subset of E with infinite ν measure so that $\nu(E) = +\infty$ as required.

In the second case we have obtained a sequence

$$E \supset E_{n_0} \supset E_{n_0+1} \supset E_{n_0+2} \ldots.$$

If $\nu(E_{n_0}) = +\infty$, we once again have a subset of E with infinite ν measure so that $\nu(E) = +\infty$ as required. If $\nu(E_{n_0}) < +\infty$, then we can use Exercise 2:3.13 to obtain

$$\nu(\lim_{n\to\infty} E_n) = \lim_{n\to\infty} \nu(E_n) \geq \lim_{n\to\infty} n = +\infty$$

and yet again have a subset of E with infinite ν measure so that $\nu(E) = +\infty$. This exhausts all possibilities and so the proof of assertion 2.24 is complete. The main theorem now follows. ■

2.6 Hahn Decomposition

The Jordan decomposition theorem is one of the primary tools of general measure theory. It can be clarified considerably by a further analysis due originally to H. Hahn (1879–1934). In fact, our proof invokes the Jordan decomposition, but Hahn's theorem could be proved first and then one can derive the Jordan decomposition from it. This decomposition is, again, one of the main tools of general measure theory; we shall have occasion to use it later in our discussion of the Radon–Nikodym theorem in Section 5.8.

Theorem 2.25 (Hahn decomposition) *Let ν be a signed measure on a σ-algebra $\mathcal{M}$. Then there exists a set $P \in \mathcal{M}$ such that $\nu(A) \geq 0$ whenever $A \subset P$, $A \in \mathcal{M}$, and $\nu(A) \leq 0$ whenever $A \subset X \setminus P$, $A \in \mathcal{M}$.*

We call the set P a *positive* set for ν, the set $N = X \setminus P$ a *negative* set for ν, and the pair (P, N) a *Hahn decomposition* for ν.

Proof. Using Exercise 2:2.8, we see that ν cannot take both the values $+\infty$ and $-\infty$. Assume for definiteness that $\nu(E) < \infty$ for all $E \in \mathcal{M}$. It follows that $\overline{V}(\nu, X)$ is finite. We construct a set P for which

$$\overline{V}\left(\nu, \widetilde{P}\right) = \underline{V}(\nu, P) = 0,$$

where $\overline{V}$ and $\underline{V}$ denote the positive and negative variations of ν as defined in Section 2.2. We know that $\overline{V}$ and $-\underline{V}$ are measures. (Recall the notation $\widetilde{P}$ for the complement of a set P.)

For each $n \in \mathbb{N}$, there exists $P_n \in \mathcal{M}$ such that

$$\nu(P_n) > \overline{V}(\nu, X) - \frac{1}{2^n}. \tag{10}$$

Define $P = \limsup_{n\to\infty} P_n$, so that $\widetilde{P} = \liminf_{n\to\infty} \widetilde{P_n}$. Then, from the inequality (10), we have

$$\overline{V}\left(\nu, \widetilde{P_n}\right) = \overline{V}(\nu, X) - \overline{V}(\nu, P_n) \le \overline{V}(\nu, X) - \nu(P_n) \le \frac{1}{2^n}.$$

Using Theorem 2.22 (i), we infer that

$$0 \le \overline{V}\left(\nu, \widetilde{P}\right) \le \liminf_{n\to\infty} \overline{V}\left(\nu, \widetilde{P_n}\right) \le \lim_{n\to\infty} \frac{1}{2^n} = 0.$$

Thus $\overline{V}\left(\nu, \widetilde{P}\right) = 0$.

It remains to show that $\underline{V}(\nu, P) = 0$. First, note that

$$-\underline{V}(\nu, P_n) = \overline{V}(\nu, P_n) - \nu(P_n) \le \overline{V}(\nu, X) - \nu(P_n) \le \frac{1}{2^n}.$$

Hence, for every $k \in \mathbb{N}$,

$$\begin{aligned} 0 \;\le\; -\underline{V}(\nu, P) &\le -\underline{V}\left(\nu, \bigcup_{n=k}^{\infty} P_n\right) \\ &\le\; -\sum_{n=k}^{\infty} \underline{V}(\nu, P_n) \le \sum_{n=k}^{\infty} \frac{1}{2^n} = \frac{1}{2^{k-1}}. \end{aligned}$$

It follows that $\underline{V}(\nu, P) = 0$ as required. ■

Note the connection with variation both in the proof of this theorem and in the decomposition itself. For any signed measure ν we shall use its Hahn decomposition (P, N) to define three further measures ν^+, ν^- and $|\nu|$ by writing for each $E \in \mathcal{M}$,

$$\nu^+(E) = \nu(E \cap P) = \overline{V}(\nu, E) \quad \text{[positive variation]}$$

$$\nu^-(E) = -\nu(E \cap N) = -\underline{V}(\nu, E) \quad \text{[− negative variation]}$$

and

$$|\nu|(E) = \nu^+(E) + \nu^-(E) \qquad \text{[total variation].}$$

Observe that the three set functions here, ν^+, ν^+, and $|\nu|$ derived from the signed measure ν are measures themselves (not merely signed measures) and that the following obvious relations hold among them:

$$\nu = \nu^+ - \nu^-$$

$$|\nu| = \nu^+ + \nu^-.$$

Two measures α and β on $\mathcal{M}$ are called *mutually singular*, written as $\alpha \perp \beta$, if there are disjoint measurable sets A and B such that $X = A \cup B$ and $\alpha(B) = \beta(A) = 0$; that is, the measures are concentrated on two different disjoint sets. The measures here ν^+ and ν^- are mutually singular, since $\nu^+(N) = \nu^-(P) = 0$.

Exercises

2:6.1 A set E is a *null set* for a signed measure ν if $|\nu|(E) = 0$. Show that if (P, N) and (P_1, N_1) are Hahn decompositions for ν then P and P_1 (and similarly N and N_1) differ by a null set [i.e., $|\nu|(P \setminus P_1) = |\nu|(P_1 \setminus P) = 0$].

2:6.2 Exhibit a Hahn decomposition for each of the signed measures μ_3 and $3\mu_1 - \mu_2$, where μ_1, μ_2, and μ_3 have been given in Example 2.16.

2:6.3 Let F be the Cantor function on [0,1] (defined in Exercise 1:22.13). Suppose that μ_F is a measure on the Borel subsets of $(0,1]$ for which $\mu_F((a,b]) = F(b) - F(a)$ for any $(a,b] \subset (0,1]$. Let λ be Lebesgue measure restricted to the Borel sets.

(a) Show that $\mu_F \perp \lambda$.

(b) Exhibit a Hahn decomposition for $\lambda - \mu_F$.

2.7 Complete Measures

Consider for a moment Lebesgue measure λ on $[0,1]$. Since λ is the restriction of λ^* to the family $\mathcal{L}$ of Lebesgue measurable sets, every subset of a zero measure set has measure zero. But, for a general measure space $(X, \mathcal{M}, \mu)$, it need not be the case that subsets of zero measure sets are necessarily measurable.

This is illustrated by the space $(X, \mathcal{B}, \lambda)$, where X is $[0,1]$ and $\mathcal{B}$ is the class of Borel sets in $[0,1]$: that is, $\mathcal{B}$ is the σ-algebra generated by the open sets. A cardinality argument (Exercise 2:7.1) shows that, while the Cantor ternary set K has 2^c subsets, only c of them are Borel sets, yet $\lambda(K) = 0$. It follows that there are Lebesgue measurable sets of measure zero that are not Borel sets. Thus $(X, \mathcal{B}, \lambda)$ is not *complete* according to the following definition.

Definition 2.26: Let $(X, \mathcal{M}, \mu)$ be a measure space. The measure μ is called *complete* if the conditions $Z \subset A$ and $\mu(A) = 0$ imply that $Z \in \mathcal{M}$. In that case, $(X, \mathcal{M}, \mu)$ is called a *complete measure space.*

Completeness of a measure refers to the domain $\mathcal{M}$ and so, properly speaking, it is $\mathcal{M}$ that might be called complete; but it is common usage to refer directly to a complete measure.

2.7.1 The completion of a measure space

It is clear from the monotonicity of μ that, when a subset of a zero measurable set is measurable, its measure must be zero. When a measure space is not complete, it possesses subsets E that intuition demands be small, but that do not happen to be in the domain of the measure μ. It may seem that such sets should have measure zero, but the measure is not defined for such sets. It would be convenient if one could always deal with a complete space. Instead of saying that a property is valid except on a "subset of a set of measure zero," we could correctly say "except on a set of measure zero." Fortunately, every measure space can be *completed* naturally by extending μ to a measure $\overline{\mu}$ defined on the σ-algebra generated by $\mathcal{M}$ and the family of subsets of sets of measure zero.

Theorem 2.27: *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let*

$$\mathcal{Z} = \{Z : \exists N \in \mathcal{M} \text{ for which } Z \subset N \text{ and } \mu(N) = 0\}.$$

Let $\overline{\mathcal{M}} = \{M \cup Z : M \in \mathcal{M}, Z \in \mathcal{Z}\}$. Define $\overline{\mu}$ on $\overline{\mathcal{M}}$ by

$$\overline{\mu}(M \cup Z) = \mu(M).$$

Then

1. *$\overline{\mathcal{M}}$ is a σ-algebra containing $\mathcal{M}$ and $\mathcal{Z}$.*
2. *$\overline{\mu}$ is a measure on $\overline{\mathcal{M}}$ and agrees with μ on $\mathcal{M}$.*
3. *$\overline{\mu}$ is complete.*

Proof. Part (i). It is clear that $\overline{\mathcal{M}}$ contains $\mathcal{M}$ and $\mathcal{Z}$. To show that $\overline{\mathcal{M}}$ is closed under complementation, let $A = M \cup Z$ with $M \in \mathcal{M}$, $Z \subset N$ and $\mu(N) = 0$. Using our usual notation for complementation, i.e., $\widetilde{E} = X \setminus E$, we can check that

$$\widetilde{A} = \widetilde{M} \cap \widetilde{Z} = (\widetilde{M} \cap \widetilde{N}) \cup (N \cap \widetilde{M} \cap \widetilde{Z}).$$

Since $\widetilde{M} \cap \widetilde{N} \in \mathcal{M}$ and $N \cap \widetilde{M} \cap \widetilde{Z} \subset N \in \mathcal{Z}$, we see from the definition of $\overline{\mathcal{M}}$ that $\widetilde{A} \in \overline{\mathcal{M}}$.

Finally, we show that $\overline{\mathcal{M}}$ is closed under countable unions. Let $\{A_n\}$ be a sequence of sets in $\overline{\mathcal{M}}$. For each $n \in \mathbb{N}$, write

$$A_n = M_n \cup Z_n$$

with $M_n \in \mathcal{M}$, $Z_n \in \mathcal{Z}$. Then

$$\bigcup A_n = \bigcup (M_n \cup Z_n) = \left(\bigcup M_n\right) \cup \left(\bigcup Z_n\right).$$

We have $M_n \in \mathcal{M}$ and $Z_n \subset N_n \in \mathcal{M} \cap \mathcal{Z}$, so $\bigcup M_n \in \mathcal{M}$ and

$$\bigcup Z_n \subset \bigcup N_n \in \mathcal{M} \cap \mathcal{Z}.$$

Thus $\bigcup A_n$ has the required representation. This completes the verification of (i).

Part (ii). We first check that $\overline{\mu}$ is well defined. Suppose that A has two different representations:

$$A = M_1 \cup Z_1 = M_2 \cup Z_2$$

for M_1, $M_2 \in \mathcal{M}$, Z_1, $Z_2 \in \mathcal{Z}$. We show $\mu(M_1) = \mu(M_2)$. Now

$$M_1 \subset A = M_2 \cup Z_2 \subset M_2 \cup N_2 \text{ with } \mu(N_2) = 0.$$

Thus

$$\mu(M_1) \leq \mu(M_2) + \mu(N_2) = \mu(M_2).$$

Similarly, $\mu(M_2) \le \mu(M_1)$, so $\overline{\mu}$ is well defined.

To show that $\overline{\mu}$ is a measure on $\overline{\mathcal{M}}$, we verify countable additivity, the remaining requirements being trivial to verify. Let $\{A_n\}$ be a sequence of pairwise disjoint sets in $\overline{\mathcal{M}}$. For every $n \in \mathbb{N}$, we can write $A_n = M_n \cup Z_n$ for sets $M_n \in \mathcal{M}$, $Z_n \in \mathcal{Z}$. Note that the union $\bigcup_{n=1}^{\infty} M_n$ belongs to $\mathcal{M}$ and that $\bigcup_{n=1}^{\infty} Z_n$ belongs to $\mathcal{Z}$. Then

$$\overline{\mu}\left(\bigcup_{n=1}^{\infty} A_n\right) = \overline{\mu}\left(\bigcup_{n=1}^{\infty} (M_n \cup Z_n)\right) = \overline{\mu}\left(\left(\bigcup_{n=1}^{\infty} M_n\right) \cup \left(\bigcup_{n=1}^{\infty} Z_n\right)\right)$$
$$= \mu\left(\bigcup_{n=1}^{\infty} M_n\right) = \sum_{n=1}^{\infty} \mu\left(M_n\right) = \sum_{n=1}^{\infty} \overline{\mu}\left(A_n\right).$$

Thus $\overline{\mu}$ is a measure on $\overline{\mathcal{M}}$. It is clear from the representation $A = M \cup Z$ and the definition of $\overline{\mu}$ that $\overline{\mu} = \mu$ on $\mathcal{M}$.

Part (iii) Let $\overline{\mu}(A) = 0$ and let $B \subset A$. We show that $\overline{\mu}(B) = 0$. Write $A = M \cup Z$, $M \in \mathcal{M}$, $Z \in \mathcal{Z}$. Since $\overline{\mu}(A) = 0$, $\mu(M) = 0$, so $A = M \cup Z \in \mathcal{Z}$. It follows that $B \in \mathcal{Z} \subset \overline{\mathcal{M}}$, and so $\overline{\mu}$ is complete as required. ■

Exercises

2:7.1 Prove each of the following assertions:

(a) The cardinality of the class $\mathcal{G}$ of open subsets of $[0,1]$ is c.

(b) The cardinality of the class $\mathcal{B}$ of Borel sets in $[0,1]$, is also c.

(c) The zero measure Cantor set has subsets that are not Borel sets.

(d) The measure space $(X, \mathcal{B}, \lambda)$ is not complete.

2:7.2 Let $\mathcal{B}$ denote the Borel sets in $[0,1]$, and let λ be Lebesgue measure on $\mathcal{B}$. Prove that

$$([0,1], \overline{\mathcal{B}}, \overline{\lambda}) = ([0,1], \mathcal{L}, \lambda).$$

2.8 Outer Measures

We turn now to the following general problem. Suppose that we have a primitive notion for some phenomenon that we wish to model in the setting of a suitable measure space. How can we construct such a space? We can abstract some ideas from Lebesgue's approach (given in Section 2.1). That procedure involved three steps. The primitive notion of the length of an open interval was the starting point. This was used to provide an outer measure defined on all subsets of $\mathbb{R}$. That, in turn, led to an inner measure and then, finally, the class of measurable sets was defined as the collection of sets on which the inner and outer measures agreed. In this section and the next we shall see that this same procedure can be used quite generally. Only one important variant is necessary—we must circumvent the use of inner measure. The reason for this will become apparent.

We begin by abstracting the essential properties of the Lebesgue outer measure. A method for constructing outer measures similar to that used to construct the Lebesgue outer measure will be developed in the next section.

Definition 2.28: Let X be a set, and let μ^* be an extended real-valued function defined on 2^X such that

1. $\mu^*(\emptyset) = 0$.
2. If $A \subset B \subset X$, then $\mu^*(A) \leq \mu^*(B)$.
3. If $\{A_n\}$ is a sequence of subsets of X, then
$$\mu^*\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu^*(A_n).$$

Then μ^* is called an *outer measure* on X.

It follows from the first two conditions that an outer measure is nonnegative. Condition 3 is called *countable subadditivity.*

Let us first address the question of how we obtain a measure from an outer measure. The simple example that follows may be instructive.

Example 2.29: Let $X = \{1, 2, 3\}$. Let $\mu^*(\emptyset) = 0$, $\mu^*(X) = 2$, and $\mu^*(A) = 1$ for every other set $A \subset X$. It is a routine matter to verify that μ^* is an outer measure. Suppose now that we wish to mimic the procedure that worked so well for the Peano–Jordan content and the Lebesgue measure. We could take our cue from the formula in assertion 2.4 and define a version of the inner measure for this example as

$$\mu_*(A) = \mu^*(X) - \mu^*(X \setminus A) = 2 - \mu^*(X \setminus A).$$

If we then call A measurable provided that $\mu_*(A) = \mu^*(A)$, and let

$$\mu(A) = \mu^*(A)$$

for such sets, our process is complete. We find that *all* eight subsets of X are measurable by this definition, but μ is clearly not additive on 2^X. The classical inner–outer measure procedure completely fails to work in this simple example!

A bit of reflection pinpoints the problem. The inner–outer measure approach puts a set A to the following test stated solely in terms of μ^*: is it true that

$$\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)?$$

In Example 2.29, every $A \subset X$ passed this test. But, for $A = \{1\}$ and $E = \{1, 2\}$, we see that

$$\mu^*(A) + \mu^*(E \setminus A) = 2 > 1 = \mu^*(E).$$

Thus, while μ^* is additive with respect to A and its complement in X, it is not with respect to A and its complement in E.

2.8.1 Measurable sets with respect to an outer measure

These considerations lead naturally to the following criterion of measurability. It is due to Constantin Carathéodory (1873–1950). We have already touched upon Lebesgue's notion of a measurable set, defined using inner and outer measures. Carathéodory's definition is more general and avoids the introduction of inner measures.

Definition 2.30: Let μ^* be an outer measure on X. A set $A \subset X$ is μ^*-*measurable* if, for all sets $E \subset X$,

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \setminus A). \tag{11}$$

This definition of the measurability of a set A requires testing the set A against every subset E of the space. In contrast the inner–outer measure approach requires only that equation (11) of Definition 2.30 be valid for the single "test set" $E = X$.

Example 2.31: Let X and μ^* be as in Example 2.29. Consider $a, b \in X$, with $a \neq b$. If $E = \{a, b\}$ is examined as the test set in (11) of Definition 2.30, we see that $\{a\}$ is not μ^*-measurable. Similarly, we find that no two-point set is μ^*-measurable. Thus only $\emptyset$ and X are μ^*-measurable. This is the best one could hope for if some kind of additivity of μ^* over the measurable sets is to occur. Note, also, that unlike Lebesgue measure, nonmeasurable sets in X have no measurable covers or measurable kernels. (See Exercise 2:1.14.)

2.8.2 The σ-algebra of measurable sets

Definition 2.30 defining measurability involves an additivity requirement of μ^*, but not any kind of σ-additivity. It may therefore be surprising that this simple modification of the inner–outer measure approach suffices to provide a σ-algebra $\mathcal{M}$ of measurable sets on which μ^* is σ-additive.

Theorem 2.32: *Let X be a set, μ^* an outer measure on X, and $\mathcal{M}$ the class of μ^*-measurable sets. Then $\mathcal{M}$ is a σ-algebra and μ^* is countably additive on $\mathcal{M}$. Thus the set function μ defined on $\mathcal{M}$ by $\mu(A) = \mu^*(A)$ for all $A \in \mathcal{M}$ is a measure.*

Proof. In applying condition (11) in Definition 2.30 for a measurability test of a set A note that it is enough, because of subadditivity, to verify that

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \setminus A) \tag{12}$$

for an arbitrary set E of finite measure. We start by checking that $\emptyset \in \mathcal{M}$. For any set $E \subset X$, obviously

$$\mu^*(E) = \mu^*(E \cap \emptyset) + \mu^*(E \setminus \emptyset),$$

verifying condition (12). If $A \in \mathcal{M}$ and $B = X \setminus A$ then the identity

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \setminus A) = \mu^*(E \setminus B) + \mu^*(E \cap B)$$

shows that $B \in \mathcal{M}$. Thus far we know that that $\emptyset \in \mathcal{M}$ and $\mathcal{M}$ is closed under complementation.

We show that any finite union of sets in $\mathcal{M}$ must also be in $\mathcal{M}$. It is sufficient to check that, if A_1, $A_2 \in \mathcal{M}$, then necessarily $A_1 \cup A_2 \in \mathcal{M}$. For any $E \subset X$, since $A_1 \in \mathcal{M}$,

$$\mu^*(E) = \mu^*(E \cap A_1) + \mu^*(E \setminus A_1).$$

For the set $A_2 \in \mathcal{M}$, we use the test set $E \setminus A_1$ to obtain

$$\mu^*(E \cap A_1) = \mu^*((E \setminus A_1) \cap A_2) + \mu^*((E \setminus A_1) \setminus A_2).$$

We need the two simple set identities

$$(E \setminus A_1) \setminus A_2 = E \setminus (A_1 \cup A_2)$$

and

$$[(E \setminus A_1) \cap A_2] \cup [E \cap A_1] = E \cap (A_1 \cup A_2).$$

Putting these together and using the subadditivity of the outer measure we obtain,

$$\mu^*(E) \geq \mu^*(E \cap A_1) + \mu^*(E \setminus A_1) \cap A_2) + \mu^*(E \setminus (A_1 \cup A_2))$$

$$\geq \mu^*(E \cap (A_1 \cup A_2)) + \mu^*(E \setminus (A_1 \cup A_2)).$$

This is exactly what we need to verify, using condition (12), to prove that $A_1 \cup A_2 \in \mathcal{M}$.

Now let $\{A_j\}$ be a sequence of measurable sets. We shall verify that the union

$$A = \bigcup_{j=1}^{\infty} A_j$$

belongs also to $\mathcal{M}$. There is no loss of generality in assuming that the sets are disjoint, for we can express A as a union of a disjoint sequence of sets from $\mathcal{M}$ using a familiar device (e.g. as in our proof of Theorem 2.21).

We let $E \subset X$ and show that Definition 2.30 is satisfied. Write $B_n = \bigcup_{j=1}^{n} A_j$. Note that $B_n \in \mathcal{M}$. We give an inductive proof that, for every $n = 1, 2, 3, \ldots$,

$$\mu^*(E \cap B_n) = \mu^*\left(E \cap \bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \mu^*(E \cap A_j). \tag{13}$$

This is certainly true for $n = 1$. Suppose that it is true for some given n. Use $E \cap B_{n+1}$ as a test set for the measurability of B_n (which we know is measurable) to obtain

$$\mu^*(E \cap B_{n+1}) = \mu^*([E \cap B_{n+1}] \cap B_n) + \mu^*([E \cap B_{n+1}] \setminus B_n)$$

and deduce, using the induction hypothesis, that

$$\mu^*(E \cap B_{n+1}) = \mu^*(E \cap B_n) + \mu^*(E \cap A_{n+1}) = \sum_{j=1}^{n+1} \mu^*(E \cap A_j).$$

Thus (13) is proved.

The monotonicity of the outer measure and the subadditivity of the outer measure now supplies the inequality

$$\sum_{j=1}^{n} \mu^*(E \cap A_j) = \mu^*(E \cap B_n) \leq \mu^*(E \cap A) \leq \sum_{j=1}^{\infty} \mu^*(E \cap A_j)$$

valid for all n. From this we see that

$$\mu^*(E \cap A) = \sum_{j=1}^{\infty} \mu^*(E \cap A_j). \tag{14}$$

It remains only to test the measurability of A using the test set E. From the monotonicity of outer measures and (13) we have

$$\mu^*(E) = \mu^*(E \cap B_n) + \mu^*(E \setminus B_n) \geq \sum_{j=1}^{n} \mu^*(E \cap A_j) + \mu^*(E \setminus A)$$

for all n. If this is true for all n, then

$$\mu^*(E) \geq \sum_{j=1}^{\infty} \mu^*(E \cap A_j) + \mu^*(E \setminus A).$$

Finally, using (14), we obtain our test for the measurability of A, that

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \setminus A).$$

This completes the proof that $\mathcal{M}$ is a σ-algebra and there remains only to observe that μ^* is countably additive on $\mathcal{M}$. But this is precisely what (14) shows, for that identity is valid for any set E and any sequence of disjoint measurable sets $\{A_j\}$. ■

Exercises

2:8.1 Let μ^* be an outer measure on X, and suppose that one of the two sets A, $B \subset X$ is measurable. Show that $\mu^*(A) + \mu^*(B) = \mu^*(A \cup B) + \mu^*(A \cap B)$.

2:8.2 Let X be an uncountable set. Let $\mu^*(A) = 0$ if A is countable and $\mu^*(A) = 1$ if A is uncountable. Show that μ^* is an outer measure, and determine the class of measurable sets.

2:8.3 Let μ^* be an outer measure on X, and let Y be a μ^*-measurable subset of X. Let $\nu^*(A) = \mu^*(A)$ for all $A \subset Y$. Show that ν^* is an outer measure on Y, and a set $A \subset Y$ is ν^*-measurable if and only if A is μ^*-measurable. Thus, for example, a subset A of $[0, 1]$ is Lebesgue measurable (as a subset of $[0, 1]$) if and only if it is Lebesgue measurable as a subset of $\mathbb{R}$.

2:8.4$\diamondsuit$ Prove that if $A \subset X$ and $\mu^*(A) = 0$ then A is μ^*-measurable. Consequently, the measure space generated by any outer measure is complete.

2.9 Method I

In Section 2.8 we have seen how one can obtain a measure μ from an outer measure μ^*. We still have the problem of determining how to obtain an outer measure μ^* so that the resulting measure μ is compatible with whatever primitive notion we wish to extend.

Once again, we can abstract this from Lebesgue's procedure. Suppose that we have a set X, a family $\mathcal{T}$ of subsets of X, and a nonnegative function $\tau : \mathcal{T} \to [0, \infty]$. We view $\mathcal{T}$ as the family of sets for which we have a primitive notion of "size" and $\tau(T)$ as a measure of that size. We shall call τ a premeasure to indicate the role that it takes in defining a measure. In order for our methods to work, we need assume no more of a premeasure τ than that it is nonnegative and vanishes on the empty set. [In the Lebesgue framework of Section 2.1, for example, we can take $X = [0,1]$, $\mathcal{T}$ as the family of open intervals, and the premeasure $\tau(T)$ as the length of the open interval T.]

Here is a more formal development of these ideas.

Definition 2.33: Let X be a set, and let $\mathcal{T}$ be a family of subsets of X such that $\emptyset \in \mathcal{T}$. A nonnegative function τ defined on $\mathcal{T}$ so that $\tau(\emptyset) = 0$ is called a *premeasure*, and we refer to the family $\mathcal{T}$ as a *covering family* for X.

Note that hardly anything is assumed about the properties of a premeasure and a covering family. The terminology is employed just to indicate the intended use: we use the members of the family to cover sets, and we use the premeasure to generate an outer measure. The process, defined in the following theorem, of constructing outer measures is often called Method I in the literature.

Method I is very useful, but it can have an important flaw when X is a metric space. In Section 3.2 we shall discuss this flaw and see how a variant, called Method II, overcomes this problem.

Theorem 2.34 (Method I construction of outer measures) *Let $\mathcal{T}$ be a covering family for a set X, and let $\tau : \mathcal{T} \to [0,\infty]$ with $\tau(\emptyset) = 0$. For $A \subset X$, let*

$$\mu^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \tau(T_n) : \ T_n \in \mathcal{T} \text{ and } A \subset \bigcup_{n=1}^{\infty} T_n \right\}, \tag{15}$$

where an empty infimum is taken as ∞. Then μ^* is an outer measure on X.

Proof. Before beginning the proof note that a set A not contained in any countable union of sets from the covering family $\mathcal{T}$ is assigned an infinite outer measure. Note too that, while the definition of the outer measure uses countable covers, finite covers are included as well since $\emptyset \in \mathcal{T}$ and $\tau(\emptyset) = 0$.

It is clear that $\mu^*(\emptyset) = 0$ and that μ^* is monotone. To verify that μ^* is countably subadditive, let $\{A_n\}$ be a sequence of subsets of X. We show that

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

If any $\mu^*(A_n) = \infty$, there is nothing to prove, so we suppose that each is finite. Let $\varepsilon > 0$. For every $n \in \mathbb{N}$, there exists a sequence $\{T_{nk}\}_{k=1}^{\infty}$ of sets from $\mathcal{T}$ such that $A_n \subset \bigcup_{k=1}^{\infty} T_{nk}$, and

$$\sum_{k=1}^{\infty} \tau(T_{nk}) \leq \mu^*(A_n) + \frac{\varepsilon}{2^n}. \tag{16}$$

Now

$$\bigcup_{n=1}^{\infty} A_n \subset \bigcup_{n=1}^{\infty} \bigcup_{k=1}^{\infty} T_{nk},$$

so by (15) and (16)

$$\mu^*(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \tau(T_{nk}) \leq \sum_{n=1}^{\infty} \left[\mu^*(A_n) + \frac{\varepsilon}{2^n}\right] = \sum_{n=1}^{\infty} \mu^*(A_n) + \varepsilon.$$

We conclude that

$$\mu^*\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$$

since ε is an arbitrary positive number. ■

2.9.1 A warning

The philosophy of the Method I construction, we recall, is to refine some premeasure τ acting on a family of sets $\mathcal{T}$ in such a way as to produce a useful outer measure. One might assume that the new outer measure reflects closely properties of the tools that were used to construct it. But the outer measure may assign different values to the sets in $\mathcal{T}$ than the premeasure τ and may not even consider the sets in $\mathcal{T}$ to be measurable.

Exercises 2:9.3 to 2:9.5 illustrate that the members of $\mathcal{T}$ need not, in general, be measurable and that $\tau(T)$ need not equal $\mu(T)$, not even when $T \in \mathcal{T}$ is in fact measurable.

For a natural and important example, suppose that we wish a measure-theoretic model for area in the Euclidean plane $\mathbb{R}^2$. We could start with $\mathcal{T}$ as the family of open squares (along with $\emptyset$) and with $\tau(T)$ as the area of the square T. We apply Method I to obtain an outer measure λ_2^* in $\mathbb{R}^2$. We then restrict λ_2^* to the class $\mathcal{L}_2$ of measurable sets, and we shall have Lebesgue's two-dimensional measure λ_2.

We would be assured at this point of having a σ-algebra of measurable sets $\mathcal{L}_2$, but we would need to do more work to show that $\mathcal{L}_2$ possesses certain desirable properties. Nothing in our general work so far guarantees, for example, that members of the original family $\mathcal{T}$ are in $\mathcal{L}_2$ (i.e., the members of $\mathcal{T}$ are measurable) or, indeed, that the measure of a square T is the original value $\tau(T)$ with which we started. In the case of $\mathcal{L}_2$, it would be unfortunate if open squares were not measurable by the criterion of Definition 2.30 and worse still if the measure of a square were not its area. We shall see later, fortunately, that no such problem exists for Lebesgue measure in $\mathbb{R}^n$ or for a variety of other important measures.

Exercises

2:9.1 Verify that the set function μ^* as defined in (15) satisfies conditions 1 and 2 of Definition 2.28.

2:9.2◇ Refer to Example 2.10. Let $\mathcal{T}$ consist of $\emptyset$ and the half-open intervals $(a, b] \subset (0, 1]$, and let $\tau = \nu_f$. Apply Method I to obtain μ^* and $\mathcal{M}$. Assuming that $\mathcal{T} \subset \mathcal{M}$ and $\mu = \tau$ on $\mathcal{T}$, this now provides a model for mass distributions on $(0, 1]$. Let $q_1, q_2, \dots$ be an enumeration of $\mathbb{Q} \cap (0, 1]$. Construct a function f, so that for all $A \subset (0, 1]$,

$$\mu(A) = \sum_{q_n \in A} \frac{1}{2^n},$$

where μ is obtained from τ by our process, and $\tau((a, b]) = f(b) - f(a)$.

2:9.3◇ Let $X = \{1, 2, 3\}$, $\mathcal{T}$ consist of $\emptyset$, X and all doubleton sets, with $\tau(\emptyset) = 0$, $\tau(\{x, y\}) = 1$, for all $x \neq y \in X$, and $\tau(X) = 2$. Show that Method I results in the outer measure μ^* of Example 2.29. How do things change if $\tau(X) = 3$?

2:9.4 Let $X = \mathbb{N}$, $\mathcal{T}$ consist of $\emptyset$, X, and all singleton sets. Let $\tau(\emptyset) = 0$, $\tau(\{x\}) = 1$, for all $x \in X$, and

(a) $\tau(X) = 2$.

(b) $\tau(X) = \infty$.

In each case, apply Method I and determine the family of measurable sets.

2:9.5 Repeat Exercise 2:9.4 with the modification that

$$\tau(\{x\}) = \frac{1}{2^{x-1}}.$$

[Note in part (b), that $X \in \mathcal{M}$, but $\tau(X) \neq \mu(X)$.] How do things change if $\tau(X) = 1$?

2:9.6 Show that if $\mathcal{T} \subset \mathcal{M}$ then $\mu(T) \leq \tau(T)$ for all $T \in \mathcal{T}$.

2.10 Regular Outer Measures

We saw in Section 2.8 that the inner–outer measure approach does not, in general, give rise to a measure on a σ-algebra. There are, however, many situations in which the class of sets whose inner and outer measures are the same is identical to the class of sets measurable according to Definition 2.30.

Definition 2.35: An outer measure μ^* is called *regular* if for every $E \subset X$ there exists a measurable set $H \supset E$ such that $\mu(H) = \mu^*(E)$. The set H is called a *measurable cover* for E.

Theorem 2.36: *Let μ^* be a regular outer measure on X and suppose that*

$$\mu^*(X) < \infty.$$

A necessary and sufficient condition that a set $A \subset X$ be measurable is that

$$\mu^*(X) = \mu^*(A) + \mu^*(X \setminus A). \tag{17}$$

Proof. The necessity is clear from Definition 2.30. To prove that the condition is sufficient, let A be a subset of X satisfying (17), let E be any subset of X, and let H be a measurable cover for E. It suffices to verify that

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \setminus A), \tag{18}$$

the reverse inequality being automatically satisfied because of the subadditivity of μ^*.

Observe first that

$$\mu^*(A \setminus H) + \mu^*((X \setminus A) \setminus H) \geq \mu^*(X \setminus H). \tag{19}$$

Since H is measurable, we have

$$\mu^*(A) = \mu^*(A \cap H) + \mu^*(A \setminus H) \tag{20}$$

and

$$\mu^*(X \setminus A) = \mu^*(H \setminus A) + \mu^*((X \setminus H) \setminus A). \tag{21}$$

Now $\mu^*(X) = \mu^*(A) + \mu^*(X \setminus A)$ by (17). Thus, from equations (20) and (21) and the subadditivity of μ^*, we infer that

$$\begin{aligned} \mu(X) &= \mu^*(A \cap H) + \mu^*(A \setminus H) + \mu^*(H \setminus A) + \mu^*((X \setminus H) \setminus A) \\ &\geq \mu(H) + \mu(X \setminus H) = \mu(X). \end{aligned}$$

It follows that the one inequality above is actually an equality. Subtracting the inequality (19) from this equality, we obtain

$$\mu^*(H \cap A) + \mu^*(H \setminus A) \leq \mu(H). \tag{22}$$

This subtraction is justified since all the quantities involved are finite. Because $E \subset H$, we see from (22) that

$$\mu^*(E \cap A) + \mu^*(E \setminus A) \leq \mu^*(H \cap A) + \mu^*(H \setminus A) \leq \mu(H) = \mu^*(E).$$

This verifies (18). ■

In Section 2.1, we gave a sketch of one-dimensional Lebesgue measure and promised there to justify those aspects of the development that we did not verify at the time. The material in Section 2.8 provides a framework for developing Lebesgue measure using the Carathéodory criterion of Definition 2.30 and Method I. But it does not justify the inner–outer measure approach of Section 2.1. For that, we need to verify that λ^* is regular and then invoke Theorem 2.36.

2.10.1 Regularity of Method I outer measures

It is not the case that every outer measure obtained by Method I is regular. Example 2.29 and Exercise 2:9.3 show this. Theorem 2.37 is useful in showing that, when Method I is invoked for the purpose of extending the primitive notions that we have already mentioned (length, area, volume, and mass) the resulting outer measures will be regular.

Theorem 2.37: *Let μ^* be constructed by Method I from $\mathcal{T}$ and τ. If all members of $\mathcal{T}$ are μ^*-measurable, then μ^* is regular.*

Proof. Let $A \subset X$. We find a measurable cover for A. If $\mu^*(A) = \infty$, then X is a measurable cover. Suppose then that $\mu^*(A) < \infty$. For each $m \in \mathbb{N}$, let $\{T_{mn}\}_{n=1}^{\infty}$ be a sequence of sets from the covering class $\mathcal{T}$ such that

$$A \subset \bigcup_{n=1}^{\infty} T_{mn} \quad \text{and} \quad \sum_{n=1}^{\infty} \tau(T_{mn}) < \mu^*(A) + \frac{1}{m}.$$

Let

$$T_m = \bigcup_{n=1}^{\infty} T_{mn} \quad \text{and} \quad H = \bigcap_{m=1}^{\infty} T_m.$$

Since each of the sets T_{mn} is measurable, so too is H. We show that H is a measurable cover for A.

Clearly, $A \subset H$ and so $\mu^*(A) \leq \mu(H)$. For the opposite inequality, we have, for each $m \in \mathbb{N}$,

$$\mu^*(T_m) \leq \sum_{n=1}^{\infty} \mu^*(T_{mn}) \leq \sum_{n=1}^{\infty} \tau(T_{mn}) \leq \mu^*(A) + \frac{1}{m}.$$

For each $m \in \mathbb{N}$, $H \subset T_m$, and so

$$\mu(H) \le \mu^*(T_m) \le \mu^*(A) + \frac{1}{m}.$$

This last inequality is true for all $m \in \mathbb{N}$, so $\mu(H) \le \mu^*(A)$. Thus $\mu(H) = \mu^*(A)$, and H is a measurable cover for A. ■

2.10.2 Regularity of Lebesgue outer measure

Corollary 2.38: *Lebesgue outer measure λ^* on $\mathbb{R}$ is regular.*

Proof. Here $\mathcal{T}$ consists of $\emptyset$ and the open intervals, and $\tau(T)$ is the length of the interval T. Because of Theorem 2.37, it suffices to show that each interval (a, b) is measurable by Carathéodory's criterion (Definition 2.30).

