Real Analysis

Modern Techniques and Their Applications

Second Edition

Gerald B. Folland

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

Content

0

Prologu

The purpose of this introductory chapter is to establish the notation and terminc that will be used throughout the book and to present a few diverse results fro theory and analysis that will be needed later. The style here is deliberately since this chapter is intended as a reference rather than a systematic exposition

0.1 THE LANGUAGE OF SET THEORY

It is assumed that the reader is familiar with the basic concepts of set theor following discussion is meant mainly to fix our terminology.

Number Systems. Our notation for the fundamental number systems follows:

$\mathbb{N}$ = the set of positive integers (not including zero)

$\mathbb{Z}$ = the set of integers

$\mathbb{Q}$ = the set of rational numbers

$\mathbb{R}$ = the set of real numbers

$\mathbb{C}$ = the set of complex numbers

Logic. We shall avoid the use of special symbols from mathematical preferring to remain reasonably close to standard English. We shall, howev the abbreviation **iff** for "if and only if."

One point of elementary logic that is often insufficiently appreciated by st

It is easily verified that the map $f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X)$ defined by the second formula commutes with union, intersections, and complements:

$$f^{-1}\Big(\bigcup_{\alpha\in A} E_\alpha\Big) = \bigcup_{\alpha\in A} f^{-1}(E_\alpha), \qquad f^{-1}\Big(\bigcap_{\alpha\in A} E_\alpha\Big) = \bigcap_{\alpha\in A} f^{-1}(E_\alpha),$$
$$f^{-1}(E^c) = \big(f^{-1}(E)\big)^c.$$

(The direct image mapping $f : \mathcal{P}(X) \to \mathcal{P}(Y)$ commutes with unions, but in general not with intersections or complements.)

If $f : X \to Y$ is a mapping, X is called the **domain** of f and $f(X)$ is called the **range** of f. f is said to be **injective** if $f(x_1) = f(x_2)$ only when $x_1 = x_2$, **surjective** if $f(X) = Y$, and **bijective** if it is both injective and surjective. If f is bijective, it has an **inverse** $f^{-1} : Y \to X$ such that $f^{-1} \circ f$ and $f \circ f^{-1}$ are the identity mappings on X and Y, respectively. If $A \subset X$, we denote by $f|A$ the restriction of f to A:

$$(f|A) : A \to Y, \qquad (f|A)(x) = f(x) \text{ for } x \in A.$$

A **sequence** in a set X is a mapping from $\mathbb{N}$ into X. (We also use the term **finite sequence** to mean a map from $\{1, \dots, n\}$ into X where $n \in \mathbb{N}$.) If $f : \mathbb{N} \to X$ is a sequence and $g : \mathbb{N} \to \mathbb{N}$ satisfies $g(n) < g(m)$ whenever $n < m$, the composition $f \circ g$ is called a **subsequence** of f. It is common, and often convenient, to be careless about distinguishing between sequences and their ranges, which are subsets of X indexed by $\mathbb{N}$. Thus, if $f(n) = x_n$, we speak of the sequence $\{x_n\}_1^\infty$; whether we mean a mapping from $\mathbb{N}$ to X or a subset of X will be clear from the context.

Earlier we defined the Cartesian product of two sets. Similarly one can define the Cartesian product of n sets in terms of ordered n-tuples. However, this definition becomes awkward for infinite families of sets, so the following approach is used instead. If $\{X_\alpha\}_{\alpha\in A}$ is an indexed family of sets, their **Cartesian product** $\prod_{\alpha\in A} X_\alpha$ is the set of all maps $f : A \to \bigcup_{\alpha\in A} X_\alpha$ such that $f(\alpha) \in X_\alpha$ for every $\alpha \in A$. (It should be noted, and then promptly forgotten, that when $A = \{1, 2\}$, the previous definition of $X_1 \times X_2$ is set-theoretically different from the present definition of $\prod_1^2 X_j$. Indeed, the latter concept depends on mappings, which are defined in terms of the former one.) If $X = \prod_{\alpha\in A} X_\alpha$ and $\alpha \in A$, we define the αth **projection** or **coordinate map** $\pi_\alpha : X \to X_\alpha$ by $\pi_\alpha(f) = f(\alpha)$. We also frequently write x and x_α instead of f and $f(\alpha)$ and call x_α the αth **coordinate** of x.

If the sets X_α are all equal to some fixed set Y, we denote $\prod_{\alpha\in A} X_\alpha$ by Y^A:

$$Y^A = \text{ the set of all mappings from } A \text{ to } Y.$$

If $A = \{1, \dots, n\}$, Y^A is denoted by Y^n and may be identified with the set of ordered n-tuples of elements of Y.

0.2 ORDERINGS

A **partial ordering** on a nonempty set X is a relation R on X with the following

- if xRy and yRz, then xRz;
- if xRy and yRx, then $x = y$;
- xRx for all x.

If R also satisfies

- if $x, y \in X$, then either xRy or yRx,

then R is called a **linear** (or **total**) ordering. For example, if E is any set, then $\mathcal{P}$ is partially ordered by inclusion, and $\mathbb{R}$ is linearly ordered by its usual orde Taking this last example as a model, we shall usually denote partial ordering $\leq$, and we write $x < y$ to mean that $x \leq y$ but $x \neq y$. We observe that a pa ordering on X naturally induces a partial ordering on every nonempty subset o Two partially ordered sets X and Y are said to be **order isomorphic** if there bijection $f : X \to Y$ such that $x_1 \leq x_2$ iff $f(x_1) \leq f(x_2)$.

If X is partially ordered by $\leq$, a **maximal** (resp. **minimal**) **element** of X i element $x \in X$ such that the only $y \in X$ satisfying $x \leq y$ (resp. $x \geq y$) is x i Maximal and minimal elements may or may not exist, and they need not be un unless the ordering is linear. If $E \subset X$, an **upper** (resp. **lower**) **bound** for E element $x \in X$ such that $y \leq x$ (resp. $x \leq y$) for all $y \in E$. An upper bound f need not be an element of E, and unless E is linearly ordered, a maximal eleme E need not be an upper bound for E. (The reader should think up some exampl

If X is linearly ordered by $\leq$ and every nonempty subset of X has a (necess unique) minimal element, X is said to be **well ordered** by $\leq$, and (in defiance c laws of grammar) $\leq$ is called a **well ordering** on X. For example, $\mathbb{N}$ is well orc by its natural ordering.

We now state a fundamental principle of set theory and derive some conseque of it.

0.1 The Hausdorff Maximal Principle. *Every partially ordered set has a ma linearly ordered subset.*

In more detail, this means that if X is partially ordered by $\leq$, there is a set E that is linearly ordered by $\leq$, such that no subset of X that properly includes linearly ordered by $\leq$. Another version of this principle is the following:

0.2 Zorn's Lemma. *If X is a partially ordered set and every linearly ordered s of X has an upper bound, then X has a maximal element.*

Clearly the Hausdorff maximal principle implies Zorn's lemma: An upper b for a maximal linearly ordered subset of X is a maximal element of X. It is als difficult to see that Zorn's lemma implies the Hausdorff maximal principle. (A Zorn's lemma to the collection of linearly ordered subsets of X, which is par ordered by inclusion.)

Proof. Let $\mathcal{W}$ be the collection of well orderings of subsets of X, and define a partial ordering on $\mathcal{W}$ as follows. If $\leq_1$ and $\leq_2$ are well orderings on the subsets E_1 and E_2, then $\leq_1$ precedes $\leq_2$ in the partial ordering if (i) $\leq_2$ extends $\leq_1$, i.e., $E_1 \subset E_2$ and $\leq_1$ and $\leq_2$ agree on E_1, and (ii) if $x \in E_2 \setminus E_1$ then $y \leq_2 x$ for all $y \in E_1$. The reader may verify that the hypotheses of Zorn's lemma are satisfied, so that $\mathcal{W}$ has a maximal element. This must be a well ordering on X itself, for if $\leq$ is a well ordering on a proper subset E of X and $x_0 \in X \setminus E$, then $\leq$ can be extended to a well ordering on $E \cup \{x_0\}$ by declaring that $x \leq x_0$ for all $x \in E$. ∎

0.4 The Axiom of Choice. *If $\{X_\alpha\}_{\alpha\in A}$ is a nonempty collection of nonempty sets, then $\prod_{\alpha\in A} X_\alpha$ is nonempty.*

Proof. Let $X = \bigcup_{\alpha\in A} X_\alpha$. Pick a well ordering on X and, for $\alpha \in A$, let $f(\alpha)$ be the minimal element of X_α. Then $f \in \prod_{\alpha\in A} X_\alpha$. ∎

0.5 Corollary. *If $\{X_\alpha\}_{\alpha\in A}$ is a disjoint collection of nonempty sets, there is a set $Y \subset \bigcup_{\alpha\in A} X_\alpha$ such that $Y \cap X_\alpha$ contains precisely one element for each $\alpha \in A$.*

Proof. Take $Y = f(A)$ where $f \in \prod_{\alpha\in A} X_\alpha$. ∎

We have deduced the axiom of choice from the Hausdorff maximal principle; in fact, it can be shown that the two are logically equivalent.

0.3 CARDINALITY

If X and Y are nonempty sets, we define the expressions

$$\text{card}(X) \leq \text{card}(Y), \qquad \text{card}(X) = \text{card}(Y), \qquad \text{card}(X) \geq \text{card}(Y)$$

to mean that there exists $f : X \to Y$ which is injective, bijective, or surjective, respectively. We also define

$$\text{card}(X) < \text{card}(Y), \qquad \text{card}(X) > \text{card}(Y)$$

to mean that there is an injection but no bijection, or a surjection but no bijection, from X to Y. Observe that we attach no meaning to the expression "card(X)" when it stands alone; there are various ways of doing so, but they are irrelevant for our purposes (except when X is finite — see below). These relationships can be extended to the empty set by declaring that

$$\text{card}(\varnothing) < \text{card}(X) \text{ and } \text{card}(X) > \text{card}(\varnothing) \text{ for all } X \neq \varnothing.$$

For the remainder of this section we assume implicitly that all sets in question are nonempty in order to avoid special arguments for $\varnothing$. Our first task is to prove that the relationships defined above enjoy the properties that the notation suggests.

0.6 Proposition. card$(X) \leq$ card(Y) *iff* card$(Y) \geq$ card(X).

Proof. If $f : X \to Y$ is injective, pick $x_0 \in X$ and define $g : Y \to X$ by $g(y) = f^{-1}(y)$ if $y \in f(X)$, $g(y) = x_0$ otherwise. Then g is surjective. Conversely, if $g : Y \to X$ is surjective, the sets $g^{-1}(\{x\})$ $(x \in X)$ are nonempty and disjoint, so any $f \in \prod_{x\in X} g^{-1}(\{x\})$ is an injection from X to Y. ∎

0.7 Proposition. *For any sets X and Y, either* card$(X) \leq$ card(Y) *or* card$(Y) \leq$ card(X).

Proof. Consider the set $\mathcal{I}$ of all injections from subsets of X to Y. The members of $\mathcal{I}$ can be regarded as subsets of $X \times Y$, so $\mathcal{I}$ is partially ordered by inclusion. It is easily verified that Zorn's lemma applies, so $\mathcal{I}$ has a maximal element f, with (say) domain A and range B. If $x_0 \in X \setminus A$ and $y_0 \in Y \setminus B$, then f can be extended to an injection from $A \cup \{x_0\}$ to $Y \cup \{y_0\}$ by setting $f(x_0) = y_0$, contradicting maximality. Hence either $A = X$, in which case card$(X) \leq$ card(Y), or $B = Y$, in which case f^{-1} is an injection from Y to X and card$(Y) \leq$ card(X). ∎

0.8 The Schröder-Bernstein Theorem. *If* card$(X) \leq$ card(Y) *and* card$(Y) \leq$ card(X) *then* card$(X) =$ card(Y).

Proof. Let $f : X \to Y$ and $g : Y \to X$ be injections. Consider a point $x \in X$: If $x \in g(Y)$, we form $g^{-1}(x) \in Y$; if $g^{-1}(x) \in f(X)$, we form $f^{-1}(g^{-1}(x))$; and so forth. Either this process can be continued indefinitely, or it terminates with an element of $X \setminus g(Y)$ (perhaps x itself), or it terminates with an element of $Y \setminus f(X)$. In these three cases we say that x is in X_∞, X_X, or X_Y; thus X is the disjoint union of X_∞, X_X, and X_Y. In the same way, Y is the disjoint union of three sets Y_∞, Y_X, and Y_Y. Clearly f maps X_∞ onto Y_∞ and X_X onto Y_X, whereas g maps Y_Y onto X_Y. Therefore, if we define $h : X \to Y$ by $h(x) = f(x)$ if $X \in X_\infty \cup X_X$ and $h(x) = g^{-1}(x)$ if $x \in X_Y$, then h is bijective. ∎

0.9 Proposition. *For any set X,* card$(X) <$ card$(\mathcal{P}(X))$.

Proof. On the one hand, the map $f(x) = \{x\}$ is an injection from X to $\mathcal{P}(X)$. On the other, if $g : X \to \mathcal{P}(X)$, let $Y = \{x \in X : x \notin g(x)\}$. Then $Y \notin g(X)$, for if $Y = g(x_0)$ for some $x_0 \in X$, any attempt to answer the question "Is $x_0 \in Y$?" quickly leads to an absurdity. Hence g cannot be surjective. ∎

A set X is called **countable** (or **denumerable**) if card$(X) \leq$ card$(\mathbb{N})$. In particular, all finite sets are countable, and for these it is convenient to interpret "card(X)" as the number of elements in X:

$$\text{card}(X) = n \text{ iff } \text{card}(X) = \text{card}(\{1, \ldots, n\}).$$

If X is countable but not finite, we say that X is **countably infinite**.