Let $E \subset \mathbb{R}$ and let $\varepsilon > 0$. There is a sequence $\{T_n\} \subset \mathcal{T}$ that covers E for which

$$\sum_{n=1}^{\infty} \tau(T_n) \le \lambda^*(E) + \frac{\varepsilon}{2}.$$

Take

$$\mathcal{U}_1 = \{T_n \cap (a, b) : n \in \mathbb{N}\},$$
$$\mathcal{U}_2 = \{T_n \cap (-\infty, a) : n \in \mathbb{N}\},$$
$$\mathcal{U}_3 = \{T_n \cap (b, \infty) : n \in \mathbb{N}\},$$

and

$$\mathcal{U}_4 = \left\{ \left(a - \tfrac{1}{8}\varepsilon, a + \tfrac{1}{8}\varepsilon\right), \left(b - \tfrac{1}{8}\varepsilon, b + \tfrac{1}{8}\varepsilon\right) \right\}.$$

Then $\mathcal{U}_1$ covers $E \cap (a,b)$ and $\mathcal{U}_2 \cup \mathcal{U}_3 \cup \mathcal{U}_4$ covers $E \setminus (a,b)$. The total length of the intervals in $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3$ is the same as for the original sequence, and the additional lengths from $\mathcal{U}_4$ have total length equal to $\varepsilon/2$. Hence

$$\lambda^*(E \cap (a,b)) + \lambda^*(E \setminus (a,b)) \leq \sum_{n=1}^{\infty} \tau(T_n) + \varepsilon/2 \leq \lambda^*(E) + \varepsilon.$$

Since ε is arbitrary, we have

$$\lambda^*(E \cap (a,b)) + \lambda^*(E \setminus (a,b)) \leq \lambda^*(E)$$

for any $E \subset \mathbb{R}$, and it follows that (a,b) must be measurable. ■

2.10.3 Summary

Let us summarize some of the ideas in Sections 2.8 and 2.10, insofar as they relate to the important case of Lebesgue measure on an interval. We start with the covering family $\mathcal{T}$ of open intervals and with the primitive notion $\tau(T)$ as the length of the interval T. Upon applying Method I, this gives rise to an outer measure μ^*. We then apply the Carathéodory process to obtain a class $\mathcal{M}$ of measurable sets and a measure μ that equals μ^* on $\mathcal{M}$. To verify that our primitive notion of length is not destroyed by the process, we show, as in the proof of Corollary 2.38, that open intervals are measurable. It is then almost trivial to verify that the measure of an interval is its length. Theorem 2.37 now tells us that μ^* is regular; thus we could have used the inner–outer measure approach of Section 2.1. This would result in the same class of measurable sets and the same measure as provided by the Carathéodory process.

Exercises

2:10.1 Prove that, if μ^* is a regular outer measure and $\{A_n\}$ is a sequence of sets in X, then $\mu^*(\liminf A_n) \leq \liminf \mu^*(A_n)$. Compare with Theorem 2.22 (i).

2:10.2◇ Prove that, if μ^* is a regular outer measure and $\{A_n\}$ is an expanding sequence of sets, then $\mu^*(\lim_n A_n) = \lim_n \mu^*(A_n)$. Compare with Theorem 2.21 (i).

2:10.3 Show that the conclusions of Exercises 2:10.1 and 2:10.2 are not valid for arbitrary outer measures.

2:10.4 Let $X = \mathbb{N}$, $\mu^*(\emptyset) = 0$, and $\mu^*(E) = 1$ for all $E \neq \emptyset$.

(a) Show that μ^* is a regular outer measure.

(b) Let $\{A_n\}$ be a sequence of subsets of X (not assumed measurable). Show that, while the analog of part (i) of Theorem 2.22 does hold (Exercise 2:10.1), the analogs of parts (ii) and (iii) do not hold.

2:10.5 Let $X = \mathbb{N}$, and let $0 = a_0, a_1 = \frac{1}{2} < a_2 < a_3 < \cdots$ with $\lim_n a_n = 1$. If E has n members, let $\mu^*(E) = a_n$. If E is infinite, let $\mu^*(E) = 1$.

(a) Show that μ^* is an outer measure, but that μ^* is not regular.

(b) Show that the conclusions of Exercise 2:10.2 and Theorem 2.36 hold.

2:10.6 Prove the following variant of Theorem 2.36: Let μ^* be a regular outer measure, let H be measurable with $\mu(H) < \infty$, and let $A \subset H$. If $\mu(H) = \mu^*(H \cap A) + \mu^*(H \setminus A)$, then A is measurable.

2:10.7◇ Let $X = (0, 1]$, $\mathcal{T}$ consist of the half-open intervals $(a, b]$ contained in $(0, 1]$, and f be increasing and right continuous on $(0, 1]$ with $\lim_{x\to 0} f(x) = 0$. Let $\tau((a, b]) = f(b) - f(a)$. Apply Method I to obtain an outer measure μ^*_f. Prove that $\mathcal{T} \subset \mathcal{M}$ and μ^*_f is regular and thus the inner–outer measure approach works here. Observe that all open sets as well as all closed sets are μ^*_f measurable.

In particular, such measures can be used to model mass distributions on $\mathbb{R}$. (See Exercise 2:4.10, and Example 2.10 and the discussion following it.)

2:10.8$\diamondsuit$ Let $\mathcal{T}$ be a covering family for X. Prove that, if Method I is applied to $\mathcal{T}$ and τ to obtain the outer measure μ^*, then for each $E \subset X$ with $\mu^*(E) < \infty$ there exists $S \in \mathcal{T}_{\sigma\delta}$ such that $E \subset S$ and $\mu^*(S) = \mu^*(E)$. (In particular, if X is a metric space and $\mathcal{T}$ consists of open sets, S can be taken to be of type G_δ.) [Hint: See the proof of Theorem 2.37.]

2.11 Nonmeasurable Sets

In any particular setting, can we determine the existence of nonmeasurable sets? Certainly, it is easy to give artificial examples where all sets are measurable or where nonmeasurable sets exist. But in important applications we would like some generally applicable methods.

The special case of Lebesgue nonmeasurable sets should be instructive. Vitali was the first to demonstrate the existence of such sets using the axiom of choice. Let $0 = r_0, r_1, r_2, \dots$ be an enumeration of $\mathbb{Q} \cap [-1, 1]$. Using this sequence, he finds a set $A \subset [-\frac{1}{2}, \frac{1}{2}]$ so that the collection of sets

$$A_k = \{x + r_k : x \in A\}$$

forms a disjoint sequence covering the interval $[-\frac{1}{2}, \frac{1}{2}]$. As Lebesgue measure is translation invariant and countably additive, the set A cannot be measurable. (See Section 1.10 for the details.) In Section 12.6 we will encounter an example of a *finitely* additive measure that extends Lebesgue measure to *all* subsets of $[0, 1]$ and is translation invariant. This set function cannot be a measure, however, because of the Vitali construction. Unfortunately, this discussion does little to help us in general as it focuses attention on the additive group structure of $\mathbb{R}$ and the invariance of λ.

Another example may help more. We have seen a proof of the existence of Bernstein sets, that is, a set of real numbers such that neither it nor its complement contains any perfect set. (See Exercises 1:22.7 and 1:22.8.) Such a set cannot be Lebesgue measurable. To see this, remember that the outer measure of any set can be approximated from above by open sets; consequently, the measure of a measurable set can be approximated from inside by closed (or perfect) sets. But a Bernstein set and its complement contain no perfect set, and so both would have to have measure zero if they were measurable.

This example does contain a clue, albeit somewhat obliquely. The example suggests that some topological property (relating to closed and open sets) of Lebesgue measure is intimately related to the existence of nonmeasurable sets. But the proof of the existence of Bernstein sets simply employed a cardinality argument and did not invoke any deep topological properties of the real line. In fact, the nonmeasurability question reduces in many cases, surprisingly, to one of cardinality.

2.11.1 Ulam's theorem

The following result of S. M. Ulam illustrates the first step in this direction. Ultimately, we wish to ask, for a set X, when is it possible to have a finite measure defined on all subsets of X, but that assigns zero measure to each singleton set?

Theorem 2.39 (Ulam) *Let Ω be the first uncountable ordinal, and let $X = [0, \Omega)$. If μ is a finite measure defined on all subsets of X and such that $\mu(\{x\}) = 0$ for each $x \in X$, then μ is the zero measure.*

Proof. For any $y \in X$, write $A_y = \{x \in X : x < y\}$, the set of all predecessors of y. Then each set A_y is countable, and so there is an injection

$$f(\cdot, y) : A_y \to \mathbb{N}.$$

Define for each $x \in X$ and $n \in \mathbb{N}$

$$B_{x,n} = \{z \in X : x < z,\ f(x,z) = n\}.$$

If x_1, x_2 are distinct points in X, then evidently the sets $B_{x_1,n}$ and $B_{x_2,n}$ are disjoint. Since μ is finite, this means that, for each integer n, $\mu(B_{x,n}) > 0$ for only countably many $x \in X$. This means, since X is uncountable, that there must be some $x_0 \in X$ for which $\mu(B_{x_0,n}) = 0$ for each integer n.

Consider the union

$$B_0 = \bigcup_{n=1}^{\infty} B_{x_0,n}$$

and observe that $\mu(B_0) = 0$. If $y > x_0$, then $f(x_0, y) = n$ for some $n \in \mathbb{N}$. Hence $\{y \in X : x_0 < y\} \subset B_0$. Thus

$$X = B_0 \cup \{y \in X : y \leq x_0\},$$

and this expresses X as the union of a set of μ measure zero and a countable set. Hence $\mu(X) = 0$ as required. ■

If we assume CH (the continuum hypothesis), it follows from Ulam's theorem that there is no finite measure defined on *all* subsets of the real line and vanishing at points except for the zero measure itself. This applies not just to the real line, then, but to any set of cardinality **c**. This is true even without invoking the continuum hypothesis, but requires other axioms of set theory. Note that this means that it is not the invariance of Lebesgue measure or its properties relative to open and closed sets that does not allow it to be defined on all subsets of the reals. There is no nontrivial finite measure defined on all subsets of an interval of the real line that vanishes on singleton sets.

These ideas can be generalized to spaces of higher cardinality. We define an *Ulam number* to be a cardinal number with the property of the theorem.

Definition 2.40: A cardinal number $\aleph$ is an *Ulam number* if whenever X is a set of cardinality $\aleph$ and μ is a finite measure defined on all subsets of X and such that $\mu(\{x\}) = 0$ for each $x \in X$ then μ is the zero measure.

Certainly, $\aleph_0$ is an Ulam number. We have seen in Theorem 2.39 that $\aleph_1$ is also an Ulam number. The class of all Ulam numbers forms a very large initial segment in the class of all cardinal numbers. It will take more set theory than we choose to develop to investigate this further,[1] but some have argued that one could consider safely that all cardinal numbers that one expects to encounter in analysis are Ulam numbers.

Exercises

2:11.1 Show that every set of real numbers that has positive Lebesgue outer measure contains a nonmeasurable set.

[1] See K. Ciesielski, "How good is Lebesgue measure?" *Math. Intelligencer* 11(2), 1989, pp. 54–58, for a discussion of material related to this section and for references to the literature. That same author's text, *Set Theory for the Working Mathematician*, Cambridge University Press, London (1997) is an excellent source for students wishing to go deeper into these ideas. In Section 12.6 we shall return to some related measure problems.

2:11.2 Show that there exist disjoint sets $\{E_k\}$ so that

$$\lambda^*\left(\bigcup_{k=1}^{\infty} E_k\right) < \sum_{k=1}^{\infty} \lambda^*\left(E_k\right).$$

2:11.3 Show that there exist sets $E_1 \supset E_2 \supset E_3 \ldots$ so that $\lambda^*(E_k) < +\infty$, for each k, and

$$\lambda^*\left(\bigcap_{k=1}^{\infty} E_k\right) < \lim_{k\to\infty} \lambda^*\left(E_k\right).$$

2:11.4 Let E be a measurable set of positive Lebesgue measure. Show that E can be written as the disjoint union of two sets $E = E_1 \cup E_2$ so that $\lambda(E) = \lambda^*(E_1) = \lambda^*(E_2)$.

2:11.5 Let H be a Hamel basis (see Exercise 1:11.3) and H_0 a nonempty finite or countable subset of H. Show that the set of rational linear combinations of elements of $H \setminus H_0$ is nonmeasurable.

2:11.6 Every totally imperfect set of real numbers contains no Cantor set but does contain an uncountable measurable set.

2:11.7 Exercise 2:11.6 suggests asking whether there can exist an uncountable set of real numbers that contains no uncountable measurable subset. Such a set (if it exists) is called a *Sierpiński set* and must clearly be nonmeasurable.

(a) Let X be a set of power $2^{\aleph_0}$ and let $\mathcal{E}$ be a family of subsets of X, also of power $2^{\aleph_0}$, with the property that X is the union of the family $\mathcal{E}$, but is not the union of any countable subfamily. Assuming CH, show that there is an uncountable subset of X that has at most countably many points in common with each member of $\mathcal{E}$.

(b) By applying (a) to the family of measure zero G_δ subsets of $\mathrm{I\!R}$, show that, assuming CH, there exists a Sierpiński set.

2:11.8 Let μ^* be an outer measure on a set X, and suppose that $E \subset X$ is not μ^*–measurable. Show that

$$\inf\{\mu^*(A \cap B) : \ A,\, B\ \mu^*\text{–measurable},\ A \supset E,\ B \supset X \setminus E\} > 0.$$

2:11.9 A cardinal number $\aleph$ is an Ulam number if and only if the following: if μ^* is an outer measure on a set X and $\mathcal{C}$ is a disjointed family of subsets of X with (i) card$(\mathcal{C}) \le \aleph$, (ii) the union of every subfamily of $\mathcal{C}$ is μ^*–measurable, (iii) $\mu^*(C) = 0$ for each $C \in \mathcal{C}$, and (iv) $\mu\left(\bigcup_{C\in\mathcal{C}} C\right) < \infty$, then

$$\mu\left(\bigcup_{C\in\mathcal{C}} C\right) = 0.$$

2:11.10 If $\mathcal{S}$ is a set of Ulam numbers and card$(\mathcal{S})$ is an Ulam number then the least upper bound of $\mathcal{S}$ is an Ulam number.

2:11.11 The successor of any Ulam number is an Ulam number. [Hint: See Federer, *Geometric Measure Theory*, Springer (1969), pp. 58–59, for a proof of these last three exercises.]

2.12 More About Method I

Let us review briefly our work to this point from the perspective of building a measure-theoretic framework for modeling some geometric or physical phenomena. In an attempt to satisfy our sense that “the whole should be the sum of its parts,” we created the structure of an algebra of sets $\mathcal{A}$ with an additive set function defined on $\mathcal{A}$. This structure had limitations—the algebra might be too small for our purposes. For example, the algebra generated by the half-open intervals on $(0, 1]$ consisted only of finite unions of such intervals (and $\emptyset$ of course). Even singletons are not in the algebra. The notion of countable additivity in place of additivity helped here—it gave rise to a σ-algebra of sets and a measure.

We then turned to the problem of how to obtain a measure space that could serve as a model for a given phenomenon for which we had a "primitive notion." We saw that we can always obtain a measure from an outer measure via the Carathéodory process and that Method I might be useful in obtaining an outer measure suitable for modeling our phenomenon. We say "might be useful" instead of "is useful" because there still are two unpleasant possibilities: our "primitive" sets T need not be measurable and, even if they are, it need not be true that

$$\tau(T) = \mu(T)$$

for all $T \in \mathcal{T}$. Such flaws might not be surprising insofar as we have placed only minimal requirements on τ and $\mathcal{T}$. What sorts of further restrictions will eliminate these two flaws?

Let us return to the family of half-open intervals on $(0, 1]$. Here we have an increasing function f defined on $[0, 1]$, and we obtain τ from f by

$$\tau((a, b]) = f(b) - f(a),$$

with τ extended to be additive on the algebra $\mathcal{T}$ generated by the half-open intervals. In this natural setting, we have some additional structure. The family $\mathcal{T}$ is an algebra of sets, and τ is additive on $\mathcal{T}$. This structure suffices to eliminate one of the unpleasant possibilities. Note that the proof is nearly identical to that for Corollary 2.38, but there, since the open intervals that were used for the covering family did not form an algebra, it was not so easy to carve up the sets.

2.12.1 Regularity for Method I outer measures

Theorem 2.41: *Let μ^* be constructed from a covering family $\mathcal{T}$ and a premeasure τ by Method I, and let $(X, \mathcal{M}, \mu)$ be the resulting measure space. If $\mathcal{T}$ is an algebra and τ is additive on $\mathcal{T}$, then $\mathcal{T} \subset \mathcal{M}$ and μ^* is regular.*

Proof. By Theorem 2.37, it is enough to check that each member of $\mathcal{T}$ is μ^*–measurable. Let $T \in \mathcal{T}$. To obtain that $T \in \mathcal{M}$, it suffices to show that, for each $E \subset X$ for which $\mu^*(E) < \infty$,

$$\mu^*(E) \geq \mu^*(E \cap T) + \mu^*(E \cap \widetilde{T}) \tag{23}$$

where we are using our usual notation for complementation, i.e., $\widetilde{T}$ denotes $X \setminus T$.

Let $\varepsilon > 0$. Choose a sequence $\{T_n\}$ from $\mathcal{T}$ such that

$$E \subset \bigcup_{n=1}^{\infty} T_n$$

and

$$\sum_{n=1}^{\infty} \tau(T_n) < \mu^*(E) + \varepsilon.$$

Since τ is additive on $\mathcal{T}$, we have, for all $n \in \mathbb{N}$

$$\tau(T_n) = \tau(T_n \cap T) + \tau(T_n \cap \widetilde{T}).$$

But

$$E \cap T \subset \bigcup_{n=1}^{\infty} (T_n \cap T) \quad \text{and} \quad E \cap \widetilde{T} \subset \bigcup_{n=1}^{\infty} (T_n \cap \widetilde{T}). \tag{24}$$

Thus

$$\begin{aligned}\mu^*(E)+\varepsilon &> \sum_{n=1}^{\infty}\tau(T_n)=\sum_{n=1}^{\infty}\tau(T_n\cap T)+\sum_{n=1}^{\infty}\tau(T_n\cap \widetilde{T})\\ &\geq \sum_{n=1}^{\infty}\mu^*(T_n\cap T)+\sum_{n=1}^{\infty}\mu^*(T_n\cap \widetilde{T})\\ &\geq \mu^*(E\cap T)+\mu^*(E\cap\widetilde{T}),\end{aligned}$$

the last inequality following from (24). Since ε is arbitrary, (23) follows. ■

Primitive notions like area, volume, and mass that are fundamentally additive might well lead to a τ, $\mathcal{T}$ combination that satisfies the hypotheses of Theorem 2.41.

2.12.2 The identity $\mu(T)=\tau(T)$ for Method I measures

We next ask whether the hypotheses of Theorem 2.41 remove the other flaw that we mentioned: $\tau(T)$ need not equal $\mu(T)$. To address this question, we look ahead.

A result of Section 12.6 enters our discussion. There is a finitely additive measure τ defined on *all* subsets of $[0,1]$ such that $\tau=\lambda$ on the class $\mathcal{L}$ of Lebesgue measurable sets. We mentioned this example in Section 2.11, where we proved too that, if μ is a finite measure on $2^{[0,1]}$ with $\mu(\{x\})=0$ for all $x\in[0,1]$, then $\mu(E)=0$ for all $E\subset[0,1]$.

Suppose now that we take $\mathcal{T}=2^{[0,1]}$ and τ the finitely additive extension of λ mentioned above and apply Method I to obtain μ^* and μ. Theorem 2.41 guarantees that all members of $\mathcal{T}$ are measurable. But this means that *every* subset of $[0,1]$ is measurable. From the material in Section 2.11 just mentioned, this implies that $\mu\equiv 0$. Since $\tau=\lambda$ on $\mathcal{L}$, τ and μ cannot agree on any set of positive Lebesgue measure. Thus, even though $\mathcal{T}$ and τ had enough structure to guarantee *all* subsets of $[0, 1]$ measurable, the measure μ did not retain anything of the primitive notion of length provided by τ!

Our development of Lebesgue measure on $[0, 1]$ actually provides a clue for removing the remaining flaw. Recall that in Section 2.1 we first extended the primitive notion of $\lambda(I)$, the length of an interval, to $\lambda(G)$, G open. This anticipated a form of σ-additivity. We then defined $\lambda(F)$, F closed. We can extend λ by additivity to the algebra $\mathcal{T}$ generated by the family of open sets (or, equivalently, by the family of closed sets). Taking $\tau = \lambda$ on $\mathcal{T}$, one can show that τ is σ-additive according to the following definition.

Definition 2.42: Let $\mathcal{A}$ be an algebra of sets, and let α be additive on $\mathcal{A}$. If

$$\alpha\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \alpha(A_n)$$

whenever $\{A_n\}$ is a sequence of pairwise disjoint sets from $\mathcal{A}$ for which

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A},$$

we say that α is *σ-additive* on $\mathcal{A}$.

Thus if $\alpha \geq 0$, it can fail to be a measure only when $\mathcal{A}$ is not a σ-algebra. It may well happen that when a concept is "fundamentally" additive, a τ, $\mathcal{T}$ combination can be found such that τ is σ-additive on $\mathcal{T}$. See Exercise 2:13.4.

Theorem 2.43: *Under the hypotheses of Theorem 2.41, if τ is σ-additive on $\mathcal{T}$, then $\mu(T) = \tau(T)$ for all $T \in \mathcal{T}$.*

Proof. We first show that, if $\{T_n\}$ is any sequence of sets in $\mathcal{T}$, $T \in \mathcal{T}$ and $T \subset \bigcup_{n=1}^\infty T_n$, then

$$\tau(T) \le \sum_{n=1}^\infty \tau(T_n). \tag{25}$$

Let $B_1 = T \cap T_1$ and, for $n \ge 2$, let

$$B_n = T \cap T_n \setminus (T_1 \cup \cdots \cup T_{n-1}).$$

Then, for all $n \in \mathbb{N}$, $B_n \subset T \cap T_n$, $B_n \in \mathcal{T}$, the sets B_n are pairwise disjoint, and $T = \bigcup_{n=1}^\infty B_n$. Since τ is σ-additive on $\mathcal{T}$,

$$\tau(T) = \sum_{n=1}^\infty \tau(B_n) \le \sum_{n=1}^\infty \tau(T_n).$$

This verifies (25). It now follows that

$$\tau(T) \le \inf\left\{ \sum_{n=1}^\infty \tau(T_n) : \bigcup_{n=1}^\infty T_n \supset T, T_n \in \mathcal{T} \right\} = \mu^*(T).$$

But since $\{T\}$ covers the set T, $\mu^*(T) \le \tau(T)$. Thus $\tau(T) = \mu^*(T)$. Since T is measurable by Theorem 2.41, $\mu^*(T) = \mu(T)$. ■

Exercises

2:12.1 Following the proof of Theorem 2.41, we gave an example of a τ, $\mathcal{T}$ combination, $\mathcal{T} = 2^{[0,1]}$ and $\tau = \lambda$ on $\mathcal{L}$, such that the μ resulting from Method I had little connection to length on $\mathcal{L}$. What would happen if we took the same τ but restricted τ to $\mathcal{T} = \mathcal{L}$?

2.13 Completions

Our presentation of Method I in Section 2.8 seemed simple and natural. It required little of τ and $\mathcal{T}$. But it had flaws that we removed in Section 2.12 by imposing additional additivity conditions on τ and $\mathcal{T}$. These conditions seemed natural because τ often represents a primitive notion of size that is intuitively additive. Exercise 2:13.4 provides a possible example of how we might naturally be led to use Theorems 2.41 and 2.43. On the other hand, these conditions seem to impose serious restrictions on the use of Method I. One might ask, what measure spaces $(X, \mathcal{M}, \mu)$ are the Method I result of a τ, $\mathcal{T}$ combination that satisfies such additivity conditions?

Such a space must be complete because any Method I measure is complete. We next show that the only other restriction on $(X, \mathcal{M}, \mu)$ is that X not be "too large."

Definition 2.44: Let $(X, \mathcal{M}, \mu)$ be a measure space. If $\mu(X) < \infty$, then we say that the measure space is *finite*. If $X = \bigcup_{n=1}^{\infty} X_n$ with $\mu(X_n) < \infty$ for all $n \in \mathbb{N}$, then we say that the space is *σ-finite.*

Theorem 2.45: *Let $(X, \mathcal{M}, \mu)$ be a σ-finite measure space. Let $\mathcal{T} = \mathcal{M}$ and $\tau = \mu$, and apply Method I to obtain an outer measure $\hat{\mu}^*$ and a measure space $(X, \widehat{\mathcal{M}}, \hat{\mu})$. Then*

1. *If $A \in \widehat{\mathcal{M}}$, then $A = M \cup Z$, where $M \in \mathcal{M}$ and $Z \subset N \in \mathcal{M}$ with $\mu(N) = 0$. Thus $(X, \widehat{\mathcal{M}}, \hat{\mu})$ is the completion of $(X, \mathcal{M}, \mu)$.*

2. *If μ is the restriction of a regular outer measure μ^* to its class of measurable sets, then $\hat{\mu}^* = \mu^*$.*

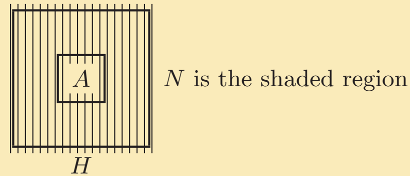


Figure 2.1. The set N is a measurable cover for $H \setminus A$.

Proof. To prove (i), assume first that $\mu(X) < \infty$. Let $A \in \widehat{\mathcal{M}}$. Now $\mathcal{M} \subset \widehat{\mathcal{M}}$ by Theorem 2.41. Thus $\hat{\mu}^*$ is regular by Theorem 2.37, so A has a $\hat{\mu}^*$-measurable cover H. Since $\mathcal{M}$ is a σ-algebra, Theorem 2.37 and Exercise 2:10.8 show that H can be taken in $\mathcal{M}$. Because $X \in \mathcal{M}$, our assumption that $\mu(X) < \infty$ implies that $\hat{\mu}^*(A) < \infty$. Since $\hat{\mu}^*$ is additive on $\widehat{\mathcal{M}}$,

$$\hat{\mu}^*(H \setminus A) = \hat{\mu}^*(H) - \hat{\mu}^*(A) = 0.$$

Now let N be a measurable cover in $\mathcal{M}$ for $H \setminus A$. See Figure 2.1.

By Theorem 2.43, $\hat{\mu}^*(N) = \mu(N)$, so $\mu(N) = \hat{\mu}^*(H \setminus A) = 0$. But

$$A = (H \setminus N) \cup (A \cap N).$$

To verify this, observe first that if $x \in A$, but $x \notin N$, then

$$x \in A \setminus N \subset H \setminus N.$$

In the other direction, since $N \supset H \setminus A$, any $x \in H \setminus N$ must be in A, and obviously $A \cap N \subset A$.

Now let $M = H \setminus N$, and let $Z = A \cap N$. Then $M \in \mathcal{M}$ and $Z \subset N$ with $\mu(N) = 0$. The equality $A = M \cup Z$ is the required one, and the proof of part (i) of the theorem is complete when $\mu(X) < \infty$. The proof when $\mu(X) = \infty$ is left as Exercise 2:13.1.

To prove (ii), let $A \subset X$. By hypothesis, μ comes from a regular outer measure μ^*. Thus there exists a measurable cover $M \in \mathcal{M}$ for A. By the definition of $\hat{\mu}^*$,

$$\hat{\mu}^*(A) \leq \mu(M) = \mu^*(A).$$

In the other direction, observe first that, since $\mathcal{M}$ is a σ-algebra,

$$\hat{\mu}^*(A) = \inf \{\mu(B) : A \subset B \in \mathcal{M}\}.$$

But if $A \subset B \in \mathcal{M}$, then $\mu^*(A) \leq \mu^*(B) = \mu(B)$, so

$$\mu^*(A) \leq \inf \{\mu(B) : A \subset B \in \mathcal{M}\}.$$

Therefore, $\hat{\mu}^*(A) = \mu^*(A)$. ■

Corollary 2.46: *Every complete σ-finite measure space $(X, \mathcal{M}, \mu)$ is its own Method I Carathéodory extension. That is, an application of Method I to $\mathcal{T} = \mathcal{M}$ and $\tau = \mu$ results in the space $(X, \mathcal{M}, \mu)$.*

Proof. Observe that the completion of a complete measure space is the space itself and apply part (i) of Theorem 2.45. ■

The hypotheses of Theorem 2.45 and Corollary 2.46 cannot be dropped. See Exercises 2:13.2 and 2:13.3.

Exercises

2:13.1 Prove part (i) of Theorem 2.45 when $\mu(X) = \infty$.

2:13.2 Let $X = \mathbb{R}$, $\mathcal{M} = \{A : A \text{ is countable or } \mathbb{R} \setminus A \text{ is countable}\}$, and define

$$\mu(A) = \begin{cases} \text{cardinality } A, & A \text{ is finite;} \\ \infty, & A \text{ is infinite.} \end{cases}$$

(a) Show that μ is a complete measure on $\mathcal{M}$.

(b) Show that $\hat{\mu}$ (See Theorem 2.45) is not the completion of μ.

(c) Show that μ is not the restriction to its measurable sets of any outer measure.

(d) Reconcile these with Theorem 2.45 and Corollary 2.46.

2:13.3 Let $(X, \mathcal{M}, \mu)$ be as in Example 2.29. Apply the process of Theorem 2.45 and determine whether $\hat{\mu}^* = \mu^*$.

2:13.4$\diamondsuit$ Suppose that we have a mass distribution on the half-open square $S = (0, 1] \times (0, 1]$ in $\mathbb{R}^2$, and we know how to compute the mass in any half-open "interval" $(a, b] \times (c, d]$. Suppose that singleton sets have zero mass. We wish to obtain a measure space $(X, \mathcal{M}, \mu)$ to model this distribution *based only on the ideas we have developed so far.*

First try: Take $\mathcal{T}$ as the half-open intervals in S, together with $\emptyset$, and let $\tau(T)$ be the mass of T for $T \in \mathcal{T}$. Apply Method I to get μ^* and then $(X, \mathcal{M}, \mu)$.

(a) Can we be sure that $\mathcal{M}$ is a σ-algebra and μ is a measure on $\mathcal{M}$? Can we be sure that $\mathcal{T} \subset \mathcal{M}$? If $T \in \mathcal{M}$, must $\mu(T) = \tau(T)$?

Second try: We note that τ is intuitively additive. So let $\mathcal{T}_1$ be the algebra generated by $\mathcal{T}$, and extend τ to τ_1 so that τ_1 is additive on $\mathcal{T}_1$.

(b) Can we do this? That is, can we be sure that $\tau_1(T_1)$, $T_1 \in \mathcal{T}_1$, does not depend on the decomposition of T_1 into a union of members of $\mathcal{T}$? If so, what are the answers to the questions posed in part (a) when we apply Method I to $\mathcal{T}_1$ and τ_1?

Third try: We believe mass is fundamentally σ-additive. But $\mathcal{T}_1$ is only an algebra. So we verify that τ_1 is σ-additive on $\mathcal{T}_1$. Can we now answer the three questions in part (a) affirmatively?

2.14 Additional Problems for Chapter 2

2:14.1 Criticize the following "argument" which is far too often seen:

> "If $G = (a,b)$ then $\overline{G} = [a,b]$. Similarly, if $G = \bigcup_{i=1}^{\infty}(a_i,b_i)$ is an open set, then $\overline{G} = \bigcup_{i=1}^{\infty}[a_i,b_i]$ so that G and $\overline{G}$ differ by a countable set. Since every countable set has Lebesgue measure zero, it follows that an open set G and its closure $\overline{G}$ have the same Lebesgue measure."(?)

2:14.2 Let A be a set of real numbers of Lebesgue measure zero. Show that the set $\{x^2 : x \in A\}$ also has measure zero.

2:14.3 Let A be the set of real numbers in the interval $(0,1)$ that have a decimal expansion that contains the number 3. Show that A is a Borel set and find its Lebesgue measure.

2:14.4 Let E be a Lebesgue measurable subset of $[0,1]$, and define

$$B = \{x \in [0,1] : \lambda(E \cap (x-\varepsilon, x+\varepsilon)) > 0 \ \text{ for all } \varepsilon > 0\}.$$

Show that B is perfect.

2:14.5 Let E be a Lebesgue measurable subset of $[0,1]$ and let $c > 0$. If $\lambda(E \cap I) \geq c\lambda(I)$ for all open intervals $I \subset [0,1]$, show that $\lambda(E) = 1$.

2:14.6 Let A_n be a sequence of Lebesgue measurable subsets of $[0,1]$ and suppose that $\limsup_{n\to\infty} \lambda(A_n) = 1$. Show that there is some subsequence with

$$\lambda\left(\bigcap_{k=1}^{\infty} A_{n_k}\right) > 0.$$

[Hint: Arrange for $\sum_{k=1}^{\infty} (1-\lambda(A_{n_k})) < 1$.]

2:14.7♢ Let $(X, \mathcal{M}, \mu)$ be a measure space. A set $A \in \mathcal{M}$ is called an *atom*, if $\mu(A) > 0$ and, for all measurable sets $B \subset A$, $\mu(B) = 0$ or $\mu(A \setminus B) = 0$. The measure space is *nonatomic* if there are no atoms.

(a) For any $x \in X$, if $\{x\} \in \mathcal{M}$ and $\mu(\{x\}) > 0$, then $\{x\}$ is an atom.

(b) Determine all atoms for the counting measure. (The counting measure is defined in Exercise 2:3.9.)

(c) Show that if $A \in \mathcal{M}$ is an atom then every subset $B \subset A$ with $B \in \mathcal{M}$ and $\mu(B) > 0$ is also an atom.

(d) Show that if A_1, $A_2 \in \mathcal{M}$ are atoms then, up to a set of μ–measure zero, either A_1 and A_2 are equal or disjoint.

(e) Suppose that μ is σ-finite. Show that there is a set $X_0 \subset X$ such that X_0 is a disjoint union of countably many atoms of $(X, \mathcal{M}, \mu)$ and $X \setminus X_0$ contains no atoms.

(f) Show that the Lebesgue measure space is nonatomic.

(g) Give an example of a nontrivial measure space $(X, \mathcal{M}, \mu)$ with $\mu(\{x\}) = 0$ for all $x \in X$ and so that every set of positive measure is an atom. [Hint: Construct a measure using Exercise 2:2.5.]

2:14.8♢ (Liaponoff's theorem) Let $\mu_1, \dots, \mu_n$ be nonatomic measures on $(X, \mathcal{M})$, with $\mu_i(X) = 1$ for all $i = 1, \dots, n$. These measures can be viewed as giving rise to a *vector measure*

$$\mu : \mathcal{M} \to [0,1]^n = [0,1] \times [0,1] \times \cdots [0,1]$$

on $(X, \mathcal{M})$ defined by

$$\mu(A) = (\mu_1(A), \dots, \mu_n(A))$$

for each $A \in \mathcal{M}$. A theorem of Liaponoff (1940) states that

The set S of n-tuples $(x_1, \dots, x_n)$ for which there exists $A \in \mathcal{M}$ such that $\mu(A) = (x_1, \dots, x_n)$ is a convex subset of $[0,1]^n$.

(a) Let $(X, \mathcal{M}, \mu)$ be a nonatomic measure space with $\mu(X) = 1$. Show that for each $\gamma \in [0,1]$ there is a set $E_\gamma \subset X$ such that $\mu(E_\gamma) = \gamma$. [Hint: Use some form of Zorn's lemma (Section 1.11) or transfinite induction.]

(b) Show that part (a) follows from Liaponoff's theorem.

(c) Show that $(1/n, 1/n, \dots, 1/n) \in S$. You may assume the validity of Liaponoff's Theorem.

(d) Interpret part (c) to obtain the following result, indicating the technical meanings of the terms in quotation marks.

Given a cake with n ingredients (e.g., butter, sugar, chocolate, garlic, etc.), each nonatomic and of unit mass and mixed together in any "reasonable" way, it is possible to "cut the cake into n pieces" such that each of the pieces contains its "share" of each of the ingredients.

2:14.9◇ Show that there exists a set $E \subset [0,1]$ such that, for every open interval $I \subset [0,1]$, $\lambda(I \cap E) > 0$ and $\lambda(I \setminus E) > 0$.

2:14.10 Let $\{E_n\}$ be a sequence of measurable sets in a measure space $(X, \mathcal{M}, \mu)$ with each $0 < \mu(E_n) < \infty$. When is it generally possible to select a set $A \in \mathcal{M}$ with each $\mu(A \cap E_n) > 0$ and each $\mu(E_n \setminus A) > 0$?

2:14.11 Let K be the Cantor set. Each point $x \in K$ has a unique ternary expansion of the form

$$x = .a_1 a_2 a_3 \dots \qquad (a_i = 0 \text{ or } a_i = 2, \quad i \in \mathbb{N}).$$

Let $b_i = a_i/2$ and let $f(x) = .b_1 b_2 b_3 \dots$, interpreted in base 2. For example, if $x = \frac{2}{9} = 0.0200\dots$ (base 3), then we would have $f(x) = \frac{1}{4} = 0.0100\dots$ (base 2). Show that if f is extended to be linear and continuous on the closure of each interval complementary to K, then the the extended function $\overline{f}$ is continuous on $[0,1]$. Determine the relationship of this function $\overline{f}$ to the Cantor function (Exercise 1:22.13).

2:14.12 Let $X=[0,1]$ and let $\tau=\lambda^*$. In each case apply Method I to the family $\mathcal{T}$ and determine μ^* and $\mathcal{M}$. How do things change if $\tau=\lambda_*$ in part (f)?

(a) $\mathcal{T}$ consists of $\emptyset$ and $[0,1]$.

(b) $\mathcal{T}$ consists of $\emptyset$ and the family of all open subintervals.

(c) $\mathcal{T}$ consists of $\emptyset$ and all nondegenerate subintervals.

(d) $\mathcal{T}$ is $\mathcal{B}$.

(e) $\mathcal{T}$ is $\mathcal{L}$.

(f) $\mathcal{T}$ is 2^X.

[Hint for (f): The nonmeasurable set A discussed in Section 1.10 has $\lambda_*(A)=0$.]

2:14.13$\diamondsuit$ Show that every set $E\subset\mathbb{R}$ with $\lambda^*(E)>0$ contains a set that is nonmeasurable. [Hint: Let $E\subset[-\frac{1}{2},\frac{1}{2}]$, and let $E_k=E\cap A_k$, where $\{A_k\}$ is the family of sets appearing in our proof in Section 1.10 of the existence of sets in $\mathbb{R}$ that are not Lebesgue measurable.]

2:14.14 Suppose that μ^* is the outer measure on X obtained by Method I from $\mathcal{T}$ and τ, and suppose that μ_1^* is any other outer measure on X satisfying $\mu_1^*(T)\le\tau(T)$ for all $T\in\mathcal{T}$. Prove that $\mu_1^*\le\mu^*$. Give an example for which $\mu_1^*(T)=\tau(T)$ for all $T\in\mathcal{T}$ and $\mu_1^*\ne\mu^*$. [Hint: Let $\mathcal{T}=\{\emptyset,[0,1]\}$ and $\mu_1^*=\lambda^*$.]

2:14.15$\diamondsuit$ Let $\mathcal{T}$ be a covering family, and let τ_1 and τ_2 be nonnegative functions on $\mathcal{T}$. Let μ_1^* and μ_2^* be the associated Method I outer measures. Prove that if $\mu_1^*(T)=\mu_2^*(T)$ for all $T\in\mathcal{T}$ then $\mu_1^*=\mu_2^*$.

2:14.16 Let $(X, \mathcal{M}, \mu)$ be a measure space with $\mu(X) = 1$, and suppose that $\mu(M) > 0$ for each nonempty $M \in \mathcal{M}$. For each $x \in X$, let

$$\alpha(x) = \inf \{\mu(E) : E \in \mathcal{M}, x \in E\}.$$

(a) Show that there is a set $A_x \in \mathcal{M}$ such that $x \in A_x$ and $\mu(A_x) = \alpha(x)$.

(b) Prove that the sets $\{A_x\}$ are either disjoint or identical.

Chapter 3

METRIC OUTER MEASURES

In Chapter 2 we studied the basic abstract structure of a measure space. The only ingredients are a set X, a σ–algebra of subsets of X, and a measure defined on the σ–algebra. In almost all cases the set X will have some other structure that is of interest. Our example of Lebesgue measure on the real line illustrates this well. While $(\mathbb{R}, \mathcal{L}, \lambda)$ is a measure space, we should remember that $\mathbb{R}$ also has a great deal of other structure and that this measure space is influenced by that other structure. For instance $\mathbb{R}$ is linearly ordered, is a metric space, and also has a number of algebraic structures. Lebesgue measure, naturally, interacts with each of these.

In this chapter we study measures in a general metric space. As it happens, the only measures that are of any genuine interest are those that interact with the metric structure in a consistent way. In Section 3.2 we introduce the concepts of metric outer measure and Borel measure, which capture this interaction in the most convenient and useful way. In Section 3.3 we give an extension of the Method I construction that allows us to obtain metric outer measures. Section 3.4 explores how the measure of sets in a metric space can be approximated by the measure of less complicated sets, notably open sets or closed sets or simple Borel sets. The remaining sections develop some applications of the theory to important special measures, the Lebesgue–Stieltjes measures on the real line and Lebesgue–Stieltjes measures and Hausdorff measures in $\mathbb{R}^n$.

We begin with a brief review of metric space theory. In this chapter, only the most rudimentary properties of a metric space need be used. Even so the reader will feel more comfortable in the ensuing discussion after obtaining some familiarity with the concepts. A full treatment of metric spaces begins in Chapter 9. Some readers may prefer to gain some expertise in that general theory before studying measures on metric spaces. Abstract theories, such as metric spaces, allow for deep and subtle generalizations. But one can also view them as simplifications in that they permit one to focus on essentials of the structure.

3.1 Metric Space

Sequence limits in $\mathbb{R}$ are defined using the metric

$$\rho(x,y) = |x-y| \quad (x,y \in \mathbb{R})$$

which describes distances between pairs of points in $\mathbb{R}$. In higher dimensions one develops a similar theory, but using for distance the familiar expression

$$\rho(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad (x,y \in \mathbb{R}^n).$$

The only properties of these distance functions that are needed to develop an adequate theory in an abstract setting are those we have listed in Section 1.1. We can take these as forming our definition.

Definition 3.1: Let X be a set and let $\rho : X \times X \to \mathbb{R}$. If ρ satisfies the following conditions, then we say ρ is a *metric* on X and call the pair (X, ρ) a *metric space*.

1. $\rho(x, y) \geq 0$ for all $x, y \in X$.
2. $\rho(x, y) = 0$ if and only if $x = y$.
3. $\rho(x, y) = \rho(y, x)$ for all $x, y \in X$.
4. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for all $x, y, z \in X$ (triangle inequality).

A metric space is a pair (X, ρ), where X is a set equipped with a metric ρ; in many cases one simply says that X is a metric space when the context makes it clear what metric is to be used. Sequence convergence in a metric space (X, ρ) means convergence relative to this distance. Thus $x_n \to x$ means that $\rho(x_n, x) \to 0$. The role that intervals on the real line play is assumed in an abstract metric space by the analogous notion of *an open ball*; that is, a set of the form

$$B(x_0, \varepsilon) = \{x : \rho(x, x_0) < \varepsilon\},$$

which can be thought of as the interior of a sphere centered at x_0 and with radius ε; avoid, however, too much geometric intuition, since "spheres" are not "round" and do not have the kind of closure properties that one may expect.

3.1.1 Metric space terminology

The language of metric space theory is just an extension of that for real numbers. Throughout (X, ρ) is a fixed metric space. For this chapter we need to understand the notions of diameter, open sets, and closed sets.

- For $x_0 \in X$ and $r > 0$, the set
$$B(x_0, r) = \{x \in X : \rho(x_0, x) < r\}$$
is called the *open ball* with center x_0 and radius r.
- For $x_0 \in X$ and $r > 0$, the set
$$B[x_0, r] = \{x \in X : \rho(x_0, x) \leq r\}$$
is called the *closed ball* with center x_0 and radius r.
- A set $G \subset X$ is called *open* if for each $x_0 \in G$ there exists $r > 0$ such that $B(x_0, r) \subset G$.
- A set F is called *closed* if its complement $\widetilde{F} = X \setminus F$ is open.
- A set is *bounded* if it is contained in some open ball.
- A *neighborhood* of x_0 is any open set G containing x_0.
- If $G = B(x_0, \varepsilon)$, we call G the *ε-neighborhood* of x_0.
- The point x_0 is called an *interior point* of a set A if x_0 has a neighborhood contained in A.

- The *interior* of A consists of all interior points of A and is denoted by A^o or, occasionally, $\mathrm{int}(A)$. It is the largest open set contained in A; it might be empty.

- A point $x_0 \in X$ is a *limit point* or *point of accumulation* of a set A if every neighborhood of x_0 contains points of A distinct from x_0.

- The *closure*, $\overline{A}$, of a set A consists of all points that are either in A or limit points of A. (It is the smallest closed set containing A.) One verifies easily that $x_0 \in \overline{A}$ if and only if there exists a sequence $\{x_n\}$ of points in A such that $x_n \to x_0$.

- A *boundary point* of A is a point x_0 such that every neighborhood of x_0 contains points of A as well as points of $\widetilde{A} = X \setminus A$.

- The *diameter* of a set $E \subset X$ is defined as
$$\mathrm{diameter}\,(E) = \sup\{\rho(x,y) : x, y \in E\}.$$
[We shall take $\mathrm{diameter}\,(\emptyset) = 0$].

- An *isolated point* of a set is a member of the set that is not a limit point of the set.

- A set is *perfect* if it is nonempty, closed, and has no isolated points.

- A set $E \subset X$ is *dense* in a set $E_0 \subset X$ if every point in E_0 is a limit point of the set E.

- The *distance* between a point $x \in X$ and a nonempty set $A \subset X$ is defined as
$$\mathrm{dist}(x, A) = \inf\{\rho(x,y) : y \in A\}.$$

- The *distance* between two nonempty sets A, $B \subset X$ is defined as

$$\operatorname{dist}(A, B) = \inf\{\rho(x, y) : x \in A,\ y \in B\}.$$

- Two nonempty sets A, $B \subset X$ are said to be *separated* if they are a positive distance apart [i.e., if $\operatorname{dist}(A, B) > 0$].

The last three of these notions play an important role in the discussion in Section 3.2, where they are discussed in more detail. Here we should note that "dist" is not itself a metric on the subsets of X since the second condition of Definition 3.1 is violated if $A \cap B \neq \emptyset$ but $A \neq B$.

3.1.2 Borel sets in a metric space

The *Borel sets* in a metric space are defined in the same manner as on the real line and have much the same properties. We shall use the following formal definition.

Definition 3.2: Let (X, ρ) be a metric space. The family of Borel subsets of (X, ρ) is the smallest σ–algebra that contains all the open sets in X.

It is convenient to have other expressions for the Borel sets. The family of Borel sets can be seen to be the smallest σ–algebra that contains all the closed sets in X. But for some applications we shall need the following characterization.

Theorem 3.3: *The family of Borel subsets of a metric space (X, ρ) is the smallest class $\mathcal{B}$ of subsets of X with the properties*

1. *If $E_1, E_2, E_3, \dots$ belong to $\mathcal{B}$, then so too does $\bigcup_{i=1}^{\infty} E_i$.*

2. *If $E_1, E_2, E_3, \dots$ belong to $\mathcal{B}$, then so too does $\bigcap_{i=1}^{\infty} E_i$.*

3. *$\mathcal{B}$ contains all the closed sets in X.*

We can also introduce the transfinite sequence of the Borel hierarchy

$$\mathcal{G} \subset \mathcal{G}_\delta \subset \mathcal{G}_{\delta\sigma} \subset \mathcal{G}_{\delta\sigma\delta} \subset \mathcal{G}_{\delta\sigma\delta\sigma} \dots$$

and

$$\mathcal{F} \subset \mathcal{F}_\sigma \subset \mathcal{F}_{\sigma\delta} \subset \mathcal{F}_{\sigma\delta\sigma} \subset \mathcal{F}_{\sigma\delta\sigma\delta} \dots,$$

just as we did in Section 1.12. Of these, we would normally not go beyond the second stage or perhaps the third stage in any of our applications.

3.1.3 Characterizations of the Borel sets

It is useful to describe the class of Borel sets in a narrower manner than that of Theorem 3.3. For easy reference we include a proof of this variant.[1]

[1] This proof has been supplied to us by R. B. Burckel.

Theorem 3.4: *The family of Borel subsets of a metric space (X, ρ) is the smallest class $\mathcal{B}$ of subsets of X with the properties:*

1. *If $E_1, E_2, E_3, \ldots$ belong to $\mathcal{B}$, and are pairwise disjoint then the union $\bigcup_{i=1}^{\infty} E_i$ also belongs to $\mathcal{B}$.*
2. *If $E_1, E_2, E_3, \ldots$ belong to $\mathcal{B}$, then the intersection $\bigcap_{i=1}^{\infty} E_i$ also belongs to $\mathcal{B}$.*
3. *$\mathcal{B}$ contains all the closed sets in X.*

Proof. It is clear that the Borel sets form a family with these properties. Thus to prove that this is a characterization we show that any family $\mathcal{B}$ with these three properties must contain all the Borel sets. We first show that every open set U in X is a member of $\mathcal{B}$. The sets

$$V = \{x \in X : 0 < \operatorname{dist}(x, X \setminus U) < 1\}$$

and

$$F = \{x \in X : \operatorname{dist}(x, X \setminus U) \geq 1\}$$

satisfy $U = V \cup F$, they are disjoint, and $F \in \mathcal{B}$ since it is closed. Thus, to prove that $U \in \mathcal{B}$, it is sufficient (because of (i) and (iii)) to prove that $V \in \mathcal{B}$.

Consider the function $f : X \to \mathbb{R}$ defined by

$$f(x) = \operatorname{dist}(x, X \setminus V).$$

This is continuous and $f^{-1}((0,1)) = V$.

We observe that the open interval $(0,1)$ in $\mathbb{R}$ can be expressed as a union and intersection of compact subsets in the following manner:

$$(0,1) = \bigcup_{m=1}^{\infty} K_m \cup \left(\bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} C_{nk} \right)$$

where K_m and C_{nk} are compact subsets of $(0,1)$ and all unions in the identity are disjoint ones. (See Exercise 3:1.1.)

Consequently

$$V = f^{-1}((0,1)) = \bigcup_{m=1}^{\infty} f^{-1}(K_m) \cup \left(\bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} f^{-1}(C_{nk}) \right)$$

expresses V in a way that allows us to see that it is a member of $\mathcal{B}$. Here we are using property (i) for the disjoint unions, property (ii) for the intersections, the continuity of f to ensure that all the sets $f^{-1}(K_m)$ and $f^{-1}(C_{nk})$ are closed, and finally property (iii) to ensure that all these sets are members of $\mathcal{B}$. Hence $\mathcal{B}$ contains V, and hence also U.

We now have a family of sets containing all open sets, all closed sets and closed under the operations (i) and (ii). Exercise 3:1.5 can be used to complete the proof, showing that $\mathcal{B}$ contains all Borel sets. ■

Exercises

3:1.1◇ Show that the open interval $(0,1)$ in $\mathbb{R}$ can be expressed as

$$(0,1)=\bigcup_{m=1}^{\infty}K_m\cup\left(\bigcap_{k=1}^{\infty}\bigcup_{n=1}^{\infty}C_{nk}\right)$$

where K_m and C_{nk} are compact subsets of $(0,1)$ and all unions in the identity are disjoint ones. [Hint: Let C be the Cantor ternary set in $[0,1]$ and let $\{I_n\}$ be the open components of the set $(0,1)\setminus C$. Use first $K_m=\overline{I_m}$. Check that these are disjoint and that

$$(0,1)\setminus\bigcup_{m=1}^{\infty}K_m=C\setminus\{c_1,c_2,c_3,\dots\}$$

where $\{c_i\}$ is a list of all the "endpoints" of the Cantor set. Fix k and describe how to construct a disjoint collection of closed subsets $\{C_{nk}\}$ of the set C so that

$$C\setminus\{c_k\}=\bigcup_{n=1}^{\infty}C_{nk}.$$

Finally verify the required identity.]

3:1.2 Prove that in a metric space every closed set is a $\mathcal{G}_\delta$.

3:1.3 Prove that in a metric space every open set is an $\mathcal{F}_\sigma$.

3:1.4 Prove Theorem 3.3.

3:1.5◇ Let $\mathcal{C}$ be a class of subsets of a metric space (X,ρ) with the following properties:

(a) If $E_1,E_2,E_3,\dots$ are disjoint and belong to $\mathcal{C}$, then so too does $\bigcup_{i=1}^{\infty}E_i$.

(b) If $E_1,E_2,E_3,\dots$ belong to $\mathcal{C}$, then so too does $\bigcap_{i=1}^{\infty}E_i$.

(c) $\mathcal{C}$ contains all the open sets in X.

Prove that $\mathcal{C}$ contains all Borel subsets of X. [Hint: Note that $\mathcal{C}$ need not itself be closed under complementation. But that should suggest a look at the family

$$\mathcal{C}_0 = \{C : C \in \mathcal{C} \text{ and } X \setminus C \in \mathcal{C}\}.$$

What properties does $\mathcal{C}_0$ have?]

3:1.6 A metric space (X, d) is said to be *separable* if there exists a countable subset of X that is dense in X. In a separable metric space, show that there are no more than $2^{\aleph_0}$ open sets and $2^{\aleph_0}$ closed sets.

3:1.7 In a separable metric space, show that there are no more than $2^{\aleph_0}$ Borel sets. [Hint: Use transfinite induction, the ideas of Section 1.12, and Exercise 3:1.6.]

3.2 Measures on Metric Spaces

We begin our discussion with an example of a Method I construction that produces a measure badly incompatible with the metric structure of $\mathbb{R}^2$. We use this to draw a number of conclusions. It will give us an insight into the conditions that we might wish to impose on measures defined on a metric space. It also gives us an important clue as to how Method I should be improved to recognize the metric structure.

Example 3.5: Take $X = \mathbb{R}^2$, let $\mathcal{T}$ be the family of open squares in X, and choose as a premeasure $\tau(T)$ to be the diameter of T. We apply Method I to obtain an outer measure μ^* and then a measure space $(\mathbb{R}^2, \mathcal{M}, \mu)$. What would we expect about the measurability of sets in $\mathcal{T}$? Since diameter is essentially a one-dimensional concept, while $\mathcal{T}$ consists of two-dimensional sets, perhaps we expect that every nonempty T has infinite measure.

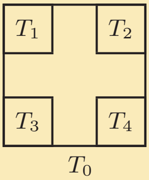


Figure 3.1. The square T_0.

Let $T_0 \in \mathcal{T}$ have side length 3, and let T_1, T_2, T_3 and T_4 be in $\mathcal{T}$, each with side length 1, and as shown in Figure 3.1. Then $\tau(T_0) = 3\sqrt{2}$, while $\tau(T_i) = \sqrt{2}$ for $i = 1, 2, 3, 4$. It is easy to verify that, for all $T \in \mathcal{T}$, $\mu^*(T) = \tau(T)$ and that

$$\mu^*\left(\bigcup_{i=1}^{4} T_i\right) \le \mu^*(T_0) = 3\sqrt{2} < 4\sqrt{2} = \sum_{i=1}^{4} \mu^*(T_i).$$

It follows that none of the sets T_i, $i = 1, 2, 3, 4$, is measurable. A moment's reflection shows that no nonempty member of $\mathcal{T}$ can be measurable.

We note two significant features of this example.

1. The squares T_i are not only pairwise disjoint, but they are also separated from each other by positive distances: if $x \in T_i$, $y \in T_j$, and $i \neq j$, then the distance between x and y exceeds 1. As we saw, μ^* is not additive on these sets. Now we know outer measures are not additive in general, but for Lebesgue outer measure, if $\mu^*(A \cup B) \neq \mu^*(A) + \mu^*(B)$ and $A \cap B = \emptyset$, then the sets A and B are badly intertwined, not separated.

2. The class $\mathcal{M}$ of measurable sets is incompatible with the topology on $\mathbb{R}^2$: open sets need not be measurable.

Indeed, these two features, we shall soon discover, are intimately linked. If we wish open sets to be measurable, we must have an outer measure which is additive on separated sets, and conversely. We take the latter requirement as our definition of a metric outer measure.

3.2.1 Metric Outer Measures

Recall that in a metric space we use

$$\operatorname{dist}(A,B) = \inf\{\rho(x,y) : x \in A \text{ and } y \in B\}$$

as a measure of the distance between two sets A and B. When $A = \{x\}$, we write $\operatorname{dist}(x,B)$ in place of $\operatorname{dist}(\{x\},B)$. Although we call $\operatorname{dist}(A,B)$ the distance between A and B, dist is not a metric on the subsets of X. Recall, too, that if $\operatorname{dist}(A,B) > 0$, then we say that A and B are *separated* sets. For example, the sets T_i appearing in Example 3.5 are pairwise separated; indeed, $\operatorname{dist}(T_i,T_j) \geq 1$ if $i \neq j$.

Definition 3.6: Let μ^* be an outer measure on a metric space X. If

$$\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$$

whenever A and B are separated subsets of X, then μ^* is called a *metric outer measure*.

Thus metric outer measures are designed to avoid the unpleasant possibility (i) that we observed for the Method I outer measure μ^* in our example. In Theorem 3.8 we show that the second unpleasant possibility of our example cannot occur: Borel sets will always be measurable for metric outer measures.

3.2.2 Measurability of Borel sets

The first step in proving that Borel sets are measurable with respect to any outer measure is supplied by the following lemma, due to Carathéodory.

Lemma 3.7: *Let μ^* be a metric outer measure on X. Let G be a proper open subset of X, let $F = X \setminus G$ be its complement in X and let $A \subset G$. Let*

$$A_n = \{x \in A : \operatorname{dist}(x, F) \geq 1/n\}.$$

Then

$$\mu^*(A) = \lim_{n\to\infty} \mu^*(A_n).$$

Proof. Recall that F denotes the set complementary to G, which in this case must be closed since G is open. The existence of the limit follows from the monotonicity of μ^* and the fact that $\{A_n\}$ is an expanding sequence of sets. Since $A_n \subset A$ for all $n \in \mathbb{N}$, $\mu^*(A) \geq \lim_{n\to\infty} \mu^*(A_n)$. It remains to verify that

$$\mu^*(A) \leq \lim_{n\to\infty} \mu^*(A_n).$$

Since G is open, $\operatorname{dist}(x, F) > 0$ for all $x \in A$, so there exists $n \in \mathbb{N}$ such that $x \in A_n$. It follows that $A = \bigcup_{n=1}^{\infty} A_n$.

For each n, let

$$B_n = A_{n+1} \setminus A_n = \left\{x : \frac{1}{n+1} \leq \operatorname{dist}(x, F) < \frac{1}{n}\right\}.$$

Then

$$A = A_{2n} \cup \bigcup_{k=2n}^{\infty} B_k = A_{2n} \cup \bigcup_{k=n}^{\infty} B_{2k} \cup \bigcup_{k=n}^{\infty} B_{2k+1}.$$

Thus

$$\mu^*(A) \leq \mu^*(A_{2n}) + \sum_{k=n}^{\infty} \mu^*(B_{2k}) + \sum_{k=n}^{\infty} \mu^*(B_{2k+1}).$$

If the series are convergent, then

$$\mu^*(A) \leq \lim_{n\to\infty} \mu^*(A_{2n}) = \lim_{n\to\infty} \mu^*(A_n),$$

as was to be proved.

The argument to this point is valid for any outer measure. We now invoke our hypothesis that μ^* is a metric outer measure. Suppose that one of the series diverges, say

$$\sum_{k=1}^{\infty} \mu^*(B_{2k}) = \infty. \tag{1}$$

It follows from the definition of the sets B_k that, for each $k \in \mathbb{N}$,

$$\operatorname{dist}(B_{2k}, B_{2k+2}) \geq \frac{1}{2k+1} - \frac{1}{2k+2} > 0,$$

so these sets are separated. Thus

$$\mu^*\left(\bigcup_{k=1}^{n-1} B_{2k}\right) = \sum_{k=1}^{n-1} \mu^*(B_{2k}). \tag{2}$$

But $A_{2n} \supset \bigcup_{k=1}^{n-1} B_{2k}$, so

$$\mu^*(A_{2n}) \geq \mu^*\left(\bigcup_{k=1}^{n-1} B_{2k}\right). \tag{3}$$

Combining (2) and (3), we see that

$$\mu^*(A_{2n}) \geq \sum_{k=1}^{n-1} \mu^*(B_{2k}).$$

It follows from our assumption (1) that $\lim_{n\to\infty} \mu^*(A_{2n}) = \infty$, so

$$\lim_{n\to\infty} \mu^*(A_n) \geq \mu^*(A).$$

Finally, if it is the series $\sum_{k=1}^{\infty} \mu^*(B_{2k+1})$ that diverges, the argument is similar. We omit the details. ■

Theorem 3.8: *Let μ^* be an outer measure on a metric space X. Then every Borel set in X is measurable if and only if μ^* is a metric outer measure.*

Proof. Assume first that μ^* is a metric outer measure. Since the class of Borel sets is the σ-algebra generated by the closed sets, it suffices to verify that every closed set is measurable. Let F be a nonempty closed set and let $G = X \setminus F$. Then G is open. We show that F satisfies the measurability condition of Definition 2.30. Let $E \subset X$, let $A = E \setminus F$, and let $\{A_n\}$ be the sequence of sets appearing in Lemma 3.7. Then $\operatorname{dist}(A_n, F) \geq 1/n$ for all $n \in \mathbb{N}$, and

$$\lim_{n\to\infty} \mu^*(A_n) = \mu^*(E \setminus F). \tag{4}$$

Since μ^* is a metric outer measure and the sets A_n are separated from F, we have, for each $n \in \mathbb{N}$,

$$\mu^*(E) \geq \mu^*((E \cap F) \cup A_n) = \mu^*(E \cap F) + \mu^*(A_n).$$

From (4) we see that

$$\mu^*(E) \geq \mu^*(E \cap F) + \mu^*(E \setminus F).$$

The reverse inequality is obvious. Thus F is measurable.

To prove the converse, assume that all Borel sets are measurable. Let A_1 and A_2 be separated sets, say $\operatorname{dist}(A_1, A_2) = \gamma > 0$. For each $x \in A_1$, let

$$G(x) = \{z : \rho(x, z) < \gamma/2\},$$

and let

$$G = \bigcup_{x \in A_1} G(x).$$

Then G is open, $A_1 \subset G$, and $G \cap A_2 = \emptyset$. Since G is measurable, it satisfies the measurability condition of Definition 2.30 for the set $E = A_1 \cup A_2$; that is,

$$\mu^*(A_1 \cup A_2) = \mu^*((A_1 \cup A_2) \cap G) + \mu^*((A_1 \cup A_2) \cap F). \tag{5}$$

But $A_1 \subset G$ and $G \cap A_2 = \emptyset$, so $(A_1 \cup A_2) \cap G = A_1$ and

$$(A_1 \cup A_2) \cap F = A_2,$$

and (5) becomes

$$\mu^*(A_1 \cup A_2) = \mu^*(A_1) + \mu^*(A_2),$$

as was to be shown. ■

Theorem 3.8 shows that metric outer measures give rise to *Borel measures*, that is, measures for which every Borel set is measurable. This does not rule out the possibility that there exist measurable sets that are not Borel sets. Some authors reserve the term Borel measure for a measure satisfying rather more. For example, one might wish compact sets to have finite measure or one might demand further approximation properties. The term Radon measure is also used in this context to denote Borel measures with special properties relative to the compact sets.

Exercises

3:2.1 Let us try to fix the problems that arose in connection with Example 3.5 that began this section. Let $\mathcal{T}$ be the family of half-open squares in $(0,1] \times (0,1]$ of the form $(a,b] \times (c,d]$, $b-a = d-c$, together with $\emptyset$, and let $\tau(T)$ be the diameter of T. Do the finite unions of elements of $\mathcal{T}$ form an algebra of sets? Can τ be extended to the algebra generated by $\mathcal{T}$ so as to be additive on this algebra? Can we use Theorem 2.41 effectively?

3:2.2 Let $X = \mathbb{R}^2$, let $\mathcal{T}$ consist of the half-open *intervals*

$$T = (a,b] \times (c,d]$$

in X, and let $\tau(T)$ be the area of T. Let μ^* be obtained from $\mathcal{T}$ and τ by Method I. Prove that μ^* is a metric outer measure. The resulting measure is called *two-dimensional Lebesgue measure.*

3.3 Method II

As we have seen, the Method I construction applied in a metric space can fail to produce a metric outer measure. We now seek to modify Method I in such a manner so as to guarantee that the resulting outer measure is metric. The modified construction will be called Method II.

Let us return to Example 3.5 involving squares in $\mathbb{R}^2$, with $\tau(T)$ the diameter of the square T. To obtain $\mu^*(T)$, we observe we can do no better than to cover T with itself. If, for example, we cover a square T of side length 1 with smaller squares, say ones of diameter no greater than $1/n$, we find that we need more than n^2 squares to do the job, and the estimate for $\mu^*(T)$ obtained from these squares exceeds $n\sqrt{2}$. The smaller the squares we use in the cover of T, the larger the estimate for $\mu^*(T)$. We do best by simply taking one square, T, for the cover. Thus the small squares are irrelevant and play no role in the construction, and yet it is precisely these that should have an influence on the size of the measure. This is the source of our problem. We now present a new method for obtaining measures from outer measures that explicitly addresses this by forcing the sets of small diameter to be taken into account.

Let $\mathcal{T}$ be a covering family on a metric space X. For each $n \in \mathbb{N}$, let

$$\mathcal{T}_n = \{T \in \mathcal{T} : \text{diameter}\,(T) \leq 1/n\}\,.$$

Then $\mathcal{T}_n$ is also a covering family for X for each $n \in \mathbb{N}$. Let τ be a premeasure defined on the family $\mathcal{T}$. For every $n \in \mathbb{N}$, we construct μ_n^* by Method I from $\mathcal{T}_n$ and τ. Since $\mathcal{T}_{n+1} \subset \mathcal{T}_n$,

$$\mu_{n+1}^*(E) \geq \mu_n^*(E)$$

for all $n \in \mathbb{N}$ and for each $E \subset X$. Thus the sequence $\{\mu_n^*(E)\}$ approaches a finite or infinite limit. We define μ_0^* as $\lim_{n\to\infty} \mu_n^*$ and refer to this as the outer measure determined by Method II from τ and $\mathcal{T}$.

3.3.1 Method II outer measures are metric outer measures

Our next theorem shows that this process that we have called Method II always gives rise to a metric outer measure.

Theorem 3.9: *Let μ_0^* be the measure determined by Method II from a premeasure τ and a family $\mathcal{T}$. Then μ_0^* is a metric outer measure.*

Proof. We first show that μ_0^* is an outer measure. That $\mu_0^*(\emptyset) = 0$, and that $\mu_0^*(A) \leq \mu_0^*(B)$ if $A \subset B$ are immediate. To verify that μ_0^* is countably subadditive, let $\{A_k\}$ be a sequence of subsets of X. Since $\mu_0^*(E) \geq \mu_n^*(E)$ for all $E \subset X$ and $n \in \mathbb{N}$, we have

$$\mu_n^*\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mu_n^*(A_k) \leq \sum_{k=1}^{\infty} \mu_0^*(A_k).$$

Thus

$$\mu_0^*\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} \mu_n^*\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mu_0^*(A_k).$$

This verifies that μ_0^* is an outer measure.

It remains to show that if A and B are separated then

$$\mu_0^*(A \cup B) = \mu_0^*(A) + \mu_0^*(B).$$

Certainly,

$$\mu_0^*(A \cup B) \leq \mu_0^*(A) + \mu_0^*(B),$$

and so it is enough to establish the opposite inequality. We may assume that $\mu_0^*(A \cup B)$ is finite. Suppose then that $\operatorname{dist}(A, B) > 0$. Choose $N \in \mathbb{N}$ such that $\operatorname{dist}(A, B) > 1/N$. Let $\varepsilon > 0$. For every $n \in \mathbb{N}$ there exists a sequence $\{T_{nk}\}$ from $\mathcal{T}_n$ such that $\bigcup_{k=1}^{\infty} T_{nk} \supset A \cup B$ and

$$\sum_{k=1}^{\infty} \tau(T_{nk}) \leq \mu_n^*(A \cup B) + \varepsilon.$$

Then, for $n \geq N$ and $k \in \mathbb{N}$, no set T_{nk} can meet both A and B and hence $T_{nk} \cap A = \emptyset$ or else $T_{nk} \cap B = \emptyset$. Let

$$\mathbb{N}_1 = \{k \in \mathbb{N} : T_{nk} \cap A \neq \emptyset\}$$

and

$$\mathbb{N}_2 = \{k \in \mathbb{N} : T_{nk} \cap B \neq \emptyset\}.$$

Then

$$\mu_n^*(A) \leq \sum_{k \in \mathbb{N}_1} \tau(T_{nk})$$

and

$$\mu_n^*(B) \leq \sum_{k \in \mathbb{N}_2} \tau(T_{nk}).$$

Therefore,

$$\mu_n^*(A) + \mu_n^*(B) \leq \sum_{k=1}^{\infty} \tau(T_{nk}) \leq \mu_n^*(A \cup B) + \varepsilon.$$

Since this is true for every $\varepsilon > 0$, we have, for $n \geq N$,

$$\mu_n^*(A) + \mu_n^*(B) \leq \mu_n^*(A \cup B).$$

Because this holds for all $n \geq N$, $\mu_0^*(A)+\mu_0^*(B) \leq \mu_0^*(A \cup B)$. Thus μ_0^* is a metric outer measure. ■

3.3.2 Agreement of Method I and Method II measures

Let us return to Example 3.5. Our previous discussion involving covers of a square T with smaller squares suggests that $\mu_0^*(T) = \infty$ for every square T. This is, in fact, the case. If T is an open square with unit side length, $\mu_n^*(T) = n\sqrt{2}$. Thus

$$\mu_0^*(T) = \lim_{n \to \infty} \mu_n^*(T) = \infty.$$

A similar argument shows that $\mu_0^*(T) = \infty$ for all $T \in \mathcal{T}$. This may be no surprise since we have used a "one-dimensional" concept (diameter) as a premeasure for a two-dimensional set T. Recall that the Method I outer measure μ^* had $\mu^*(T) = \tau(T)$, since we could efficiently cover T by itself. In this example, small squares cannot cover large squares efficiently, and the Method I outcome differs from that of Method II. Our next result, Theorem 3.10, shows that if "small squares *can* cover large squares efficiently" then the Method I and Method II measures do agree.

Theorem 3.10: *Let μ_0^* be the measure determined by Method II from a premeasure τ and a family $\mathcal{T}$ and let μ^* be the Method I measure constructed from τ and $\mathcal{T}$. A necessary and sufficient condition that $\mu_0^* = \mu^*$ is that for each choice of $\varepsilon > 0$, $T \in \mathcal{T}$, and $n \in \mathbb{N}$, there is a sequence $\{T_k\}$ from $\mathcal{T}_n$ such that $T \subset \bigcup_{k=1}^{\infty} T_k$ and*

$$\sum_{k=1}^{\infty} \tau(T_k) \le \tau(T) + \varepsilon.$$

Proof. Necessity is clear. If the condition fails for some ε, T, and n, then $\mu_0^*(T) > \mu^*(T)$. To prove sufficiency, observe first that, since $\mathcal{T}_n \subset \mathcal{T}$ for all $n \in \mathbb{N}$,

$$\mu^* \le \mu_n^* \le \mu_0^*. \tag{6}$$

To verify the reverse inequality, let $A \subset X$ and let $\varepsilon > 0$. We may assume that $\mu^*(A) < \infty$. Let $\{T_i\}$ be a sequence of sets from $\mathcal{T}$ such that $A \subset \bigcup_{i=1} T_i$ and

$$\sum_{i=1}^{\infty} \tau(T_i) \le \mu^*(A) + \frac{\varepsilon}{2}. \tag{7}$$

Let $n \in \mathbb{N}$. Using our hypotheses, we have, for each $i \in \mathbb{N}$, a sequence $\{S_{ik}\}$ of sets from $\mathcal{T}_n$ covering T_i such that

$$\sum_{k=1}^{\infty} \tau(S_{ik}) \leq \tau(T_i) + \frac{\varepsilon}{2^{i+1}}. \tag{8}$$

Now $A \subset \bigcup_{i=1}^{\infty} \bigcup_{k=1}^{\infty} S_{ik}$, so by (7) and (8) we have

$$\mu_n^*(A) \leq \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \tau(S_{ik}) \leq \sum_{i=1}^{\infty} \left[\tau(T_i) + \frac{\varepsilon}{2^{i+1}}\right] \leq \mu^*(A) + \varepsilon.$$

Since ε is arbitrary, $\mu_n^*(A) \leq \mu^*(A)$. This is true for every $n \in \mathbb{N}$, so

$$\mu_0^*(A) = \lim_{n\to\infty} \mu_n^*(A) \leq \mu^*(A). \tag{9}$$

From (6) and (9), we see that $\mu^* = \mu_0^*$. ■

Corollary 3.11: *Under the hypotheses of Theorem 3.10, Method I results in a metric outer measure.*

Method II also has a regularity result identical to Theorem 2.37. We leave the details as Exercise 3:3.4.

Theorem 3.12: *Let μ_0^* be constructed from $\mathcal{T}$ and τ by Method II. If all members of $\mathcal{T}$ are measurable, then μ_0^* is regular. In particular, if each $T \in \mathcal{T}$ is an open set, the measurable covers can be chosen to be Borel sets of type $\mathcal{G}_\delta$.*

Exercises

3:3.1 In the proof of Theorem 3.9, verify that $\mu_0^*(\emptyset) = 0$ and $\mu_0^*(A) \le \mu_0^*(B)$ if $A \subset B$.

3:3.2 Let $\mathcal{T}$ consist of $\emptyset$ and the open intervals in $X = (-1, 1)$, and let $\tau((a, b)) = |b^2 - a^2|$. Apply Method I to obtain μ^* and Method II to obtain μ_0^*.

(a) Determine the class of μ^*-measurable sets.

(b) Calculate $\mu^*((0, 1))$ and $\mu_0^*((0, 1))$.

3:3.3 Let $X = \mathbb{R}$, $\mathcal{T}$ consist of $\emptyset$ and the open intervals in $\mathbb{R}$. Let $\tau(\emptyset) = 0$ and let $\tau((a, b)) = (b - a)^{-1}$ for all other $(a, b) \in \mathcal{T}$. Let μ_1 and μ_2 be the measures obtained from $\mathcal{T}$ and τ by Methods I and II, respectively.

(a) Show that $\mu_1(E) = 0$ for all $E \subset X$.

(b) Show that $\mu_2(E) = \infty$ for every nonempty set $E \subset X$.

Note $\tau(T)$, $\mu_1(T)$, and $\mu_2(T)$ are all different in this example. While Method I always results in $\mu^*(T) \le \tau(T)$, this inequality is not valid in general when Method II is used. We had already seen this in our example with squares.

3:3.4 Prove Theorem 3.12.

3:3.5 Verify that in Theorem 3.12, if we do not assume that the sets in $\mathcal{T}$ are measurable, we can still conclude that each set $A \subset X$ with finite measure has a cover in $\mathcal{T}_{\sigma\delta}$. (Compare with Exercise 2:10.8.)

3.4 Approximations

In most settings the measure of a measurable set can be approximated from inside or outside by simpler sets, perhaps open sets or $\mathcal{G}_\delta$ sets, as we were able to do on $\mathrm{I\!R}$ with Lebesgue measure. By the use of Theorems 2.36 and 3.12, one can obtain such approximations from sets that were used in the first place to construct the measure. The approximation theorem that follows is of a different sort, however, in that it does not involve Methods I or II, or outer measures. We show how to approximate the measure of any Borel set first from inside by closed sets and then from outside by open sets for any Borel measure. Recall that for μ to be a Borel measure requires merely that μ be a measure whose σ–algebra of measurable sets includes all Borel sets.

3.4.1 Approximation from inside

The first approximation theorem asserts conditions under which we can be sure of approximating the measure of a Borel set by using a closed subset of the Borel set.

Theorem 3.13: *Let X be a metric space, μ a Borel measure on X, $\varepsilon > 0$ and B_0 a Borel set with $\mu(B_0) < \infty$. Then B_0 contains a closed set F for which $\mu(B_0 \setminus F) < \varepsilon$.*

Proof. We assume first that $\mu(X) < \infty$ and show that *all* Borel sets have the stated property. Let $\mathcal{E}$ consist of those sets $E \subset X$ that have the property that for any $\gamma > 0$ there is a closed subset K of E for which $\mu(E \setminus K) < \gamma$. We claim that every Borel set $B \subset X$ is a member of $\mathcal{E}$.

We show that $\mathcal{E}$ contains the closed sets and that it is closed under countable unions and closed under countable intersections. By Theorem 3.3, it follows that $\mathcal{E}$ must contain all the Borel sets.

It is clear that $\mathcal{E}$ contains the closed sets. Suppose now that E_1, E_2, ... belong to $\mathcal{E}$. There must exist closed sets $K_i \subset E_i$ with $\mu(E_i \setminus K_i) < \varepsilon 2^{-i}$. We get immediately that

$$\mu\left(\bigcap_{i=1}^{\infty} E_i \setminus \bigcap_{i=1}^{\infty} K_i\right) \leq \mu\left(\bigcup_{i=1}^{\infty} (E_i \setminus K_i)\right) < \sum_{i=1}^{\infty} \varepsilon 2^{-i} = \varepsilon.$$

Since $\bigcap_{i=1}^{\infty} K_i$ is a closed subset of $\bigcap_{i=1}^{\infty} E_i$, we see that the intersection of the sequence $\{E_i\}$ belongs to $\mathcal{E}$.