0.10 Proposition.

a. *If X and Y are countable, so is $X \times Y$.*

b. *If A is countable and X_α is countable for every $\alpha \in A$, then $\bigcup_{\alpha \in A} X_\alpha$ is countable.*

c. *If X is countably infinite, then* $\text{card}(X) = \text{card}(\mathbb{N})$.

Proof. To prove (a) it suffices to prove that $\mathbb{N}^2$ is countable. But we can define a bijection from $\mathbb{N}$ to $\mathbb{N}^2$ by listing, for n successively equal to $2, 3, 4, \ldots$, those elements $(j, k) \in \mathbb{N}^2$ such that $j + k = n$ in order of increasing j, thus:

$$(1,1),\ (1,2),\ (2,1),\ (1,3),\ (2,2),\ (3,1),\ (1,4),\ (2,3),\ (3,2),\ (4,1), \ldots$$

As for (b), for each $\alpha \in A$ there is a surjective $f_\alpha : \mathbb{N} \to X_\alpha$, and then the map $f : \mathbb{N} \times A \to \bigcup_{\alpha \in A} X_\alpha$ defined by $f(n, \alpha) = f_\alpha(n)$ is surjective; the result therefore follows from (a). Finally, for (c) it suffices to assume that X is an infinite subset of $\mathbb{N}$. Let $f(1)$ be the smallest element of X, and define $f(n)$ inductively to be the smallest element of $E \setminus \{f(1), \ldots, f(n-1)\}$. Then f is easily seen to be a bijection from $\mathbb{N}$ to X. ∎

0.11 Corollary. *$\mathbb{Z}$ and $\mathbb{Q}$ are countable.*

Proof. $\mathbb{Z}$ is the union of the countable sets $\mathbb{N}$, $\{-n : n \in \mathbb{N}\}$, and $\{0\}$, and one can define a surjection $f : \mathbb{Z}^2 \to \mathbb{Q}$ by $f(m, n) = m/n$ if $n \neq 0$ and $f(m, 0) = 0$. ∎

A set X is said to have the **cardinality of the continuum** if $\text{card}(X) = \text{card}(\mathbb{R})$. We shall use the letter $\mathfrak{c}$ as an abbreviation for $\text{card}(\mathbb{R})$:

$$\text{card}(X) = \mathfrak{c} \text{ iff } \text{card}(X) = \text{card}(\mathbb{R}).$$

0.12 Proposition. $\text{card}(\mathcal{P}(\mathbb{N})) = \mathfrak{c}$.

Proof. If $A \subset \mathbb{N}$, define $f(A) \in \mathbb{R}$ to be $\sum_{n \in A} 2^{-n}$ if $\mathbb{N} \setminus A$ is infinite and $1 + \sum_{n \in A} 2^{-n}$ if $\mathbb{N} \setminus A$ is finite. (In the two cases, $f(A)$ is the number whose base-2 decimal expansion is $0.a_1a_2\cdots$ or $1.a_1a_2\cdots$, where $a_n = 1$ if $n \in A$ and $a_n = 0$ otherwise.) Then $f : \mathcal{P}(\mathbb{N}) \to \mathbb{R}$ is injective. On the other hand, define $g : \mathcal{P}(\mathbb{Z}) \to \mathbb{R}$ by $g(A) = \log(\sum_{n \in A} 2^{-n})$ if A is bounded below and $g(A) = 0$ otherwise. Then g is surjective since every positive real number has a base-2 decimal expansion. Since $\text{card}(\mathcal{P}(\mathbb{Z})) = \text{card}(\mathcal{P}(\mathbb{N}))$, the result follows from the Schröder-Bernstein theorem. ∎

0.13 Corollary. *If* $\text{card}(X) \geq \mathfrak{c}$, *then X is uncountable.*

Proof. Apply Proposition 0.9. ∎

The converse of this corollary is the so-called continuum hypothesis, whose ...

0.14 Proposition.

a. *If* $\text{card}(X) \leq \mathfrak{c}$ *and* $\text{card}(Y) \leq \mathfrak{c}$, *then* $\text{card}(X \times Y) \leq \mathfrak{c}$.

b. *If* $\text{card}(A) \leq \mathfrak{c}$ *and* $\text{card}(X_\alpha) \leq \mathfrak{c}$ *for all $\alpha \in A$, then* $\text{card}(\bigcup_{\alpha \in A} X_\alpha) \leq$

Proof. For (a) it suffices to take $X = Y = \mathcal{P}(\mathbb{N})$. Define $\phi, \psi : \mathbb{N} \to$ ... $\phi(n) = 2n$ and $\psi(n) = 2n - 1$. It is then easy to check that the map $f : \mathcal{P}(\mathbb{N})$ ... $\mathcal{P}(\mathbb{N})$ defined by $f(A, B) = \phi(A) \cup \psi(B)$ is bijective. (b) follows from (a) as i... proof of Proposition 0.10.

0.4 MORE ABOUT WELL ORDERED SETS

The material in this section is optional; it is used only in a few exercises and in ... notes at the ends of chapters.

Let X be a well ordered set. If $A \subset X$ is nonempty, A has a minimal ele... which is its maximal lower bound or **infimum**; we shall denote it by $\inf A$. If ... bounded above, it also has a minimal upper bound or **supremum**, denoted by s... If $x \in X$, we define the **initial segment** of x to be

$$I_x = \{y \in X : y < x\}.$$

The elements of I_x are called **predecessors** of x.

The principle of mathematical induction is equivalent to the fact that $\mathbb{N}$ is ... ordered. It can be extended to arbitrary well ordered sets as follows:

0.15 The Principle of Transfinite Induction. *Let X be a well ordered set. I... a subset of X such that $x \in A$ whenever $I_x \subset A$, then $A = X$.*

Proof. If $X \neq A$, let $x = \inf(X \setminus A)$. Then $I_x \subset A$ but $x \notin A$.

0.16 Proposition. *If X is well ordered and $A \subset X$, then $\bigcup_{x \in A} I_x$ is either an ... segment or X itself.*

Proof. Let $J = \bigcup_{x \in A} I_x$. If $J \neq X$, let $b = \inf(X \setminus J)$. If there existed ... with $y > b$, we would have $y \in I_x$ for some $x \in A$ and hence $b \in I_x$, contr... construction. Hence $J \subset I_b$, and it is obvious that $I_b \subset J$.

0.17 Proposition. *If X and Y are well ordered, then either X is order isom... to Y, or X is order isomorphic to an initial segment in Y, or Y is order isom... to an initial segment in X.*

Proof. Consider the set $\mathcal{F}$ of order isomorphisms whose domains are ... segments in X or X itself and whose ranges are initial segments in Y or Y ... $\mathcal{F}$ is nonempty since the unique $f : \{\inf X\} \to \{\inf Y\}$ belongs to $\mathcal{F}$, an...

An application of Zorn's lemma shows that $\mathcal{F}$ has a maximal element f, with (say) domain A and range B. If $A = I_x$ and $B = I_y$, then $A \cup \{x\}$ and $B \cup \{y\}$ are again initial segments of X and Y, and f could be extended by setting $f(x) = y$, contradicting maximality. Hence either $A = X$ or $B = Y$ (or both), and the result follows. ∎

0.18 Proposition. *There is an uncountable well ordered set Ω such that I_x is countable for each $x \in \Omega$. If Ω' is another set with the same properties, then Ω and Ω' are order isomorphic.*

Proof. Uncountable well ordered sets exist by the well ordering principle; let X be one. Either X has the desired property or there is a minimal element x_0 such that I_{x_0} is uncountable, in which case we can take $\Omega = I_{x_0}$. If Ω' is another such set, Ω' cannot be order isomorphic to an initial segment of Ω or vice versa, because Ω and Ω' are uncountable while their initial segments are countable, so Ω and Ω' are order isomorphic by Proposition 0.17. ∎

The set Ω in Proposition 0.18, which is essentially unique *qua* well ordered set, is called the **set of countable ordinals**. It has the following remarkable property:

0.19 Proposition. *Every countable subset of Ω has an upper bound.*

Proof. If $A \subset \Omega$ is countable, $\bigcup_{x\in A} I_x$ is countable and hence is not all of Ω. By Proposition 0.16, there exists $y \in \Omega$ such that $\bigcup_{x\in A} I_x = I_y$, and y is thus an upper bound for A. ∎

The set $\mathbb{N}$ of positive integers may be identified with a subset of Ω as follows. Set $f(1) = \inf \Omega$, and proceeding inductively, set $f(n) = \inf(\Omega \setminus \{f(1), \ldots, f(n-1)\})$. The reader may verify that f is an order isomorphism from $\mathbb{N}$ to I_ω, where ω is the minimal element of Ω such that I_ω is infinite.

It is sometimes convenient to add an extra element ω_1 to Ω to form a set $\Omega^* = \Omega \cup \{\omega_1\}$ and to extend the ordering on Ω to Ω^* by declaring that $x < \omega_1$ for all $x \in \Omega$. ω_1 is called the **first uncountable ordinal**. (The usual notation for ω_1 is Ω, since ω_1 is generally taken to be the set of countable ordinals itself.)

0.5 THE EXTENDED REAL NUMBER SYSTEM

It is frequently useful to adjoin two extra points ∞ ($= +\infty$) and $-\infty$ to $\mathbb{R}$ to form the **extended real number system** $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, and to extend the usual ordering on $\mathbb{R}$ by declaring that $-\infty < x < \infty$ for all $x \in \mathbb{R}$. The completeness of $\mathbb{R}$ can then be stated as follows: Every subset A of $\overline{\mathbb{R}}$ has a least upper bound, or **supremum**, and a greatest lower bound, or **infimum**, which are denoted by $\sup A$ and $\inf A$. If $A = \{a_1, \ldots a_n\}$, we also write

From completeness it follows that every sequence $\{x_n\}$ in $\overline{\mathbb{R}}$ has a **limit superior** and a **limit inferior**:

$$\limsup x_n = \inf_{k\geq 1}\Big(\sup_{n\geq k} x_n\Big), \qquad \liminf x_n = \sup_{k\geq 1}\Big(\inf_{n\geq k} x_n\Big).$$

The sequence $\{x_n\}$ converges (in $\mathbb{R}$) iff these two numbers are equal (and finite), in which case its limit is their common value. One can also define lim sup and lim inf for functions $f : \mathbb{R} \to \overline{\mathbb{R}}$, for instance:

$$\limsup_{x\to a} f(x) = \inf_{\delta>0}\Big(\sup_{0<|x-a|<\delta} f(x)\Big).$$

The arithmetical operations on $\mathbb{R}$ can be partially extended to $\overline{\mathbb{R}}$:

$$x \pm \infty = \pm\infty \ (x \in \mathbb{R}), \qquad \infty + \infty = \infty, \qquad -\infty - \infty = -\infty,$$
$$x \cdot (\pm\infty) = \pm\infty \ (x > 0), \qquad x \cdot (\pm\infty) = \mp\infty \ (x < 0).$$

We make no attempt to define $\infty - \infty$, but we abide by the convention that, unless otherwise stated,

$$0 \cdot (\pm\infty) = 0.$$

(The expression $0 \cdot \infty$ turns up now and then in measure theory, and for various reasons its proper interpretation is almost always 0.)