The union can be handled similarly but requires an extra step, since countable unions of closed sets are not necessarily closed. Note that

$$\lim_{n\to\infty} \mu\left(\bigcup_{i=1}^{\infty} E_i \setminus \bigcup_{i=1}^{n} K_i\right) = \mu\left(\bigcup_{i=1}^{\infty} E_i \setminus \bigcup_{i=1}^{\infty} K_i\right)$$
$$\leq \mu\left(\bigcup_{i=1}^{\infty} (E_i \setminus K_i)\right) < \sum_{i=1}^{\infty} \varepsilon 2^{-i} = \varepsilon.$$

(It is here that we are using the finiteness assumption, since to invoke the limit requires Theorem 2.21.) Thus, for sufficiently large n, we must have

$$\mu\left(\bigcup_{i=1}^{\infty} E_i \setminus \bigcup_{i=1}^{n} K_i\right) < \varepsilon,$$

and this set, $\bigcup_{i=1}^{n} K_i$, is a closed subset of $\bigcup_{i=1}^{\infty} E_i$.

To complete the proof we need to address the case $\mu(X) = \infty$. Define a new measure μ_0 by setting $\mu_0(E) = \mu(B_o \cap E)$ for all $E \subset X$. Then μ_0 is a finite Borel measure on X. By what we have just proved, all Borel sets can be approximated from inside by closed sets. In particular, there is a closed set $F \subset B_0$ for which $\mu_0(B_0 \setminus F) < \varepsilon$. Since $\mu_0(B_0 \setminus F) = \mu(B_0 \setminus F)$ we are done. ■

We mention that the discussion following Theorem 3.20 will show that the σ-algebra $\mathcal{E}$ in the proof just given need not consist of all measurable sets. See also Exercise 3:6.3. We now turn to the approximation from the outside by open sets.

3.4.2 Approximation from outside

The second approximation theorem asserts conditions under which we can be sure of approximating the measure of a Borel set by using a larger open set that contains the Borel set.

Theorem 3.14: *Let X be a metric space, μ a Borel measure on X, $\varepsilon > 0$, and B a Borel set. If $\mu(X) < \infty$ or, more generally, if B is contained in the union of countably many open sets O_i each of finite μ-measure, then B is contained in an open set G with $\mu(G \setminus B) < \varepsilon$.*

Proof. This theorem follows from the preceding. Choose each closed set $C_i \subset O_i \setminus B$ in such a way that

$$\mu\left((O_i \setminus C_i) \setminus B\right) = \mu\left((O_i \setminus B) \setminus C_i\right) < \varepsilon 2^{-i}.$$

Here $B \cap O_i$ is a subset of the open set $O_i \setminus C_i$. Define

$$G = \bigcup_{i=1}^{\infty} (O_i \setminus C_i).$$

Then G is open, G contains B, and $\mu(G \setminus B) < \varepsilon$. ■

For reference let us put the two theorems together to derive a corollary, valid in spaces of finite measure.

Corollary 3.15: *Let X be a metric space and μ a Borel measure with $\mu(X) < \infty$. For every $\varepsilon > 0$ and every Borel set B, there is a closed set F and an open set G such that*

$$F \subset B \subset G,$$

with

$$\mu(B) - \varepsilon < \mu(F) \leq \mu(B) \leq \mu(G) < \mu(B) + \varepsilon.$$

3.4.3 Approximation using $\mathcal{F}_\sigma$ and $\mathcal{G}_\delta$ sets

From these two theorems we easily derive a further approximation theorem that uses slightly larger classes of sets than the open and closed sets.

Theorem 3.16: *Let X be a metric space, and μ a Borel measure on X such that $\mu(X)$ is finite. Then every Borel set $B \subset X$ has a subset K of type $\mathcal{F}_\sigma$ and a superset H of type $\mathcal{G}_\delta$, such that*

$$\mu(K) = \mu(B) = \mu(H).$$

In terms of the language of Exercise 2:1.14, every Borel set in X has a measurable cover of type $\mathcal{G}_\delta$ and a measurable kernel of type $\mathcal{F}_\sigma$. The requirement that $\mu(X) < \infty$ in the statement of Theorem 3.16 cannot be dropped. See Exercise 3:4.3.

Corollary 3.15 and Theorem 3.16 involve approximations of Borel sets by simpler sets. If we know that measurable sets can be approximated by Borel sets, then the conclusions of 3.15 and 3.16 can be sharpened. For example, under the hypotheses of Theorem 3.12, if $\mathcal{T}$ consists of Borel sets, every measurable set M has a cover $H \in \mathcal{B}$. If $\mu(X) < \infty$, H has a cover H' of type $\mathcal{G}_\delta$. Thus H' is a cover for M as well. If one wished, one could combine the hypotheses of 3.12, 3.15, and 3.16 suitably to obtain various results concerning approximations of measurable sets by Borel sets, sets of type $\mathcal{G}_\delta$, open sets, and so on.

Exercises

3:4.1 Prove Theorem 3.14 in the simplest case where $\mu(X) < \infty$.

3:4.2 Prove Theorem 3.16.

3:4.3 Let $\mathcal{B}$ denote the Borel sets in $\mathbb{R}$. Recall that part of the Baire category theorem for $\mathbb{R}$ that asserts that a set of type $\mathcal{G}_\delta$ that is dense in some interval cannot be expressed as a countable union of nowhere dense sets. For $E \in \mathcal{B}$, let $\mu(E) = \lambda(E)$ if E is a countable union of nowhere dense sets, $\mu(E) = \infty$ otherwise. Show that $(\mathbb{R}, \mathcal{B}, \mu)$ is a measure space for which the conclusion of Theorem 3.16 fails.

3:4.4 Let μ be a finite Borel measure on a metric space X. Prove that, for every Borel set $B \subset X$,

$$\mu(B) = \inf\{\mu(G) : B \subset G,\ G \text{ open}\}$$

and

$$\mu(B) = \sup\{\mu(F) : F \subset B,\ F \text{ closed}\}.$$

3.5 Construction of Lebesgue–Stieltjes Measures

The most important class of Borel measures on $\mathbb{R}^n$ are those that are finite on bounded sets. Often these are called Lebesgue–Stieltjes measures after the Dutch mathematician, T. J. Stieltjes (1856–1894), whose integral (see Section 1.19) played a key role in the development of measure theory by J. Radon (1887–1956) in the second decade of the last century. For the same reason, they have also been called Radon measures. Certain of the Hausdorff measures that we discuss in Section 3.8 are, in contrast, examples of important Borel measures that are infinite on *every* open set.

Lebesgue–Stieltjes measures are Borel measures in $\mathbb{R}^n$ that can serve to model mass distributions. Some previews can be found in Example 2.10 and Exercises 2:2.14, 2:9.2, and 2:10.7. We can now use the machinery we have developed to obtain such models rigorously and compatibly with our intuition. We consider the one-dimensional situation in detail here and then outline the construction for $\mathbb{R}^n$ in Section 3.7.

Suppose, for each $x \in \mathbb{R}$, that we know the mass of intervals of the form $(0, x]$ or of the form $(x, 0]$ and that all such masses are finite. Let

$$f(x) = \begin{cases} \text{mass } (0, x], & \text{if } x > 0; \\ 0, & \text{if } x = 0; \\ -\text{mass } (x, 0], & \text{if } x < 0. \end{cases} \tag{10}$$

Then f is a nondecreasing function on $\mathbb{R}$. While f need not be continuous, we require f to be right continuous. Since monotonic functions have left and right limits at every point, this just fixes the value of f at its countably many points of discontinuity in a particular way.

We now carry out a program similar to the one we outlined in Exercise 2:13.4. Here we are dealing with intervals in $\mathbb{R}$, rather than in $\mathbb{R}^2$. Let $\mathcal{T}$ consist of the half-open intervals of the form $(a, b]$, the empty set, and the unbounded intervals of the form $(-\infty, b]$ and (a, ∞). For a premeasure $\tau : \mathcal{T} \to [0,\infty]$, we shall use

$$\tau(T) = \begin{cases} 0, & \text{if } T = \emptyset; \\ f(b) - f(a), & \text{if } T = (a,b]; \\ f(b) - \lim_{a\to-\infty} f(a), & \text{if } T = (-\infty, b]; \\ \lim_{b\to\infty} f(b) - f(a), & \text{if } T = (a,\infty). \end{cases} \tag{11}$$

The limits involved exist, finite or infinite, because f is nondecreasing.

Continuing the program, we let $\mathcal{T}_1$ be the algebra generated by $\mathcal{T}$. One sees immediately that $\mathcal{T}_1$ consists of all finite unions of elements of $\mathcal{T}$. We wish to extend the premeasure τ to an additive function $\tau_1 : \mathcal{T}_1 \to [0,\infty]$. For $T \in \mathcal{T}_1$, write

$$T = T_1 \cup T_2 \cup \cdots \cup T_n,$$

with $T_i \in \mathcal{T}$ for each $i = 1, \ldots, n$, and $T_i \cap T_j = \emptyset$ if $i \neq j$. We "define"

$$\tau_1(T) = \tau(T_1) + \tau(T_2) + \cdots + \tau(T_n). \tag{12}$$

The quotes indicate that we must verify that (12) is unambiguous. (Recall our example of squares in Section 3.2 when τ was the diameter of the square.)

3.17: *The set function τ_1 is well defined on $\mathcal{T}_1$.*

Proof. Consider first the case that $T \in \mathcal{T}$. Let

$$T = (a,b] = \bigcup_{i=1}^{n} (a_i, b_i]$$

with $a_1 = a$, $b_n = b$, and $a_{i+1} = b_i$ for all $i = 1, \ldots, n-1$. Thus

$$\tau((a,b]) = f(b) - f(a) = \sum_{i=1}^{n} (f(b_i) - f(a_i)) = \sum_{i=1}^{n} \tau((a_i, b_i]).$$

A similar argument shows that if an unbounded interval $T \in \mathcal{T}$ is decomposed into finitely many members of $\mathcal{T}$ then (12) holds. Finally, any $T \in \mathcal{T}_1$ is a finite union of members of $\mathcal{T}$. These members can be appropriately combined, if necessary, to become a disjoint collection

$$\{(a_i, b_i]\}_{i=1}^{n} \text{ with } b_i < a_{i+1}. \tag{13}$$

Here it is possible that $a_1 = -\infty$ or $b_n = \infty$. Suppose that T is decomposed into a finite disjoint union of sets in $\mathcal{T}$, say $T = \bigcup_{j=1}^{m} T_j$. Let

$$A_i = \{j : T_j \subset (a_i, b_i]\}.$$

Then, $(a_i, b_i] = \bigcup_{j \in A_i} T_j$. We have already seen that, for all $i = 1, \ldots, n$,

$$\tau((a_i, b_i]) = \sum_{j \in A_i} \tau(T_j).$$

Since any representation of T as a finite disjoint union of members of $\mathcal{T}$ heads to the same collection (13), the sum in (12) does not depend on the representation for T. ■

Because of Theorem 2.41, we now know that an application of Method I would lead to a measure space in which every member of $\mathcal{T}$ is measurable. This implies that every Borel set is measurable. To see this, note that an open interval is a countable union of half-open intervals,

$$(a, b) = \bigcup_{n=1}^{\infty} (a, b_n],$$

where $a < b_1 < b_2 < \cdots < b$ and $\lim_{n\to\infty} b_n = b$. It follows from Theorem 3.8 that μ^* is a metric outer measure. From Theorem 2.37 we see that μ^* is also regular and from Exercise 2:10.8 that each set $A \subset \mathbb{R}$ has a Borel set B as a measurable cover. It now follows readily from Theorem 3.16 that B can be taken to be of type $\mathcal{G}_\delta$ (left as Exercise 3:5.1). What we do not yet know is that the members of $\mathcal{T}_1$, or even of $\mathcal{T}$, have the right measure; that is, that $\mu^*(T) = \tau(T)$. To obtain this result, it suffices to show that τ_1 is σ-additive on $\mathcal{T}_1$. We can then invoke Theorem 2.43.

3.18: *The set function τ_1 is σ-additive on $\mathcal{T}_1$.*

Proof. To show that τ_1 is σ-additive on $\mathcal{T}_1$, we must show that, if $\{T_n\}$ is a sequence of pairwise disjoint sets in $\mathcal{T}_1$ whose union T is also in $\mathcal{T}_1$, then

$$\tau_1(T) = \sum_{n=1}^{\infty} \tau_1(T_n).$$

Observe that it is sufficient to consider only the case that T is a single interval, either a finite half-open interval $(a, b]$, an infinite interval $(-\infty, b]$, or an infinite interval (a, ∞). Every other set in $\mathcal{T}_1$ is a finite disjoint union of intervals of these three types.

We address only the case where $T = (a, b]$, a bounded interval; the other cases can be handled similarly. For *finite* additivity, our work was simplified by the fact that if $(a, b] = \bigcup_{i=1}^{n} (a_i, b_i]$, with the sets $\{(a_i, b_i]\}$ pairwise disjoint,

$$f(b) - f(a) = \sum_{i=1}^{n} (f(b_i) - f(a_i)),$$

because the intervals must form a partition of $(a, b]$.

This telescoping of the sum is not always possible when dealing with an infinite decomposition of the form $(a,b] = \bigcup_{i=1}^{\infty}(a_i,b_i]$ with the sets $\{(a_i,b_i]\}$ pairwise disjoint. For example, consider

$$(-1,1] = (-1,0] \cup \bigcup_{n=1}^{\infty} \left((n+1)^{-1}, n^{-1}\right].$$

Here 0 is a right endpoint of an interval in the collection, but not a left endpoint of any other interval. It is still true that

$$f(1) - f(-1) = f(0) - f(-1) + \sum_{n=1}^{\infty} \left[f(n^{-1}) - f((n+1)^{-1})\right],$$

but this requires handling right-hand limits at 0. In general, if for some $i \in \mathbb{N}$, b_i is a limit point of the set $\{a_j\}_{j=1}^{\infty}$, then $b_i \neq a_j$ for any $j \in \mathbb{N}$. Thus we do not get the cancelations from which we benefited when we had telescoping sums. Moreover, there can be infinitely many points of this type to handle. Note that it is only the right endpoints that have this feature.

Let us look at the situation in some detail. Let $A = \{a_i\}$ and $B = \{b_i\}$. Then $A \subset B \cup \{a\}$, but B is not necessarily contained in A. A simple diagram can illustrate that $B \setminus A$ can be infinite. Now

$$[a,b] = \bigcup (a_k, b_k) \cup B \cup \{a\}.$$

It follows that $B \cup \{a\}$ is a countable closed set. Let $J_0 = [f(a), f(b)]$ and, for $k \in \mathbb{N}$, let $J_k = [f(a_k), f(b_k)]$. Since f is nondecreasing, $\bigcup_{k=1}^{\infty} J_k \subset J_0$, and the intervals J_k have no interior points in common. Because f is right continuous at $x = a$,

$$J_0 \subset \bigcup_{k=1}^{\infty} J_k \cup f(B) \cup \{f(a)\}.$$

B is countable, so $f(B)$ is also countable, and hence

$$\lambda(f(B) \cup \{f(a)\}) = 0,$$

where, as usual, λ denotes the Lebesgue measure. It follows that

$$\begin{aligned}\sum_{k=1}^{\infty}(f(b_k) - f(a_k)) &= \lambda\left(\bigcup_{k=1}^{\infty} J_k\right) \leq \lambda(J_0) \\ &\leq \lambda\left(\bigcup_{k=1}^{\infty} J_k \cup f(B) \cup \{f(a)\}\right) \\ &= \sum_{k=1}^{\infty} \lambda(J_k) = \sum_{k=1}^{\infty}(f(b_k) - f(a_k)).\end{aligned}$$

Thus $f(b) - f(a) = \lambda(J_0) = \sum_{k=1}^{\infty}(f(b_k) - f(a_k))$, so that

$$\tau_1((a,b]) = \sum_{k=1}^{\infty} \tau_1((a_k, b_k])$$

as required. ■

We have now completed the program. We can finally conclude that an application of Method I will give rise to an outer measure μ_f^* and then to a measure space $(X, \mathcal{M}_f, \mu_f)$ with

$$\mu_f((a,b]) = f(b) - f(a).$$

We call μ_f the *Lebesgue–Stieltjes measure with distribution function* f. We shall also use such phrases as μ_f is the measure "induced by" f or "associated with" f. Observe that for $c \in \mathbb{R}$ the function $f + c$ can also serve as a distribution function for μ_f. When dealing with finite

Lebesgue–Stieltjes measures, it is often convenient to choose f so that $\lim_{x\to-\infty} f(x) = 0$. Moreover, when all the measure is located in some interval I, it may be convenient merely to specify f only on I itself (as, for example, we do in Exercise 3:5.5). Technically, this amounts to extending f to all of $\mathbb{R}$ in such a way that $\mu_f(\mathbb{R} \setminus I) = 0$. (Such an extension would be required for Exercise 3:13.5.)

Example 3.19: A *probability space* is a measure space of total measure 1. If $X = \mathbb{R}$, the distribution function can be chosen so that $\lim_{x\to-\infty} f(x) = 0$ and will then satisfy $\lim_{x\to\infty} f(x) = 1$. For a measurable set A, $\mu_f(A)$ represents the probability that a random variable lies in A. As a concrete example, if ϕ is the standard normal density (bell-shaped curve),

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad (-\infty < x < \infty),$$

then $\int_{-\infty}^{\infty} \phi\,(x)\; dx = 1$, and one can take $f(x) = \int_{-\infty}^{x} \phi\,(t)\; dt$ as an associated distribution function.

In the setting of probability, the "mass" of a Borel set A is interpreted as the probability of the "event" A occurring. Thus the probability that a standard normal random variable Z satisfies $a < Z \le b$ is

$$\Pr(a < Z \le b) = f(b) - f(a) = \int_a^b \phi\,(x)\; dx.$$

More generally, for any Borel set A we would have

$$\Pr(Z \in A) = \mu_f(A) = \int_A \phi(x)\, dx,$$

where the integral must be interpreted in the Lebesgue sense. (We will have to wait until Chapter 5 for this.)

Exercises

3:5.1 Prove that, for any Lebesgue–Stieltjes measure μ, every $A \subset X$ has a measurable cover of type $\mathcal{G}_\delta$ and a measurable kernel of type $\mathcal{F}_\sigma$.

3:5.2 Use Theorems 3.9 and 3.10 to give another proof that a Lebesgue–Stieltjes outer measure μ_f^* is a metric outer measure.

3:5.3 Let

$$f(x) = \begin{cases} 0, & \text{if } x < 0; \\ 1, & \text{if } 0 \leq x < 1; \\ 2, & \text{if } x \geq 1. \end{cases}$$

Show that $\mu_f((0,1)) < \mu_f((0,1]) < \mu_f([0,1])$.

3:5.4 Let $X = \mathbb{R}$ and

$$\mu(A) = \begin{cases} n, & \text{if card } A \cap \mathbb{N} = n; \\ \infty, & \text{if } A \cap \mathbb{N} \text{ is infinite.} \end{cases}$$

Construct a distribution function f such that $\mu_f = \mu$.

3:5.5 Let f be the Cantor function, and let μ_f be the associated Lebesgue–Stieltjes measure. Calculate $\mu_f((\frac{1}{3}, \frac{2}{3}))$ and $\mu_f((K \cap (\frac{2}{9}, \frac{1}{3}))$, where K is the Cantor ternary set.

3:5.6 Let μ_f be a Lebesgue–Stieltjes measure. Show that

$$\mu_f((a,b)) = \lim_{x \to b-} (f(x) - f(a))$$

and calculate $\mu_f(\{b\})$.

3:5.7$\diamondsuit$ The term Lebesgue–Stieltjes measure is often used to apply to what would more properly be called a "*Lebesgue–Stieltjes signed measure.*"

(a) What should we mean by a Lebesgue–Stieltjes signed measure associated with a function f that is not nondecreasing? [Hint: If $f = f_1 - f_2$ where f_1 and f_2 are nondecreasing what should one use?]

(b) Let

$$f(x) = \begin{cases} 1, & \text{if } x < -1; \\ x^2, & \text{if } -1 \leq x \leq 1; \\ 1, & \text{if } 1 < x. \end{cases}$$

Let μ_f be the associated Lebesgue–Stieltjes measure. Calculate the Jordan decomposition for the signed measure μ_f, and compute

$$\mu_f((-1,1)) \text{ and } V(\mu_f,(-1,1)).$$

(c) Let $f : [a,b] \to \mathbb{R}$ be a function of bounded variation and let

$$F(x) = V(f;[a,x]) \quad (a \leq x \leq b)$$

be its total variation function. What is the relation between the Lebesgue-Stieltjes measure for F and the signed Lebesgue-Stieltjes measure for f? [Hint: Compare $|\mu_f|$ and μ_F.]

[Note that functions of bounded variation give rise to Lebesgue–Stieltjes signed measures via their decomposition into a difference of two nondecreasing functions.]

3:5.8◇ Let $(X, \mathcal{M}, \mu)$ be a measure space. A set $A \in \mathcal{M}$ is called an *atom* if $\mu(A) > 0$ and for all measurable sets $B \subset A$, $\mu(B) = 0$ or $\mu(A \setminus B) = 0$. (See Exercise 2:14.7.)

(a) Give an example of a space $(\mathbb{R}, \mathcal{M}, \mu)$ for which $[0,1]$ is an atom.

(b) Let $(\mathbb{R}, \mathcal{M}_f, \mu_f)$ be a Lebesgue–Stieltjes measure space. Prove that, if A is an atom in this space, A contains a singleton atom with the same measure. That is, there exists $a \in A$ for which $\mu_f(\{a\}) = \mu_f(A)$. One also uses the term “point mass” to describe a singleton atom of μ_f.

(c) A measure μ is *nonatomic* if there are no atoms. Prove that a Lebesgue–Stieltjes measure is nonatomic if and only if its distribution function is continuous.

3.6 Properties of Lebesgue–Stieltjes Measures

We investigate now some of the important properties of Lebesgue–Stieltjes measures in one dimension. The first theorem provides a sense of the generality of such measures.

Theorem 3.20: *Let f be nondecreasing and right continuous on $\mathbb{R}$. Let μ_f^* be the associated Method I outer measure, and let $(\mathbb{R}, \mathcal{M}_f, \mu_f)$ be the resulting measure space. Then*

1. *μ_f^* is a metric outer measure and thus all Borel sets are μ_f^*-measurable.*
2. *If A is a bounded Borel set, then $\mu_f(A) < \infty$.*
3. *Each set $A \subset \mathbb{R}$ has a measurable cover of type $\mathcal{G}_\delta$.*
4. *For every half-open interval $(a, b]$, $\mu_f((a, b]) = f(b) - f(a)$.*

Conversely, let μ^* be an outer measure on $\mathbb{R}$ with $(X, \mathcal{M}, \mu)$ the resulting measure space. If conditions (i), (ii), and (iii) are satisfied by μ^* and μ, then there exists a nondecreasing, right-continuous function f defined on $\mathbb{R}$ such that $\mu_f^*(A) = \mu^*(A)$ for all $A \subset \mathbb{R}$. In particular, $\mu_f(A) = \mu(A)$ for all $A \in \mathcal{M}$.

Proof. Most of the proof of the first half of the theorem is contained in our development. The converse direction needs some justification, since our concept of "mass" was not made precise. Define f on $\mathbb{R}$ by

$$f(x) = \begin{cases} \mu((0,x]), & \text{if } x > 0; \\ 0, & \text{if } x = 0; \\ -\mu((x,0]), & \text{if } x < 0. \end{cases}$$

It is clear that f is nondecreasing. To verify that f is right continuous, let $x \in \mathbb{R}$ and let $\{\delta_n\}$ be a sequence of positive numbers decreasing to zero. Suppose, without loss of generality, that $x > 0$. Then

$$(0,x] = \bigcap_{n=1}^{\infty} (0, x + \delta_n].$$

Since $\mu((0, x + \delta_1]) < \infty$ by (ii), we see from Theorem 2.21, part (ii), that

$$\mu((0,x]) = \lim_{n\to\infty} \mu((0, x + \delta_n]),$$

that is, $f(x) = \lim_{n\to\infty} f(x + \delta_n)$, and f is right continuous.

To show that $\mu_f^* = \mu^*$, we proceed in stages. We start by showing agreement on half-open intervals, then open intervals, open sets, bounded $\mathcal{G}_\delta$ sets, bounded sets, and finally arbitrary sets.

First, it follows from the definition of f that

$$\mu_f((a,b]) = \mu((a,b])$$

for every finite half-open interval $(a,b]$. Next, observe that, since both μ and μ_f are σ-additive, and every open interval is a countable disjoint union of half-open intervals, $\mu(G) = \mu_f(G)$ for every open interval G. This extends immediately to all open sets G. Now let H be any bounded set of type $\mathcal{G}_\delta$. Write $H = \bigcap_{n=1}^{\infty} G_n$, where $\{G_n\}$ is a decreasing sequence of bounded open sets. That the sequence $\{G_n\}$ can be chosen decreasing follows from the fact that the intersection of a finite number of open sets containing H is also an open set containing H. Since $\mu_f(G_n) = \mu(G_n)$ for every $n \in \mathbb{N}$, it follows from (ii) and Theorem 2.21, part (ii), that $\mu_f(H) = \mu(H)$. Thus μ_f and μ agree on all bounded sets of type $\mathcal{G}_\delta$. (We needed these sets to be bounded so that we could apply the limit theorem.)

Now let A be any bounded subset of $\mathbb{R}$. By (iii), there exist sets H_1 and H_2 of type $\mathcal{G}_\delta$ such that $H_1 \supset A$, $H_2 \supset A$, $\mu_f(H_1) = \mu_f^*(A)$, and $\mu(H_2) = \mu^*(A)$. Let $H = H_1 \cap H_2$. Then $A \subset H$. It follows that

$$\mu_f^*(A) = \mu(H) = \mu^*(A).$$

Finally, let A be any subset of $\mathbb{R}$. For $n \in \mathbb{N}$, let

$$A_n = A \cap [-n, n].$$

Then $\mu_f^*(A_n) = \mu^*(A_n)$. Since both μ_f^* and μ^* are regular outer measures, we obtain

$$\mu_f^*(A) = \lim_{n\to\infty} \mu_f^*(A_n) = \lim_{n\to\infty} \mu^*(A_n) = \mu^*(A)$$

from Exercise 2:10.2. ■

3.6.1 How regular are Borel measures?

We should add here a word about regularity of Borel measures. One might expect, given the nice approximation properties of Borel measures, that in any setting in which the Borel sets are measurable one would find a Borel regular measure. This is not the case; a Borel measure may behave quite weirdly on the non-Borel sets. Our next example gives such a construction that shows in particular that condition (iii) in Theorem 3.20 cannot be dropped.

Example 3.21: Let $(\mathbb{R}, \mathcal{M}, \mu)$ be an extension of Lebesgue measure λ to a σ-algebra larger than $\mathcal{L}$. (See Exercise 3:13.13.) Thus $\mathcal{L}$ is a proper subset of $\mathcal{M}$, and $\mu = \lambda$ on $\mathcal{L}$. Let $A \in \mathcal{M}$, and suppose that A is bounded, say $A \subset I = [a, b]$. Suppose further that A and $I \setminus A$ have Borel covers with respect to μ. Let H_1 and H_2 be such covers. Thus $A \subset H_1$, $I \setminus A \subset H_2$, $\mu(H_1) = \mu(A)$, and $\mu(H_2) = \mu(I \setminus A)$. We may assume that H_1 and H_2 are also λ^*-covers of A and $I \setminus A$, respectively, since we could intersect H_1 and H_2 with such Borel covers. Since $\mu = \lambda$ on $\mathcal{L}$,

$$\mu(A) = \mu(H_1) = \lambda(H_1) = \lambda^*(A)$$

and

$$\mu(I \setminus A) = \mu(H_2) = \lambda(H_2) = \lambda^*(I \setminus A).$$

Then

$$\mu(I) = \mu(A) + \mu(I \setminus A) = \lambda^*(A) + \lambda^*(I \setminus A).$$

We see from the regularity of λ^* that $A \in \mathcal{L}$. It follows that there are μ-measurable sets A without Borel covers: if $A \subset B \in \mathcal{B}$, then $\mu(B) > \mu(A)$.

We can apply this discussion to the converse part of Theorem 3.20 to show that the regularity condition (iii) cannot be dropped. Let us first apply the machinery of Theorem 2.45. We arrive at the complete measure space $(\mathbb{R}, \widehat{\mathcal{M}}, \hat{\mu})$. It is clear that $\hat{\mu}$ is a Borel measure that is finite on bounded Borel sets, but not every $A \in \widehat{\mathcal{M}}$ has a Borel cover with respect to $\hat{\mu}$. We show that there is no nondecreasing, right-continuous function f such that

$$\mu_f = \hat{\mu} \text{ on } \widehat{\mathcal{M}}. \tag{14}$$

Thus, for all such functions, $\mu_f^* \neq \mu^*$.

Suppose, by way of contradiction, that there is a function f so that $\mu_f = \hat{\mu}$ on $\widehat{\mathcal{M}}$. Since $\hat{\mu} = \lambda$ on $\mathcal{L}$, the function f must be of the form $f(x) = x + c$, $c \in \mathbb{R}$. Otherwise, there would be an interval $(a, b]$ such that

$$\mu_f((a, b]) = f(b) - f(a) \neq b - a = \lambda((a, b]).$$

It follows that μ_f is Lebesgue measure. But $\widehat{\mathcal{M}}$ contains sets that are not Lebesgue measurable, so μ_f is not defined on all of $\widehat{\mathcal{M}}$, contradicting (14).

3.6.2 A characterization of finite Borel measures on the real line

We do, however, have the following theorem that illustrates the generality of Lebesgue–Stieltjes measures. In particular, every finite Borel measure on $\mathbb{R}$ agrees with some Lebesgue–Stieltjes measure on the class of Borel sets. This is of interest in certain disciplines, such as probability, in which measure space models have finite measure. See Exercise 3:13.4 for an improvement of Theorem 3.22.

Theorem 3.22: *Let μ be a Borel measure on $\mathbb{R}$ with $\mu(B) < \infty$ for every bounded Borel set B. Then there exists a nondecreasing, right-continuous function f such that $\mu_f(B) = \mu(B)$ for every Borel set $B \subset \mathbb{R}$.*

Proof. We leave the proof as Exercise 3:6.1. ■

3.6.3 Measuring the growth of a continuous function on a set

Let us return to Theorem 3.20. From condition (iv) we see that

$$\mu_f((a, b]) = f(b) - f(a)$$

for every half-open interval $(a, b]$. If f is continuous, $\mu_f(\{x\}) = 0$ for all $x \in \mathbb{R}$ (see Exercise 3:5.8), and the four intervals with endpoints a and b have the same μ_f-measure. We can interpret that measure as the "growth" of f on the interval:

$$\mu_f(I) = \lambda(f(I)).$$

If one replaces the intervals by arbitrary sets E, one might expect $\mu_f^*(E) = \lambda^*(f(E))$; the outer measure of E is the amount of "growth" of f on E. This is, in fact, the case.

Theorem 3.23: *Let f be continuous and nondecreasing on $\mathbb{R}$, and let μ_f^* be the associated Lebesgue–Stieltjes outer measure. For every set $E \subset \mathbb{R}$,*

$$\mu_f^*(E) = \lambda^*(f(E)).$$

Proof. Let $E \subset \mathbb{R}$ and let $\varepsilon > 0$. Cover E with a sequence of intervals $\{(a_n, b_n]\}$ so that

$$\sum_{n=1}^{\infty}(f(b_n) - f(a_n)) \le \mu_f^*(E) + \varepsilon.$$

Let $J_n = f((a_n, b_n])$. Since f is continuous and nondecreasing, each interval J_n has endpoints $f(a_n)$ and $f(b_n)$. Now

$$f(E) \subset \bigcup_{n=1}^{\infty} J_n$$

so,

$$\lambda^*(f(E)) \le \lambda^*\left(\bigcup_{n=1}^{\infty} J_n\right) \le \sum_{n=1}^{\infty}(f(b_n) - f(a_n)) \le \mu_f^*(E) + \varepsilon.$$

Since ε is arbitrary,

$$\lambda^*(f(E)) \leq \mu_f^*(E). \tag{15}$$

To prove the reverse inequality, let G be an open set containing $f(E)$ so that

$$\lambda(G) \leq \lambda^*(f(E)) + \varepsilon.$$

Let $\{J_n\}$ be the sequence of open component intervals of G. For each $n \in \mathbb{N}$, let $I_n = f^{-1}(J_n)$. Since f is continuous, each I_n is open and, since f is nondecreasing, I_n is an interval. It is clear that $E \subset \bigcup_{n=1}^{\infty} I_n$. Thus, for $I_n = (a_n, b_n)$, we have

$$\mu_f^*(E) \leq \mu_f\left(\bigcup_{n=1}^{\infty} I_n\right) \leq \sum_{n=1}^{\infty} \mu_f(I_n)$$
$$= \sum_{n=1}^{\infty}(f(b_n) - f(a_n)) = \sum_{n=1}^{\infty} \lambda(J_n) = \lambda(G) \leq \lambda^*(f(E)) + \varepsilon.$$

Since ε is arbitrary,

$$\mu_f^*(E) \leq \lambda^*(f(E)). \tag{16}$$

The desired conclusion follows from (15) and (16). ■

The hypothesis that f be continuous is essential in the statement of Theorem 3.23. Exercise 3:6.4 provides a version that handles discontinuities.

Exercises

3:6.1 Prove Theorem 3.22. [Hint: Follow the proof of Theorem 3.20 to the point that a measurable cover of type $\mathcal{G}_\delta$ is not available.]

3:6.2 Give an example of a σ-finite measure μ on the Borel sets in $\mathbb{R}$ for which no Lebesgue–Stieltjes measure agrees with μ on the Borel sets. [Hint: Let $\mu(\{x\}) = 1$ for all $x \in \mathbb{Q}$.]

3:6.3 Show that there exists a measure space $(X, \mathcal{M}, \mu)$ with $\mu(X) < \infty$ and all Borel sets measurable, which also meets the following condition. There exists a measurable set M and an $\varepsilon > 0$ such that if G is open and $G \supset M$ then $\mu(G) > \mu(M) + \varepsilon$. Compare with Corollary 3.15. [Hint: See the discussion following Theorem 3.20.]

3:6.4 Let f be nondecreasing, and let μ_f denote its associated Lebesgue–Stieltjes measure.

(a) Prove that the set of atoms of μ_f is at most countable.

(b) Let A be the set of atoms of μ_f. Prove that, for every $E \subset X$,

$$\mu_f^*(E) = \lambda^*(f(E)) + \sum_{a \in A \cap E} \mu_f(\{a\}).$$

[Hint: See Exercise 3:5.8 and Theorem 3.23.]

3.7 Lebesgue–Stieltjes Measures in $\mathbb{R}^n$

We turn now to a brief, simplified discussion of Lebesgue–Stieltjes measures in n-dimensional Euclidean space $\mathbb{R}^n$. As before, we are interested in Borel measures that assume finite values on bounded sets.

For ease of exposition, we limit our discussion to the case $n = 2$. We wish to model a mass distribution or probability distribution on $\mathbb{R}^2$. As a further concession to simplification, let us assume finite total mass, all contained in the half-open square

$$T_0 = (0, 1] \times (0, 1] = \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_1 \leq 1,\ 0 < x_2 \leq 1\}.$$

Let $\mathcal{T}$ denote the family of half-open intervals $(a_1, b_1] \times (a_2, b_2]$ contained in T_0; that is, sets of the form

$$(a, b] = \{(x_1, x_2) : 0 < a_1 < x_1 \leq b_1 \leq 1,\ 0 < a_2 < x_2 \leq b_2 \leq 1\},$$

where $a = (a_1, a_2)$, $b = (b_1, b_2)$. Since $\emptyset = (a, a]$ for any $a \in T_0$, $\emptyset \in \mathcal{T}$.

Suppose now that for all $b \in T_0$ we know the mass "up to b"; more precisely, we have a function $f : T_0 \to \mathbb{R}$ such that $f(b)$ represents the mass of $(0, b]$. We wish to obtain τ from f as we did in the one-dimensional setting. This will provide a means of measuring our primitive notion of mass. Since two or more intervals can be pieced together to form a single interval, τ must be additive on such intervals. We achieve this in the following way. Let $T = (a, b] \in \mathcal{T}$. Two of the corners of T are $a = (a_1, a_2)$ and $b = (b_1, b_2)$. The other two corners are (a_1, b_2) and (b_1, a_2). Define a premeasure τ on the covering family $\mathcal{T}$ by

$$\tau(T) = f(b_1, b_2) - f(a_1, b_2) - f(b_1, a_2) + f(a_1, a_2). \tag{17}$$

Figure 3.2 illustrates.

We can now proceed as we did before. We extend τ to the algebra $\mathcal{T}_1$ generated by $\mathcal{T}$. This algebra consists of all finite unions of half-open intervals contained in T_0. We then extend τ to τ_1 by additivity and verify that τ_1 is actually σ-additive on $\mathcal{T}_1$. The ideas are the same as those in the one-dimensional case, but the details are messy. Method I leads to a metric outer measure μ_f^*, and each $A \subset T_0$ has a measurable cover of type $\mathcal{G}_\delta$. Furthermore, every interval $(a, b]$ is measurable, and

$$\mu_f((a, b]) = \tau((a, b]),$$

with $\tau((a, b])$ as given in (17).

In our preceding discussion, we chose the function f to satisfy our intuitive notion of "the mass up to b." Suppose that we turn the problem around. We ask which functions f can serve

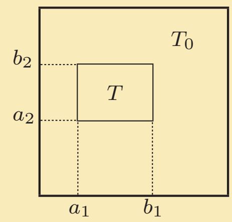


Figure 3.2. Define $\tau(T) = f(b_1, b_2) - f(a_1, b_2) - f(b_1, a_2) + f(a_1, a_2)$.

as such distributions. In the one-dimensional case, it sufficed to require that f be nondecreasing and right continuous. The monotonicity of f guaranteed that μ_f be nonnegative, and right continuity followed from Theorem 2.21 and the equality

$$(0, x] = \bigcap_{\delta>0} (0, x + \delta].$$

In the present setting, f must lead to $\tau(T) \geq 0$ in expression (17). This replaces the monotonicity requirement in the one-dimensional case. Right continuity is needed for the same reason that it is needed in one dimension. Here this means right continuity of f in each variable separately.

Exercises 3:7.1 to 3:7.3 provide illustrations of Lebesgue–Stieltjes measures on T_0.