We employ the following notation for intervals in $\overline{\mathbb{R}}$: if $-\infty \leq a < b \leq \infty$,

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

We shall occasionally encounter uncountable sums of nonnegative numbers. If X is an arbitrary set and $f : X \to [0,\infty]$, we define $\sum_{x\in X} f(x)$ to be the supremum of its finite partial sums:

$$\sum_{x\in X} f(x) = \sup\Big\{\sum_{x\in F} f(x) : F \subset X,\ F \text{ finite}\Big\}.$$

(Later we shall recognize this as the integral of f with respect to counting measure on X.)

0.20 Proposition. *Given $f : X \to [0,\infty]$, let $A = \{x : f(x) > 0\}$. If A is uncountable, then $\sum_{x\in X} f(x) = \infty$. If A is countably infinite, then $\sum_{x\in X} f(x) = \sum_1^\infty f(g(n))$ where $g : \mathbb{N} \to A$ is any bijection and the sum on the right is an ordinary infinite series.*

Proof. We have $A = \bigcup_1^\infty A_n$ where $A_n = \{x : f(x) > 1/n\}$. If A is uncountable, then some A_n must be uncountable, and $\sum_{x\in F} f(x) > \text{card}(F)/n$ for ... If A is countably infinite,

$g : \mathbb{N} \to A$ is a bijection, and $B_N = g(\{1, \ldots, N\})$, then every finite subset F of A is contained in some B_N. Hence

$$\sum_{x \in F} f(x) \le \sum_1^N f(g(n)) \le \sum_{x \in X} f(x).$$

Taking the supremum over N, we find

$$\sum_{x \in F} f(x) \le \sum_1^\infty f(g(n)) \le \sum_{x \in X} f(x),$$

and then taking the supremum over F, we obtain the desired result. ∎

Some terminology concerning (extended) real-valued functions: A relation between numbers that is applied to functions is understood to hold pointwise. Thus $f \le g$ means that $f(x) \le g(x)$ for every x, and $\max(f, g)$ is the function whose value at x is $\max(f(x), g(x))$. If $X \subset \overline{\mathbb{R}}$ and $f : X \to \overline{\mathbb{R}}$, f is called **increasing** if $f(x) \le f(y)$ whenever $x \le y$ and **strictly increasing** if $f(x) < f(y)$ whenever $x < y$; similarly for **decreasing**. A function that is either increasing or decreasing is called **monotone**.

If $f : \mathbb{R} \to \mathbb{R}$ is an increasing function, then f has right- and left-hand limits at each point:

$$f(a+) = \lim_{x \searrow a} f(x) = \inf_{x > a} f(x), \qquad f(a-) = \lim_{x \nearrow a} f(x) = \sup_{x < a} f(x).$$

Moreover, the limiting values $f(\infty) = \sup_{a \in \mathbb{R}} f(x)$ and $f(-\infty) = \inf_{a \in \mathbb{R}} f(x)$ exist (possibly equal to $\pm\infty$). f is called **right continuous** if $f(a) = f(a+)$ for all $a \in \mathbb{R}$ and **left continuous** if $f(a) = f(a-)$ for all $a \in \mathbb{R}$.

For points x in $\mathbb{R}$ or $\mathbb{C}$, $|x|$ denotes the ordinary absolute value or modulus of x, $|a + ib| = \sqrt{a^2 + b^2}$. For points x in $\mathbb{R}^n$ or $\mathbb{C}^n$, $|x|$ denotes the Euclidean norm:

$$|x| = \left[\sum_1^n |x_j|^2\right]^{1/2}.$$

We recall that a set $U \subset \mathbb{R}$ is **open** if, for every $x \in U$, U includes an interval centered at x.

0.21 Proposition. *Every open set in $\mathbb{R}$ is a countable disjoint union of open intervals.*

Proof. If U is open, for each $x \in U$ consider the collection $\mathcal{I}_x$ of all open intervals I such that $x \in I \subset U$. It is easy to check that the union of any family of open intervals containing a point in common is again an open interval, and hence $J_x = \bigcup_{I \in \mathcal{I}_x} I$ is an open interval; it is the largest element of $\mathcal{I}_x$. If $x, y \in U$ then either $J_x = J_y$ or $J_x \cap J_y = \varnothing$, for otherwise $J_x \cup J_y$ would be a larger open interval than J_x in $\mathcal{I}_x$. Thus if $\mathcal{J} = \{J_x : x \in U\}$, the (distinct) members of $\mathcal{J}$ are disjoint, and $U = \bigcup_{J \in \mathcal{J}} J$. For each $J \in \mathcal{J}$, pick a rational number $f(J) \in J$. The map $f : \mathcal{J} \to \mathbb{Q}$ thus defined is injective, for if $J \neq J'$ then $J \cap J' = \varnothing$; therefore $\mathcal{J}$ is countable.

0.6 METRIC SPACES

A **metric** on a set X is a function $\rho : X \times X \to [0, \infty)$ such that

- $\rho(x, y) = 0$ iff $x = y$;
- $\rho(x, y) = \rho(y, x)$ for all $x, y \in X$;
- $\rho(x, z) \le \rho(x, y) + \rho(y, z)$ for all $x, y, z \in X$.

(Intuitively, $\rho(x, y)$ is to be interpreted as the distance from x to y.) A set equipped with a metric is called a **metric space**. Some examples:

i. The Euclidean distance $\rho(x, y) = |x - y|$ is a metric on $\mathbb{R}^n$.

ii. $\rho_1(f, g) = \int_0^1 |f(x) - g(x)|\,dx$ and $\rho_\infty(f, g) = \sup_{0 \le x \le 1} |f(x) - g(x)|$ are metrics on the space of continuous functions on $[0, 1]$.

iii. If ρ is a metric on X and $A \subset X$, then $\rho|(A \times A)$ is a metric on A.

iv. If (X_1, ρ_1) and (X_2, ρ_2) are metric spaces, the **product metric** ρ on $X_1 \times X_2$ is given by

$$\rho((x_1, x_2), (y_1, y_2)) = \max\left(\rho_1(x_1, y_1), \rho_2(x_2, y_2)\right).$$

Other metrics are sometimes used on $X_1 \times X_2$, for instance,

$$\rho_1(x_1, y_1) + \rho_2(x_2, y_2) \quad \text{or} \quad \left[\rho_1(x_1, y_1)^2 + \rho_2(x_2, y_2)^2\right]^{1/2}.$$

These, however, are equivalent to the product metric in the sense that we shall define at the end of this section.

Let (X, ρ) be a metric space. If $x \in X$ and $r > 0$, the (open) **ball** of radius r about x is

$$B(r, x) = \{y \in X : \rho(x, y) < r\}.$$

A set $E \subset X$ is **open** if for every $x \in E$ there exists $r > 0$ such that $B(r, x) \subset E$, and **closed** if its complement is open. For example, every ball $B(r, x)$ is open, for if $y \in B(r, x)$ and $\rho(x, y) = s$ then $B(r - s, y) \subset B(r, x)$. Also, X and $\varnothing$ are both open and closed. Clearly the union of any family of open sets is open, and hence the intersection of any family of closed sets is closed. Also, the intersection (resp. union) of any finite family of open (resp. closed) sets is open (resp. closed). Indeed, if $U_1, \ldots U_n$ are open and $x \in \bigcap_1^n U_j$, for each j there exists $r_j > 0$ such that $B(r_j, x) \subset U_j$, and then $B(r, x) \subset \bigcap_1^n U_j$ where $r = \min(r_1, \ldots, r_n)$, so $\bigcap_1^n U_j$ is open.

If $E \subset X$, the union of all open sets $U \subset E$ is the largest open set contained in E; it is called the **interior** of E and is denoted by E^o. Likewise, the intersection of all closed sets $F \supset E$ is the smallest closed set containing E; it is called the **closure** of

$\overline{E}$ has empty interior. X is called **separable** if it has a countable dense subset. (For example, $\mathbb{Q}^n$ is a countable dense subset of $\mathbb{R}^n$.) A sequence $\{x_n\}$ in X **converges** to $x \in X$ (symbolically: $x_n \to x$ or $\lim x_n = x$) if $\lim_{n\to\infty} \rho(x_n, x) = 0$.

0.22 Proposition. *If X is a metric space, $E \subset X$, and $x \in X$, the following are equivalent:*

a. $x \in \overline{E}$.

b. $B(r,x) \cap E \neq \varnothing$ *for all* $r > 0$.

c. There is a sequence $\{x_n\}$ in E that converges to x.

Proof. If $B(r,x) \cap E = \varnothing$, then $B(r,x)^c$ is a closed set containing E but not x, so $x \notin \overline{E}$. Conversely, if $x \notin \overline{E}$, since $(\overline{E})^c$ is open there exists $r > 0$ such that $B(r,x) \subset (\overline{E})^c \subset E^c$. Thus (a) is equivalent to (b). If (b) holds, for each $n \in \mathbb{N}$ there exists $x_n \in B(n^{-1}, x) \cap E$, so that $x_n \to x$. On the other hand, if $B(r,x) \cap E = \varnothing$, then $\rho(y,x) \geq r$ for all $y \in E$, so no sequence of E can converge to x. Thus (b) is equivalent to (c). ∎

If (X_1, ρ_1) and (X_2, ρ_2) are metric spaces, a map $f : X_1 \to X_2$ is called **continuous** at $x \in X$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\rho_2(f(y), f(x)) < \epsilon$ whenever $\rho_1(x,y) < \delta$ — in other words, such that $f^{-1}(B(\epsilon, f(x))) \supset B(\delta, x)$. The map f is called **continuous** if it is continuous at each $x \in X_1$ and **uniformly continuous** if, in addition, the δ in the definition of continuity can be chosen independent of x.

0.23 Proposition. *$f : X_1 \to X_2$ is continuous iff $f^{-1}(U)$ is open in X_1 for every open $U \subset X_2$.*

Proof. If the latter condition holds, then for every $x \in X_1$ and $\epsilon > 0$, the set $f^{-1}(B(\epsilon, f(x)))$ is open and contains x, so it contains some ball about x; this means that f is continuous at x. Conversely, suppose that f is continuous and U is open in X_2. For each $y \in U$ there exists $\epsilon_y > 0$ such that $B(\epsilon_y, y) \subset U$, and for each $x \in f^{-1}(\{y\})$ there exists $\delta_x > 0$ such that $B(\delta_x, x) \subset f^{-1}(B(\epsilon_y, y)) \subset f^{-1}(U)$. Thus $f^{-1}(U) = \bigcup_{x \in f^{-1}(U)} B(\delta_x, x)$ is open. ∎

A sequence $\{x_n\}$ in a metric space (X, ρ) is called **Cauchy** if $\rho(x_n, x_m) \to 0$ as $n, m \to \infty$. A subset E of X is called **complete** if every Cauchy sequence in E converges and its limit is in E. For example, $\mathbb{R}^n$ (with the Euclidean metric) is complete, whereas $\mathbb{Q}^n$ is not.

0.24 Proposition. *A closed subset of a complete metric space is complete, and a complete subset of an arbitrary metric space is closed.*

Proof. If X is complete, $E \subset X$ is closed, and $\{x_n\}$ is a Cauchy sequence in E, $\{x_n\}$ has a limit in X. By Proposition 0.22, $x \in \overline{E} = E$. If $E \subset X$ is complete and $x \in \overline{E}$, by Proposition (0.22) there is a sequence $\{x_n\}$ in E converging to x. $\{x_n\}$

In a metric space (X, ρ) we can define the distance from a point to a set and the distance between two sets. Namely, if $x \in X$ and $E, F \subset X$,

$$\rho(x, E) = \inf\{\rho(x,y) : y \in E\},$$

$$\rho(E, F) = \inf\{\rho(x,y) : x \in E,\ y \in F\} = \inf\{\rho(x, F) : x \in E\}.$$

Observe that, by Proposition 0.22, $\rho(x, E) = 0$ iff $x \in \overline{E}$. We also define the **diameter** of $E \subset X$ to be

$$\operatorname{diam} E = \sup\{\rho(x,y) : x, y \in E\}.$$

E is called **bounded** if $\operatorname{diam} E < \infty$.

If $E \subset X$ and $\{V_\alpha\}_{\alpha \in A}$ is a family of sets such that $E \subset \bigcup_{\alpha \in A} V_\alpha$, $\{V_\alpha\}_{\alpha \in A}$ is called a **cover** of E, and E is said to be **covered** by the V_α's. E is called **totally bounded** if, for every $\epsilon > 0$, E can be covered by finitely many balls of radius ϵ. Every totally bounded set is bounded, for if $x, y \in \bigcup_1^n B(\epsilon, z_j)$, say $x \in B(\epsilon, z_1)$ and $y \in B(\epsilon, z_2)$, then

$$\rho(x,y) \leq \rho(x, z_1) + \rho(z_1, z_2) + \rho(z_2, y) \leq 2\epsilon + \max\{\rho(z_j, z_k) : 1 \leq j, k \leq n\}.$$

(The converse is false in general.) If E is totally bounded, so is $\overline{E}$, for it is easily seen that if $E \subset \bigcup_1^n B(\epsilon, z_j)$, then $\overline{E} \subset \bigcup_1^n B(2\epsilon, z_j)$.

0.25 Theorem. *If E is a subset of the metric space (X, ρ), the following are equivalent:*

a. E is complete and totally bounded.

b. **(The Bolzano-Weierstrass Property)** *Every sequence in E has a subsequence that converges to a point of E.*

c. **(The Heine-Borel Property)** *If $\{V_\alpha\}_{\alpha \in A}$ is a cover of E by open sets, there is a finite set $F \subset A$ such that $\{V_\alpha\}_{\alpha \in F}$ covers E.*

Proof. We shall show that (a) and (b) are equivalent, that (a) and (b) together imply (c), and finally that (c) implies (b).

(a) implies (b): Suppose that (a) holds and $\{x_n\}$ is a sequence in E. E can be covered by finitely many balls of radius 2^{-1}, and at least one of them must contain x_n for infinitely many n: say, $x_n \in B_1$ for $n \in N_1$. $E \cap B_1$ can be covered by finitely many balls of radius 2^{-2}, and at least one of them must contain x_n for infinitely many $n \in N_1$: say, $x_n \in B_2$ for $n \in N_2$. Continuing inductively, we obtain a sequence of balls B_j of radius 2^{-j} and a decreasing sequence of subsets N_j of $\mathbb{N}$ such that $x_n \in B_j$ for $n \in N_j$. Pick $n_1 \in N_1$, $n_2 \in N_2, \ldots$ such that $n_1 < n_2 < \cdots$. Then $\{x_{n_j}\}$ is a Cauchy sequence, for $\rho(x_{n_j}, x_{n_k}) < 2^{1-j}$ if $k > j$, and since E is complete, it has a limit in E.

(b) implies (a): We show that if either condition in (a) fails, then so does (b). If E is not complete, there is a Cauchy sequence $\{x_n\}$ in E with no limit in E. No subsequence of $\{x_n\}$ can converge in E, for otherwise the whole squence would be such that E cannot be covered by finitely many balls of radius ϵ. Choose $x_n \in E$ inductively as follows. Begin with any $x_1 \in E$, and having chosen $x_1, \ldots, x_n$, pick $x_{n+1} \in E \setminus \bigcup_1^n B(\epsilon, x_j)$. Then $\rho(x_n, x_m) > \epsilon$ for all n, m, so $\{x_n\}$ has no convergent subsequence.

(a) and (b) imply (c): It suffices to show that if (b) holds and $\{V_\alpha\}_{\alpha \in A}$ is a cover of E by open sets, there exists $\epsilon > 0$ such that every ball of radius ϵ that intersects E is contained in some V_α, for E can be covered by finitely many such balls by (a). Suppose to the contrary that for each $n \in \mathbb{N}$ there is a ball B_n of radius 2^{-n} such that $B_n \cap E \neq \varnothing$ and B_n is contained in no V_α. Pick $x_n \in B_n \cap E$; by passing to a subsequence we may assume that $\{x_n\}$ converges to some $x \in E$. We have $x \in V_\alpha$ for some α, and since V_α is open, there exists $\epsilon > 0$ such that $B(\epsilon, x) \subset V_\alpha$. But if n is large enough so that $\rho(x_n, x) < \epsilon/3$ and $2^{-n} < \epsilon/3$, then $B_n \subset B(\epsilon, x) \subset V_\alpha$, contradicting the assumption on B_n.

(c) implies (b): If $\{x_n\}$ is a sequence in E with no convergent subsequence, for each $x \in E$ there is a ball B_x centered at x that contains x_n for only finitely many n (otherwise some subsequence would converge to x). Then $\{B_x\}_{x \in E}$ is a cover of E by open sets with no finite subcover. ∎

A set E that possesses the properties (a)–(c) of Theorem 0.25 is called **compact**. Every compact set is closed (by Proposition 0.24) and bounded; the converse is false in general but true in $\mathbb{R}^n$.

0.26 Proposition. *Every closed and bounded subset of $\mathbb{R}^n$ is compact.*

Proof. Since closed subsets of $\mathbb{R}^n$ are complete, it suffices to show that bounded subsets of $\mathbb{R}^n$ are totally bounded. Since every bounded set is contained in some cube

$$Q = [-R, R]^n = \{x \in \mathbb{R}^n : \max(|x_1|, \ldots, |x_n|) \leq R\},$$

it is enough to show that Q is totally bounded. Given $\epsilon > 0$, pick an integer $k > R\sqrt{n}/\epsilon$, and express Q as the union of k^n congruent subcubes by dividing the interval $[-R, R]$ into k equal pieces. The side length of these subcubes is $2R/k$ and hence their diameter is $\sqrt{n}(2R/k) < 2\epsilon$, so they are contained in the balls of radius ϵ about their centers. ∎

Two metrics ρ_1 and ρ_2 on a set X are called **equivalent** if

$$C\rho_1 \leq \rho_2 \leq C'\rho_1 \text{ for some } C, C' > 0.$$

It is easily verified that equivalent metrics define the same open, closed, and compact sets, the same convergent and Cauchy sequences, and the same continuous and uniformly continuous mappings. Consequently, most results concerning metric spaces depend not on the particular metric chosen but only on its equivalence class.

0.7 NOTES AND REFERENCES

§§0.1–0.4: The best exposition of set theory for beginners is Halmos [62], and

also contains a concise account of of basic axiomatic set theory. All of these books present a deduction of the Hausdorff maximal principle from the axiom of choice, as does Hewitt and Stromberg [76].

The axiom of choice (or one of the propositions equivalent to it) is generally taken as one of the basic postulates in the axiomatic formulations of set theory. Some mathematicians of the intuitionist or constructivist persuasion reject it on the grounds that one has not proved the existence of a mathematical object until one has shown how to construct it in some reasonably explicit fashion, whereas the whole point of the axiom of choice is to provide existence theorems when constructive methods fail (or are too cumbersome for comfort). People who are seriously bothered by such objections belong to a minority that does not include the present writer; in this book the axiom of choice is used sparingly but freely.

The **continuum hypothesis** is the assertion that if $\text{card}(X) < \mathfrak{c}$, then X is countable. (Since it follows easily from the construction of Ω, the set of countable ordinals, that $\text{card}(\Omega) \leq \text{card}(X)$ for any uncountable X, an equivalent assertion is that $\text{card}(\Omega) = \mathfrak{c}$.) It is known, thanks to Gödel and Cohen, that the continuum hypothesis and its negation are both consistent with the standard axioms of set theory including the axiom of choice, assuming that those axioms are themselves consistent. (An exposition of the consistency and independence theorems for the axiom of choice and the continuum hypothesis can be found in Smullyan and Fitting [135].) Some mathematicians are willing to accept the continuum hypothesis as true, seemingly as a matter of convenience, but Gödel [56] and Cohen [26, p. 151] have both expressed suspicions that it should be false, and as of this writing no one has found any really compelling evidence on one side or the other. My own feeling, subject to revision in the event of a major breakthrough in set theory, is that if the answer to one's question turns out to depend on the continuum hypothesis, one should give up and ask a different question.

§0.6: A more detailed discussion of metric spaces can be found in Loomis and Sternberg [95] and DePree and Swartz [32].

1
Measures

In this chapter we set forth the basic concepts of measure theory, develop a general procedure for constructing nontrivial examples of measures, and apply this procedure to construct measures on the real line.

1.1 INTRODUCTION

One of the most venerable problems in geometry is to determine the area or volume of a region in the plane or in 3-space. The techniques of integral calculus provide a satisfactory solution to this problem for regions that are bounded by "nice" curves or surfaces but are inadequate to handle more complicated sets, even in dimension one. Ideally, for $n \in \mathbb{N}$ we would like to have a function μ that assigns to each $E \subset \mathbb{R}^n$ a number $\mu(E) \in [0, \infty]$, the n-dimensional measure of E, such that $\mu(E)$ is given by the usual integral formulas when the latter apply. Such a function μ should surely possess the following properties:

i. If $E_1, E_2, \ldots$ is a finite or infinite sequence of disjoint sets, then

$$\mu(E_1 \cup E_2 \cup \cdots) = \mu(E_1) + \mu(E_2) + \cdots.$$

ii. If E is congruent to F (that is, if E can be transformed into F by translations, rotations, and reflections), then $\mu(E) = \mu(F)$.

iii. $\mu(Q) = 1$, where Q is the unit cube

Unfortunately, these conditions are mutually inconsistent. Let us see why this is true for $n = 1$. (The argument can easily be adapted to higher dimensions.) To begin with, we define an equivalence relation on $[0, 1)$ by declaring that $x \sim y$ iff $x - y$ is rational. Let N be a subset of $[0, 1)$ that contains precisely one member of each equivalence class. (To find such an N, one must invoke the axiom of choice.) Next, let $R = \mathbb{Q} \cap [0, 1)$, and for each $r \in R$ let

$$N_r = \{x + r : x \in N \cap [0, 1 - r)\} \cup \{x + r - 1 : x \in N \cap [1 - r, 1)\}.$$

That is, to obtain N_r, shift N to the right by r units and then shift the part that sticks out beyond $[0, 1)$ one unit to the left. Then $N_r \subset [0, 1)$, and every $x \in [0, 1)$ belongs to precisely one N_r. Indeed, if y is the element of N that belongs to the equivalence class of x, then $x \in N_r$ where $r = x - y$ if $x \geq y$ or $r = x - y + 1$ if $x < y$; on the other hand, if $x \in N_r \cap N_s$, then $x - r$ (or $x - r + 1$) and $x - s$ (or $x - s + 1$) would be distinct elements of N belonging to the same equivalence class, which is impossible.

Suppose now that $\mu : \mathcal{P}(\mathbb{R}) \to [0, \infty]$ satisfies (i), (ii), and (iii). By (i) and (ii),

$$\mu(N) = \mu(N \cap [0, 1 - r)) + \mu(N \cap [1 - r, 1)) = \mu(N_r)$$

for any $r \in R$. Also, since R is countable and $[0, 1)$ is the disjoint union of the N_r's,

$$\mu([0, 1)) = \sum_{r \in R} \mu(N_r)$$

by (i) again. But $\mu([0, 1)) = 1$ by (iii), and since $\mu(N_r) = \mu(N)$, the sum on the right is either 0 (if $\mu(N) = 0$) or ∞ (if $\mu(N) > 0$). Hence no such μ can exist.

Faced with this discouraging situation, one might consider weakening (i) so that additivity is required to hold only for finite sequences. This is not a very good idea, as we shall see: The additivity for countable sequences is what makes all the limit and continuity results of the theory work smoothly. Moreover, in dimensions $n \geq 3$, even this weak form of (i) is inconsistent with (ii) and (iii). Indeed, in 1924 Banach and Tarski proved the following amazing result:

> Let U and V be arbitrary bounded open sets in $\mathbb{R}^n$, $n \geq 3$. There exist $k \in \mathbb{N}$ and subsets $E_1, \ldots, E_k, F_1, \ldots, F_k$ of $\mathbb{R}^n$ such that
>
> - the E_j's are disjoint and their union is U;
> - the F_j's are disjoint and their union is V;
> - E_j is congruent to F_j for $j = 1, \ldots, k$.

Thus one can cut up a ball the size of a pea into a finite number of pieces and rearrange them to form a ball the size of the earth! Needless to say, the sets E_j and F_j are *very* bizarre. They cannot be visualized accurately, and their construction depends on the axiom of choice. But their existence clearly precludes the construction of any $\mu : \mathcal{P}(\mathbb{R}^n) \to [0, \infty]$ that assigns positive, finite values to bounded open sets and

The moral of these examples is that $\mathbb{R}^n$ contains subsets which are so strangely put together that it is impossible to define a geometrically reasonable notion of measure for them, and the remedy for the situation is to discard the requirement that μ should be defined on *all* subsets of $\mathbb{R}^n$. Rather, we shall content ourselves with constructing μ on a class of subsets of $\mathbb{R}^n$ that includes all the sets one is likely to meet in practice unless one is deliberately searching for pathological examples. This construction will be carried out for $n = 1$ in §1.5 and for $n > 1$ in §2.6.