Exercises

3:7.1 Let f be defined on T_0 by

$$f(x,y) = \begin{cases} x\sqrt{2}, & \text{for } y > x; \\ y\sqrt{2}, & \text{for } y \le x. \end{cases}$$

Let μ_f be the associated Lebesgue–Stieltjes measure. Prove that for every Borel set $B \subset T_0$,

$$\mu_f(B) = \lambda(B \cap L),$$

where L is the line with equation $y = x$ and λ is one-dimensional Lebesgue measure on L. Observe that f is continuous, yet certain closed rectangles with one side on L have larger measures than their interiors.

3:7.2 Let f be defined on T_0 by

$$f(x,y) = \begin{cases} x, & \text{if } y \ge \frac{1}{2}; \\ 0, & \text{if } y < \frac{1}{2}. \end{cases}$$

Let μ_f be the associated Lebesgue–Stieltjes measure. Show that μ_f represents a mass all of which is located on the line $y = \frac{1}{2}$.

3:7.3 Let f be defined on T_0 by

$$f(x,y) = \begin{cases} x+y, & \text{if } x+y < 1; \\ 1, & \text{if } x+y \ge 1. \end{cases}$$

Show f is increasing in each variable separately, but that the resulting τ takes on some negative values. In particular, $\tau(T_0) = -1$.

3.8 Hausdorff Measures and Hausdorff Dimension

The measures and dimensional concepts that we shall describe here go back to the work of Felix Hausdorff in 1919, based on earlier work of Carathéodory, who had developed a notion of "length" for sets in $\mathbb{R}^N$. In our language, the length of a set $E \subset \mathbb{R}^N$ will be its Hausdorff one-dimensional outer measure, $\mu^{*(1)}$. Considerable advances were made in the years following, particularly by A. S. Besicovitch and his students. In recent years, the subject has attracted a large number of researchers because of its fundamental importance in the study of fractal geometry. A development of this subject would take us too far afield. For such developments, we refer the reader to the many excellent recent books on the subject.[2] Here we give only an indication of how to construct the Hausdorff measures, how the dimensional ideas arise, and an indication of how the dimension of a set can provide a more delicate sense of the size of a set in $\mathbb{R}^N$ than Lebesgue measure provides.

Let us return once again to our illustration with squares in Section 3.2. This time, however, in anticipation of our needs, we change the covering family $\mathcal{T}$. We take $\mathcal{T}$ to consist of *all* open sets in $\mathbb{R}^2$, with $\tau(T) = \text{diameter}\,(T)$, the diameter of the set $T \in \mathcal{T}$. Method II gives rise to a metric outer measure μ_0^* such that $\mu_0^*(T) = \infty$ for all open squares $T \in \mathcal{T}$. This might have been expected, since diameter is a one-dimensional notion and open squares are two-dimensional.

Suppose that we take, instead, a different premeasure

$$\tau(T) = (\text{diameter}\,(T))^3$$

which is smaller for sets of diameter smaller than 1. Perhaps, now, Method II will give rise to an outer measure for which squares will have *zero* measure, a two-dimensional object being measured by a "three-dimensional" measure. Let T_0 be a square of unit diameter, and let m, $n \in \mathbb{N}$.

[2]For example, C. A. Rogers, *Hausdorff Measures,* Cambridge (1970) and K. J. Falconer, *The Geometry of Fractal Sets,* Cambridge (1985).

We cover T_0 with $(n+1)^2$ open squares T_i $(i = 1, 2, \ldots, (n+1)^2)$, each of diameter $1/n$, and find for all $m \leq n$ that

$$\mu_m^*(T_0) \leq \sum_{i=1}^{(n+1)^2} \tau(T_i) = \frac{(n+1)^2}{n^3}. \tag{18}$$

Consequently, each measure has $\mu_m^*(T_0) = 0$ and

$$\mu_0^*(T_0) = \lim_{m\to\infty} \mu_m^*(T_0) = 0.$$

The same is true of any open square. In fact, $\mu_0^*(\mathbb{R}^2) = 0$.

Consider now a further choice of premeasure

$$\tau(T) = (\text{diameter}\,(T))^2,$$

which is intermediate between the two preceding examples. A similar analysis shows that

$$\mu_m^*(T_0) \leq \frac{(n+1)^2}{n^2} \qquad (m \leq n),$$

so

$$\mu_0^*(T_0) \leq 1 = \tau(T_0) = 2\lambda_2(T_0), \tag{19}$$

where λ_2 denotes two-dimensional Lebesgue measure. On the other hand, if $T_0 \subset \bigcup_{k=1}^{\infty} T_k$ and $T_k \in \mathcal{T}_n$, then

$$\begin{aligned}\sum_{k=1}^{\infty} \tau(T_k) \;&=\; \sum_{k=1}^{\infty} (\text{diameter}\,(T_k))^2 \geq \sum_{k=1}^{\infty} \lambda_2(T_k)\\ &\geq\; \lambda_2\Big(\bigcup_{k=1}^{\infty} T_k\Big) \geq \lambda_2(T_0),\end{aligned}$$

the first inequality following from the fact that any set of finite diameter δ is contained in a square of the side length δ. It follows that $\lambda_2(T_0) \leq \mu_0^*(T_0)$. Combine this inequality with (19) and recognize that T_0 is μ_0^*-measurable to obtain

$$\lambda_2(T_0) \leq \mu_0(T_0) \leq 2\lambda_2(T_0).$$

Let us take a more general viewpoint. Let $\mathcal{T}$ consist of the open sets in $\mathbb{R}^N$. For each real $s > 0$, let

$$\tau(T) = (\text{diameter}\,(T))^s,$$

and let $\mu^{(s)}$ be the measure obtained from τ and $\mathcal{T}$ by Method II. A bit of reflection suggests several facts. In the space $\mathbb{R}^2$ ($N = 2$), we have

$$\mu^{*(s)}(T) = \begin{cases} 0, & \text{if } s > 2; \\ \infty, & \text{if } s < 2. \end{cases} \quad \text{for every } T \in \mathcal{T},$$

and

$$2 = \sup\{s : \mu^{(s)}(\mathbb{R}^2) = \infty\} = \inf\{s : \mu^{(s)}(\mathbb{R}^2) = 0\}.$$

Similarly, for arbitrary N, we have

$$N = \sup\{s : \mu^{(s)}(\mathbb{R}^N) = \infty\} = \inf\{s : \mu^{(s)}(\mathbb{R}^N) = 0\}.$$

The proofs of the last three assertions are not difficult. One can actually show that if λ_N is Lebesgue N-dimensional measure in $\mathbb{R}^N$ and if we use

$$\tau(T) = (\text{diameter}\,(T))^N$$

then $\mu^{(N)}$ is a multiple of λ_N, a multiple that depends on the dimension N. For example, in $\mathbb{R}^2$ ($N = 2$), this multiple can be proved to be $4/\pi$. In the special case on the real line $\mathbb{R}$ ($N = 1$), we are using as premeasure

$$\tau(T) = \text{diameter}\,(T),$$

which is just the length if T is an open interval. Method II reduces to Method I in this case and we have $\mu^{(1)} = \lambda$. Thus the multiple connecting Lebesgue one-dimensional measure and $\mu^{(1)}$ is 1.

3.8.1 Hausdorff dimension

These concepts can be extended to a more general setting and will allow us to define a notion of dimension for subsets of a metric space.

Definition 3.24: Let X be a metric space, let $\mathcal{T}$ denote the family of all open subsets of X, and let $s > 0$. Define a premeasure τ on $\mathcal{T}$ by

$$\tau(T) = (\text{diameter}\,(T))^s.$$

Then the outer measure $\mu^{*(s)}$ obtained from τ and $\mathcal{T}$ by Method II is called the *Hausdorff s-dimensional outer measure*, and the resulting measure $\mu^{(s)}$, the *Hausdorff s-dimensional measure.*

We know that $\mu^{*(s)}$ is a metric outer measure by Theorem 3.9 and that it is regular, with covers in $\mathcal{G}_\delta$ by Theorem 3.12. These measures are all translation invariant, since the premeasures are easily seen to be so. We could have taken $\mathcal{T} = 2^X$ in the definition, but our work in Section 3.2 indicates advantages to having $\mathcal{T}$ consist of open sets. Furthermore, for $E \subset X$, $s > 0$, and $\varepsilon > 0$, there exists an open set $G \supset E$ such that

$$(\text{diameter}\,(G))^s < (\text{diameter}\,(E))^s + \varepsilon.$$

It follows (Exercise 3:8.1) that the outer measures $\mu^{*(s)}$ that we obtain do not depend on whether we take for our covering family $\mathcal{T} = 2^X$ or $\mathcal{T} = \mathcal{G}$, the family of open sets in X.

Our first theorem shows that, in general, the behavior we have seen in $\mathbb{R}^N$ using $s = 1$, 2, 3 must occur. For any set $E \subset X$, there is a number s_0 so that for $s > s_0$ the assigned s-dimensional measure is zero, while for $s < s_0$ the s-dimensional measure is infinite.

Theorem 3.25: *If $\mu^{*(s)}(E) < \infty$ and $t > s$, then $\mu^{*(t)}(E) = 0$.*

Proof. Write $\delta(T)$ for diameter$\,(T)$, where T is any subset of our metric space X. Let $n \in \mathbb{N}$ and let $\{T_i\}$ be a sequence from $\mathcal{T}$ such that $E \subset \bigcup_{i=1}^{\infty} T_i$ and $\delta(T_i) \leq 1/n$, for all $i \in \mathbb{N}$. Then, for all $i \in \mathbb{N}$,

$$\frac{(\delta(T_i))^t}{(\delta(T_i))^s} = (\delta(T_i))^{t-s} \leq \left(\frac{1}{n}\right)^{t-s},$$

and

$$\mu_n^{*(t)}(E) \leq \sum_{i=1}^{\infty} (\delta(T_i))^t \leq \left(\frac{1}{n}\right)^{t-s} \sum_{i=1}^{\infty} (\delta(T_i))^s. \tag{20}$$

Since (20) is valid for every covering of E by sets in $\mathcal{T}_n$,

$$\mu_n^{*(t)}(E) \leq \left(\frac{1}{n}\right)^{t-s} \mu_n^{*(s)}(E).$$

Now let $n \to \infty$ to obtain $\lim_{n\to\infty} \mu_n^{*(t)}(E) = \mu^{*(t)}(E) = 0$. ∎

Note that this theorem shows that, for $s < 1$, $\mu^{(s)}$ is a Borel measure on $\mathbb{R}$ that assigns infinite measure to every open set. In fact, $\mu^{(s)}$ is not even σ-finite on $\mathbb{R}$ (Exercise 3:8.8). Thus we have an important example of regular Borel measures on $\mathbb{R}$ that are not Lebesgue–Stieltjes measures.

Theorem 3.25 justifies Definition 3.26.

Definition 3.26: Let E be a subset of a metric space X, and let $\mu^{*(s)}(E)$ denote the Hausdorff s-dimensional outer measure of E. If there is no value $s > 0$ for which $\mu^{*(s)}(E) = \infty$, then we let $\dim(E) = 0$. Otherwise, let

$$\dim(E) = \sup\{s : \mu^{*(s)}(E) = \infty\}.$$

Then $\dim(E)$ is called the *Hausdorff dimension* of E.

Suppose that K is a Cantor set, that is a nonempty, bounded nowhere dense perfect set in $\mathbb{R}$. It is possible that $\lambda(K) > 0$, in which case $\mu^{(1)}(K) = \lambda(K)$, but if $\lambda(K) = 0$, Lebesgue measure can contribute no additional information as to its size. Hausdorff dimension, however, provides a more delicate sense of size. Exercises 3:8.2 and 3:8.3 show that there exists Cantor sets in $[0,1]$ of dimension 1 and 0 respectively. Exercise 3:8.11 shows that the Cantor ternary set has dimension $\log 2/\log 3$. Moreover, one can show that for every $s \in [0,1]$ there exists a Cantor set of dimension s. If

$$\dim(K_1) = s_1 < s_2 = \dim(K_2),$$

then for $t \in (s_1, s_2)$, $\mu^{(t)}(K_1) = 0$, while $\mu^{(t)}(K_2) = \infty$. Thus the measure $\mu^{(t)}$ distinguishes between the sizes of K_1 and K_2.

3.8.2 Hausdorff dimension of a curve

Hausdorff dimension has an intuitive appeal when familiar objects are under consideration. We have noted, for example, that $\dim(\mathbb{R}^N) = n$. What about $\dim(\mathcal{C})$, where $\mathcal{C}$ is a curve, say in $\mathbb{R}^3$? Before we jump to the conclusion that $\dim(\mathcal{C}) = 1$, we should recall that there are curves in $\mathbb{R}^3$ that fill a cube.[3] Such curves must have dimension 3. And there are curves in $\mathbb{R}^2$, even graphs of continuous functions $f : \mathbb{R} \to \mathbb{R}$, that are of dimension strictly between 1 and 2. But for rectifiable curves, that is, curves of finite arc length, we have the expected result, which we present in Theorem 3.27.

First, we review a definition of the length of a curve. By a curve in a metric space (X, ρ), we mean a continuous function $f : [0,1] \to X$. The *length* of the curve is

$$\sup \sum_{i=1}^{m} \rho(f(x_{i-1}), f(x_i)),$$

where the supremum is taken over all partitions

$$0 = x_0 < x_1 < \cdots < x_m = 1$$

of $[0,1]$. The set of points $\mathcal{C} = f[0,1]$ is a subset of X, and it is the dimension of the set $\mathcal{C}$ that is our concern. The proof uses elementary knowledge of compact sets in metric spaces. The continuous image of a compact set is again compact. Also, the diameter of a compact set K is attained; that is, there are points x, $y \in K$ so that $\rho(x, y)$ is the diameter of K.

[3]See, for example, G. Edgar, *Measure, Topology and Fractal Geometry*, Springer (1990), for the construction of such curves.

Theorem 3.27: *Let $f : [0,1] \to X$ be a continuous, nonconstant curve in a metric space X, and let ℓ be its arc length. Then, for $\mathcal{C} = f([0,1])$,*

1. $0 < \mu^{(1)}(\mathcal{C}) \le \ell$.
2. *If f is one to one, then $\mu^{(1)}(\mathcal{C}) = \ell$.*

Thus, if $\ell < \infty$, $\dim(\mathcal{C}) = 1$.

Proof. Write $\delta(T)$ for diameter (T) for any set $T \subset X$. We prove first that $\mu^{(1)}(\mathcal{C}) \le \ell$. If $\ell = \infty$, there is nothing to prove, so that assume $\ell < \infty$. It is convenient here to use the result of Exercise 3:8.1 and to use coverings of $\mathcal{C}$ by arcs of $\mathcal{C}$. If $A_1, \dots, A_m$ is a collection of subarcs of $\mathcal{C}$ such that $\mathcal{C} = \bigcup_{i=1}^m A_i$, and $\delta(A_i) \le 1/n$, for all $i = 1, \dots, m$, then

$$\mu_n^{*(1)}(\mathcal{C}) \le \sum_{i=1}^m \delta(A_i). \tag{21}$$

We wish to relate the right side of (21) to the definition of ℓ.

First, let us obtain the arcs A_i formally. Let $n \in \mathbb{N}$. Since f is uniformly continuous, there exists $\gamma > 0$ such that

$$\rho(f(x), f(y)) < \frac{1}{n}$$

whenever $x, y \in [0,1]$ and $|x - y| < \gamma$. Let

$$0 = x_0 < x_1 \cdots < x_m = 1$$

be a partition of $[0,1]$ with $|x_i - x_{i-1}| < \gamma$ for all $i = 1, \dots, m$. Then the arcs $A_i = f([x_{i-1}, x_i])$

cover $\mathcal{C}$, and
$$\frac{1}{n} > \delta(A_i) \geq \rho(f(x_{i-1}), f(x_i))$$
for all $i = 1, \dots, m$. It follows from the compactness of $[x_{i-1}, x_i]$ that A_i is compact for each i. Thus the diameter of A_i is actually achieved by points $f(y_i)$ and $f(z_i)$, with
$$x_{i-1} \leq y_i \leq z_i \leq x_i.$$
This means that
$$\delta(A_i) = \rho(f(y_i), f(z_i)).$$
We now use the partition
$$0 \leq y_1 \leq z_1 \leq y_2 \leq z_2 \leq \cdots \leq y_m \leq z_m \leq 1$$
to obtain a lower estimate for ℓ. Continuing (21), we have
$$\mu_n^{*(1)}(\mathcal{C}) \leq \sum_{i=1}^{m} \delta(A_i) = \sum_{i=1}^{m} \rho(f(y_i), f(z_i)) \leq \ell. \tag{22}$$
Letting $n \to \infty$, we infer that
$$\mu^{*(1)}(\mathcal{C}) = \lim_{n\to\infty} \mu_n^{*(1)}(\mathcal{C}) \leq \ell.$$
That
$$\mu^{(1)}(\mathcal{C}) \leq \ell$$
now follows from the fact that $\mathcal{C}$ is $\mu^{*(1)}$-measurable. That $0 < \mu^{(1)}(\mathcal{C})$ follows from the fact that if $0 \leq a < b \leq 1$ then
$$\mu^{(1)}(f[a,b]) \geq \rho(f(a), f(b)). \tag{23}$$
(See Exercise 3:8.7.) This completes the proof of part (i).

Suppose now that f is one-one. Let

$$0 = x_0 < x_1 < \cdots < x_m = 1$$

be a partition of $[0,1]$, and note that the sets $f([x_{i-1}, x_i))$ are pairwise disjoint Borel sets. Thus, using (23) on each arc,

$$\begin{aligned}
\sum_{i=1}^{m} \rho(f(x_{i-1}), f(x_i)) &\leq \sum_{i=1}^{m} \mu^{(1)}(f([x_{i-1}, x_i))) \\
&= \mu^{(1)}\left(\bigcup_{i=1}^{m} f([x_{i-1}, x_i))\right) \\
&= \mu^{(1)}(f([0,1))) \\
&= \mu^{(1)}(f([0,1])) = \mu^{(1)}(\mathcal{C}).
\end{aligned}$$

This is valid for all partitions, and so $\ell \leq \mu^{(1)}(\mathcal{C})$. In view of part (i), $\ell = \mu^{(1)}(\mathcal{C})$. ■

3.8.3 Exceptional sets

We end this section with a comment about "exceptional sets". Consider the following statements about a nondecreasing function f defined on an interval I. Let

$$\begin{aligned}
D &= \{x : f \text{ is discontinuous at } x\}, \\
N &= \{x : f \text{ is nondifferentiable at } x\}, \\
N' &= \{x : f \text{ has no derivative, finite or infinite, at } x\}.
\end{aligned}$$

Then

1. D is countable.

2. $\lambda(N) = 0$.

3. $\mu^{(1)}(G) = 0$ where $G \subset \mathbb{R}^2$ consists of the points on the graph of f corresponding to points of continuity in N'.

Each of these statements indicates that a nondecreasing function has some desirable property outside some small exceptional set. The notion of smallness differs in these three statements. Observe that statement (iii) involves a subset of $\mathbb{R}^2$. The weaker statement, that $\lambda_2(G) = 0$, provides much less information than statement (iii). We shall prove a theorem corresponding to assertion (ii) later in Chapter 7.

We shall encounter a number of theorems involving exceptional sets. Cardinality and measure are only two of the many frameworks for expressing a sense in which a set may be small. The notion of first category set is another such framework; we study this intensively in Chapter 10. We mention another sense of smallness involving "porosity" in Exercise 7:8.12.

Exceptional sets of measure zero are encountered so frequently that we employ special terminology for dealing with them. Suppose that a function-theoretic property is valid except, perhaps, on a set of μ–measure zero. We then say that this property holds *almost everywhere*, or perhaps μ–*almost everywhere* or even *for almost all* members of X. This is frequently abbreviated as *a.e.* For example, statement (ii) above could be expressed as "f is differentiable a.e."

Exercises

3:8.1 Verify that, for all $s > 0$ and $E \subset X$, $\mu^{*(s)}(E)$ has the same value when $\mathcal{T} = \mathcal{G}$ as when $\mathcal{T} = 2^X$.

3:8.2 Let P be a Cantor set in $\mathbb{R}$ with $\lambda(P) > 0$. What is $\dim(P)$?

3:8.3 Construct a Cantor set in $\mathbb{R}$ of dimension zero. [Hint: Control the sizes of the intervals comprising the sets A_n in Example 2.1.]

3:8.4 Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is called a *Lipschitz function* if there exists $M > 0$ such that $|f(y)-f(x)| \leq M|y-x|$ for all $x, y \in \mathbb{R}$. Show that if f is a Lipschitz function then, for all $E \subset \mathbb{R}$, $\dim(f(E)) \leq \dim(E)$.

3:8.5 Show how to construct a set A in $\mathbb{R}$ such that $\lambda(A) = 0$, but $\dim(A) = 1$.

3:8.6 Give an example of a continuous curve $\mathcal{C}$ of finite length such that $\mu^{(1)}(\mathcal{C}) < \ell$.

3:8.7 Prove that if $f : [0,1] \to X$ is a continuous curve in X and $0 \leq a < b \leq 1$ then

$$\mu^{(1)}(f[a,b]) \geq \rho(f(a), f(b)).$$

[Hint: Define $g : f([a,b]) \to \mathbb{R}$ by $g(w) = \rho(f(a), w)$. Use g to obtain a comparison between $\mu^{(1)}(f[a,b])$ and the length of the interval $[0, \rho(f(a), f(b))]$.]

3:8.8 Show that, for $s < 1$, $(\mathbb{R}, \mathcal{B}, \mu^{*(s)})$ is not a σ-finite measure space.

3:8.9 Let $X = \mathbb{R}$ but supplied with the metric $\rho(x,y) = 1$ if $x \neq y$. What is the result of applying Method II to $\mathcal{T} = \mathcal{G}$, $\tau = \text{diameter}(T)$. (What are the families $\mathcal{T}_n$?)

3:8.10 Suppose that we were trying to measure the length of a hike. We count our steps, each of which is exactly 1 meter long, and arrive at a distance that we publish in our trail guide. A mouse does the same thing, but its steps are only 1 centimeter long. Since it must walk around rocks and other objects that we ignore, it will report a longer length. An insect's measurement would be still longer, and a germ, noticing every tiny undulation, would measure the distance as enormous. Probably, the actual distance along an ideal curve covering the trail is infinite. A better sense of "size of the trail" can be given by its Hausdorff dimension. Benoit Mandelbrot[4] discusses the differences in reported lengths of borders between countries. He also provides estimates of the dimensions (between 1 and 2) of these borders. Express our fanciful discussion of trail length in the more precise language of coverings, Hausdorff measures, and Hausdorff dimension.

3:8.11 Let K be the Cantor set, and let $s = \log 2/\log 3$. Cover K with 2^n intervals, each of length 3^{-n}. Show that
$$\mu^{*(s)}_{3^n}(K) \leq 1.$$
Show that these intervals are the most economical ones with which to cover K. Deduce that $\dim(K) = \log 2/\log 3$ and $\mu^{*(s)}(K) = 1$.

3.9 Methods III and IV

In applications of measure theory to analysis, one may need to construct an appropriate measure to serve as a tool in the investigation. We have already seen the usefulness of Methods I and II, both of which were developed by Carathéodory. In this section we extend this collection of methods by adopting a new approach, but one again built on the same theme of refining some crude premeasure into a useful outer measure. These methods can also be used to develop Lebesgue–Stieltjes and Hausdorff measures. We shall use them in Section 7.6 to construct total variation measures for arbitrary continuous functions.

We assume that X is a metric space and that a covering relation $\mathbb{H}$, in the sense of the following definition, has been specified.

Definition 3.28: By a *covering relation* on a metric space X we mean a collection of pairs (C, x) where $C \subset X$ and $x \in X$.

[4]B. Mandelbrot, *The Fractal Geometry of Nature,* W. H. Freeman and Co. (1982).

Most often in applications, when $(C, x) \in \mathbb{H}$ then the point x would be a member of C and it is in this sense that such collections have been named as "covering" relations. This covering relation $\mathbb{H}$ establishes a relationship under which points x are "attached" to certain selected sets C by the requirement that $(C, x) \in \mathbb{H}$. For example, a simple and useful such relation would be to take x is attached to C if $x \in C$; a slight variant would have $x \in \overline{C}$. If the sets in $\mathcal{T}$ are balls, then a useful version is to have x is attached to C to mean that C is centered at x. In general, the geometry and the application dictate how this can be interpreted. No special assumptions are needed on the relationship in general.

We suppose that a premeasure τ is defined on $\mathbb{H}$, i.e., $\tau : \mathbb{H} \to \mathbb{R}$ with

$$0 \leq \tau(C, x) \leq \infty \quad [(C, x) \in \mathbb{H}].$$

We assume, just as for Methods I and II, that there is no structure on τ, and we will require that $\tau(\emptyset, x) = 0$ if this happens to be defined (although it usually is not). As before, this crude premeasure will be refined into a genuine outer measure by some kind of approximation process.

Here, however, we shall use *packings* rather than coverings. The idea of a covering estimate, we recall, is to approximate the measure of a set E by some minimal covering of E using sets from a covering. Naturally, overlapping of sets would occur even in a good covering. For a packing, we allow no overlap.

Definition 3.29: A finite subset π of $\mathbb{H}$ is said to be a *packing* if

$$\pi = \{(I_i, x_i) : i = 1, 2, 3, \ldots, n\}$$

and $I_i \cap I_j = \emptyset$ for $i \neq j$.

Definition 3.30: Let τ be a premeasure on $\mathbb{H}$. Then the *variation* of τ on a packing $\pi \subset \mathbb{H}$ is defined as

$$V(\tau,\pi) = \sum_{i=1}^{n} \tau(C_i, x_i)$$

where

$$\pi = \{(I_i, x_i) : i = 1, 2, 3, \ldots, n\}$$

is the packing.

Definition 3.31: Let τ be a premeasure on $\mathbb{H}$. Then the *variation* of τ on a covering relation $\beta \subset \mathbb{H}$ is defined as

$$V(\tau,\beta) = \sup\{V(\tau,\pi) :\ \pi \subset \beta \text{ and } \pi \text{ is a packing}\}.$$

We shall find ways of using the variational estimates $V(\tau,\beta)$ to obtain our measures. The first step is to define the collection of covers that will play a role in the computation. (Recall the notation $B(x,\delta)$ for the open ball in the metric space X centered at x and with radius δ.)

Definition 3.32: Let $\mathbb{H}$ be a covering relation on a metric space X. Let $E \subset X$.

1. A family $\beta \subset \mathbb{H}$ is said to be a *full cover* of E (relative to $\mathbb{H}$) if for every $x \in E$ there is a $\delta > 0$ so that
$$(C, x) \in \mathbb{H} \ \text{ and } \ C \subset B(x, \delta) \ \Rightarrow \ (C, x) \in \beta.$$

2. A family $\beta \subset \mathbb{H}$ is said to be a *fine cover* of E (relative to $\mathbb{H}$) if, for every $x \in E$ and every $\varepsilon > 0$, either
$$\text{there exists at least one pair } (C, x) \in \beta \text{ for which } C \subset B(x, \varepsilon)$$
or else no such pair (C, x) exists in all of $\mathbb{H}$.

The fine covers are closely related to the notion of a Vitali cover in the literature (see Section 7.1). They play a key role in the study of differentiation of functions and integrals.

3.9.1 Constructing measures using full and fine covers

We now define our two methods of constructing outer measures.

Definition 3.33: Let $\mathbb{H}$ be a covering relation on a metric space X and τ a premeasure on $\mathbb{H}$. For every $E \subset X$, we define

1. $\tau^{\bullet}(E) = \inf\{V(\tau, \beta) : \ \beta \subset \mathbb{H} \text{ a full cover of } E\}$.
2. $\tau^{\circ}(E) = \inf\{V(\tau, \beta) : \ \beta \subset \mathbb{H} \text{ a fine cover of } E\}$.

The set functions $\tau^{\bullet}$ and τ° will be called the Method III and Method IV outer measures (respectively) generated by τ.

Theorem 3.34: *Let $\mathbb{H}$ be a covering relation on a metric space X and τ a premeasure on $\mathbb{H}$. Then $\tau^\bullet$ and τ° are metric outer measures on X and $\tau^\circ \le \tau^\bullet$.*

Proof. Most of the details of the proof are either elementary or routine. Here are two details that may not be seen immediately.

First, let us check the countable subadditivity of $\tau^\bullet$. Suppose that E is contained in a union $\bigcup_{n=1}^\infty E_n$ and that each $\tau^\bullet(E_n)$ is finite. Then for any $\varepsilon > 0$ and $n = 1, 2, 3, \dots$ there is a full cover $\beta_n \subset \mathbb{H}$ of E_n so that

$$V(\tau, \beta_n) \le \tau^\bullet(E_n) + \varepsilon 2^{-n}.$$

Observe that $\beta = \bigcup_{n=1}^\infty \beta_n$ is a full cover of E. Hence

$$\tau^\bullet(E) \le V(\tau, \beta) \le \sum_{n=1}^\infty V(\tau, \beta_n) \le \sum_{n=1}^\infty \left(\tau^\bullet(E_n) + \varepsilon 2^{-n}\right).$$

From this one sees that

$$\tau^\bullet(E) \le \sum_{n=1}^\infty \tau^\bullet(E_n).$$

Second, let us consider how to prove that $\tau^\bullet$ is a *metric* outer measure. Suppose that A, B are subsets of X a positive distance apart. Let β be a full cover of $A \cup B$ with

$$V(\tau, \beta) \le \tau^\bullet(A \cup B) + \varepsilon.$$

Because of this separation, we may choose two disjoint open sets G_1 and G_2 covering A and B, respectively. Consider the families

$$\beta_1 = \{(C, x) \in \beta : C \subset G_1\}$$

and

$$\beta_2 = \{(C,x) \in \beta : C \subset G_2\}.$$

Then β_1 is a full cover of A and β_2 is a full cover of B. If $(C_1, x_1) \in \beta_1$ and $(C_2, x_2) \in \beta_2$ then certainly the sets C_1 and C_2 are disjoint. This means that

$$\tau^\bullet(A) + \tau^\bullet(B) \leq V(\tau, \beta_1) + V(\tau, \beta_2) \leq V(\tau, \beta) \leq \tau^\bullet(A \cup B) + \varepsilon.$$

From this inequality and the subadditivity of $\tau^\bullet$ the identity,

$$\tau^\bullet(A \cup B) = \tau^\bullet(A) + \tau^\bullet(B)$$

can be readily obtained. The remaining details of the proof are left as exercises. ■

Example 3.35: Let $\mathbb{H}_0$ denote the set of all pairs $([u,v], w)$ where $[u,v]$ is a compact interval on the real line and w a point in $[u,v]$. This is a covering relation on $\mathbb{R}$; we take for $\ell([u,v],w)$ the length of the interval $[u,v]$, i.e.,

$$\ell([u,v],w) = v - u.$$

Then ℓ is a premeasure and it is possible to prove that

$$\ell^\circ = \ell^\bullet = \lambda^*.$$

This is known as the *Vitali covering theorem.* That is, both measures recover the Lebesgue outer measure. This will be discussed in greater detail in Section 7.6.

Example 3.36: Let $\mathbb{H}_0$ be as in Example 3.35 and define the premeasure τ_g by requiring that $\tau([u,v],w) = g(v) - g(u)$, where g is a continuous nondecreasing function on the real line. Then an extension of the Vitali covering theorem can be proved asserting that

$$\tau_g^\circ = \tau_g^\bullet = \mu_g^*.$$

That is, both measures recover the Lebesgue–Stieltjes outer measure μ_g^* generated by the monotonic function g. This too will be discussed in greater detail in Section 7.6.

Example 3.37: Again let $\mathbb{H}_0$ be as in Examples 3.35 and 3.36. For a premeasure take, $\tau_s([u,v],w) = (v-u)^s$, where $0 < s < 1$. Then τ_s° can be shown to be exactly the s-dimensional Hausdorff measure, and the larger measure $\tau_s^\bullet$ is indeed larger and plays a role in many investigations under the name "packing measure."

3.9.2 A regularity theorem

Here is a simple regularity theorem that illustrates some methods that can be used in the study of these measures. In any application, one would need to adjust the ideas to the geometry of the situation.

Theorem 3.38: *Let $\mathcal{C}$ be a collection of subsets of a metric space X and define the covering relation*

$$\mathbb{H} = \{(C,x) : C \in \mathcal{C} \text{ and } x \text{ is an interior point of } C\}.$$

Let τ be any premeasure on $\mathbb{H}$. Let $E \subset X$ with $\tau^\bullet(E) < \infty$ and let $\varepsilon > 0$. Then there is an $\mathcal{F}_\sigma$ set $C_1 \supset E$ and there is an $\mathcal{F}_{\sigma\delta}$ set $C_2 \supset E$ such that

$$\tau^\bullet(C_1) < \tau^\bullet(E) + \varepsilon \text{ and } \tau^\bullet(C_2) = \tau^\bullet(E).$$

Proof. There is a full cover $\beta \subset \mathbb{H}$ of E so that

$$V(\tau,\beta) < \tau^\bullet(E) + \varepsilon.$$

Choose $\delta(x) > 0$ for each $x \in E$ so that

$$C \in \mathcal{C},\ x \in \operatorname{int}(C), \text{ and } C \subset B(x,\delta) \Rightarrow (C,x) \in \beta.$$

Define
$$E_n = \{x \in E : \delta(x) > 1/n\}$$
and consider the closed sets $\{\overline{E_n}\}$. One checks, directly from the definition, that β is a full cover (relative to $\mathbb{H}$) of each set $\overline{E_n}$. Thus
$$\tau^\bullet\left(\overline{E_n}\right) \le V(\tau,\beta) < \tau^\bullet(E) + \varepsilon$$
and so also, since this is a metric outer measure,
$$\tau^\bullet\left(\bigcup_{n=1}^\infty \overline{E_n}\right) \le \tau^\bullet(E) + \varepsilon.$$
The set $C_1 = \bigcup_{n=1}^\infty \overline{E_n}$ is an $\mathcal{F}_\sigma$ set that contains E and affords our desired approximation to $\tau^\bullet(E)$. The set C_2 of the theorem can now be obtained by taking an intersection of an appropriate sequence of closed sets. ■

Exercises

3:9.1◊ Let $\mathbb{H}$ be a covering relation on a metric space X.

(a) Show that every full cover of a set is also a fine cover of that set.

(b) Let β be a full (fine) cover of E and suppose that G is an open set containing E. Then
$$\beta_1 = \{(C,x) \in \beta : C \subset G\}$$
is also a full (fine) cover of E.

(c) Let β_n be a full (fine) cover of E_n for each $n = 1, 2, 3, \dots$. Show that $\bigcup_{n=1}^\infty \beta_n$ is a full (fine) cover of $\bigcup_{n=1}^\infty E_n$.

(d) Suppose that β_1, β_2, ... are subsets of $\mathbb{H}$ and τ any premeasure. Show that

$$V\left(\tau, \bigcup_{n=1}^{\infty} \beta_n\right) \leq \sum_{n=1}^{\infty} V(\tau, \beta_n).$$

(e) If β_1 is a full cover of E and β_2 is a full cover of E then $\beta_1 \cap \beta_2$ is a full cover of E.

(f) If β_1 is a fine cover of E and β_2 is a full cover of E, then $\beta_1 \cap \beta_2$ is a fine cover of E.

(g) If β_1 is a fine cover of E and β_2 is a fine cover of E, then $\beta_1 \cap \beta_2$ need not be a fine cover of E.

3:9.2 Complete the proof of Theorem 3.34 by verifying that $\tau^\circ \leq \tau^\bullet$. [Hint: In the preceding exercise we checked that every full cover was also a fine cover.]

3:9.3$\diamondsuit$ Let $\mathbb{H}_r$ be the covering relation consisting of all pairs $([u, v], u)$ (u, $v \in \mathbb{R}$). Suppose that f is a real function. Show that the collection

$$\beta = \{([x, y], x) : f(y) - f(x) > c\}$$

is a full cover (relative to $\mathbb{H}_r$) of the set

$$\left\{x : \liminf_{y \to x+} [f(y) - f(x)] > c\right\}$$

and a fine cover of the (larger) set

$$\left\{x : \limsup_{y \to x+} [f(y) - f(x)] > c\right\}.$$

3:9.4$\diamondsuit$ Define *upper and lower derivates* for a function $F : \mathbb{R} \to \mathbb{R}$ as

$$\overline{D}F(x) = \inf_{\delta > 0} \sup \left\{ \frac{F(v) - F(u)}{v - u} : x \in [u, v],\ 0 < v - u < \delta \right\}$$

and

$$\underline{D}F(x) = \sup_{\delta>0} \inf \left\{ \frac{F(v)-F(u)}{v-u} : x \in [u,v],\ 0 < v-u < \delta \right\}$$

Let $\mathbb{H}_0$ be the covering relation consisting of all pairs $([u,v],w)$, u, $v \in \mathrm{I\!R}$, $u \leq w \leq v$. Let $\alpha \in \mathrm{I\!R}$, and let

$$\beta = \left\{ ([u,v],w) : \frac{F(v)-F(u)}{v-u} > \alpha,\ w \in [u,v] \subset [a,b] \right\}.$$

Prove that β is a full cover (relative to $\mathbb{H}_0$) of the set

$$E_1 = \{x \in (a,b) : \underline{D}F(x) > \alpha\}$$

and a fine cover of the larger set

$$E_2 = \{x \in (a,b) : \overline{D}F(x) > \alpha\}.$$

3:9.5 In the proof of Theorem 3.38, show in detail that β is a full cover of each set $\overline{E_n}$.

3.10 Mini-Vitali Theorem

Let us return to Example 3.35. The covering relation we used, $\mathbb{H}_0$, denotes the set of all pairs $([u,v],w)$ where $[u,v]$ is a compact interval on the real line and w a point in $[u,v]$. We employ the premeasure

$$\ell([u,v],w) = v-u \quad (u \leq w \leq v).$$

The choice of letter ℓ here suggests "length." The classical Vitali covering theorem asserts that

$$\ell^\circ = \ell^\bullet = \lambda^*.$$

Thus Lebesgue outer measure λ^* can be realized by Methods I and II using coverings, and equally well realized by Methods III and IV using packings from full and fine covers. We will discuss the Vitali covering theorem in depth in Section 7.1.

The Mini-Vitali theorem, rather easier to prove, is the same assertion about sets of measure zero. A set of Lebesgue measure can be characterized in three different ways, using coverings, using full covers, and using fine covers. In the language of the measures above, we shall prove that $\lambda(E) = 0$ if and only if either $\ell^{\circ}(E) = 0$ or $\ell^{\bullet}(E) = 0$. We state and prove this theorem now and use it, in Section 3.11 to obtain our first proof of the celebrated Lebesgue differentiation theorem.