It is worthwhile, and not much extra work, to develop the theory in much greater generality. The conditions (ii) and (iii) are directly related to Euclidean geometry, but set functions satisfying (i), called *measures*, arise also in a great many other situations. For example, in a physics problem involving mass distributions, $\mu(E)$ could represent the total mass in the region E. For another example, in probability theory one considers a set X that represents the possible outcomes of an experiment, and for $E \subset X$, $\mu(E)$ is the probability that the outcome lies in E. We therefore begin by studying the theory of measures on abstract sets.

1.2 σ-ALGEBRAS

In this section we discuss the families of sets that serve as the domains of measures.

Let X be a nonempty set. An **algebra** of sets on X is a nonempty collection $\mathcal{A}$ of subsets of X that is closed under finite unions and complements; in other words, if $E_1, \ldots, E_n \in \mathcal{A}$, then $\bigcup_1^n E_j \in \mathcal{A}$; and if $E \in \mathcal{A}$, then $E^c \in \mathcal{A}$. A **σ-algebra** is an algebra that is closed under countable unions. (Some authors use the terms **field** and **σ-field** instead of algebra and σ-algebra.)

We observe that since $\bigcap_j E_j = (\bigcup_j E_j^c)^c$, algebras (resp. σ-algebras) are also closed under finite (resp. countable) intersections. Moreover, if $\mathcal{A}$ is an algebra, then $\varnothing \in \mathcal{A}$ and $X \in \mathcal{A}$, for if $E \in \mathcal{A}$ we have $\varnothing = E \cap E^c$ and $X = E \cup E^c$.

It is worth noting that an algebra $\mathcal{A}$ is a σ-algebra provided that it is closed under countable *disjoint* unions. Indeed, suppose $\{E_j\}_1^\infty \subset \mathcal{A}$. Set

$$F_k = E_k \setminus \left[\bigcup_1^{k-1} E_j\right] = E_k \cap \left[\bigcup_1^{k-1} E_j\right]^c.$$

Then the F_k's belong to $\mathcal{A}$ and are disjoint, and $\bigcup_1^\infty E_j = \bigcup_1^\infty F_k$. This device of replacing a sequence of sets by a disjoint sequence is worth remembering; it will be used a number of times below.

Some examples: If X is any set, $\mathcal{P}(X)$ and $\{\varnothing, X\}$ are σ-algebras. If X is uncountable, then

$$\mathcal{A} = \{E \subset X : E \text{ is countable or } E^c \text{ is countable}\}$$

is a σ-algebra, called the **σ-algebra of countable or co-countable sets**. (The point here is that if $\{E_j\}_1^\infty \subset \mathcal{A}$, then $\bigcup_1^\infty E_j$ is countable if all E_j are countable and is

It is trivial to verify that the intersection of any family of σ-algebras on X is again a σ-algebra. It follows that if $\mathcal{E}$ is any susbset of $\mathcal{P}(X)$, there is a unique smallest σ-algebra $\mathcal{M}(\mathcal{E})$ containing $\mathcal{E}$, namely, the intersection of all σ-algebras containing $\mathcal{E}$. (There is always at least one such, namely, $\mathcal{P}(X)$.) $\mathcal{M}(\mathcal{E})$ is called the σ-algebra **generated** by $\mathcal{E}$. The following observation is often useful:

1.1 Lemma. *If $\mathcal{E} \subset \mathcal{M}(\mathcal{F})$ then $\mathcal{M}(\mathcal{E}) \subset \mathcal{M}(\mathcal{F})$.*

Proof. $\mathcal{M}(\mathcal{F})$ is a σ-algebra containing $\mathcal{E}$; it therefore contains $\mathcal{M}(\mathcal{E})$. ∎

If X is any metric space, or more generally any topological space (see Chapter 4), the σ-algebra generated by the family of open sets in X (or, equivalently, by the family of closed sets in X) is called the **Borel σ-algebra** on X and is denoted by $\mathcal{B}_X$. Its members are called **Borel sets**. $\mathcal{B}_X$ thus includes open sets, closed sets, countable intersections of open sets, countable unions of closed sets, and so forth.

There is a standard terminology for the levels in this hierarchy. A countable intersection of open sets is called a G_δ set; a countable union of closed sets is called an F_σ set; a countable union of G_δ sets is called a $G_{\delta\sigma}$ set; a countable intersection of F_σ sets is called an $F_{\sigma\delta}$ set; and so forth. (δ and σ stand for the German *Durchschnitt* and *Summe*, that is, intersection and union.)

The Borel σ-algebra on $\mathbb{R}$ will play a fundamental role in what follows. For future reference we note that it can be generated in a number of different ways:

1.2 Proposition. *$\mathcal{B}_{\mathbb{R}}$ is generated by each of the following:*

a. *the open intervals: $\mathcal{E}_1 = \{(a,b) : a < b\}$,*
b. *the closed intervals: $\mathcal{E}_2 = \{[a,b] : a < b\}$,*
c. *the half-open intervals: $\mathcal{E}_3 = \{(a,b] : a < b\}$ or $\mathcal{E}_4 = \{[a,b) : a < b\}$,*
d. *the open rays: $\mathcal{E}_5 = \{(a,\infty) : a \in \mathbb{R}\}$ or $\mathcal{E}_6 = \{(-\infty,a) : a \in \mathbb{R}\}$,*
e. *the closed rays: $\mathcal{E}_7 = \{[a,\infty) : a \in \mathbb{R}\}$ or $\mathcal{E}_8 = \{(-\infty,a] : a \in \mathbb{R}\}$.*

Proof. The elements of $\mathcal{E}_j$ for $j \neq 3,4$ are open or closed, and the elements of $\mathcal{E}_3$ and $\mathcal{E}_4$ are G_δ sets — for example, $(a,b] = \bigcap_1^\infty (a, b+n^{-1})$. All of these are Borel sets, so by Lemma 1.1, $\mathcal{M}(\mathcal{E}_j) \subset \mathcal{B}_{\mathbb{R}}$ for all j. On the other hand, every open set in $\mathbb{R}$ is a countable union of open intervals, so by Lemma 1.1 again, $\mathcal{B}_{\mathbb{R}} \subset \mathcal{M}(\mathcal{E}_1)$. That $\mathcal{B}_{\mathbb{R}} \subset \mathcal{M}(\mathcal{E}_j)$ for $j \geq 2$ can now be established by showing that all open intervals lie in $\mathcal{M}(\mathcal{E}_j)$ and applying Lemma 1.1. For example, $(a,b) = \bigcup_1^\infty [a+n^{-1}, b-n^{-1}] \in \mathcal{M}(\mathcal{E}_2)$. Verification of the other cases is left to the reader (Exercise 2). ∎

Let $\{X_\alpha\}_{\alpha\in A}$ be an indexed collection of nonempty sets, $X = \prod_{\alpha\in A} X_\alpha$, and $\pi_\alpha : X \to X_\alpha$ the coordinate maps. If $\mathcal{M}_\alpha$ is a σ-algebra on X_α for each α, the **product σ-algebra** on X is the σ-algebra generated by

$$\{\pi_\alpha^{-1}(E_\alpha) : E_\alpha \in \mathcal{M}_\alpha,\ \alpha \in A\}.$$

We denote this σ-algebra by $\bigotimes_{\alpha\in A} \mathcal{M}_\alpha$. (If $A = \{1,\ldots,n\}$ we also write $\bigotimes_1^n \mathcal{M}_j$ for the moment we give an alternative, and perhaps more intuitive, characterization of product σ-algebras in the case of countably many factors.

1.3 Proposition. *If A is countable, then $\bigotimes_{\alpha\in A} \mathcal{M}_\alpha$ is the σ-algebra generated by $\{\prod_{\alpha\in A} E_\alpha : E_\alpha \in \mathcal{M}_\alpha\}$.*

Proof. If $E_\alpha \in \mathcal{M}_\alpha$, then $\pi_\alpha^{-1}(E_\alpha) = \prod_{\beta\in A} E_\beta$ where $E_\beta = X_\beta$ for $\beta \neq \alpha$; on the other hand, $\prod_{\alpha\in A} E_\alpha = \bigcap_{\alpha\in A} \pi_\alpha^{-1}(E_\alpha)$. The result therefore follows from Lemma 1.1. ∎

1.4 Proposition. *Suppose that $\mathcal{M}_\alpha$ is generated by $\mathcal{E}_\alpha$, $\alpha \in A$. Then $\bigotimes_{\alpha\in A} \mathcal{M}_\alpha$ is generated by $\mathcal{F}_1 = \{\pi_\alpha^{-1}(E_\alpha) : E_\alpha \in \mathcal{E}_\alpha,\ \alpha \in A\}$. If A is countable and $X_\alpha \in \mathcal{E}_\alpha$ for all α, $\bigotimes_{\alpha\in A} \mathcal{M}_\alpha$ is generated by $\mathcal{F}_2 = \{\prod_{\alpha\in A} E_\alpha : E_\alpha \in \mathcal{E}_\alpha\}$.*

Proof. Obviously $\mathcal{M}(\mathcal{F}_1) \subset \bigotimes_{\alpha\in A} \mathcal{M}_\alpha$. On the other hand, for each α, the collection $\{E \subset X_\alpha : \pi_\alpha^{-1}(E) \in \mathcal{M}(\mathcal{F}_1)\}$ is easily seen to be a σ-algebra on X_α that contains $\mathcal{E}_\alpha$ and hence $\mathcal{M}_\alpha$. In other words, $\pi_\alpha^{-1}(E) \in \mathcal{M}(\mathcal{F}_1)$ for all $E \in \mathcal{M}_\alpha$, $\alpha \in A$, and hence $\bigotimes_{\alpha\in A} \mathcal{M}_\alpha \subset \mathcal{M}(\mathcal{F}_1)$. The second assertion follows from the first as in the proof of Proposition 1.3. ∎

1.5 Proposition. *Let $X_1, \ldots, X_n$ be metric spaces and let $X = \prod_1^n X_j$, equipped with the product metric. Then $\bigotimes_1^n \mathcal{B}_{X_j} \subset \mathcal{B}_X$. If the X_j's are separable, then $\bigotimes_1^n \mathcal{B}_{X_j} = \mathcal{B}_X$.*

Proof. By Proposition 1.4, $\bigotimes_1^n \mathcal{B}_{X_j}$ is generated by the sets $\pi_j^{-1}(U_j)$, $1 \leq j \leq n$, where U_j is open in X_j. Since these sets are open in X, Lemma 1.1 implies that $\bigotimes_1^n \mathcal{B}_{X_j} \subset \mathcal{B}_X$. Suppose now that C_j is a countable dense set in X_j, and let $\mathcal{E}_j$ be the collection of balls in X_j with rational radius and center in C_j. Then every open set in X_j is a union of members of $\mathcal{E}_j$ — in fact, a countable union since $\mathcal{E}_j$ itself is countable. Moreover, the set of points in X whose jth coordinate is in C_j for all j is a countable dense subset of X, and the balls of radius r in X are merely products of balls of radius r in the X_j's. It follows that $\mathcal{B}_{X_j}$ is generated by $\mathcal{E}_j$ and $\mathcal{B}_X$ is generated by $\{\prod_1^n E_j : E_j \in \mathcal{E}_j\}$. Therefore $\mathcal{B}_X = \bigotimes_1^n \mathcal{B}_{X_j}$ by Proposition 1.4. ∎

1.6 Corollary. $\mathcal{B}_{\mathbb{R}^n} = \bigotimes_1^n \mathcal{B}_{\mathbb{R}}$.

We conclude this section with a technical result that will be needed later. We define an **elementary family** to be a collection $\mathcal{E}$ of subsets of X such that

- $\varnothing \in \mathcal{E}$,
- if $E, F \in \mathcal{E}$ then $E \cap F \in \mathcal{E}$,
- if $E \in \mathcal{E}$ then E^c is a finite disjoint union of members of $\mathcal{E}$.