Theorem 3.39 (Mini-Vitali Covering Theorem) *The following are three equivalent statements that assert that a set E of real numbers has Lebesgue measure zero:*

1. *For every $\varepsilon > 0$ there is an open set G containing E for which $\lambda(G) < \varepsilon$.*

2. *For every $\varepsilon > 0$ there is a full cover (relative to $\mathbb{H}_0$) β of E for which $V(\ell, \beta) < \varepsilon$.*

3. *For every $\varepsilon > 0$ there is a fine cover (relative to $\mathbb{H}_0$) β of E for which $V(\ell, \beta) < \varepsilon$.*

The proof follows after we establish some simple covering lemmas.

3.10.1 Covering lemmas

We begin with an elementary covering lemma for finite families of compact intervals on the real line. Recall that we are using throughout the fixed covering relation

$$\mathbb{H}_0 = \{([u, v], w) : u < v,\ u \leq w \leq v\}.$$

All statements about our covers concern subsets of $\mathbb{H}_0$.

Lemma 3.40: *Let β be a finite subset of the covering relation $\mathbb{H}_0$. Then there is a packing $\pi \subset \beta$,*

$$\pi = \{([c_i, d_i], e_i) : i = 1, 2, \ldots, m\},$$

for which[a]

$$\bigcup_{([u,v],w)\in\beta} [u, v] \subset \bigcup_{i=1}^{m} 3 * [c_i, d_i].$$

[a]By $3 * [u, v]$ we mean the interval centered at the same point as $[u, v]$ but with three times the length.

Proof. For $[c_1, d_1]$ simply choose the largest interval available. Note that $3 * [c_1, d_1]$ will then include any other interval $[u, v]$ for which $([u, v], w) \in \beta$ and for which $[u, v]$ intersects $[c_1, d_1]$. See Figure 3.3.

For $[c_2, d_2]$ choose the largest interval from among those that do not intersect $[c_1, d_1]$. Note that together $3 * [c_1, d_1]$ and $3 * [c_2, d_2]$ include any interval of the family that intersects either $[c_1, d_1]$ or $[c_2, d_2]$. Continue inductively, choosing $([c_{k+1}, d_{k+1}], e_{k+1}) \in \beta$ so that $[c_{k+1}, d_{k+1}]$ is the largest interval available that does not intersect one the previously chosen intervals $[c_1, d_1]$, $[c_2, d_2]$, …, $[c_k, d_k]$. Stop when you run out of intervals-pairs to select. ■

Our second covering lemma is nearly as elementary, and is just an observation about the structure of open sets.

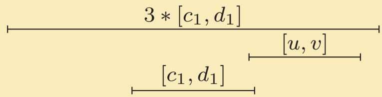


Figure 3.3. Note that $3 * [c_1, d_1]$ will include any shorter interval $[u, v]$ that intersects $[c_1, d_1]$.

Lemma 3.41: *Let β be any subset of $\mathbb{H}_0$. Then the set*

$$G = \bigcup_{([u,v],w)\in\beta} (u, v)$$

is an open set that contains all but countably many points of the set

$$E = \bigcup_{([u,v],w)\in\beta} [u, v].$$

Proof. Certainly G is open, since it is a union of a family of open intervals. Any point that is in E but not in G must be an endpoint of a component interval of G. For example if $a \in E$ but not in G then there must be an element $([a, b], c)$ or $([b, a], c)$ in β. In the former case $(a, b) \subset G$ but $[a, b) \not\subset G$. In the latter case $(a, b) \subset G$ but $(a, b] \not\subset G$. But the collection of endpoints of the component intervals of G is countable. Consequently $E \setminus G$ is countable. ■

Our next lemma is the key to the argument for our proof of the mini version of the Vitali covering theorem.

Lemma 3.42: *Let $\beta \subset \mathbb{H}_0$. Write*

$$G = \bigcup_{([u,v],w\in\beta} (u,v).$$

Then G is an open set and, if $g = \lambda(G)$ is finite, then there must exist a packing $\pi \subset \beta$,

$$\pi = \{([x_k, y_k], z_k) : k = 1, 2, \dots, p\}$$

for which

$$\sum_{k=1}^{p} (y_k - x_k) \geq g/6. \tag{24}$$

In particular

$$G' = G \setminus \bigcup_{([u,v],w)\in\pi} [u,v]$$

is an open subset of G and $\lambda(G') \leq 5g/6$.

Proof. It is clear that the set G of the lemma, expressed as the union of a family of open intervals, must be an open set. Let $\{(a_i, b_i)\}$ be the sequence of component intervals of G. We know then that the Lebesgue measure of G must be

$$g = \lambda(G) = \sum_{i=1}^{\infty} (b_i - a_i).$$

Choose an integer N large enough that

$$\sum_{i=1}^{N} (b_i - a_i) > 3g/4.$$

Inside each open interval (a_i, b_i), for $i = 1, 2, \ldots, N$, choose a compact interval $[c_i, d_i]$ so that

$$\sum_{i=1}^{N}(d_i - c_i) > g/2.$$

Write

$$K = \bigcup_{i=1}^{N}[c_i, d_i]$$

and note that K is a compact set covered by the family

$$\{(u, v) : ([u, v], w) \in \beta\}.$$

Consequently there must be, by the Heine-Borel theorem, a finite subset

$$\beta_1 = \{([u_1, v_1], w_1),\ ([u_2, v_2], w_2),\ ([u_3, v_3], w_3),\ \ldots, ([u_m, v_m], w_m)\}$$

from β for which

$$K \subset \bigcup_{i=1}^{m}(u_i, v_i).$$

By the simple covering Lemma 3.40 there is a packing $\pi \subset \beta_1$,

$$\pi = \{([x_k, y_k], z_k) : k = 1, 2, \ldots, p\}$$

for which

$$\bigcup_{i=1}^{N}[c_i, d_i] \subset \bigcup_{i=1}^{m}(u_i, v_i) \subset \bigcup_{k=1}^{p} 3 * [x_k, y_k].$$

Thus

$$\sum_{k=1}^{p} 3(y_k - x_k) \geq \sum_{i=1}^{N} (d_i - c_i) > g/2.$$

Statement (24) and the estimate on $\lambda(G')$ then follow. ■

3.10.2 Proof of the Mini-Vitali covering theorem

We use these elementary covering lemmas now to complete our proof of Theorem 3.39. To sort out the three concepts being compared in the statement of the theorem define:

1. A set E is of *measure zero* if, for every $\varepsilon > 0$, there is an open set $G \supset E$ for which $\lambda(G) < \varepsilon$.

2. E is *full null* if for every $\varepsilon > 0$ there is a full cover $\beta \subset \mathbb{H}_0$ for which $V(\ell, \beta) < \varepsilon$.

3. E is *fine null* if for every $\varepsilon > 0$ there is a fine cover $\beta \subset \mathbb{H}_0$ for which $V(\ell, \beta) < \varepsilon$.

Every full null set is clearly a fine null set; this is because every full cover is also a fine cover. Every set of measure zero is a full null set by a simple covering argument. Let $\varepsilon > 0$ and, supposing that E is a set of measure zero, choose an open set G containing E for which $\lambda(G) < \varepsilon/2$. Then

$$\beta = \{([u,v], w) : w \in E,\ [u,v] \subset G\}$$

is a full cover (relative to $\mathbb{H}_0$) of E. If π is any packing contained in β, then certainly

$$V(\ell, \pi) = \sum_{([u,v],w\in\pi} (v - u) \leq \lambda(G) < \varepsilon/2.$$

Consequently $V(\ell,\beta) \le \varepsilon/2 < \varepsilon$. Hence E is full null.

To complete the proof we show that every fine null set is a set of measure zero. Let us suppose that E is not a set of measure zero. We show that it is not fine full then. Define

$$\varepsilon_0 = \inf\{\lambda(G) : G \text{ open and } G \supset E\}.$$

Since E is not measure zero, $\varepsilon_0 > 0$.

Let β be an arbitrary fine cover of E. Define

$$G = \bigcup_{([u,v],w)\in\beta} (u,v).$$

This is an open set and, by Lemma 3.41, G covers all of E except for a countable set. It follows that $\lambda(G) \ge \varepsilon_0$, since if $\lambda(G) < \varepsilon_0$ we could add to G a small open set G' that contains the missing countable set of points and for which the combined set $G \cup G'$ is an open set containing E but with measure smaller than ε_0.

By Lemma 3.42 there must exist a packing $\pi \subset \beta$ for which

$$\sum_{([u,v],w)\in\pi} (v-u) \ge \varepsilon_0/6. \tag{25}$$

In particular $V(\ell,\beta) \ge \varepsilon_0/6$. But that means that E is not a fine null set, since this is true for *every* fine cover β.

3.11 Lebesgue differentiation theorem

Using the mini-Vitali theorem, we can prove that functions of bounded variation on the real line are differentiable at every point excepting possibly a set of Lebesgue measure zero. We will return to this theorem in Section 7.2 and give, at that time, another rather more revealing proof that uses the full Vitali covering theorem. Here we shall need only the mini-version.

Definition 3.43: The *total variation* of a function $F : [a,b] \to \mathbb{R}$ on that interval is the number $V(F;[a,b])$ defined as the supremum of the values

$$\sum_{i=1}^{n} |F(s_i) - F(s_{i-1})|$$

taken over all choices of points

$$a = s_0 < s_1 < \cdots < s_{n-1} < s_n = b.$$

Definition 3.44: A function $F : [a,b] \to \mathbb{R}$ is said to have *bounded variation* on $[a,b]$ provided that $V(F;[a,b]) < \infty$.

Note that, should F be monotonic on $[a,b]$ then

$$V(F;[a,b]) = |F(b) - F(a)|.$$

Thus all monotonic functions have bounded variation.

Theorem 3.45 (Lebesgue differentiation theorem) *Suppose that $F : [a,b] \to \mathbb{R}$ is a function of bounded variation. Then F is differentiable at almost every point in (a,b).*

Corollary 3.46: *Let $F : [a,b] \to \mathbb{R}$ be a monotonic function. Then F is differentiable at almost every point in (a,b).*

The proof uses the *upper and lower derivates.* To analyze how a derivative $F'(x)$ may fail to exist we split that failure into two pieces, an upper and a lower derivative, defined already in Exercise 3:9.4. We will show that, for almost every point x in (a,b),

$$\overline{D}F(x) > -\infty, \qquad \underline{D}F(x) < \infty,$$

and

$$\overline{D}F(x) = \underline{D}F(x).$$

From these three assertions it follows that F has a finite derivative $F'(x)$ at almost every point x in (a,b).

3.11.1 A geometrical lemma

The proof employs an elementary geometric lemma that Donald Austin[5] used in 1965 to give a simple proof of this theorem. Our proof of the differentiation theorem is essentially his, but written in different language. See also the version of Michael Botsko[6].

[5]D. Austin, *A geometric proof of the Lebesgue differentiation theorem.* Proc. Amer. Math. Soc. 16 (1965) 220–221.

[6]M. W. Botsko, *An elementary proof of Lebesgue's differentiation theorem.* Amer. Math. Monthly 110 (2003), no. 9, 834–838.

Lemma 3.47 (Austin's lemma) *Let $G : [a,b] \to I\!R$, $\alpha > 0$ and suppose that $G(a) \leq G(b)$. Let*
$$\beta = \left\{ ([u,v],w) : \frac{G(v)-G(u)}{v-u} < -\alpha,\ w \in [u,v] \subset [a,b] \right\}.$$
Then, for any nonempty packing $\pi \subset \beta$,
$$\alpha \left(\sum_{([u,v],w)\in\pi} (v-u) \right) < V(G;[a,b]) - |G(b)-G(a)|.$$

Proof. To prove the lemma, let π_1 be a partition of $[a,b]$ that contains the packing π. By a *partition* we mean a finite collection of interval-point pairs $\{([c_i,d_i],e_i)\}$ with nonoverlapping (not disjoint) intervals that exhausts the interval $[a,b]$. This is clearly possible.

Now write
$$|G(b)-G(a)| = G(b)-G(a) = \sum_{([u,v],w)\in\pi_1} [G(v)-G(u)]$$
$$= \sum_{([u,v],w)\in\pi} [G(v)-G(u)] + \sum_{([u,v],w)\in\pi_1\setminus\pi} [G(v)-G(u)]$$
$$< -\alpha \left(\sum_{([u,v],w)\in\pi} [v-u] \right) + V(G;[a,b]).$$
The statement of the lemma follows. ■

As a corollary we can replace G with $-G$ to obtain a similar statement.

Corollary 3.48: *Let $G : [a,b] \to I\!\!R$, $\alpha > 0$ and suppose that $G(b) \le G(a)$. Let*

$$\beta = \left\{([u,v],w) : \frac{G(v)-G(u)}{v-u} > \alpha,\ w \in [u,v] \subset [a,b]\right\}.$$

Then, for any nonempty packing $\pi \subset \beta$,

$$\alpha \left(\sum_{([u,v],w)\in\pi} (v-u)\right) < V(G;[a,b]) - |G(b)-G(a)|.$$

3.11.2 Proof of the Lebesgue differentiation theorem

We now prove Theorem 3.45. The first step in the proof is to show that, at almost every point t in (a,b),

$$\underline{D}F(t) = \overline{D}F(t).$$

If this is not true then there must exist a pair of rational numbers r and s for which the set

$$E_{rs} = \{t \in (a,b) : \underline{D}F(t) < r < s < \overline{D}F(t)\}$$

is not a set of measure zero. This is because the union of the countable collection of sets E_{rs} contains all points t for which $\underline{D}F(t) \neq \overline{D}F(t)$.

Let us show that each such set E_{rs} is fine null in the language of the Mini-Vitali theorem; we then know that E_{rs} is a set of Lebesgue measure zero. Write $\alpha = (s-r)/2$, $B = (r+s)/2$, $G(t) = F(t) - Bt$. Note that

$$E_{rs} = \{t \in (a,b) : \underline{D}G(t) < -\alpha < 0 < \alpha < \overline{D}G(t)\}.$$

Since F has bounded variation on $[a,b]$, so too does the function G. In fact
$$V(G;[a,b]) \leq V(F;[a,b]) + B(b-a).$$
Let $\varepsilon > 0$ and select points
$$a = s_0 < s_1 < \cdots < s_{n-1} < s_n = b$$
so that
$$\sum_{i=1}^{n} |G(s_i) - G(s_{i-1})| > V(G;[a,b]) - \alpha\varepsilon.$$

Let $E'_{rs} = E_{rs} \setminus \{s_1, s_2, \ldots, s_{n-1}\}$. Let us call an interval $[s_{i-1}, s_i]$ *black* if $G(s_i) - G(s_{i-1}) \geq 0$ and call it *red* if $G(s_i) - G(s_{i-1}) < 0$.

For each $i = 1, 2, 3, \ldots, n$ we define a covering relation β_i as follows. If $[s_{i-1}, s_i]$ is a black interval then
$$\beta_i = \left\{ ([u,v],w) : \frac{G(v) - G(u)}{v-u} < -\alpha,\ w \in [u,v] \subset [s_{i-1}, s_i] \right\}.$$
If, instead, $[s_{i-1}, s_i]$ is a red interval then
$$\beta_i = \left\{ ([u,v],w) : \frac{G(v) - G(u)}{v-u} > \alpha,\ w \in [u,v] \subset [s_{i-1}, s_i] \right\}.$$
Let $\beta = \bigcup_{i=1}^{n} \beta_i$. It is easy to check that this collection β is a fine cover of E'_{rs} (cf. Exercise 3:9.4).

Let π be any nonempty packing contained in β. Write $\pi_i = \pi \cap \beta_i$. By Lemma 3.47 applied to the black intervals and Corollary 3.48 applied to the red intervals we obtain that
$$\alpha \left(\sum_{([u,v],w) \in \pi_i} (v-u) \right) < V(G;[s_{i-1}, s_i]) - |G(s_i) - G(s_{i-1})|.$$

Consequently

$$\alpha\left(\sum_{([u,v],w)\in\pi}(v-u)\right)=\alpha\left(\sum_{i=1}^{n}\sum_{([u,v],w)\in\pi_i}(v-u)\right)$$
$$\leq\sum_{i=1}^{n}V(G;[s_{i-1},s_i])-\sum_{i=1}^{n}|G(s_i-G(s_{i-1})|$$
$$\leq V(G;[a,b])-[V(G;[a,b])-\alpha\varepsilon]=\alpha\varepsilon.$$

We have proved that β is a fine cover of E'_{rs} with the property that

$$\sum_{([u,v],w)\in\pi}(v-u)<\varepsilon$$

for every packing $\pi\subset\beta$. In the earlier language we have shown that $V(\ell,\beta)\leq\varepsilon$. It follows that E'_{rs} is fine null, and hence a set of Lebesgue measure zero. So too then is E_{rs} since the two sets differ by only a finite number of points.

We know now that the function F has a derivative, finite or infinite, almost everywhere in (a,b). We wish to exclude the possibility of the infinite derivative, except on a set of measure zero.

Let

$$E_\infty=\{t\in(a,b):\underline{D}F(t)=\infty\}.$$

Choose any B so that $F(b)-F(a)\leq B(b-a)$ and set $G(t)=F(t)-Bt$. Note that $G(b)\leq G(a)$ which will allow us to apply Corollary 3.48.

Let $\varepsilon>0$ and choose a positive number α large enough so that

$$V(G;[a,b])-|G(b)-G(a)|<\alpha\varepsilon.$$

Define

$$\beta = \left\{ ([u,v],w) : \frac{G(v)-G(u)}{v-u} > \alpha, [u,v] \subset [a,b] \right\}.$$

This is a fine cover of E_∞. Let π be any packing $\pi \subset \beta$. By our corollary then

$$\alpha \sum_{([u,v],w)\in\pi} (v-u) < V(G;[a,b]) - |G(b)-G(a)| < \alpha\varepsilon.$$

We have proved that β is a fine cover of E_∞ with the property that

$$\sum_{([u,v],w)\in\pi_i} (v-u) < \varepsilon$$

for every packing $\pi \subset \beta$. It follows that E_∞ is fine null, and hence a set of measure zero. The same arguments will handle the set

$$E_{-\infty} = \{t \in (a,b) : \overline{D}F(t) = -\infty\}.$$

Exercises

3:11.1 Suppose that F, F_1 and F_2 are real-valued functions defined on an interval $[a,b]$.

(a) Compute $V(F;[a,b])$ if F is monotonic on $[a,b]$.

(b) Estimate $V(F_1+F_2;[a,b])$.

(c) Estimate $V(rF_1+sF_2;[a,b])$.

(d) Estimate $V(F_1 \cdot F_2;[a,b])$.

3:11.2 Compute $V(F;[0,1])$ where F is given by the formula $F(x) = x\sin(1/x)$.

3:11.3 Show that $V(F;[0,1]) < \infty$ if F is the continuous function given by the formula $F(x) = x^2\sin(1/x)$.

3:11.4 Show that every function that has bounded variation on an interval is bounded there.

3:11.5 Let $\{F_k\}$ be a sequence of functions on a compact interval $[a,b]$ such that

$$\sup_k V(F_k,[a,b]) < \infty.$$

If $F(x) = \lim_{k\to\infty} F_k(x)$ for all x in $[a,b]$ show that F has bounded variation on $[a,b]$.

3:11.6 Give an example of a sequence of functions $\{F_k\}$ such that $V(F_k;[a,b]) < \infty$ for each k and for which $F(x) = \lim_{k\to\infty} F_k(x)$ exists at every point, but for which F does not have bounded variation on $[a,b]$.

3.12 Additional Remarks on Special Sets

We end this chapter with some additional remarks concerning monotonic functions, Cantor sets, and nonatomic measures. Any subset of the real line that is nonempty, bounded, perfect, and nowhere dense is said to be a *Cantor set*.[7]

3.12.1 Cantor sets

We have already discussed Cantor-like functions in Exercise 1:22.13. These are continuous, nondecreasing functions that map a Cantor set onto an interval. Speaking loosely, we can say that Cantor functions do all their rising on a Cantor set.

Our first theorem gives an indication of the role of Cantor sets in the rising of a nondecreasing function.

[7]More generally, a set K in a metric space X is said to be a Cantor set if K is homeomorphic to the classical Cantor set (cf. Exercise 3:12.2).

Theorem 3.49: *Let $A \subset [0,1]$, and let $f : A \to I\!R$ be a nondecreasing function. If $\lambda_*(f(A)) > 0$, then A contains a Cantor set.*

Proof. We may assume that f is bounded on A. Otherwise, we do our work on an appropriate smaller interval I. We begin by extending f to a nondecreasing function $\hat{f}$ defined on all of $[0,1]$. Let

$$\hat{f}(x) = \begin{cases} \inf\ f, & \text{for } 0 \le x \le \inf\ A; \\ \sup\{f(t) : t \in A, t \le x\}, & \text{for } \inf\ A < x \le 1. \end{cases}$$

Then $\hat{f}$ is nondecreasing on $[0,1]$.

Our objective is to find a Cantor set P of positive measure such that $P \subset f(A)$ and f^{-1} maps P homeomorphically into A. To do this, we first remove from consideration any points of discontinuity of $\hat{f}$, as well as any intervals on which $\hat{f}$ is constant. Since $\hat{f}$ is nondecreasing, its set D of points of discontinuity is countable. Thus

$$\lambda(\hat{f}(D)) = 0. \tag{26}$$

Now, for each $y \in f(A)$, the set $\hat{f}^{-1}(y)$ is an interval, since $\hat{f}$ is nondecreasing. Let $\mathcal{I}$ be the family of such intervals that are not degenerate. The intervals in $\mathcal{I}$ are pairwise disjoint and each has positive length. Thus $\mathcal{I}$ is countable, say $\mathcal{I} = \{I_k\}$. Let $G = \bigcup_{k=1}^{\infty} I_k$. Since $\hat{f}$ is constant on each member of $\mathcal{I}$, $\hat{f}(G)$ is countable and

$$\lambda(\hat{f}(G)) = 0. \tag{27}$$

Let $M = f(A) \setminus \hat{f}(D \cup G)$. It follows from (26) and (27) that $\lambda_*(M) > 0$. Let $y \in M$. There exists $x \in A$ such that $f(x) = y$. We see from the definition of the set M that

$$\hat{f}(t) < y \text{ for } t < x \quad \text{and} \quad \hat{f}(t) > y \text{ for } t > x.$$

Thus $\hat{f}^{-1}(y) = \{x\}$. It follows that $\hat{f}^{-1}$ is strictly increasing on the set M, and $\hat{f}^{-1}(M) \subset A$. Note that, since $M \subset f(A)$ and $\hat{f}^{-1}(M) \subset A$, $\hat{f}^{-1} = f^{-1}$ on M.

The set E of points of discontinuity of $f^{-1} : M \to A$ is countable. Thus there is a Cantor set P of positive measure contained in $M \setminus E$. Since f^{-1} is continuous and strictly increasing on P, the set $K = f^{-1}(P)$ is also a Cantor set (see Exercise 3:12.1), and K is a subset of A. It is clear that f maps the Cantor set K onto the set P of positive measure. ■

Exercise 3:13.14 at the end of this chapter shows that we cannot replace the monotonicity hypothesis with one of continuity in Theorem 3.49.

3.12.2 Bernstein sets

We observed in Section 2.1 how nineteenth century misconceptions about nowhere dense subsets of $\mathbb{R}$ may have delayed the development of measure theory. Cantor sets were not part of the mathematical repertoire until late in the nineteenth century. Nowadays, Cantor sets appear in diverse areas of mathematics. Our familiarity with them makes it difficult to visualize an uncountable set that does not contain a Cantor set, though this is, in fact, possible. We have earlier (e.g., Exercises 1:22.7 and 1:22.8) discussed totally imperfect sets; that is, an uncountable set of real numbers that contains no Cantor set. We have shown the existence of Bernstein sets (a set such that neither it nor its complement contains a Cantor set). The existence can be obtained by a cardinality argument (which is especially simple under the continuum hypothesis).

Bernstein sets have a number of interesting properties relative to Lebesgue measure and Lebesgue–Stieltjes measures. Let f be continuous and nondecreasing on $[0, 1]$, with $f([0, 1]) =$

$[0,1]$. Suppose that neither A nor its complement $\widetilde{A}$ contains a Cantor set. Then

$$\lambda_*(A) = \lambda_*(\widetilde{A}) = 0.$$

It follows that

$$\lambda^*(A) = \lambda^*(\widetilde{A}) = 1.$$

Now $f(A) \cup f(\widetilde{A}) = [0,1]$. By Theorem 3.49,

$$\lambda_*(f(A)) = \lambda_*(f(\widetilde{A})) = 0.$$

Thus

$$\lambda^*(f(A)) = \lambda^*(f(\widetilde{A})) = 1$$

and the set A cannot be measurable with respect to any nonatomic Lebesgue–Stieltjes measure except the zero measure. We know, by Exercise 3:13.13, that there are extensions $\overline{\overline{\lambda}}$ of λ for which the set A is $\overline{\overline{\lambda}}$-measurable. Similarly, there are extensions $\overline{\overline{\mu}}_f$ of any given Lebesgue–Stieltjes measure for which A is $\overline{\overline{\mu}}_f$-measurable. But such extensions are not Lebesgue–Stieltjes measures. See the discussion following the proof of Theorem 3.20.

Arguments similar to the ones we have given show that if A is totally imperfect then, for every nonatomic Lebesgue–Stieltjes measure μ_f, either $\mu_f(A) = 0$ or A is not μ_f-measurable. Which alternative applies depends on whether $\lambda(f(A)) = 0$ or $\lambda^*(f(A)) > 0$.

3.12.3 Lusin sets

We turn now to the opposite phenomenon. Are there sets that are measurable with respect to *every* nonatomic Lebesgue–Stieltjes measure? Since Lebesgue–Stieltjes measures are Borel measures, the question should be asked about non-Borel sets.

To address this question, we construct another example of an unusual set of real numbers (cf. Exercise 1:22.9), called occasionally a Lusin set.

Lemma 3.50: *Assuming the continuum hypothesis, there exists a set X of real numbers such that X has cardinality c, yet every nowhere dense subset of X is countable.*

Proof. We shall construct a set $X \subset [0,1]$ so that, if A is a nowhere dense subset of the space X using the Euclidean metric, then A is countable. To construct the set X, arrange the nowhere dense closed subsets of [0,1] into a transfinite sequence $\{F_\alpha\}$, $0 \le \alpha < \Omega$, where Ω is the first uncountable ordinal. For each $\alpha < \Omega$, consider the difference

$$F_\alpha \setminus \bigcup_{\beta<\alpha} F_\beta.$$

Since the interval [0,1] is complete, uncountably many of these differences must be nonempty. Let X be a set that contains exactly one point from each such difference. Then X has cardinality c.

We now show that if N is a nowhere dense subset of [0,1] then $N \cap X$ is countable. Since $\overline{N}$ is also nowhere dense in [0,1], there exists $\alpha < \Omega$ such that $\overline{N} = F_\alpha$. The construction of X implies that, for $\gamma > \alpha$, $X \cap F_\gamma \cap F_\alpha = \emptyset$. Thus

$$\overline{N} \cap X \subset \bigcup_{\beta\le\alpha} F_\beta,$$

so $\overline{N} \cap X$ is countable. The same is true of $N \cap X$. Since any set that is nowhere dense in X is also nowhere dense in [0,1], we infer that every nowhere dense subset of X is countable. ■

For this space X, we have the following.

Theorem 3.51: *The space X has the following properties.*

1. *The only finite nonatomic Borel measure μ on X is the zero measure.*

2. *Any nondecreasing function f on X maps X onto a set of measure zero.*

3. *For every nonatomic Lebesgue–Stieltjes measure μ_f on the real line, X is μ_f-measurable and $\mu_f(X) = 0$.*

Proof. Let D be a countable dense subset of X, and let $\varepsilon > 0$. Since μ is nonatomic, $\mu(D) = 0$. By Corollary 3.15, there exists an open set G containing D such that $\mu(G) < \varepsilon$. The set G is a dense and open subset of X. Thus $X \setminus G$ is nowhere dense in X. But for this space X, this implies that $X \setminus G$ is countable. Since μ is nonatomic, $\mu(X \setminus G) = 0$. It follows that

$$\mu(X) = \mu(G) + \mu(X \setminus G) < \varepsilon.$$

Since ε is arbitrary, $\mu(X) = 0$. This proves (i). The proof of (ii) is similar. We leave it as Exercise 3:12.5. Part (iii) follows directly from part (ii) and Theorem 3.23. ■

It is a fact (proved later in Theorem 11.11) that every uncountable analytic set in $\mathbb{R}$ contains a Cantor set. Since all Borel sets are analytic, it follows that every uncountable Borel set in $\mathbb{R}$ has positive measure with respect to some nonatomic Lebesgue–Stieltjes measure. The space X is not a Borel subset of $\mathbb{R}$. It has cardinality c, yet has *universal measure* zero. This means every finite, nonatomic Lebesgue–Stieltjes measure gives X measure zero. The space X can be used to show that there is no nontrivial nonatomic measure defined on all subsets of $[0, 1]$. This gives another proof of Theorem 2.39 of Ulam, here using the continuum hypothesis.

Theorem 3.52: *If μ is a nonatomic, finite measure defined on all subsets of $[0,1]$, then $\mu([0,1]) = 0$.*

Proof. Let h be a one-to-one function mapping X onto $[0,1]$. Define ν on 2^X by

$$\nu(E) = \mu(h(E)).$$

Then ν is a finite, nonatomic measure on 2^X. By Theorem 3.51 (i), $\nu(X) = 0$. In particular, $\mu([0,1]) = \mu(h(X)) = \nu(X) = 0$. ■

There is nothing special about the interval $[0,1]$. The proof of Theorem 3.52 works equally well for any set of cardinality c. Nontrivial finite, nonatomic measures cannot be defined for all subsets of any set Y of cardinality c. It is perhaps curious that this statement is one of pure set theory: no metric or topological conditions are imposed on Y. The proof here, however, did make heavy use of a strange property of the metric space X.

Exercises

3:12.1◇ Let $P \subset \mathbb{R}$ be a Cantor set and suppose that $f : P \to \mathbb{R}$ is continuous and strictly increasing. Show that $f(P)$ is also a Cantor set.

3:12.2 Show that any two Cantor sets on the real line are homeomorphic.

3:12.3 In this exercise we introduce the concepts of a connected set and a totally disconnected set in our context of Cantor sets. We will return to connectedness in Exercise 10:8.6.

> **Definition** A metric space X is *connected* if it cannot be expressed as a disjoint union of two nonempty open sets. A subset S of X is connected if S is a connected metric space.

Definition A metric space X is *totally disconnected* if it contains no connected subsets apart from the empty set and singleton sets. A subset S of X is totally disconnected if S is a totally disconnected metric space.

(a) Prove that the only connected sets in $\mathbb{R}$ are intervals, singleton sets (i.e., sets containing only one point), and the empty set.

(b) Prove that a set of real numbers is totally disconnected if and only if it contains no interval.

(c) Prove that a nonempty set of real numbers is a Cantor set if and only if it is compact, has no isolated points, and is totally disconnected.

(d) Prove that a Cantor set in any metric space[8] is compact, has no isolated points, and is totally disconnected.

(e) Prove, in fact, that a set in a metric space is a Cantor set if and only if it is a nonempty compact set that has no isolated points and is totally disconnected.

3:12.4 Show that if $A \subset [0,1]$ is totally imperfect then, for every Lebesgue–Stieltjes measure μ_f, either $\mu_f(A) = 0$ or A is not μ_f-measurable. [Hint: For the second alternative, apply Theorem 3.23 to A and its complement $\widetilde{A}$.]

3:12.5 Verify part (ii) of Theorem 3.51.

3:12.6 The only finite, nonatomic Borel measure on the space X appearing in Theorem 3.51 is the zero measure. If one tries to imitate the proof of Theorem 3.52 to show that every nonatomic, finite Borel measure on $[0,1]$ is the zero measure, one step fails. Which is it?

[8] As mentioned earlier, a Cantor set in a metric space is one that is homeomorphic to the classical Cantor set.

3:12.7$\diamondsuit$ Let h be continuous and strictly increasing on $\mathbb{R}$. Prove that $h(B)$ is a Borel set if and only if B is a Borel set. [Hint: Let $\mathcal{S}$ be the family of all sets $A \subset \mathbb{R}$ such that $h(A)$ is a Borel set. Show that $\mathcal{S}$ is a σ-algebra that contains the closed sets. For the "only if" part, consider h^{-1}.]

3.13 Additional Problems for Chapter 3

3:13.1 Let μ be a regular Borel measure on a compact metric space X such that $\mu(X) = 1$, and let $\mathcal{E}$ be the family of all closed subsets F of X such that $\mu(F) = 1$.

(a) Prove that the intersection of any finite subcollection of $\mathcal{E}$ also belongs to $\mathcal{E}$.

(b) Prove that the intersection H of the sets in $\mathcal{E}$ is a nonempty compact set.

(c) Prove that $\mu(H) = 1$.

(d) Prove that $\mu(H \cap V) > 0$ for each open set V with $H \cap V \neq \emptyset$.

(e) Prove that if K is a compact subset of X such that $\mu(K) = 1$ and $\mu(K \cap V) > 0$ for each open set V with $K \cap V \neq \emptyset$ then $H = K$.

3:13.2 Let X be a well-ordered set that has a last element Ω such that if $x \in X$ then the set of predecessors of x,

$$\{y \in X : y < x\},$$

is countable. Let $Y = \{y \in X : y < \Omega\}$, and let $\mathcal{M}$ be a σ–algebra of subsets of Y that contains at least all singleton sets. Prove that for any measure on $\mathcal{M}$ the following assertions are equivalent:

(a) For every $a \in Y$, $\mu(\{x \in Y : x \leq a\}) < \infty$.

(b) The set $P = \{x \in Y : \mu(\{x\}) > 0\}$ is countable and $\mu(P) < \infty$.

3:13.3 Let A and B be sets. The set

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$$

is called the *symmetric difference* of A and B.

Prove that there exists a countable family $\mathcal{A}$ of open sets in $[0,1]$ with the following property: For every $\varepsilon > 0$ and $E \in \mathcal{L}$, there exists $A \in \mathcal{A}$ with $\lambda(A \triangle E) < \varepsilon$. Thus the countable family $\mathcal{A}$ can be used to approximate all members of $\mathcal{L}$. We shall see later that $\lambda(A \triangle B)$ is "almost" a metric on $\mathcal{L}$.

3:13.4 Let $\mathcal{E}$ be defined as in the proof of Theorem 3.13. Let $(X, \hat{\mathcal{B}}, \hat{\mu})$ be the completion of $(X, \mathcal{B}, \mu)$.

(a) Show that $\mathcal{E} \supset \hat{\mathcal{B}}$. [Hint: Use Theorems 2.37, 2.45, and 3.16.]

(b) Use part (a) to improve Theorem 3.22 to give the conclusion $\mu_f(E) = \hat{\mu}(E)$ for all $E \in \mathcal{E}$.

3:13.5 Let I be an interval in $\mathbb{R}$. Show how one can reduce a theory of Lebesgue–Stieltjes measures on I to the theory that we developed for Lebesgue–Stieltjes measures on $\mathbb{R}$.

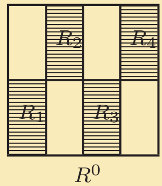
3:13.6$\diamondsuit$ Let f be continuous on $[0,1]$. Let $\mathcal{T}$ consist of $\emptyset$ and the closed intervals in $[0,1]$. Let $\tau([a,b]) = |f(b) - f(a)|$, and let μ_1^* and μ_2^* be the associated Method I and Method II outer measures, respectively.

(a) Is μ_1^* equal to μ_2^*?

(b) What relationship exists between the measure μ_2 and the variation of f?

(c) What is the answer to (b) if f is piecewise monotonic?

3:13.7 Let R^0 be the unit square. Divide R^0 into 8 rectangles of height $\frac{1}{2}$ and width $\frac{1}{4}$, as indicated in Figure 3.4. Now divide each of the rectangles R_i into 8 or 10 rectangles, giving rise to the situation depicted in Figure 3.4 for R^2. Continue this process by cutting heights in half and widths into 4 or 5 parts in such a way that $R^{k+1} \subset R^k$, and R^k is compact and connected. Let $R = \bigcap_{k=1}^{\infty} R^k$.



$R^1 = R_1 \cup R_2 \cup R_3 \cup R_4.$

Figure 3.4. The rectangles R^0, and R_i $(i = 1 \ldots 4)$ in Exercise 3:13.7.

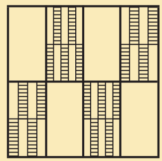


Figure 3.5. The rectangles R^2 (the shaded region).

(a) Show that this intersection R is the graph of a continuous function g. (The construction of this function is due to James Foran.)

(b) Show that for each $c \in [0,1]$ the set $\{x : g(x) = c\}$ is a Cantor set.

(c) Let $\mathcal{T}$ consist of $\emptyset$ and the closed intervals in $[0,1]$, and let $\tau([a,b]) = |g(b) - g(a)|$. Let μ_0^* be the Method II outer measure obtained from $\mathcal{T}$ and τ. Calculate $\mu_0^*(E)$ for $E \subset [0,1]$. [Hint: Calculate $\mu_0^*([0,1])$.]

(d) Compare your answer to part (c) with your answer to part (b) of Exercise 3:13.6.

3:13.8♢ Prove that there exists a set $E \subset [0,1]$ with $E \in \mathcal{L}$, but $F(E) \notin \mathcal{L}$, where F is the Cantor function. [Hint: Use Exercise 2:14.13.]

3:13.9♢ Let f be continuous on $[a,b]$. Prove that the following statements are equivalent.

(a) There exists $E \subset [a,b]$ such that $E \in \mathcal{L}$, but $f(E) \notin \mathcal{L}$.

(b) There exists $E \subset [a,b]$ such that $\lambda(E) = 0$, but $\lambda^*(f(E)) \neq 0$.

3:13.10♢ Let μ_1 and μ_2 be measures defined on a common σ-algebra $\mathcal{M}$. We say that μ_1 is *absolutely continuous* with respect to μ_2, written $\mu_1 \ll \mu_2$, if $\mu_1(E) = 0$ whenever $\mu_2(E) = 0$, $E \in \mathcal{M}$. Let $\mathcal{M} = \mathcal{B}$, and let F be the Cantor function. Is $\mu_F \ll \lambda$? Is $\lambda \ll \mu_F$?

3:13.11♢ Refer to Exercise 3:13.10. Let μ_g be a continuous Lebesgue–Stieltjes measure on $\mathcal{B}$.

(a) Prove that $\mu_g \ll \lambda$ if and only if, for $E \in \mathcal{B}$ and $\lambda(E) = 0$, $\lambda(g(E)) = 0$.

(b) Prove that if $\lambda \ll \mu_g$ then g is strictly increasing.

3:13.12 Let $\{L_n\}$ be a sequence of pairwise disjoint Lebesgue measurable sets in $\mathbb{R}$, let $L = \bigcup_{n=1}^{\infty} L_n$, and let $E \subset \mathbb{R}$.

(a) Prove that $\lambda^*(L \cap E) = \sum_{n=1}^{\infty} \lambda^*(L_n \cap E)$. [Hint: Let H be a measurable cover for $L \cap E$, H_n for $L_n \cap E$ with the sets H_n pairwise disjoint.]

(b) Prove that $\lambda_*(L \cap E) = \sum_{n=1}^{\infty} \lambda_*(L_n \cap E)$.
[Outline of proof: Let K be a measurable kernel for $L \cap E$. Justify the inequalities

$$\begin{aligned}\lambda_*(L \cap E) \;&=\; \lambda(K) = \sum_{n=1}^{\infty} \lambda(L_n \cap K) \\ &\leq\; \sum_{n=1}^{\infty} \lambda_*(L_n \cap E) \leq \lambda_*(L \cap E).\end{aligned}$$

3:13.13◇ (Extending $\mathcal{L}$ and λ) Let $X = [0,1]$.

(a) Prove that, for each $E \subset X$ and $L \in \mathcal{L}$,

$$\lambda(L) = \lambda_*(L \cap E) + \lambda^*(L \cap \widetilde{E}).$$

(b) Let $E \subset X$, $E \notin \mathcal{L}$. Let $\overline{\mathcal{L}}$ be the algebra generated by $\mathcal{L}$ and $\{E\}$. Show that $\overline{\mathcal{L}}$ consists of all sets of the form

$$\overline{L} = (L_1 \cap E) \cup (L_2 \cap \widetilde{E}) \text{ with } L_1, L_2 \in \mathcal{L}.$$

(c) Define $\overline{\lambda}$ on $\overline{\mathcal{L}}$ by

$$\overline{\lambda}(\overline{L}) = \lambda^*(\overline{L} \cap E) + \lambda_*(\overline{L} \cap \widetilde{E}).$$

Let $\mathcal{T} = \overline{\mathcal{L}}$, $\tau = \overline{\lambda}$ and let $(X, \overline{\overline{\mathcal{L}}}, \overline{\overline{\lambda}})$ be the measure space obtained by an application of Method I. Prove that $\overline{\overline{\lambda}} = \lambda$ on $\mathcal{L}$. Thus $(X, \overline{\overline{\mathcal{L}}}, \overline{\overline{\lambda}})$ is an extension of $(X, \mathcal{L}, \lambda)$ and contains sets not in $\mathcal{L}$.

(d) Show that $\overline{\overline{\lambda}}(E) = \lambda^*(E)$. Thus E has a $\mathcal{G}_\delta$ cover with respect to $\overline{\overline{\lambda}}$. That is, there exists $H \in \mathcal{G}_\delta$ such that $H \supset E$ and $\overline{\overline{\lambda}}(H) = \overline{\overline{\lambda}}(E) = \lambda^*(E)$. Does $\widetilde{E}$ also have such a cover in $\mathcal{G}_\delta$?

3:13.14 We stated Theorem 3.49 for nondecreasing functions. That hypothesis cannot be dropped. Let $g : [0,1] \to \mathbb{R}$ be a continuous function all of whose level sets are uncountable[9] (e.g., the continuous function g of Exercise 3:13.7). Show that there exists a totally imperfect set A such that $g(A) = [0,1]$. This exercise shows that, unlike monotonic functions, continuous functions *can* rise on totally imperfect sets.

[Hint: A proof can be based on the continuum hypothesis and transfinite induction. Let $\{y_\alpha\}$, $\alpha < \Omega$, be a well-ordering of the interval $[0,1]$, and let $\{P_\alpha\}$, $\alpha < \Omega$, be a well-ordering of the Cantor

[9]One might imagine that such functions are rare, but see Exercise 10:8.4.

sets in $[0,1]$. Choose a_1 such that $f(a_1) = y_1$. Now choose $b_1 \in P_1 \setminus \{a_1\}$. Proceed inductively. If we have $\{a_\beta\} \subset [0,1]$ and $\{b_\beta\} \subset [0,1]$ for all $\beta < \alpha$, choose

$$a_\alpha \in [0,1] \setminus \bigcup_{\beta<\alpha} (\{a_\beta\} \cup \{b_\beta\})$$

such that $f(a_\alpha) = y_\alpha$. Then choose

$$b_\alpha \in [0,1] \setminus \left(\bigcup_{\beta\leq\alpha} \{a_\beta\} \cup \bigcup_{\beta<\alpha} \{b_\beta\} \right)$$

such that $b_\alpha \in P_\alpha$. Let

$$A = \bigcup_{\alpha<\Omega} \{a_\alpha\} \text{ and } B = \bigcup_{\alpha<\Omega} \{b_\alpha\}.$$

Then $f(A) = [0,1]$. If P is a Cantor set in $[0,1]$, there exists α such that $P = P_\alpha$. By construction, $b_\alpha \in P_\alpha$ and $A \cap B = \emptyset$. Thus $b_\alpha \notin A$, so A does not contain P.]

3:13.15 Use the continuum hypothesis to prove the existence of a set A of real numbers such that A and its complement $\widetilde{A}$ are both totally imperfect. [Hint: Modify the argument in Exercise 3:13.14 to choose points a_α and b_α from P_α.]

Chapter 4

MEASURABLE FUNCTIONS

We saw in Section 1.20 that the definition of the Lebesgue integral of a function f involves the measure of sets such as

$$\{x : \alpha \leq f(x) < \beta\}.$$

We devote this chapter to the study of functions for which these sets, and others defined by similar inequalities, are necessarily measurable. These will be called *measurable functions*. We shall see that, for a given measure space $(X, \mathcal{M}, \mu)$, the class of μ-measurable functions is well behaved with respect to the elementary algebraic operations and with respect to various operations involving limits. The proofs here will follow readily from our requirement that $\mathcal{M}$ be a σ-algebra, together with a bit of set-theoretic algebra. We provide the necessary development in Sections 4.1 and 4.2.

In Chapters 2 and 3 we saw that, while measurable sets can be quite complicated, one can under certain circumstances approximate measurable sets, and even nonmeasurable sets, by simpler sets. For example, when dealing with the Lebesgue–Stieltjes measure space $(\mathbb{R}, \mathcal{M}_f, \mu_f)$, we know that every set $M \in \mathcal{M}_f$ has a $\mathcal{G}_\delta$ cover and an $\mathcal{F}_\sigma$ kernel, and we know that for every $\varepsilon > 0$ there exists an open set G and a closed set F such that $F \subset M \subset G$ and $\mu_f(G \setminus F) < \varepsilon$. Such approximations have allowed us to deal with measurable sets that might be unwieldy to combine or manipulate by replacing them with simpler sets that we can handle. Similar simplifications are also available when dealing with measurable functions. In Sections 4.4 and 4.5 we see that, under suitable hypotheses on a measure space $(X, \mathcal{M}, \mu)$, measurable functions can be approximated by simpler functions in several ways. In particular, for many important classes of measure spaces, the approximating functions can be taken to be continuous.

We also need to discuss convergence of sequences of measurable functions. Of the several notions of convergence that we encounter in Section 4.2, the "preferred" notion may be uniform convergence. It became apparent in the middle of the nineteenth century that a number of theorems that are easy to prove when uniform convergence is assumed in appropriate places are either false or more difficult to prove when weaker forms of convergence are hypothesized. In Section 4.3 we show that, on a finite measure space, a sequence $\{f_n\}$ of measurable functions that is known to converge in some weaker sense actually converges "almost uniformly," that is, uniformly when one ignores a set of small measure.

Thus three fundamental concepts in analysis—set, convergence, and function—allow approximations by more tractable objects. Although one gives up a bit at the stages where one makes the approximation, the conclusion reached at the end of the argument is still often the best possible.

4.1 Definitions and Basic Properties

We begin with Lebesgue's original definition of a measurable function.

Definition 4.1: Let $(X, \mathcal{M}, \mu)$ be a measure space, and let

$$f : X \to [-\infty, \infty].$$

The function f is *measurable* if for every $\alpha \in \mathbb{R}$ the set

$$E_\alpha(f) = \{x : f(x) > \alpha\}$$

is a measurable set.

A special case of this definition has its own terminology.

Definition 4.2: Let X be a metric space, and let $f : X \to [-\infty, \infty]$. The function f is a *Borel function* or is *Borel measurable* if the set

$$E_\alpha(f) = \{x : f(x) > \alpha\}$$

is a Borel set for every $\alpha \in \mathbb{R}$.

Observe that measurability of f depends on the σ-algebra $\mathcal{M}$ under consideration, but not on the measure μ. Nonetheless, one often sees phrases asserting that a function f is μ-measurable with no specific mention of the measure space that is assumed.

Example 4.3: Take $(\mathbb{R}, \mathcal{L}, \lambda)$ as the measure space. Let f be a continuous function, g be a discontinuous increasing function, and $h = \chi_A$ for some set $A \subset \mathbb{R}$. Then, for every $\alpha \in \mathbb{R}$, $E_\alpha(f)$ is open and $E_\alpha(g)$ is an interval. Thus both f and g are measurable. For h we find that

$$E_\alpha(h) = \begin{cases} \emptyset, & \text{if } \alpha \geq 1, \\ A, & \text{if } 0 \leq \alpha < 1, \\ \mathbb{R}, & \text{if } \alpha < 0. \end{cases}$$

Hence h is λ-measurable if and only if $A \in \mathcal{L}$. If we had taken $(\mathbb{R}, \mathcal{B}, \lambda)$ as our measure space, then f and g are measurable (and hence Borel functions) because open sets and arbitrary intervals are Borel sets, and h is measurable if and only if A is a Borel set.

Example 4.4: If $\mathcal{M} = \{\emptyset, X\}$, only constant functions are measurable, while if $\mathcal{M} = 2^X$, all functions are measurable. In particular, if X is countable and each singleton set is measurable, then every function on X is measurable.

Theorem 4.5 shows that there is nothing special about the specific inequality we chose in Definition 4.1.

Theorem 4.5: *Let $(X, \mathcal{M}, \mu)$ be a measure space. The following conditions on a function f are equivalent.*

1. *f is measurable.*
2. *For all $\alpha \in \mathbb{R}$, the set $\{x : f(x) \geq \alpha\} \in \mathcal{M}$.*
3. *For all $\alpha \in \mathbb{R}$, the set $\{x : f(x) < \alpha\} \in \mathcal{M}$.*
4. *For all $\alpha \in \mathbb{R}$, the set $\{x : f(x) \leq \alpha\} \in \mathcal{M}$.*

Proof. Suppose that f is measurable and let $\alpha \in \mathbb{R}$. Observe that

$$\{x : f(x) \geq \alpha\} = \bigcap_{n=1}^{\infty} \left\{x : f(x) > \alpha - \frac{1}{n}\right\}.$$

Since f is measurable, each set in the intersection is measurable and, hence, so is the intersection itself. This proves that (i) $\Rightarrow$ (ii). The implication (ii) $\Rightarrow$ (iii) follows directly from the equality

$$\{x : f(x) < \alpha\} = X - \{x : f(x) \geq \alpha\}\,.$$

The implication (iii) $\Rightarrow$ (iv) follows from the equality

$$\{x : f(x) \leq \alpha\} = \bigcap_{n=1}^{\infty} \left\{x : f(x) < \alpha + \frac{1}{n}\right\}.$$

Finally, the implication (iv) $\Rightarrow$ (i) follows by complementation in (iii). It now follows that all four statements are equivalent. ■

Simple arguments show that various other sets associated with a measurable function f are measurable, for example, the sets

$$\{x : f(x) = \alpha\} \;\text{ and }\; \{x : \alpha \leq f(x) \leq \beta\}.$$

Note that measurability of a function f is related to the mapping properties of f^{-1}. In fact, measurability of f is equivalent to the condition that f^{-1} take Borel sets to measurable sets. (The proof is left as Exercise 4:1.2.)

Theorem 4.6: *Let $(X, \mathcal{M}, \mu)$ be a measure space and f a real-valued function on X. Then f is measurable if and only if $f^{-1}(B) \in \mathcal{M}$ for every Borel set $B \subset \mathbb{R}$.*

Our next example shows that we cannot replace Borel sets with arbitrary measurable sets in this theorem. It also shows that the mapping properties of f (as opposed to f^{-1}) may be quite different for measurable functions. (The reader may wish to consult Exercises 2:14.13 and 3:13.8 to 3:13.10 before proceeding with this example.)

Example 4.7: We work with the Lebesgue measure space $(\mathbb{R}, \mathcal{L}, \lambda)$. Let K be the Cantor ternary set, and let P be a Cantor set of positive measure. Write $a = \min\{x : x \in P\}$ and $b = \max\{x : x \in P\}$. Exercise 4:1.13 shows that there exists a strictly increasing continuous function h that maps $[a, b]$ onto $[0, 1]$ and maps P onto K.

Let A be a nonmeasurable subset of P, and let $E = h(A)$. Since $E \subset K$, $\lambda(E) = 0$ and, in particular, E is Lebesgue measurable. It follows that

1. $h^{-1}(E) = A$. Thus, even for the strictly increasing continuous function h, the inverse image of a measurable set need not be measurable.

2. The function h^{-1} is also continuous and strictly increasing. It maps the zero measure set E onto a nonmeasurable set.

3. Let $f = h^{-1}$ and let μ_f be the associated Lebesgue–Stieltjes measure on $[0, 1]$. Then μ_f is not absolutely continuous with respect to λ, since $\lambda(K) = 0$, but by Theorem 3.23
$$\mu_f(K) = \lambda(f(K)) = \lambda(P) > 0.$$

Observe that part (i) offers another proof that there are Lebesgue measurable sets that are not Borel sets. The set E is Lebesgue measurable. If it were a Borel set, then $A = h^{-1}(E)$ would also be measurable by Theorem 4.6.

4.1.1 Combining measurable functions

We next consider various ways that measurable functions combine to give rise to other measurable functions. Note first that, because we are allowing infinite values for our functions, expressions such as $f + g$ and fg require some caution. We cannot interpret $\infty - \infty$ nor $0 \times \infty$. Thus we will require some comment assuring us that the functions are defined. Often this is taken for granted.

Theorem 4.8: *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f and g be measurable functions on X. Let $\phi : \mathbb{R} \to \mathbb{R}$ be continuous, and let $c \in \mathbb{R}$. Then, provided the following functions are defined,*

1. *cf is measurable.*
2. *$f + g$ is measurable.*
3. *$\phi \circ f$ is measurable (where f must be finite-valued).*
4. *fg is measurable.*

Proof. The proof of (i) is trivial for finite-valued functions. We interpret $c \times \pm\infty = \pm\infty$ for $c > 0$ and $c \times \pm\infty = \mp\infty$ for $c < 0$. Then cf is easily shown to be measurable for all $c \neq 0$ and any measurable function f. Just work separately on the (measurable) sets $X_1 = \{x \in X : -\infty < f(x) < \infty\}$, $X_2 = \{x \in X : -\infty = f(x)\}$, and $X_3 = \{x \in X : f(x) = \infty\}$.

To verify (ii) for finite-valued functions, observe first that for, any $\alpha \in \mathbb{R}$, the function $\alpha - g$ is measurable. Now let $\{q_k\}$ be an enumeration of the rational numbers. Then

$$\{x : f(x) + g(x) > \alpha\} = \{x : f(x) > \alpha - g(x)\}$$

$$= \bigcup_{k=1}^{\infty} (\{x : f(x) > q_k\} \cap \{x : g(x) > \alpha - q_k\}).$$

This set is clearly measurable. Since this is true for all $\alpha \in \mathbb{R}$, $f+g$ is measurable. For functions that are permitted to assume values $\pm\infty$ some extra bookkeeping would be needed, left as Exercise 4:1.7.

In statement (iii) we cannot allow infinite values since $\phi(f(x))$ would not be defined if $f(x) = \pm\infty$. To verify (iii) then, let $\alpha \in \mathbb{R}$, and observe that

$$(\phi \circ f)^{-1}((\alpha, \infty)) = f^{-1}(\phi^{-1}((\alpha, \infty))).$$

Since ϕ is continuous, the set $G = \phi^{-1}((\alpha, \infty))$ is open, and since f is measurable, $f^{-1}(G) \in \mathcal{M}$. This verifies (iii).

Let us prove part (iv). Suppose first that f and g are finite-valued. Then

$$4f(x)g(x) = (f(x) + g(x))^2 - (f(x) - g(x))^2 \tag{1}$$

at every point $x \in X$. From parts (i) and (ii) we see that $f+g$ and $f-g$ are measurable. Take the continuous function $\phi(t) = t^2$ and apply part (iii) to conclude that both $(f+g)^2$ and $(f-g)^2$ are measurable. Finally, then, parts (i) and (ii) applied again to the identity (1) shows that the product fg is measurable. The case for functions that are permitted to assume values $\pm\infty$ is left as Exercise 4:1.8. ∎

In part (iii) of Theorem 4.8, note the order of composition: the function f maps X to $\mathbb{R}$ and the continuous function ϕ maps $\mathbb{R}$ to $\mathbb{R}$, thus $\phi \circ f$ is defined, while $f \circ \phi$ may not be. If $X = \mathbb{R}$ then the latter composition $f \circ \phi$ would also be defined. Must it, too, be measurable? Exercise 4:1.10 shows that it may not be.

Exercises

4:1.1 Let $(X, \mathcal{M}, \mu)$ be a measure space. Show that for an arbitrary function f on X the class $\{A \subset \mathbb{R} : f^{-1}(A) \in \mathcal{M}\}$ is a σ-algebra.

4:1.2$\diamondsuit$ Let $(X, \mathcal{M}, \mu)$ be a measure space. Show that a function $f : X \to \mathbb{R}$ is measurable if and only if $\{A \subset \mathbb{R} : f^{-1}(A) \in \mathcal{M}\}$ contains all Borel sets.

4:1.3 Let $(X, \mathcal{M}, \mu)$ be a measure space and suppose that $f : X \to \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. Show that it is possible for f to fail to be measurable and yet the family

$$\{A \subset \mathbb{R} : f^{-1}(A) \in \mathcal{M}\}$$

contains all Borel sets. Compare with Exercise 4:1.2. Give a correct formulation of the statement in Exercise 4:1.2 that permits f to assume infinite values.

4:1.4 Suppose that, for each rational number q, the set $\{x : f(x) > q\}$ is measurable. Can we conclude that f is measurable?

4:1.5 Let $\mathcal{S}_0$ be a family of subsets of $\mathbb{R}$ such that all open sets belong to the smallest σ-algebra containing $\mathcal{S}_0$. If $f^{-1}(E)$ is measurable for all $E \in \mathcal{S}_0$ then f is measurable. Apply this to obtain another proof of the preceding exercise and another proof of Theorem 4.5.

4:1.6 Show that there exists a function $f : \mathbb{R} \to \mathbb{R}$ such that, for each $\alpha \in \mathbb{R}$, the set $\{x : f(x) = \alpha\}$ is in $\mathcal{L}$, but f is not Lebesgue measurable.

[Hint: Map a nonmeasurable set onto $(0, 1)$ and its complement onto $(1, 2)$ in an appropriate manner.]

4:1.7 Complete the proof of Theorem 4.8, part (ii) by discussing the case where f and g are permitted to have infinite values.

4:1.8 Complete the proof of Theorem 4.8, part (iv) by discussing the case where f and g are permitted to have infinite values.

4:1.9 Provide conditions under which a quotient of measurable functions is measurable.

4:1.10 Give an example of a continuous function ϕ and a Lebesgue measurable function f, both defined on $[0,1]$, such that $f \circ \phi$ is not measurable. Give an example of a nonmeasurable function f on $[0,1]$ such that $|f|$ is measurable. [Hint: See Example 4.7.]

4:1.11 Let $(X, \mathcal{M}, \mu)$ be a measure space. Suggest conditions under which there can exist a nonmeasurable function f on X for which $|f|$ is measurable.

4:1.12 Show that a measurable function f defined on $[0,1]$ has the property that for every $\varepsilon > 0$ there is a $M_\varepsilon > 0$ so that

$$\lambda\left(\{x \in [0,1] : |f(x)| \leq M_\varepsilon\}\right) \geq 1 - \varepsilon$$

if and only if f is finite almost everywhere.

4:1.13$\diamondsuit$ Let E and F be any two Cantor sets in $\mathbb{R}$. Let $\mathcal{I} = \{I_k\}$ and $\mathcal{J} = \{J_k\}$ be the sequences of intervals complementary to E and F, respectively.

(a) Show that to each pair of distinct intervals I_i and I_k in $\mathcal{I}$ there exists an interval $I_j \in \mathcal{I}$ between I_i and I_k.

(b) Use part (a) to show that there exists an order-preserving correspondence between $\mathcal{I}$ and $\mathcal{J}$. That is, there exists a function γ mapping $\mathcal{I}$ onto $\mathcal{J}$ such that if $I, I' \in \mathcal{I}$ and $J = \gamma(I)$, while $J' = \gamma(I')$, then J is to the left of J' if and only if I is to the left of I'.

(c) For each $I_i \in \mathcal{I}$, let f_i be continuous and strictly increasing on I_i, and map I_i onto the interval $\gamma(I_i)$. Use the functions f_i to obtain a strictly increasing continuous function f mapping $\bigcup_{i=1}^{\infty} I_i$ onto $\bigcup_{i=1}^{\infty} J_i$.

(d) Extend f to be a continuous strictly increasing function mapping $\mathbb{R}$ onto $\mathbb{R}$ and E onto F.

4:1.14 Let $\mathcal{T}$ consist of $\emptyset$ and the open squares in $\mathbb{R}^2$, and let $\tau(T)$ be the diameter of T. Use Method I to obtain an outer measure μ^* and a measure space $(\mathbb{R}^2, \mathcal{M}, \mu)$. Is every continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ measurable with respect to $\mathcal{M}$? What would your answer be if we had used Method II instead of Method I?

4:1.15 Let $f : \mathbb{R} \to \mathbb{R}$ be continuous.

(a) Show that f maps compact sets to compact sets.

(b) Show that f maps sets of type F_σ to sets of the same type.

(c) Show that, if f is also one-one, then f maps Borel sets to Borel sets.

(d) Show that, if f is also Lipschitz, then f maps sets of Lebesgue measure zero to sets of the same type.

(e) Show that, if f is Lipschitz, then f maps Lebesgue measurable sets to sets of the same type.

(We have seen in Example 4.7 that a one-to-one continuous function $f : \mathbb{R} \to \mathbb{R}$ need not map Lebesgue measurable sets to Lebesgue measurable sets. We mention that, without the assumption that f be one to one, we cannot be sure that f maps Borel sets to Borel sets. It is true that a continuous function f maps Borel sets onto Lebesgue measurable sets. Proofs appear in Chapter 11.)

4:1.16 Let $(X, \mathcal{M}, \mu)$ be a complete measure space with X a metric space.

(a) Prove that if all Borel sets are measurable each function f that is continuous a.e. is measurable.

(b) Prove that if every continuous function $f : X \to \mathbb{R}$ is measurable then $\mathcal{M} \supset \mathcal{B}$.
[Hint: Let G be open in X. Let $f(x) = \operatorname{dist}(x, X \setminus G)$. See Section 3.2. Show that f is continuous and $f^{-1}((0,\infty)) = G$.]

(c) Let $X = [0,1]$, $\mathcal{M} = \{\emptyset, X\}$, and let $f(x) = x$. Is f measurable?

4:1.17 Suppose that there exists[1] a Lebesgue nonmeasurable subset E of $\mathbb{R}^2$ such that E intersects every horizontal or vertical line in exactly one point. Use this set to show that there exists a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that f is Borel measurable in each variable separately, yet f is not Lebesgue measurable. Note also that the restriction of f to any horizontal or vertical line has only one point of discontinuity. Compare with Exercise 4:1.16 (a).

4:1.18 In part (iii) of Theorem 4.8 we had to assume f finite. Otherwise the function $\phi \circ f$ is not defined on the set $\{x : f(x) = \pm\infty\}$. Suppose that $(X, \mathcal{M}, \mu)$ is complete. Since the measurability of a function does not depend on its values on a set of measure zero, one can discuss the measurability of functions defined only a.e. Formulate how this can be done, and then prove part (iii) of Theorem 4.8 under the assumption that f is finite a.e.

4:1.19 Let $(X, \mathcal{M}, \mu)$ be a measure space and Y a metric space. Give a reasonable definition for a function $f : X \to Y$ to be measurable. How much of the theory of this section and the next can be done in this generality?

4.2 Sequences of Measurable Functions

Several forms of convergence of a sequence of functions are important in the theory of integration. Two of these forms, *pointwise convergence* and *uniform convergence*, form part of the standard material of courses in elementary analysis. We assume that the reader is familiar with these forms of convergence. We discuss two other forms in this section: *almost everywhere convergence* and *convergence in measure.* We first show that the class of measurable functions is closed under certain operations on sequences.

[1]This can be proved under the assumption of the continuum hypothesis. For the construction of such a set without assuming CH, see E. K. van Douwen, *Fubini's theorem for null sets*, Amer. Math. Monthly 96 (1989), no. 8, 718–721.

Theorem 4.9: *Let $(X, \mathcal{M}, \mu)$ be a measure space, and let $\{f_n\}$ be a sequence of measurable functions on X. Then each of the functions $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$ and $\liminf_n f_n$ is measurable.*

Proof. Since

$$\left\{x : \sup_n f_n(x) \leq \alpha\right\} = \bigcap_{n=1}^{\infty} \{x : f_n(x) \leq \alpha\},$$

the function $\sup_n f_n$ is measurable. That $\inf_n f_n$ is measurable follows from the identity

$$\inf_n f_n = -\sup_n(-f_n).$$

The identities

$$\limsup_n f_n = \inf_k \sup_{n\geq k} f_n \quad \text{and} \quad \liminf_n f_n = \sup_k \inf_{n\geq k} f_n$$

supply the measurability of the other two functions. ■

It follows that the set

$$\left\{x : \limsup_n f_n(x) = \liminf_n f_n(x)\right\}$$

is a measurable set. This is precisely the set of convergence of the sequence $\{f_n\}$. Here one must allow the possibility that $f_n(x) \to \pm\infty$. It is also true that the set on which $\{f_n\}$ converges to a finite limit is measurable. See Exercise 4:2.4. It follows readily that if $\{f_n(x)\}$ converges for all $x \in X$ then the limit function $f(x) = \lim_n f_n(x)$ is measurable.

4.2.1 Convergence almost everywhere

We shall see in Chapter 5 that the integral of a function f does not depend on the values that f assumes on a set of measure zero. It is also true that one can often assert no more than that the sequence $\{f_n\}$ converges for *almost every* $x \in X$. This form of convergence suffices in many applications. We present a formal definition.

Definition 4.10: Let $\{f_n\}$ be a sequence of finite a.e., measurable functions on a measurable set $E \subset X$. If there exists a function f such that

$$\lim_{n\to\infty} |f_n(x) - f(x)| = 0$$

for almost all $x \in E$, we say that $\{f_n\}$ *converges* to f *almost everywhere* on E, and we write

$$\lim_n f_n = f \text{ [a.e.]} \quad \text{or} \quad f_n \to f \text{ [a.e.] on } E.$$

The usual slight variation in language applies when $E = X$.

It is now clear that if $f_n \to f$ [a.e.] then f is measurable. A bit of care is needed in interpreting this statement if the measure space is not complete. Removing the set of measure zero on which $\{f_n\}$ does not converge to f leaves a measurable set on which the sequence converges pointwise, and f is measurable on that set.

We mention that some authors provide slightly different definitions for convergence [a.e.]. For example, the concept makes sense without the functions being measurable or finite a.e., so more inclusive definitions are possible. We shall rarely deal with nonmeasurable functions or with functions that take on infinite values on sets of positive measure. By imposing the extra restrictions in our definition, we focus on the way convergence [a.e.] actually arises in our development. Observe that if $f_n \to f$ [a.e.] then our definition guarantees that f is finite a.e.

4.2.2 Convergence in measure

We turn now to another form of convergence, closely related to pointwise convergence.

Definition 4.11: Let $(X, \mathcal{M}, \mu)$ be a measure space, and let $E \in \mathcal{M}$. Let $\{f_n\}$ be a sequence of finite a.e., measurable functions on E. We say that $\{f_n\}$ *converges in measure* on E to a measurable function f and we write

$$\lim_n f_n = f \text{ [meas] } \text{ or } f_n \to f \text{ [meas] on } E$$

if for any pair (ε, η) of positive numbers there corresponds $N \in \mathbb{N}$ such that, if $n \geq N$, then

$$\mu(\{x \in E : |f_n(x) - f(x)| \geq \eta\}) < \varepsilon.$$

Equivalently, $f_n \to f$ [meas] if, for every $\eta > 0$,

$$\lim_n \mu(\{x \in E : |f_n(x) - f(x)| \geq \eta\}) = 0.$$

These notions of convergence are used, too, in probability theory. There convergence a.e. is called "convergence almost surely" and convergence in measure is called "convergence in probability." We shall see in Section 4.3 that, when $\mu(X) < \infty$, convergence [a.e.] implies convergence [meas]. Thus in probability theory where the space has measure 1, almost sure convergence always implies convergence in probability. In general, this is not so, as the next example shows.

Example 4.12: Let

$$f_n(x) = \frac{x}{n}.$$

Each function f_n is finite and Lebesgue measurable on $\mathbb{R}$. One verifies easily that $f_n \to 0$ [a.e.], but $\{f_n\}$ does not converge in measure to any function on $\mathbb{R}$.

Our next example shows that it is possible for $f_n \to 0$ [meas] without $\{f_n(x)\}$ converging for any x. This example also illustrates a feature of this convergence that will play a role in integration theory. Even though the sequence has no pointwise limit, we can still write

$$\lim_{m\to\infty} \int_0^1 f_m \, d\lambda = 0 = \int_0^1 \lim_{m\to\infty} f_m \, d\lambda,$$

provided that $\lim_{m\to\infty} f_m$ is taken in the sense of convergence in measure.

Example 4.13: (A sliding sequence of functions) For nonnegative integers n, k, with $0 \leq k < 2^n$ and $m = 2^n + k$, let

$$E_m = \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right].$$

Let $f_1 = \chi_{[0,1]}$ and, for $n > 1$, $f_m = \chi_{E_m}$. We see that

$$\begin{aligned} f_2 &= \chi_{\left[0,\frac{1}{2}\right]}, \quad f_3 = \chi_{\left[\frac{1}{2},1\right]}, \\ f_4 &= \chi_{\left[0,\frac{1}{4}\right]}, \quad f_5 = \chi_{\left[\frac{1}{4},\frac{1}{2}\right]}, \quad f_6 = \chi_{\left[\frac{1}{2},\frac{3}{4}\right]}, \quad f_7 = \chi_{\left[\frac{3}{4},1\right]}, \\ f_8 &= \chi_{\left[0,\frac{1}{8}\right]}, \ \ldots \end{aligned}$$

Every point $x \in [0,1]$ belongs to infinitely many of the sets E_m, and so

$$\limsup_m f_m(x) = 1,$$

while

$$\liminf_m f_m(x) = 0.$$

Thus $\{f_m\}$ converges at no point in $[0,1]$, yet $\lambda(E_m) = 2^{-n}$ for $m = 2^n + k$. As $m \to \infty$, $n \to \infty$ also. For every $\eta > 0$,

$$\lambda(\{x : f_m(x) \geq \eta\}) \leq \frac{1}{2^n}.$$

It follows that $f_m \to 0$ [meas] on the interval $[0,1]$.

4.2.3 Pointwise convergence and convergence in measure

If we study Example 4.13 further, we might note that, while the sequence $\{f_m\}$ converges at no point, suitable subsequences converge [a.e.]. For example, $f_{2^n}(x) \to 0$ for each $x \neq 0$. It is true, in general, that such convergent subsequences exist. This is the first of our attempts at finding relations among the various notions of convergence.

Theorem 4.14: *If $f_n \to f$ [meas], there exists a subsequence $\{f_{n_k}\}$ such that $f_{n_k} \to f$ [a.e.].*

Proof. For each $k \in \mathbb{N}$, choose $n_k \in \mathbb{N}$ such that

$$\mu\left(\left\{x : |f_n(x) - f(x)| \geq \frac{1}{2^k}\right\}\right) < \frac{1}{2^k}$$

for every $n \geq n_k$. We choose the sequence $\{n_k\}$ to be increasing. Let

$$A_k = \left\{x : |f_{n_k}(x) - f(x)| \geq \frac{1}{2^k}\right\},$$

and let $A = \limsup_k A_k$. Since $\sum_{k=1}^{\infty} \mu(A_k) < 1 < \infty$, it follows that $\mu(A) = 0$ by the Borel–Cantelli lemma (Exercise 2:4.8). Let $x \notin A$. Then x is a member of only finitely many of the sets A_k. Thus there exists K such that, if $k \geq K$,

$$|f_{n_k}(x) - f(x)| < \frac{1}{2^k}.$$

It follows that $\{f_{n_k}\} \to f$ [a.e.]. ■

In Section 4.3 we shall introduce yet another form of convergence and obtain some more relations that exist among the various modes of convergence.

Exercises

4:2.1 Let $\{f_n\}$ be a sequence of finite functions on a space X, and let $\alpha \in \mathbb{R}$. Prove that

$$\left\{x : \liminf_n f_n(x) > \alpha\right\} = \bigcup_{m=1}^{\infty} \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left\{x : f_n(x) - \alpha \geq \frac{1}{m}\right\}.$$

Use this to provide another proof of the fact that a pointwise limit of measurable functions is measurable.

4:2.2 Let $\{A_n\}$ be a sequence of measurable sets, and write $f_n(x) = \chi_{A_n}(x)$. Describe in terms of the sets $\{A_n\}$ what it means for the sequence of functions $\{f_n\}$ (a) to converge pointwise, (b) to converge uniformly, (c) to converge almost everywhere, and (d) to converge in measure.

4:2.3 Characterize convergence in measure in the case where the measure is the counting measure.

4:2.4 Show that if $\{f_n\}$ is a sequence of measurable functions then the set of points x at which $\{f_n(x)\}$ converges to a finite limit is measurable.

4:2.5 Prove that if, for each $n \in \mathbb{N}$, f_n is finite a.e. and if $f_n \to f$ [a.e.] then f is finite [a.e.].

[Hint: This is a feature of Definition 4.10 and may not be true for other definitions of a.e. convergence.]

4:2.6 Verify that the sequence $\{f_n\}$ in Example 4.12 converges to 0 [a.e.], but does not converge [meas].

4:2.7 Prove that if $f_n \to f$ and $g_n \to g$ both in measure then $f_n + g_n \to f + g$ in measure.

4:2.8 (a) Prove that if $f_n \to f$ [meas], $g_n \to g$ [meas], and $\mu(X) < \infty$ then $f_n g_n \to fg$ [meas].

[Hint. Consider first the case that $f_n \to 0$ [meas] and $g_n \to 0$ [meas].]

(b) Use $f_n(x) = x$ and $g_n(x) = 1/n$ to show that the finiteness assumption in part (a) cannot be dropped.

4:2.9 Let $X = \mathbb{N}$, $\mathcal{M} = 2^{\mathbb{N}}$, and $\mu(\{n\}) = 2^{-n}$. Determine which of the four modes of convergence coincide in this case. [Hint: Uniform and pointwise do not coincide here.]

4:2.10 Let $(X, \mathcal{M}, \mu)$ be a measure space with $\mu(X) < \infty$. Prove that $f_n \to f$ [meas] if and only if every subsequence $\{f_{n_k}\}$ of $\{f_n\}$ has a subsequence $\{f_{n_{k_j}}\}$ such that $f_{n_{k_j}} \to f$ [a.e.].

4:2.11 Let $\{f_n\}$ be a sequence of measurable functions on a finite measure space $(X, \mathcal{M}, \mu)$, and let α_n be a sequence of positive numbers. Suppose that

$$\sum_{n=1}^{\infty} \mu\left(\{x \in X : |f_n(x)| > \alpha_n\}\right) < \infty.$$

Prove that

$$-1 \leq \liminf_{n\to\infty} \frac{f_n(x)}{\alpha_n} \leq \limsup_{n\to\infty} \frac{f_n(x)}{\alpha_n} \leq 1$$

for μ–almost every $x \in X$.

4.3 Egoroff's Theorem

We saw in Section 4.2 that neither of the two forms of convergence, convergence a.e. and convergence in measure, implies the other. We now develop a third form of convergence that is stronger than these two, but weaker than uniform convergence. If $\{f_n\}$ converges *uniformly* to f on X, we write

$$\lim_n f_n = f \text{ [unif] or } f_n \to f \text{ [unif].}$$

Almost uniform convergence is just uniform convergence off a set of arbitrarily small measure.

Definition 4.15: Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{f_n\}$ be a sequence of finite a.e., measurable functions on X. We say that $\{f_n\}$ converges *almost uniformly* to f on X if for every $\varepsilon > 0$ there exists a measurable set E such that $\mu(X \setminus E) < \varepsilon$ and $\{f_n\}$ converges uniformly to f on E. We then write

$$\lim_n f_n = f \text{ [a.u.] or } f_n \to f \text{ [a.u.].}$$

It is instructive to compare convergence [a.u.] with convergence [meas]. Suppose that $f_n \to f$ [meas] on X. Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that, for all $n \geq N$,

$$|f_n(x) - f(x)| < \varepsilon$$

for all x in a set A_n with $\mu(X \setminus A_n) < \varepsilon$. The sets A_n can vary with n. In Example 4.13, the sets $X \setminus A_n$ "slide" so much that $\{f_n(x)\}$ converge for no $x \in [0, 1]$. Convergence [a.u.] requires that a *single* set E suffice for all sufficiently large n: the set E does not depend on n.

Almost uniform convergence, in general, implies both convergence [a.e.] and convergence [meas]. (We leave verification of these facts as Exercise 4:3.1.) Neither converse is true. Example 4.13 and the functions $f_n(x) = x/n$, $x \in \mathbb{R}$, show this.

On a finite measure space convergence [a.u.] and convergence [a.e.] are equivalent. This is a form of a theorem due to D. Egoroff (1869–1931) (also transliterated sometimes as Egorov). One obtains the immediate corollary that, when $\mu(X) < \infty$, convergence [a.e.] implies convergence [meas]. If the measure space is not finite then different conditions are needed[2].