1.7 Proposition. *If $\mathcal{E}$ is an elementary family, the collection $\mathcal{A}$ of finite disjoint*

Proof. If $A, B \in \mathcal{E}$ and $B^c = \bigcup_1^J C_j$ ($C_j \in \mathcal{E}$, disjoint), then $A \setminus B = \bigcup_1^J (A \cap C_j)$ and $A \cup B = (A \setminus B) \cup B$, where these unions are disjoint, so $A \setminus B \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$. It now follows by induction that if $A_1 \ldots, A_n \in \mathcal{E}$, then $\bigcup_1^n A_j \in \mathcal{A}$; indeed, by inductive hypothesis we may assume that $A_1, \ldots, A_{n-1}$ are disjoint, and then $\bigcup_1^n A_j = A_n \cup \bigcup_1^{n-1}(A_j \setminus A_n)$, which is a disjoint union. To see that $\mathcal{A}$ is closed under complements, suppose $A_1, \ldots A_n \in \mathcal{E}$ and $A_m^c = \bigcup_{j=1}^{J_m} B_m^j$ with $B_m^1, \ldots, B_m^{J_m}$ disjoint members of $\mathcal{E}$. Then

$$\Big(\bigcup_{m=1}^n A_m\Big)^c = \bigcap_{m=1}^n \Big(\bigcup_{j=1}^{J_m} B_m^j\Big) = \bigcup\{B_1^{j_1} \cap \cdots \cap B_n^{j_n} : 1 \le j_m \le J_m,\ 1 \le m \le n\},$$

which is in $\mathcal{A}$. ∎

Exercises

1. A family of sets $\mathcal{R} \subset \mathcal{P}(X)$ is called a **ring** if it is closed under finite unions and differences (i.e., if $E_1, \ldots, E_n \in \mathcal{R}$, then $\bigcup_1^n E_j \in \mathcal{R}$, and if $E, F \in \mathcal{R}$, then $E \setminus F \in \mathcal{R}$). A ring that is closed under countable unions is called a **σ-ring**.

- **a.** Rings (resp. σ-rings) are closed under finite (resp. countable) intersections.
- **b.** If $\mathcal{R}$ is a ring (resp. σ-ring), then $\mathcal{R}$ is an algebra (resp. σ-algebra) iff $X \in \mathcal{R}$.
- **c.** If $\mathcal{R}$ is a σ-ring, then $\{E \subset X : E \in \mathcal{R} \text{ or } E^c \in \mathcal{R}\}$ is a σ-algebra.
- **d.** If $\mathcal{R}$ is a σ-ring, then $\{E \subset X : E \cap F \in \mathcal{R} \text{ for all } F \in \mathcal{R}\}$ is a σ-algebra.

2. Complete the proof of Proposition 1.2.

3. Let $\mathcal{M}$ be an infinite σ-algebra.

- **a.** $\mathcal{M}$ contains an infinite sequence of disjoint sets.
- **b.** card$(\mathcal{M}) \ge \mathfrak{c}$.

4. An algebra $\mathcal{A}$ is a σ-algebra iff $\mathcal{A}$ is closed under countable increasing unions (i.e., if $\{E_j\}_1^\infty \subset \mathcal{A}$ and $E_1 \subset E_2 \subset \cdots$, then $\bigcup_1^\infty E_j \in \mathcal{A}$).

5. If $\mathcal{M}$ is the σ-algebra generated by $\mathcal{E}$, then $\mathcal{M}$ is the union of the σ-algebras generated by $\mathcal{F}$ as $\mathcal{F}$ ranges over all countable subsets of $\mathcal{E}$. (Hint: Show that the latter object is a σ-algebra.)

1.3 MEASURES

Let X be a set equipped with a σ-algebra $\mathcal{M}$. A **measure** on $\mathcal{M}$ (or on $(X, \mathcal{M})$, or simply on X if $\mathcal{M}$ is understood) is a function $\mu : \mathcal{M} \to [0, \infty]$ such that

i. $\mu(\varnothing) = 0$,

ii. if $\{E_j\}_1^\infty$ is a sequence of disjoint sets in $\mathcal{M}$, then $\mu(\bigcup_1^\infty E_j) = \sum_1^\infty \mu(E_j)$.

ii′. if $E_1, \ldots E_n$ are disjoint sets in $\mathcal{M}$, then $\mu(\bigcup_1^n E_j) = \sum_1^n \mu(E_j)$,

because one can take $E_j = \varnothing$ for $j > n$. A function μ that satisfies (i) and (ii′) but not necessarily (ii) is called a **finitely additive measure**.

If X is a set and $\mathcal{M} \subset \mathcal{P}(X)$ is a σ-algebra, $(X, \mathcal{M})$ is called a **measurable space** and the sets in $\mathcal{M}$ are called **measurable sets**. If μ is a measure on $(X, \mathcal{M})$, then $(X, \mathcal{M}, \mu)$ is called a **measure space**.

Let $(X, \mathcal{M}, \mu)$ be a measure space. Here is some standard terminology concerning the "size" of μ. If $\mu(X) < \infty$ (which implies that $\mu(E) < \infty$ for all $E \in \mathcal{M}$ since $\mu(X) = \mu(E) + \mu(E^c)$), μ is called **finite**. If $X = \bigcup_1^\infty E_j$ where $E_j \in \mathcal{M}$ and $\mu(E_j) < \infty$ for all j, μ is called **σ-finite**. More generally, if $E = \bigcup_1^\infty E_j$ where $E_j \in \mathcal{M}$ and $\mu(E_j) < \infty$ for all j, the set E is said to be **σ-finite** for μ. (It would be correct but more cumbersome to say that E is of σ-finite measure.) If for each $E \in \mathcal{M}$ with $\mu(E) = \infty$ there exists $F \in \mathcal{M}$ with $F \subset E$ and $0 < \mu(F) < \infty$, μ is called **semifinite**.

Every σ-finite measure is semifinite (Exercise 13), but not conversely. Most measures that arise in practice are σ-finite, which is fortunate since non-σ-finite measures tend to exhibit pathological behavior. The properties of non-σ-finite measures will be explored from time to time in the exercises.

Let us examine a few examples of measures. These examples are of a rather trivial nature, although the first one is of practical importance. The construction of more interesting examples is a task to which we shall turn in the next two sections.

- Let X be any nonempty set, $\mathcal{M} = \mathcal{P}(X)$, and f any function from X to $[0, \infty]$. Then f determines a measure μ on $\mathcal{M}$ by the formula $\mu(E) = \sum_{x \in E} f(x)$. (For the definition of such possibly uncountable sums, see §0.5.) The reader may verify that μ is semifinite iff $f(x) < \infty$ for every $x \in X$, and μ is σ-finite iff μ is semifinite and $\{x : f(x) > 0\}$ is countable. Two special cases are of particular significance: If $f(x) = 1$ for all x, μ is called **counting measure**; and if, for some $x_0 \in X$, f is defined by $f(x_0) = 1$ and $f(x) = 0$ for $x \ne x_0$, μ is called the **point mass** or **Dirac measure** at x_0. (The same names are also applied to the restrictions of these measures to smaller σ-algebras on X.)

- Let X be an uncountable set, and let $\mathcal{M}$ be the σ-algebra of countable or co-countable sets. The function μ on $\mathcal{M}$ defined by $\mu(E) = 0$ if E is countable and $\mu(E) = 1$ if E is co-countable is easily seen to be a measure.

- Let X be an infinite set and $\mathcal{M} = \mathcal{P}(X)$. Define $\mu(E) = 0$ if E is finite, $\mu(E) = \infty$ if E is infinite. Then μ is a finitely additive measure but not a measure.

The basic properties of measures are summarized in the following theorem.

1.8 Theorem. *Let $(X, \mathcal{M}, \mu)$ be a measure space.*

a. (**Monotonicity**) *If $E, F \in \mathcal{M}$ and $E \subset F$, then $\mu(E) \le \mu(F)$.*

c. **(Continuity from below)** *If* $\{E_j\}_1^\infty \subset \mathcal{M}$ *and* $E_1 \subset E_2 \subset \cdots$, *then* $\mu(\bigcup_1^\infty E_j) = \lim_{j\to\infty} \mu(E_j)$.

d. **(Continuity from above)** *If* $\{E_j\}_1^\infty \subset \mathcal{M}$, $E_1 \supset E_2 \supset \cdots$, *and* $\mu(E_1) < \infty$, *then* $\mu(\bigcap_1^\infty E_j) = \lim_{j\to\infty} \mu(E_j)$.

Proof. (a) If $E \subset F$, then $\mu(F) = \mu(E) + \mu(F \setminus E) \geq \mu(E)$.

(b) Let $F_1 = E_1$ and $F_k = E_k \setminus (\bigcup_1^{k-1} E_j)$ for $k > 1$. Then the F_k's are disjoint and $\bigcup_1^n F_j = \bigcup_1^n E_j$ for all n. Therefore, by (a),

$$\mu\Big(\bigcup_1^\infty E_j\Big) = \mu\Big(\bigcup_1^\infty F_j\Big) = \sum_1^\infty \mu(F_j) \leq \sum_1^\infty \mu(E_j).$$

(c) Setting $E_0 = \varnothing$, we have

$$\mu\Big(\bigcup_1^\infty E_j\Big) = \sum_1^\infty \mu(E_j \setminus E_{j-1}) = \lim_{n\to\infty} \sum_1^n \mu(E_j \setminus E_{j-1}) = \lim_{n\to\infty} \mu(E_n).$$

(d) Let $F_j = E_1 \setminus E_j$; then $F_1 \subset F_2 \subset \cdots$, $\mu(E_1) = \mu(F_j) + \mu(E_j)$, and $\bigcup_1^\infty F_j = E_1 \setminus (\bigcap_1^\infty E_j)$. By (c), then,

$$\mu(E_1) = \mu\Big(\bigcap_1^\infty E_j\Big) + \lim_{j\to\infty} \mu(F_j) = \mu\Big(\bigcap_1^\infty E_j\Big) + \lim_{j\to\infty} [\mu(E_1) - \mu(E_j)].$$

Since $\mu(E_1) < \infty$, we may subtract it from both sides to yield the desired result. ∎

We remark that the condition $\mu(E_1) < \infty$ in part (d) could be replaced by $\mu(E_n) < \infty$ for some $n > 1$, as the first $n-1$ E_j's can be discarded from the sequence without affecting the intersection. However, some finiteness assumption is necessary, as it can happen that $\mu(E_j) = \infty$ for all j but $\mu(\bigcap_1^\infty E_j) < \infty$. (For example, let μ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ and let $E_j = \{n : n \geq j\}$; then $\bigcap_1^\infty E_j = \varnothing$.)

If $(X, \mathcal{M}, \mu)$ is a measure space, a set $E \in \mathcal{M}$ such that $\mu(E) = 0$ is called a **null** set. By subadditivity, any countable union of null sets is a null set, a fact which we shall use frequently. If a statement about points $x \in X$ is true except for x in some null set, we say that it is true **almost everywhere** (abbreviated **a.e.**), or for **almost every** x. (If more precision is needed, we shall speak of a **μ-null** set, or **μ-almost everywhere**).

If $\mu(E) = 0$ and $F \subset E$, then $\mu(F) = 0$ by monotonicity provided that $F \in \mathcal{M}$, but in general it need not be true that $F \in \mathcal{M}$. A measure whose domain includes all subsets of null sets is called **complete**. Completeness can sometimes obviate annoying technical points, and it can always be achieved by enlarging the domain of μ, as follows.

1.9 Theorem. *Suppose that* $(X, \mathcal{M}, \mu)$ *is a measure space. Let* $\mathcal{N} = \{N \in \mathcal{M} : \mu(N) = 0\}$ *and* $\overline{\mathcal{M}} = \{E \cup F : E \in \mathcal{M}$ *and* $F \subset N$ *for some* $N \in \mathcal{N}\}$. *Then* $\overline{\mathcal{M}}$ *is*

Proof. Since $\mathcal{M}$ and $\mathcal{N}$ are closed under countable unions, so is $\overline{\mathcal{M}}$. If $E \cup F \in \overline{\mathcal{M}}$ where $E \in \mathcal{M}$ and $F \subset N \in \mathcal{N}$, we can assume that $E \cap N = \varnothing$ (otherwise, replace F and N by $F \setminus E$ and $N \setminus E$). Then $E \cup F = (E \cup N) \cap (N^c \cup F)$, so $(E \cup F)^c = (E \cup N)^c \cup (N \setminus F)$. But $(E \cup N)^c \in \mathcal{M}$ and $N \setminus F \subset N$, so that $(E \cup F)^c \in \overline{\mathcal{M}}$. Thus $\overline{\mathcal{M}}$ is a σ-algebra.

If $E \cup F \in \overline{\mathcal{M}}$ as above, we set $\overline{\mu}(E \cup F) = \mu(E)$. This is well defined, since if $E_1 \cup F_1 = E_2 \cup F_2$ where $F_j \subset N_j \in \mathcal{N}$, then $E_1 \subset E_2 \cup N_2$ and so $\mu(E_1) \leq \mu(E_2) + \mu(N_2) = \mu(E_2)$, and likewise $\mu(E_2) \leq \mu(E_1)$. It is easily verified that $\overline{\mu}$ is a complete measure on $\overline{\mathcal{M}}$, and that $\overline{\mu}$ is the only measure on $\overline{\mathcal{M}}$ that extends μ; details are left to the reader (Exercise 6). ∎

The measure $\overline{\mu}$ in Theorem 1.9 is called the **completion** of μ, and $\overline{\mathcal{M}}$ is called the **completion** of $\mathcal{M}$ with respect to μ.

Exercises

6. Complete the proof of Theorem 1.9.

7. If $\mu_1, \ldots, \mu_n$ are measures on $(X, \mathcal{M})$ and $a_1, \ldots, a_n \in [0, \infty)$, then $\sum_1^n a_j\mu_j$ is a measure on $(X, \mathcal{M})$.

8. If $(X, \mathcal{M}, \mu)$ is a measure space and $\{E_j\}_1^\infty \subset \mathcal{M}$, then $\mu(\liminf E_j) \leq \liminf \mu(E_j)$. Also, $\mu(\limsup E_j) \geq \limsup \mu(E_j)$ provided that $\mu(\bigcup_1^\infty E_j) < \infty$.

9. If $(X, \mathcal{M}, \mu)$ is a measure space and $E, F \in \mathcal{M}$, then $\mu(E) + \mu(F) = \mu(E \cup F) + \mu(E \cap F)$.

10. Given a measure space $(X, \mathcal{M}, \mu)$ and $E \in \mathcal{M}$, define $\mu_E(A) = \mu(A \cap E)$ for $A \in \mathcal{M}$. Then μ_E is a measure.

11. A finitely additive measure μ is a measure iff it is continuous from below as in Theorem 1.8c. If $\mu(X) < \infty$, μ is a measure iff it is continuous from above as in Theorem 1.8d.

12. Let $(X, \mathcal{M}, \mu)$ be a finite measure space.

a. If $E, F \in \mathcal{M}$ and $\mu(E \Delta F) = 0$, then $\mu(E) = \mu(F)$.

b. Say that $E \sim F$ if $\mu(E \Delta F) = 0$; then $\sim$ is an equivalence relation on $\mathcal{M}$.

c. For $E, F \in \mathcal{M}$, define $\rho(E, F) = \mu(E \Delta F)$. Then $\rho(E, G) \leq \rho(E, F) + \rho(F, G)$, and hence ρ defines a metric on the space $\mathcal{M}/\sim$ of equivalence classes.

13. Every σ-finite measure is semifinite.

14. If μ is a semifinite measure and $\mu(E) = \infty$, for any $C > 0$ there exists $F \subset E$ with $C < \mu(F) < \infty$.

15. Given a measure μ on $(X, \mathcal{M})$, define μ_0 on $\mathcal{M}$ by $\mu_0(E) = \sup\{\mu(F) : F \subset E$ and $\mu(F) < \infty\}$.

a. μ_0 is a semifinite measure. It is called the **semifinite part** of μ.

c. There is a measure ν on $\mathcal{M}$ (in general, not unique) which assumes only the values 0 and ∞ such that $\mu = \mu_0 + \nu$.

16. Let $(X, \mathcal{M}, \mu)$ be a measure space. A set $E \subset X$ is called **locally measurable** if $E \cap A \in \mathcal{M}$ for all $A \in \mathcal{M}$ such that $\mu(A) < \infty$. Let $\widetilde{\mathcal{M}}$ be the collection of all locally measurable sets. Clearly $\mathcal{M} \subset \widetilde{\mathcal{M}}$; if $\mathcal{M} = \widetilde{\mathcal{M}}$, then μ is called **saturated.**

a. If μ is σ-finite, then μ is saturated.

b. $\widetilde{\mathcal{M}}$ is a σ-algebra.

c. Define $\widetilde{\mu}$ on $\widetilde{\mathcal{M}}$ by $\widetilde{\mu}(E) = \mu(E)$ if $E \in \mathcal{M}$ and $\widetilde{\mu}(E) = \infty$ otherwise. Then $\widetilde{\mu}$ is a saturated measure on $\widetilde{\mathcal{M}}$, called the **saturation** of μ.

d. If μ is complete, so is $\widetilde{\mu}$.

e. Suppose that μ is semifinite. For $E \in \widetilde{\mathcal{M}}$, define $\underline{\mu}(E) = \sup\{\mu(A) : A \in \mathcal{M}$ and $A \subset E\}$. Then $\underline{\mu}$ is a saturated measure on $\widetilde{\mathcal{M}}$ that extends μ.

f. Let X_1, X_2 be disjoint uncountable sets, $X = X_1 \cup X_2$, and $\mathcal{M}$ the σ-algebra of countable or co-countable sets in X. Let μ_0 be counting measure on $\mathcal{P}(X_1)$, and define μ on $\mathcal{M}$ by $\mu(E) = \mu_0(E \cap X_1)$. Then μ is a measure on $\mathcal{M}$, $\widetilde{\mathcal{M}} = \mathcal{P}(X)$, and in the notation of parts (c) and (e), $\widetilde{\mu} \neq \underline{\mu}$.

1.4 OUTER MEASURES

In this section we develop the tools we shall use to construct measures. To motivate the ideas, it may be useful to recall the procedure used in calculus to define the area of a bounded region E in the plane $\mathbb{R}^2$. One draws a grid of rectangles in the plane and approximates the area of E from below by the sum of the areas of the rectangles in the grid that are subsets of E, and from above by the sum of the areas of the rectangles in the grid that intersect E. The limits of these approximations as the grid is taken finer and finer give the "inner area" and "outer area" of E, and if they are equal, their common value is the "area" of E. (We shall discuss these matters in more detail in §2.6.) The key idea here is that of outer area, since if R is a large rectangle containing E, the inner area of E is just the area of R minus the outer area of $R \setminus E$.

The abstract generalization of the notion of outer area is as follows. An **outer measure** on a nonempty set X is a function $\mu^* : \mathcal{P}(X) \to [0, \infty]$ that satisfies

- $\mu^*(\varnothing) = 0$,
- $\mu^*(A) \le \mu^*(B)$ if $A \subset B$,
- $\mu^*(\bigcup_1^\infty A_j) \le \sum_1^\infty \mu^*(A_j)$.

The most common way to obtain outer measures is to start with a family $\mathcal{E}$ of "elementary sets" on which a notion of measure is defined (such as rectangles in the plane) and then to approximate arbitrary sets "from the outside" by countable unions

1.10 Proposition. *Let $\mathcal{E} \subset \mathcal{P}(X)$ and $\rho : \mathcal{E} \to [0, \infty]$ be such that $\varnothing \in \mathcal{E}$, $X \in \mathcal{E}$, and $\rho(\varnothing) = 0$. For any $A \subset X$, define*

$$\mu^*(A) = \inf\Big\{\sum_1^\infty \rho(E_j) : E_j \in \mathcal{E} \text{ and } A \subset \bigcup_1^\infty E_j\Big\}.$$

Then μ^* is an outer measure.

Proof. For any $A \subset X$ there exists $\{E_j\}_1^\infty \subset \mathcal{E}$ such that $A \subset \bigcup_1^\infty E_j$ (take $E_j = X$ for all j) so the definition of μ^* makes sense. Obviously $\mu^*(\varnothing) = 0$ (take $E_j = \varnothing$ for all j), and $\mu^*(A) \le \mu^*(B)$ for $A \subset B$ because the set over which the infimum is taken in the definition of $\mu^*(A)$ includes the corresponding set in the definition of $\mu^*(B)$. To prove the countable subadditivity, suppose $\{A_j\}_1^\infty \subset \mathcal{P}(X)$ and $\epsilon > 0$. For each j there exists $\{E_j^k\}_{k=1}^\infty \subset \mathcal{E}$ such that $A_j \subset \bigcup_{k=1}^\infty E_j^k$ and $\sum_{k=1}^\infty \rho(E_j^k) \le \mu^*(A_j) + \epsilon 2^{-j}$. But then if $A = \bigcup_1^\infty A_j$, we have $A \subset \bigcup_{j,k=1}^\infty E_j^k$ and $\sum_{j,k} \rho(E_j^k) \le \sum_j \mu^*(A_j) + \epsilon$, whence $\mu^*(A) \le \sum_j \mu^*(A_j) + \epsilon$. Since ϵ is arbitrary, we are done. ∎

The fundamental step that leads from outer measures to measures is as follows. If μ^* is an outer measure on X, a set $A \subset X$ is called **μ^*-measurable** if

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c) \text{ for all } E \subset X.$$

Of course, the inequality $\mu^*(E) \le \mu^*(E \cap A) + \mu^*(E \cap A^c)$ holds for any A and E, so to prove that A is μ^*-measurable, it suffices to prove the reverse inequality. The latter is trivial if $\mu^*(E) = \infty$, so we see that A is μ^*-measurable iff

$$\mu^*(E) \ge \mu^*(E \cap A) + \mu^*(E \cap A^c) \text{ for all } E \subset X \text{ such that } \mu^*(E) < \infty.$$

Some motivation for the notion of μ^*-measurability can be obtained by referring to the discussion at the beginning of this section. If E is a "well-behaved" set such that $E \supset A$, the equation $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$ says that the outer measure of A, $\mu^*(A)$, is equal to the "inner measure" of A, $\mu^*(E) - \mu^*(E \cap A^c)$. The leap from "well-behaved" sets containing A to arbitrary subsets of X is a large one, but it is justified by the following theorem.

1.11 Carathéodory's Theorem. *If μ^* is an outer measure on X, the collection $\mathcal{M}$ of μ^*-measurable sets is a σ-algebra, and the restriction of μ^* to $\mathcal{M}$ is a complete measure.*

Proof. First, we observe that $\mathcal{M}$ is closed under complements since the definition of μ^*-measurability of A is symmetric in A and A^c. Next, if $A, B \in \mathcal{M}$ and $E \subset X$,

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap A) + \mu^*(E \cap A^c)\\ &= \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B) + \mu^*(E \cap A^c \cap B^c).\end{aligned}$$

But $(A \cup B) = (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$, so by subadditivity, and hence

$$\mu^*(E) \geq \mu^*(E \cap (A \cup B)) + \mu^*(E \cap (A \cup B)^c).$$

It follows that $A \cup B \in \mathcal{M}$, so $\mathcal{M}$ is an algebra. Moreover, if $A, B \in \mathcal{M}$ and $A \cap B = \varnothing$,

$$\mu^*(A \cup B) = \mu^*((A \cup B) \cap A) + \mu^*((A \cup B) \cap A^c) = \mu^*(A) + \mu^*(B),$$

so μ^* is finitely additive on $\mathcal{M}$.

To show that $\mathcal{M}$ is a σ-algebra, it will suffice to show that $\mathcal{M}$ is closed under countable disjoint unions. If $\{A_j\}_1^\infty$ is a sequence of disjoint sets in $\mathcal{M}$, let $B_n = \bigcup_1^n A_j$ and $B = \bigcup_1^\infty A_j$. Then for any $E \subset X$,

$$\begin{aligned}\mu^*(E \cap B_n) &= \mu^*(E \cap B_n \cap A_n) + \mu^*(E \cap B_n \cap A_n^c) \\ &= \mu^*(E \cap A_n) + \mu^*(E \cap B_{n-1}),\end{aligned}$$

so a simple induction shows that $\mu^*(E \cap B_n) = \sum_1^n \mu^*(E \cap A_j)$. Therefore,

$$\mu^*(E) = \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) \geq \sum_1^n \mu^*(E \cap A_j) + \mu^*(E \cap B^c),$$

and letting $n \to \infty$ we obtain

$$\begin{aligned}\mu^*(E) &\geq \sum_1^\infty \mu^*(E \cap A_j) + \mu^*(E \cap B^c) \geq \mu^*\Big(\bigcup_1^\infty (E \cap A_j)\Big) + \mu^*(E \cap B^c) \\ &= \mu^*(E \cap B) + \mu^*(E \cap B^c) \geq \mu^*(E).\end{aligned}$$

All the inequalities in this last calculation are thus equalities. It follows that $B \in \mathcal{M}$ and — taking $E = B$ — that $\mu^*(B) = \sum_1^\infty \mu^*(A_j)$, so μ^* is countably additive on $\mathcal{M}$. Finally, if $\mu^*(A) = 0$, for any $E \subset X$ we have

$$\mu^*(E) \leq \mu^*(E \cap A) + \mu^*(E \cap A^c) = \mu^*(E \cap A^c) \leq \mu^*(E),$$

so that $A \in \mathcal{M}$. Therefore $\mu^*|\mathcal{M}$ is a complete measure. ∎

Our first applications of Carathéodory's theorem will be in the context of extending measures from algebras to σ-algebras. More precisely, if $\mathcal{A} \subset \mathcal{P}(X)$ is an algebra, a function $\mu_0 : \mathcal{A} \to [0, \infty]$ will be called a **premeasure** if

- $\mu_0(\varnothing) = 0$,
- if $\{A_j\}_1^\infty$ is a sequence of disjoint sets in $\mathcal{A}$ such that $\bigcup_1^\infty A_j \in \mathcal{A}$, then $\mu_0(\bigcup_1^\infty A_j) = \sum_1^\infty \mu_0(A_j)$.