Theorem 4.16 (Egoroff) *Let $(X, \mathcal{M}, \mu)$ be a measure space with $\mu(X) < \infty$. Let $\{f_n\}$ be a sequence of finite a.e., measurable functions such that $f_n \to f$ [a.e.]. Then $f_n \to f$ [a.u.].*

Proof. For every $n, k \in \mathbb{N}$, let

$$A_{nk} = \bigcap_{m=n}^{\infty} \left\{x : |f_m(x) - f(x)| < \frac{1}{k}\right\}.$$

The function f is measurable, from which it follows that each of the sets A_{nk} is measurable. Let

$$E = \left\{x : \lim_n |f_n(x) - f(x)| = 0\right\}.$$

Since $f_n \to f$ [a.e.], E is measurable, $\mu(E) = \mu(X)$, and for each $k \in \mathbb{N}$, $E \subset \bigcup_{n=1}^{\infty} A_{nk}$. For fixed k, the sequence $\{A_{nk}\}_{n=1}^{\infty}$ is expanding, so that

$$\lim_n \mu(A_{nk}) = \mu\left(\bigcup_{n=1}^{\infty} A_{nk}\right) \geq \mu(E) = \mu(X).$$

Since $\mu(X) < \infty$,

$$\lim_n \mu(X \setminus A_{nk}) = 0. \tag{2}$$

[2]See R. G. Bartle, *An extension of Egorov's theorem.* Amer. Math. Monthly 87 (1980), no. 8, 628–633.

Now let $\varepsilon > 0$. It follows from (2) that there exists $n_k \in \mathbb{N}$ such that

$$\mu(X \setminus A_{n_k k}) < \varepsilon 2^{-k}. \tag{3}$$

We have shown that for each $\varepsilon > 0$ there exists $n_k \in \mathbb{N}$ such that inequality (3) holds. Let

$$A = \bigcap_{k=1}^{\infty} A_{n_k k}.$$

We now show that $\mu(X \setminus A) < \varepsilon$ and that $f_n \to f$ [unif] on A. It is clear that A is measurable. Furthermore,

$$\mu(X \setminus A) = \mu\left(\bigcup_{k=1}^{\infty} (X \setminus A_{n_k k})\right) \leq \sum_{k=1}^{\infty} \mu\left(X \setminus A_{n_k k}\right) < \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} = \varepsilon.$$

We see from the definition of the sets A_{nk} that, for $m \geq n_k$,

$$|f_m(x) - f(x)| < \frac{1}{k}$$

for every $x \in A_{n_k k}$ and therefore for every $x \in A$. Thus $f_n \to f$ [unif] on A as we wished to show. ■

One often restricts one's attention to some measurable subset E of X. It is clear how the concepts and results of this section apply to this setting. For example, if $\mu(E) < \infty$, then $f_n \to f$ [a.u.] on E whenever $f_n \to f$ [a.e.] on E, even if $\mu(X) = \infty$.

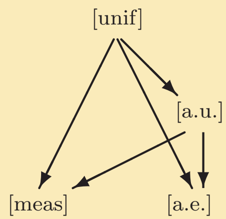


Figure 4.1. Comparison of modes of convergence in a general measure space.

4.3.1 Comparisons

We summarize our comparison of the modes of convergence with two figures.[3] In each case, we assume that $\{f_n\}$ is a sequence of finite a.e., measurable functions on X. Figure 4.1 shows the situation in a general measure space. Figure 4.2 gives the implications that are valid when $\mu(X) < \infty$. Where an arrow is missing, a counterexample is needed. The sliding sequence of Example 4.13 shows that convergence in measure does not imply any of the other forms of convergence, even when $\mu(X) < \infty$. The sequence $\{x^n\}$ shows uniform convergence is not implied by any other form of convergence. Finally, the sequence $\{x/n\}$ shows that convergence [a.e.] does not in general imply convergence [a.u.] or convergence [meas].

We view the implications given in the figures as preliminary comparisons of four forms of convergence. In Chapter 5, we shall study a fifth form of convergence, called *mean convergence*,

[3]These figures have been popular for many years, since appearing in M. E. Munroe, *Introduction to Measure and Integration*, Addison-Wesley (1953).

$$
\begin{tikzcd}
& \text{[unif]} \arrow[ddl] \arrow[ddr] \arrow[dr] & \\
& & \text{[a.u.]} \arrow[dll] \arrow[d, shift right] \\
\text{[meas]} & & \text{[a.e.]} \arrow[u, shift right] \arrow[ll]
\end{tikzcd}
$$

Figure 4.2. Comparison of modes of convergence in a *finite* measure space.

and indicate its "place" in the diagrams. We shall also provide a third diagram that applies even when $\mu(X) = \infty$ if functions in the sequence are suitably dominated by some integrable function. Exercise 4:3.4 provides an example in this spirit, but not expressed in the language of integration.

Exercises

4:3.1 Prove that if $f_n \to f$ [a.u.] on X then $f_n \to f$ [a.e.] on X and $f_n \to f$ [meas] on X.

4:3.2 By quoting results of this section or by other means, verify each implication appearing in the figures. Also verify that no additional implications can be added to the diagrams.

4:3.3 Let α_n be a sequence of positive numbers converging to zero. If f is continuous, then certainly $f(x - \alpha_n)$ converges to $f(x)$. Find a bounded measurable function on $[0, 1]$ such that the sequence of functions $f_n(x) = f(x - \alpha_n)$ is not a.e. convergent to f.

[Hint: Take the characteristic function of a Cantor set of positive measure.]

4:3.4◇ Let $\{f_n\}$ be a sequence of Lebesgue measurable functions on $[0,\infty)$ such that $|f_n(x)| \leq e^{-x}$ for all $x \in [0,\infty)$. If $f_n \to 0$ [a.e.], then $f_n \to 0$ [a.u.].

[Hint: The only place where we used our assumption that $\mu(X) < \infty$ in the proof of Theorem 4.16 was to obtain the limit in equation (2).]

4:3.5 Prove another version of Egoroff's theorem:

> **Theorem** *Let $(X, \mathcal{M}, \mu)$ be a finite or σ-finite measure space. Let $\{f_n\}$ be a sequence of finite a.e., measurable functions such that $f_n \to f$ [a.e.]. Then there is a partition of X into a sequence E_0, E_1, E_2, ... of disjoint measurable sets such that $\mu(E_0) = 0$ and $f_n \to f$ uniformly on each E_i, $i \geq 1$.*

4.4 Approximations by Simple Functions

A recurring theme in our development has been to find approximations to complicated objects by simpler ones. Naturally, we wish to do the same for measurable functions. The simplest measurable functions in a general space are those that are linear combinations of characteristic functions of measurable sets. In this section we show that these *simple* functions can be used to approximate general measurable functions. The simplest measurable functions in a metric space are continuous. In the next section we show that all measurable functions in a metric space furnished with an appropriate measure can be approximated by continuous functions.

We have not seen many examples of measurable functions and may not appreciate just how they come about or just how complicated they may, at first, appear. Thus it is instructive to begin with an example that exhibits some interesting features.

Example 4.17: We work on the interval $I_0 = (0,1)$. Each $x \in I_0$ has a unique base 2 expansion $x = .a_1a_2\ldots$ that does not end in a string of 1's. For each $i \in \mathbb{N}$, a_i is a function of x with only a finite number of discontinuities. Thus a_i is Borel measurable. For each $n \in \mathbb{N}$, let

$$f_n(x) = \frac{a_1(x) + a_2(x) + \cdots + a_n(x)}{n}.$$

Finally, let

$$f(x) = \limsup_n f_n(x).$$

One verifies easily that f is Borel measurable. Observe that, while $f_n(x)$ depends only on the first n bits in the binary expansion of x, f depends only on the "tail" of the expansion. If

$$x = .a_1a_2\ldots \quad \text{and} \quad y = .b_1b_2\ldots$$

and if there exists $j, N \in \mathbb{N}$ such that

$$b_{k+j} = a_k \quad \text{for all} \quad k \geq N,$$

then $f(x) = f(y)$. One can also verify that, for every nondegenerate interval $I \subset I_0$, f maps I onto the interval [0,1]. For example, any x whose expansion has the tail $.\overline{1000}$ will map onto $\frac{1}{4}$ (decimal), and the set of all such x is dense in I_0. (Some other features of f and related functions appear in Exercise 4:4.2.)

Notice one remarkable feature of the Borel measurable function f: it takes every one of its values on a dense set.[4] Despite this apparent complexity, we can still approximate such a function by much simpler functions, indeed by a continuous function as we will see in the next section.

[4]Functions with this property can arise quite naturally. See Exercise 7:8.15.

Definition 4.18: Let $E_1, E_2, \dots, E_n$ be pairwise disjoint measurable sets, and let c_1, c_2, $\dots$, c_n be real numbers. Let

$$f = c_1\chi_{E_1} + \cdots + c_n\chi_{E_n}.$$

Then f is called a *simple* function.

We can deduce that a simple function is one that takes on only finitely many values, all real. Each value is assumed on a measurable set. Our restriction that the sets E_i be measurable guarantees that simple functions are measurable. If the sets $E_1, E_2, \dots, E_n$ are measurable but not assumed to be pairwise disjoint the definition would be equivalent, but it is less transparent then exactly what values the function assumes.

Theorem 4.19 (Approximation by simple functions) *Let $(X, \mathcal{M}, \mu)$ be a measure space, and let f be measurable on X. Then there exists a sequence $\{f_n\}$ of simple functions such that*

$$\lim_n f_n(x) = f(x) \ \ \textit{for all} \ \ x \in X.$$

If $f(x) \geq 0$ for all $x \in X$, the sequence $\{f_n\}$ can be chosen to be a nondecreasing sequence, so that $f_n(x) \leq f_{n+1}(x)$ for all $n \in \mathbb{N}$ and $x \in X$. If f is bounded on X, then $f_n \to f$ [unif].

Proof. Suppose first that f is nonnegative. Fix $n \in \mathbb{N}$. For each $k = 1, 2, \dots, n2^n$, let

$$J_k = \left[\frac{k-1}{2^n}, \frac{k}{2^n}\right).$$

Let

$$f_n(x) = \begin{cases} \frac{k-1}{2^n}, & \text{if } f(x) \in J_k; \\ n, & \text{otherwise.} \end{cases}$$

The intervals J_k are pairwise disjoint, and

$$\bigcup_{k=1}^{n2^n} J_k = [0,n).$$

Since f is measurable, so is the function f_n. It is clear that f_n is a simple function and that $f_n(x) \leq f(x)$ for all $x \in X$. It is also clear for every $x \in X$, that $f_{n+1}(x) \geq f_n(x)$. Also

$$f_{n+1}(x) - f_n(x) \leq \frac{1}{2^{n+1}}$$

if $f(x) \leq n$, and

$$f_{n+1}(x) - f_n(x) \leq 1$$

if $n < f(x)$. It follows that

$$\lim f_n(x) = f(x)$$

and that the convergence is uniform if f is bounded. [Indeed, if $0 \leq f(x) \leq M$ for all $x \in X$, then

$$f_{n+1}(x) - f(x) \leq \frac{1}{2^n}$$

for all $n \geq M$, so that the convergence is uniform.]

In the general case, f need not be nonnegative. Let

$$f^+(x) = \begin{cases} f(x), & \text{if } f(x) \geq 0; \\ 0, & \text{if } f(x) < 0; \end{cases}$$

and let

$$f^-(x) = \begin{cases} -f(x), & \text{if } f(x) < 0; \\ 0, & \text{if } f(x) \geq 0. \end{cases}$$

Then $f = f^+ - f^-$. Each of the functions f^+ and f^- is measurable and nonnegative. Thus there exist sequences $\{g_n\}$ and $\{h_n\}$ of simple functions having the desired properties with respect to f^+ and f^-, respectively. For each $n \in \mathbb{N}$, let

$$f_n = g_n - h_n.$$

The sequence $\{f_n\}$ has all the required properties. ■

4.4.1 Approximation by bounded, measurable functions

Our next result provides a sense of how measurable functions that are finite a.e. can be approximated by bounded measurable functions. Since these, in turn, can be approximated uniformly by simple functions, we are close to an understanding of the structure of arbitrary measurable functions.

Theorem 4.20: *Suppose that f is finite a.e. and measurable on X with $\mu(X) < \infty$. Let $\varepsilon > 0$. Then there exists a bounded measurable function g such that*

$$\mu(\{x : g(x) \neq f(x)\}) < \varepsilon.$$

Proof. Let

$$A_\infty = \{x : |f(x)| = \infty\},$$

and for every $k \in \mathbb{N}$ let

$$A_k = \{x : |f(x)| > k\}.$$

By hypothesis, $\mu(A_\infty) = 0$. The sequence $\{A_k\}$ is a descending sequence of measurable sets, and $A_\infty = \bigcap_{k=1}^\infty A_k$. Since $\mu(X) < \infty$, it follows from Theorem 2.21 (ii) that

$$\lim_k \mu(A_k) = \mu(A_\infty) = 0.$$

Thus there exists $K \in \mathbb{N}$ such that $\mu(A_K) < \varepsilon$. Let

$$g(x) = \begin{cases} f(x), & \text{if } x \notin A_K; \\ 0, & \text{if } x \in A_K. \end{cases}$$

Then g is measurable, and $|g(x)| \leq K$ for all $x \in X$. Now

$$\{x : g(x) \neq f(x)\} = A_K$$

and $\mu(A_K) < \varepsilon$, so g is the required function. ■

Exercises

4:4.1 Show that the following statement is equivalent to (but different from Definition 4.18): A function f is a *simple function* if there exists collections $E_1, E_2, \ldots, E_n$ of measurable sets and real numbers $c_1, c_2, \ldots, c_n$ such that

$$f = c_1\chi_{E_1} + \cdots + c_n\chi_{E_n}.$$

4:4.2◇ Let f be the function on $(0,1)$ defined in Example 4.17.

(a) Prove that $f(I) = [0,1]$ for every open interval $I \subset I_0$. That is, for every $c \in [0,1]$, the set $f^{-1}(c)$ is dense in I_0.

(b) Prove that the graph of f is dense in $I_0 \times [0,1]$.

(c) Let

$$g(x) = \begin{cases} f(x), & \text{if } f(x) \neq x; \\ 0, & \text{if } f(x) = x. \end{cases}$$

Show that g has the properties of f given in (a) and (b).

(d) Show that the graph of g is not a connected subset of $\mathbb{R}^2$.

(e) Show that $h(x) = g(x) - x$ does not have the Darboux property.

We have mentioned that some nineteenth century mathematicians believed that the Darboux property (intermediate-value property) should be taken as a definition of continuity. They obviously were not aware of functions such as f and g above, nor of the function $h(x) = g(x) - x$. The function h is the sum of a Darboux function with a genuinely continuous function.

4:4.3 Show that the class of simple functions on a measure space is closed under linear combinations and products.

4:4.4 Characterize those functions that can be expressed as uniform limits of simple functions.

4:4.5 Let $I_1, I_2, \ldots, I_n$ be pairwise disjoint intervals with $[a,b] = \bigcup_{k=1}^n I_k$, and let c_1, c_2, ..., c_n be real numbers. Let $f = \sum_{k=1}^n c_k \chi_{I_k}$. Then f is called a *step function*.

(a) Show that every step function is a simple function for Lebesgue measure.

(b) Show that the proof of Theorem 4.19 applied to the function $f(x) = x$ on $[a,b]$ shows that f can be expressed as a uniform limit of step functions.

(c) Can every bounded measurable function on $[a,b]$ be expressed as a uniform limit of step functions?

(d) Characterize those functions that can be expressed as uniform limits of step functions. (This is harder.)

4:4.6 Let $f : X \to [0, +\infty]$ be measurable, and let $\{r_k\}$ be any sequence of positive numbers for which $r_k \to 0$ and $\sum_{k=1}^\infty r_k = +\infty$. Then there are measurable sets $\{A_k\}$ so that

$$f(x) = \sum_{k=1}^\infty r_k \chi_{A_k}(x)$$

at every $x \in X$.

Hint: Inductively define the sets $A_k = \left\{ x \in X : f(x) \geq r_k + \sum_{j<k} r_j \chi_{A_j}(x) \right\}$.

4.5 Approximation by Continuous Functions

We turn now to the problem of approximating a measurable function by a continuous one. We shall show that, under suitable hypotheses, we can redefine a measurable function f on a small set so that the new function g is continuous.

Throughout this section we take X to be a metric space and μ to be a Borel measure with $\mu(X) < \infty$. We also assume the following.

4.21: *If E is measurable and $\varepsilon > 0$, then there exists a closed set $F \subset E$ such that $\mu(E \setminus F) < \varepsilon$.*

We recall that when E is also a Borel set this inner approximation by a closed set is always available (see Corollary 3.15). The force of this assumption is that all *measurable* sets are assumed to have the same property. For example, if μ is a Lebesgue–Stieltjes measure on $\mathbb{R}$ with $\mu(\mathbb{R}) < \infty$, Theorem 3.20 (iii) can be used to show that assertion 4.21 applies.

Before we embark on our program of approximating measurable functions, even badly behaved ones like the function f of Example 4.17, by continuous functions, we discuss briefly the notions of *relative continuity* and *extendibility*.

Suppose that X is a metric space, and $S \subset X$. Let $f : X \to \mathbb{R}$, and let $s_0 \in S$. The statement that f is continuous at s_0 means that

$$\lim_{x \to s_0} f(x) = f(s_0).$$

It may be that f is discontinuous at s_0, but continuous at s_0 relative to the set S, that is

$$\lim_{x \to s_0,\ x \in S} f(x) = f(s_0).$$

In other words, the restriction of the function f to the set S is continuous at s_0. It is possible that $f|S$ is continuous, but cannot be extended to a function continuous on all of X. For example, $f(x) = \sin x^{-1}$ is continuous on $S = (0, 1]$, but cannot be extended to a continuous function on $[0, 1]$. For that, one needs f to be uniformly continuous on S.

4.5.1 Tietze extension theorem

We make use of the Tietze extension theorem that we will establish in Chapter 9 in greater generality for functions defined on metric spaces. We prove it here only for the case of functions on the real line.

Theorem 4.22 (Tietze extension theorem) *Let S be a closed subset of a metric space X and suppose that $f : S \to \mathbb{R}$ is continuous. Then f can be extended to a continuous function g defined on all of X. Furthermore, if $|f(x)| \le M$ on S, then $|g(x)| \le M$ on X.*

Proof. For $X = \mathbb{R}$, this is easy to prove. Let $\{(a_n, b_n)\}$ be the sequence of intervals complementary to S. Define g to be equal to f on S, and to be linear and continuous on each interval $[a_n, b_n]$ if $-\infty < a_n < b_n < \infty$. If $a_n = -\infty$ or $b_n = \infty$, we define g to be the appropriate constant on $(-\infty, b_n]$ or $[a_n, \infty)$. One verifies easily that g is continuous on $\mathbb{R}$. Note also that if $|f(x)| \leq M$ on S then $|g(x)| \leq M$ on $\mathbb{R}$. ■

We shall use the Tietze extension theorem in conjunction with "inside" approximation of measurable sets by closed sets. For this we shall use Corollary 3.15. We approximate X by closed sets. On these closed sets we shall obtain continuous functions that approximate our measurable function f. These functions can, in turn, be extended to functions continuous on all of X. We shall obtain a succession of theorems, each improving the sense of approximation of f by continuous functions. Each of these theorems is of interest in itself.

4.5.2 Lusin's theorem

The theorems just discussed culminate in an important theorem discovered independently by Guiseppe Vitali (1875–1932) and Nikolai Lusin (1883–1950). It is almost universally called Lusin's theorem. It asserts that for every $\varepsilon > 0$ there is a continuous function g defined on X such that $g = f$ except on a set of measure less than ε. (Lusin, often transliterated as Luzin, was a student of Egoroff, who is known mainly for the theorem on almost uniform convergence that we have just seen in the preceding section.)

Since we have not yet proved the Tietze extension theorem in a general metric space, the reader may wish to take X in the theorem to be an interval $[a, b]$ in $\mathbb{R}$.

Theorem 4.23: *Let $(X, \mathcal{M}, \mu)$ be a finite measure space with X a metric space and μ a Borel measure. Suppose that $\mathcal{M}$ satisfies condition 4.21. Let f be finite a.e. and measurable on X. Then to each pair (ε, η) of positive numbers corresponds a bounded, continuous function g such that*

$$\mu(\{x : |f(x) - g(x)| \geq \eta\}) < \varepsilon.$$

Furthermore, if $|f(x)| \leq M$ on X, then one can choose g so that $|g(x)| \leq M$ on X.

Proof. Suppose first that $|f(x)| \leq M$ on X. By Theorem 4.19 there exists a simple function h, also bounded by M, such that

$$|h(x) - f(x)| < \eta \qquad (x \in X).$$

Let $c_1, \ldots, c_m$ be the values that h assumes on X, and for each $i = 1, \ldots, m$ let

$$E_i = \{x : h(x) = c_i\}.$$

The sets E_i are pairwise disjoint and cover X. Choose closed sets $F_1, \ldots, F_m$ such that, for each $i = 1, \ldots, m$, $F_i \subset E_i$ and

$$\mu(E_i \setminus F_i) < \frac{\varepsilon}{m}.$$

Let

$$F = F_1 \cup \cdots \cup F_m.$$

Then F is closed, $F \subset X$ and $\mu(X \setminus F) < \varepsilon$. Furthermore, the restriction of h to F_i, $h|F_i$, is constant for $i = 1, \ldots, m$. It follows that $h|F$ is continuous.

To see this, we need only note that, if $x_0 \in F_i$ and $x_n \to x_0$ with $x_n \in F$ for all n, then for n sufficiently large $x_n \in F_i$, a set on which h is constant. By the Tietze extension theorem the function $h|F$ can be extended to a function g continuous on X with $|g(x)| \le M$ on X. Since

$$\mu(X \setminus F) < \varepsilon,$$

g is the desired function.

The general case in which we do not assume f bounded now follows readily from Theorem 4.20. ■

Theorem 4.24: *Let $(X, \mathcal{M}, \mu)$ be a finite measure space with X a metric space and μ a Borel measure. Suppose that $\mathcal{M}$ satisfies condition 4.21. Let f be finite a.e. and measurable on X. There exists a sequence $\{g_k\}$ of bounded, continuous functions for which $g_k \to f$ [a.u.].*

Proof. It follows immediately from Theorem 4.23 that there exists a sequence $\{f_n\}$ of continuous functions for which $f_n \to f$[meas]. By Theorem 4.14, there exists a subsequence $\{f_{n_k}\}$ such that $f_{n_k} \to f$ [a.e.]. The desired conclusion now follows from Egoroff's theorem, by defining $g_k = f_{n_k}$. ■

We are now ready to state and prove the main theorem of this section.

Theorem 4.25 (Lusin) *Let $(X, \mathcal{M}, \mu)$ be a finite measure space with X a metric space and μ a Borel measure. Suppose that $\mathcal{M}$ satisfies condition 4.21. Let f be finite a.e. and measurable on X, and let $\varepsilon > 0$. There exists a continuous function g on X such that $f(x) = g(x)$ for all x in a closed set F with $\mu(X \setminus F) < \varepsilon$. If $|f(x)| \le M$ for all $x \in X$, we can choose g to satisfy $|g(x)| \le M$ for all $x \in X$.*

Proof. By Theorem 4.24, there exists a measurable set E such that $\mu(X \setminus E) < \varepsilon/2$ and a sequence $\{g_k\}$ of continuous functions on X such that $g_k \to f$ [unif] on E. By condition 4.21, there exists a closed set $F \subset E$ such that $\mu(X \setminus F) < \varepsilon$. Since $g_k \to f$ [unif] on E, the restriction $f|F$ of f to F is continuous. By Tietze's theorem, this function can be extended to a function g continuous on all of X, so that g and f have the same bounds on X. ■

4.5.3 Further discussion

Let us return for a moment to Example 4.17. How complicated must a continuous function g be to approximate the function f of that example in the Lusin sense? A theorem in number theory asserts that almost every number in $[0,1]$ is "normal".[5] This means that for almost all $x \in [0,1]$ the binary expansion of x has, in the limit, half the bits equaling zero and half equaling one. More precisely, for almost every x in the interval $[0,1]$ with $x = .a_1a_2a_3\ldots$ the binary expansion of x, it is true that

$$\lim_n \frac{a_1 + \cdots + a_n}{n} = \tfrac{1}{2}.$$

Thus the function f in Example 4.17 satisfies $f(x) = \frac{1}{2}$ a.e. In other words, we can choose $g \equiv \frac{1}{2}$ and conclude that $f = g$ a.e. The approximation was not so difficult in this case! Here we have a much stronger result than Lusin's theorem guarantees. The exceptional set has measure zero.

When we approximate measurable sets by simpler sets, we get the following results. If we are willing to ignore sets of arbitrarily small measure, we can take the approximating sets to be open or closed. If we are willing to ignore only zero measure sets, we must give up a bit of the

[5]See Hardy and Wright, *An Introduction to the Theory of Numbers*, Oxford (1938).

regularity of the approximating sets—we can use sets of type $\mathcal{G}_\delta$ on the outside and sets of type $\mathcal{F}_\sigma$ on the inside.

The analogous situation for the approximation of measurable functions would suggest something similar. If we are willing to ignore sets of arbitrarily small measure, we can choose the approximating functions to be continuous. This is Lusin's theorem. Observe that for a continuous function g the associated sets

$$\{x : \alpha < g(x) < \beta\} \text{ and } \{x : \alpha \leq g(x) \leq \beta\}$$

are open and closed, respectively. One might expect that, if one is willing to ignore only sets of measure zero, we can choose the approximating functions g in the *first Borel class*; that is, one for which the corresponding associated sets are of type $\mathcal{F}_\sigma$ and $\mathcal{G}_\delta$, respectively. This is not quite the case. Instead, g can be taken from the *second Borel class* where the associated sets are of type $\mathcal{G}_{\delta\sigma}$ and $\mathcal{F}_{\sigma\delta}$, respectively. Exercise 4:6.2 at the end of the chapter deals with the Borel and Baire classes of functions and with how one can approximate measurable functions by functions from these classes.

Exercises

4:5.1 Complete the proof of Theorem 4.23 for the case f unbounded.

4:5.2 Show that Lusin's theorem is valid on $(\mathbb{R}, \mathcal{M}, \mu_f)$, where μ_f is a Lebesgue–Stieltjes measure, even if $\mu_f(\mathbb{R}) = \infty$.

4:5.3 Let $X = \mathbb{Q} \cap [0, 1]$ and $\mathcal{M} = 2^X$.

(a) Let μ be the counting measure on X, let Q_1 and Q_2 be complementary dense subsets of X, and let $f = \chi_{Q_1}$. Show that the conclusion of Lusin's theorem fails. What hypotheses in Lusin's theorem fail here?

(b) Let r_1, r_2, r_3, ... be an enumeration of the rationals, and let μ be the measure that assigns value 2^{-i} to the singleton set $\{r_i\}$. Let f be as in (a). Show how to construct the function g called for in the conclusion of Lusin's theorem.

4:5.4 The purpose of this exercise is to show the essential role that the regularity condition 4.21 plays in the hypotheses of Lusin's theorem. Let E be a subset of $[0, 1]$ such that both E and its complement $\widetilde{E}$ are totally imperfect (see Section 3.12). Let $f = \chi_E$. Let g be Lebesgue measurable, and suppose that $L = \{x : f(x) = g(x)\} \in \mathcal{L}$.

(a) Show that $\lambda_*(E) = 0$ and $\lambda^*(E) = 1$.

(b) Show that

$$E \cap L = \{x : f(x) = 1\} \cap L = \{x : g(x) = 1\} \cap L$$

and hence that $E \cap L \in \mathcal{L}$. Similarly, show that $\widetilde{E} \cap L \in \mathcal{L}$.

(c) Show that $E \cap L \subset E$ and $\lambda_*(E) = 0$, and hence that $\lambda(E \cap L) = 0$. Similarly show that $\lambda(\widetilde{E} \cap L) = 0$ and $\lambda(L) = 0$. (Recall that $\widetilde{E}$ denotes the complement of E.)

We have shown that if $\lambda_*(E) = 0$ and $\lambda^*(E) = 1$, for $E \subset [0, 1]$, then the function χ_E is not λ-measurable on any set of positive Lebesgue measure. We now use this fact to show that Lusin's theorem can fail dramatically when the condition 4.21 is not hypothesized.

Refer to Exercise 3:13.13. Let $\overline{\overline{\lambda}}$ be the extension of λ to the σ-algebra generated by $\mathcal{L}$ and $\{E\}$. Note that the measure space $([0, 1], \mathcal{M}, \overline{\overline{\lambda}})$ does not satisfy the assertion 4.21.

(d) Show that $\overline{\overline{\lambda}}(L) = \lambda(L) = 0$.

Thus the $\overline{\overline{\lambda}}$-measurable function f does not agree with any function that is λ-measurable even on a set of positive Lebesgue measure. In particular, if g is continuous and $f(x) = g(x)$ for all x in a closed set F, then $\overline{\overline{\lambda}}(F) = \lambda(F) = 0$.

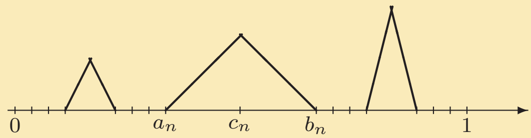


Figure 4.3. Construction of f in Exercise 4:6.1.

(e) Give an example of a λ-measurable function g (even a continuous one) such that $\overline{\overline{\lambda}}(\{x : f(x) = g(x$
1.

4.6 Additional Problems for Chapter 4

4:6.1 Let K be the Cantor ternary set, and let $\{(a_n, b_n)\}$ be the sequence of intervals complementary to K in $(0, 1)$. For each $n \in \mathbb{N}$, let $c_n = (a_n + b_n)/2$. Let $f = 0$ on K be linear and continuous on $[a_n, c_n]$ and on $[c_n, b_n]$, with the values $f(c_n)$ as yet unspecified (see Figure 4.3). What conditions on the values $f(c_n)$ are necessary and sufficient (a) for f to be continuous, (b) for f to be a Baire 1 function, or (c) for f to be of bounded variation? (See Exercise 4:6.2).

4:6.2$\diamondsuit$ **(Baire functions and Borel functions)** For this problem, all functions are assumed finite unless explicitly stated otherwise. Let $\mathcal{B}_0$ consist of the continuous functions on an interval $X \subset \mathbb{R}$. We do not assume X bounded.

(a) For $n \in \mathbb{N}$, let $\mathcal{B}_n$ consist of those functions that are pointwise limits of sequences of functions in $\mathcal{B}_{n-1}$. The class $\mathcal{B}_n$ is called the *Baire functions of class* n or *the Baire-*n *functions.* Prove that if $f \in \mathcal{B}_1$ then, for all $\alpha \in \mathbb{R}$, the sets $\{x : f(x) > \alpha\}$ and $\{x : f(x) < \alpha\}$ are of type $\mathcal{F}_\sigma$.

(b) Show that, if $f \in \mathcal{B}_2$, then for all $\alpha \in \mathbb{R}$ the sets

$$\{x : f(x) > \alpha\} \quad \text{and} \quad \{x : f(x) < \alpha\}$$

are of type $\mathcal{G}_{\delta\sigma}$.

(c) Show that a function $f : X \to \mathbb{R}$ that is continuous except on a countable set is in $\mathcal{B}_1$. (Compare with Exercise 4:1.16.)

(d) Let $f = \chi_{\mathbb{Q}}$. Show that $f \in \mathcal{B}_2 \setminus \mathcal{B}_1$.

(e) Prove that $\mathcal{B}_1$ is closed under addition and multiplication.

(f) Let $\{M_n\}$ be a sequence of positive numbers and suppose that $\sum_{n=1}^{\infty} M_n < \infty$. Let $\{f_n\} \subset \mathcal{B}_1$ with $|f_n(x)| \leq M_n$ for all $n \in \mathbb{N}$ and all $x \in X$. Prove that $\sum_{n=1}^{\infty} f_n \in \mathcal{B}_1$.

(g) Prove that if $f_n \to f$ [unif] and $f_n \in \mathcal{B}_1$ for all $n \in \mathbb{N}$ then $f \in \mathcal{B}_1$.

[Hint: Choose an increasing sequence $\{n_k\}$ of positive integers such that $\lim_k n_k = \infty$ and $|f_{n_k}(x) - f(x)| < 2^{-k}$ on X. Then apply (f) appropriately.]

(h) Prove that the composition of a function $f \in \mathcal{B}_1$ with a continuous function is in $\mathcal{B}_1$.

(i) Prove the converse to part (a): If for every $\alpha \in \mathbb{R}$ the sets $\{x : f(x) > \alpha\}$ and $\{x : f(x) < \alpha\}$ are of type $\mathcal{F}_\sigma$, then $f \in \mathcal{B}_1$.

(j) Prove that if f is differentiable then $f' \in \mathcal{B}_1$.

(k) Prove that if $\{f_n\} \subset \mathcal{B}_1$ then $\sup f_n \in \mathcal{B}_2$. [This assumes that $\sup f_n$ is a finite function.]

(l) Prove that if $\{f_n\} \subset \mathcal{B}_0$ then $\limsup_n f_n \in \mathcal{B}_2$. [This assumes that $\limsup f_n$ is a finite function.]

(m) Prove that if f is finite a.e. and measurable on X then there exists $g \in \mathcal{B}_2$ such that $f = g$ a.e..

(n) Give an example of a finite Lebesgue measurable function on $\mathbb{R}$ that agrees with no $g \in \mathcal{B}_1$ a.e..

[Hint: Using Exercise 2:14.9, let $f = \chi_A$ where $\lambda(I \cap A) > 0$ and $\lambda(I \setminus A) > 0$ for every open interval I. Show that if $g \in \mathcal{B}_1$ and $g = f$ a.e. then $\{x : g(x) = 0\}$ and $\{x : g(x) = 1\}$ are disjoint, dense subsets of $\mathbb{R}$ of type $\mathcal{G}_\delta$. This violates the Baire category theorem for $\mathbb{R}$.]

(o) The smallest class of functions that contains $\mathcal{B}_0$ and is closed under the operation of taking pointwise limits is called the class of *Baire functions*. It is true, though difficult to prove, that for each $n \in \mathbb{N}$ there exists $f \in \mathcal{B}_{n+1} \setminus \mathcal{B}_n$. Show that there exists a Baire function g on $X = [0, \infty)$ that is not in any of the classes $\mathcal{B}_n$.

[Hint: Let $g \in \mathcal{B}_{n+1} \setminus \mathcal{B}_n$ on $[n, n+1)$.]

This function is in the class $\mathcal{B}_\omega$, where ω is the first infinite ordinal. One then defines $\mathcal{B}_{\omega+1}$ as those functions that are limits of sequences of functions in $\mathcal{B}_\omega$. Using transfinite induction, one obtains classes $\mathcal{B}_\gamma$ for every countable ordinal. One can show that for every countable ordinal γ there exist functions $f \in \mathcal{B}_\gamma \setminus \bigcup_{\beta<\gamma} \mathcal{B}_\beta$. One can also show that the class of Baire functions on the interval X is exactly the class of Borel measurable functions.

(p) Use the fact that there are Lebesgue measurable sets that are not Borel sets to show that there are Lebesgue measurable functions that are not Baire functions.

4:6.3 Show that a function $f : \mathbb{R}^2 \to \mathbb{R}$ that is continuous in each variable separately is a Baire 1 function. (This is the original problem that led Baire to this line of research.)

[Hint: Define

$$F_n(x, y) = f((i+1)2^{-n}, y)[x - i2^{-n}] - f(i2^{-n}, y)[x - (i+1)2^{-n}]$$

if $i2^{-n} \le x < (i+1)2^{-n}$ for some integer i. Show that F_n is continuous on $\mathbb{R}^2$ and $2^n F_n \to f$ pointwise.]

4:6.4 Construct a function $f : [0,1] \to [0,1]$ as follows. Let $\{I_n\}$ be an enumeration of the open intervals in $[0,1]$ having rational endpoints. For each $n \in \mathbb{N}$, let $K_n \subset I_n$ be a Cantor set of positive Lebesgue measure such that the sequence $\{K_n\}$ is pairwise disjoint and $\sum_{n=1}^{\infty} \lambda(K_n) = 1$. Define f_n on K_n to be continuous on K_n, nondecreasing, and such that $f_n(K_n) = [0,1]$. Let

$$f(x) = \begin{cases} f_n(x), & \text{if } x \in K_n; \\ 0, & \text{if } x \in [0,1] \setminus \bigcup_{n=1}^{\infty} K_n. \end{cases}$$

(a) Show that f is Lebesgue measurable.

(b) Show that $f(I) = [0,1]$ for every open interval $I \subset [0,1]$.

(c) Using the sets K_n, find continuous functions on $[0,1]$ that approximate f in the Lusin sense.

(d) Refer to Exercise 4:6.2. Does there exist $g \in \mathcal{B}_1$ such that $g = f$ [a.e.]?

(e) Give an example of a function $g \in \mathcal{B}_2$ for which $f = g$ [a.e.]. [Hint: Easy.]

4:6.5 Measurability can be expressed as a separation property. Let μ^* be an outer measure on a space X. Show that a function $f : X \to [-\infty, +\infty]$ is measurable with respect to μ^* if and only if

$$\mu^*(T) \geq \mu^*(T \cap \{x \in X : f(x) \leq a\}) + \mu^*(T \cap \{x \in X : f(x) \geq b\})$$

for all $T \subset X$ and all $-\infty < a < b < +\infty$.

4:6.6 Let $(X, \mathcal{M}, \mu)$ be a measure space and, for every measurable function $f : X \to [-\infty, +\infty]$, define

$$\|f\|_\mu = \inf\{r : \mu(\{x : |f(x)| > r\}) \leq r\}.$$

(a) Show that $\mu\left(\{x : |f(x)| > \|f\|_\mu\}\right) \leq \|f\|_\mu$.

(b) Check the triangle inequality $\|f + g\|_\mu \leq \|f\|_\mu + \|g\|_\mu$.

(c) Show that $f_n \to f$ in μ–measure if and only if $\|f_n - f\|_\mu \to 0$.

(d) Show that, if $f = \chi_A$, then $\|cf\|_\mu = \inf\{c, \mu(A)\}$ for any $0 \le c < \infty$. In particular, it is not true in general that $\|cf\|_\mu = c\,\|f\|_\mu$.

(e) Show that, for $c > 0$,

$$\|cf\|_\mu \le \max\left\{\|f\|_\mu, c\,\|f\|_\mu\right\}$$

and hence that $\|cf\|_\mu \to 0$ as $\|f\|_\mu \to 0$.

(f) Show that if $\mu(\{x : f(x) \ne 0\}) < \infty$ and $\mu\{x : |f(x)| = \infty\} = 0$ then $\|cf\|_\mu \to 0$ as $c \to 0$.

(g) Show that every Cauchy sequence $\{g_k\}$ in measure has a subsequence that converges both μ–almost everywhere and in measure.

[Hint: Pick an increasing sequence $N(k)$ so that

$$\|g_i - g_j\|_\mu \le 2^{-n}$$

whenever $i \ge j \ge N(k)$.]

(h) Show that if

$$\sum_{k=1}^{\infty} \|g_{k+1} - g_k\|_\mu < \infty$$

then $\{g_k\}$ converges to some function g μ–almost everywhere, and $\|g_k - g\|_\mu$ converges to 0.

SUBJECT INDEX