In particular, a premeasure is finitely additive since one can take $A_j = \varnothing$ for j large.

is a premeasure on $\mathcal{A} \subset \mathcal{P}(X)$, it induces an outer measure on X in accordance with Proposition 1.10, namely,

$$\mu^*(E) = \inf\Big\{\sum_1^\infty \mu_0(A_j) : A_j \in \mathcal{A},\ E \subset \bigcup_1^\infty A_j\Big\}. \tag{1.12}$$

1.13 Proposition. *If μ_0 is a premeasure on $\mathcal{A}$ and μ^* is defined by (1.12), then*

a. $\mu^*|\mathcal{A} = \mu_0$;

b. every set in $\mathcal{A}$ is μ^*-measurable.

Proof. (a) Suppose $E \in \mathcal{A}$. If $E \subset \bigcup_1^\infty A_j$ with $A_j \in \mathcal{A}$, let $B_n = E \cap (A_n \setminus \bigcup_1^{n-1} A_j)$. Then the B_n's are disjoint members of $\mathcal{A}$ whose union is E, so $\mu_0(E) = \sum_1^\infty \mu_0(B_j) \leq \sum_1^\infty \mu_0(A_j)$. It follows that $\mu_0(E) \leq \mu^*(E)$, and the reverse inequality is obvious since $E \subset \bigcup_1^\infty A_j$ where $A_1 = E$ and $A_j = \varnothing$ for $j > 1$.

(b) If $A \in \mathcal{A}$, $E \subset X$, and $\epsilon > 0$, there is a sequence $\{B_j\}_1^\infty \subset \mathcal{A}$ with $E \subset \bigcup_1^\infty B_j$ and $\sum_1^\infty \mu_0(B_j) \leq \mu^*(E) + \epsilon$. Since μ_0 is additive on $\mathcal{A}$,

$$\mu^*(E) + \epsilon \geq \sum_1^\infty \mu_0(B_j \cap A) + \sum_1^\infty \mu_0(B_j \cap A^c) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c).$$

Since ϵ is arbitrary, A is μ^*-measurable. ∎

1.14 Theorem. *Let $\mathcal{A} \subset \mathcal{P}(X)$ be an algebra, μ_0 a premeasure on $\mathcal{A}$, and $\mathcal{M}$ the σ-algebra generated by $\mathcal{A}$. There exists a measure μ on $\mathcal{M}$ whose restriction to $\mathcal{A}$ is μ_0 — namely, $\mu = \mu^*|\mathcal{M}$ where μ^* is given by (1.12). If ν is another measure on $\mathcal{M}$ that extends μ_0, then $\nu(E) \leq \mu(E)$ for all $E \in \mathcal{M}$, with equality when $\mu(E) < \infty$. If μ_0 is σ-finite, then μ is the unique extension of μ_0 to a measure on $\mathcal{M}$.*

Proof. The first assertion follows from Carathéodory's theorem and Proposition 1.13 since the σ-algebra of μ^*-measurable sets includes $\mathcal{A}$ and hence $\mathcal{M}$. As for the second assertion, if $E \in \mathcal{M}$ and $E \subset \bigcup_1^\infty A_j$ where $A_j \in \mathcal{A}$, then $\nu(E) \leq \sum_1^\infty \nu(A_j) = \sum_1^\infty \mu_0(A_j)$, whence $\nu(E) \leq \mu(E)$. Also, if we set $A = \bigcup_1^\infty A_j$, we have

$$\nu(A) = \lim_{n\to\infty} \nu\Big(\bigcup_1^n A_j\Big) = \lim_{n\to\infty} \mu\Big(\bigcup_1^n A_j\Big) = \mu(A).$$

If $\mu(E) < \infty$, we can choose the A_j's so that $\mu(A) < \mu(E) + \epsilon$, hence $\mu(A \setminus E) < \epsilon$, and

$$\mu(E) \leq \mu(A) = \nu(A) = \nu(E) + \nu(A \setminus E) \leq \nu(E) + \mu(A \setminus E) \leq \nu(E) + \epsilon.$$

Since ϵ is arbitrary, $\mu(E) = \nu(E)$. Finally, suppose $X = \bigcup_1^\infty A_j$ with $\mu_0(A_j) < \infty$, where we can assume that the A_j's are disjoint. Then for any $E \in \mathcal{M}$,

$$\mu(E) = \sum_1^\infty \mu(E \cap A_j) = \sum_1^\infty \nu(E \cap A_j) = \nu(E),$$

The proof of this theorem yields more than the statement. Indeed, μ_0 may be extended to a measure on the algebra $\mathcal{M}^*$ of all μ^*-measurable sets. The relation between $\mathcal{M}$ and $\mathcal{M}^*$ is explored in Exercise 22 (along with Exercise 20b, which ensures that the outer measures induced by μ_0 and μ are the same).

Exercises

17. If μ^* is an outer measure on X and $\{A_j\}_1^\infty$ is a sequence of disjoint μ^*-measurable sets, then $\mu^*(E \cap (\bigcup_1^\infty A_j)) = \sum_1^\infty \mu^*(E \cap A_j)$ for any $E \subset X$.

18. Let $\mathcal{A} \subset \mathcal{P}(X)$ be an algebra, $\mathcal{A}_\sigma$ the collection of countable unions of sets in $\mathcal{A}$, and $\mathcal{A}_{\sigma\delta}$ the collection of countable intersections of sets in $\mathcal{A}_\sigma$. Let μ_0 be a premeasure on $\mathcal{A}$ and μ^* the induced outer measure.

a. For any $E \subset X$ and $\epsilon > 0$ there exists $A \in \mathcal{A}_\sigma$ with $E \subset A$ and $\mu^*(A) \leq \mu^*(E) + \epsilon$.

b. If $\mu^*(E) < \infty$, then E is μ^*-measurable iff there exists $B \in \mathcal{A}_{\sigma\delta}$ with $E \subset B$ and $\mu^*(B \setminus E) = 0$.

c. If μ_0 is σ-finite, the restriction $\mu^*(E) < \infty$ in (b) is superfluous.

19. Let μ^* be an outer measure on X induced from a finite premeasure μ_0. If $E \subset X$, define the **inner measure** of E to be $\mu_*(E) = \mu_0(X) - \mu^*(E^c)$. Then E is μ^*-measurable iff $\mu^*(E) = \mu_*(E)$. (Use Exercise 18.)

20. Let μ^* be an outer measure on X, $\mathcal{M}^*$ the σ-algebra of μ^*-measurable sets, $\overline{\mu} = \mu^*|\mathcal{M}^*$, and μ^+ the outer measure induced by $\overline{\mu}$ as in (1.12) (with $\overline{\mu}$ and $\mathcal{M}^*$ replacing μ_0 and $\mathcal{A}$).

a. If $E \subset X$, we have $\mu^*(E) \leq \mu^+(E)$, with equality iff there exists $A \in \mathcal{M}^*$ with $A \supset E$ and $\mu^*(A) = \mu^*(E)$.

b. If μ^* is induced from a premeasure, then $\mu^* = \mu^+$. (Use Exercise 18a.)

c. If $X = \{0, 1\}$, there exists an outer measure μ^* on X such that $\mu^* \neq \mu^+$.

21. Let μ^* be an outer measure induced from a premeasure and $\overline{\mu}$ the restriction of μ^* to the μ^*-measurable sets. Then $\overline{\mu}$ is saturated. (Use Exercise 18.)

22. Let $(X, \mathcal{M}, \mu)$ be a measure space, μ^* the outer measure induced by μ according to (1.12), $\mathcal{M}^*$ the σ-algebra of μ^*-measurable sets, and $\overline{\mu} = \mu^*|\mathcal{M}^*$.

a. If μ is σ-finite, then $\overline{\mu}$ is the completion of μ. (Use Exercise 18.)

b. In general, $\overline{\mu}$ is the saturation of the completion of μ. (See Exercises 16 and 21.)

23. Let $\mathcal{A}$ be the collection of finite unions of sets of the form $(a, b] \cap \mathbb{Q}$ where $-\infty \leq a < b \leq \infty$.

a. $\mathcal{A}$ is an algebra on $\mathbb{Q}$. (Use Proposition 1.7.)

b. The σ-algebra generated by $\mathcal{A}$ is $\mathcal{P}(\mathbb{Q})$.

c. Define μ_0 on $\mathcal{A}$ by $\mu_0(\varnothing) = 0$ and $\mu_0(A) = \infty$ for $A \neq \varnothing$. Then μ_0 is a premeasure on $\mathcal{A}$, and there is more than one measure on $\mathcal{P}(\mathbb{Q})$ whose restriction to $\mathcal{A}$ is μ_0.

24. Let μ be a finite measure on $(X, \mathcal{M})$, and let μ^* be the outer measure induced by μ. Suppose that $E \subset X$ satisfies $\mu^*(E) = \mu^*(X)$ (but not that $E \in \mathcal{M}$).

a. If $A, B \in \mathcal{M}$ and $A \cap E = B \cap E$, then $\mu(A) = \mu(B)$.

b. Let $\mathcal{M}_E = \{A \cap E : A \in \mathcal{M}\}$, and define the function ν on $\mathcal{M}_E$ defined by $\nu(A \cap E) = \mu(A)$ (which makes sense by (a)). Then $\mathcal{M}_E$ is a σ-algebra on E and ν is a measure on $\mathcal{M}_E$.

1.5 BOREL MEASURES ON THE REAL LINE

We are now in a position to construct a definitive theory for measuring subsets of $\mathbb{R}$ based on the idea that the measure of an interval is its length. We begin with a more general (but only slightly more complicated) construction that yields a large family of measures on $\mathbb{R}$ whose domain is the Borel σ-algebra $\mathcal{B}_\mathbb{R}$; such measures are called **Borel measures** on $\mathbb{R}$.

To motivate the ideas, suppose that μ is a finite Borel measure on $\mathbb{R}$, and let $F(x) = \mu((-\infty, x])$. (F is sometimes called the **distribution function** of μ.) Then F is increasing by Theorem 1.8a and right continuous by Theorem 1.8d since $(-\infty, x] = \bigcap_1^\infty (-\infty, x_n]$ whenever $x_n \searrow x$. (Recall the discussion of increasing functions in §0.5.) Moreover, if $b > a$, $(-\infty, b] = (-\infty, a] \cup (a, b]$, so $\mu((a, b]) = F(b) - F(a)$. Our procedure will be to turn this process around and construct a measure μ starting from an increasing, right-continuous function F. The special case $F(x) = x$ will yield the usual "length" measure.

The building blocks for our theory will be the left-open, right-closed intervals in $\mathbb{R}$ — that is, sets of the form $(a, b]$ or (a, ∞) or $\varnothing$, where $-\infty \leq a < b < \infty$. In this section we shall refer to such sets as **h-intervals** (h for "half-open"). Clearly the intersection of two h-intervals is an h-interval, and the complement of an h-interval is an h-interval or the disjoint union of two h-intervals. By Proposition 1.7, the collection $\mathcal{A}$ of finite disjoint unions of h-intervals is an algebra, and by Proposition 1.2, the σ-algebra generated by $\mathcal{A}$ is $\mathcal{B}_\mathbb{R}$.

1.15 Proposition. *Let $F : \mathbb{R} \to \mathbb{R}$ be increasing and right continuous. If $(a_j, b_j]$ ($j = 1, \ldots, n$) are disjoint h-intervals, let*

$$\mu_0\Big(\bigcup_1^n (a_j, b_j]\Big) = \sum_1^n [F(b_j) - F(a_j)],$$

and let $\mu_0(\varnothing) = 0$. Then μ_0 is a premeasure on the algebra $\mathcal{A}$.

Proof. First we must check that μ_0 is well defined, since elements of $\mathcal{A}$ can be represented in more than one way as disjoint unions of h-intervals. If $\{(a_j, b_j]\}_1^n$ are disjoint and $\bigcup_1^n (a_j, b_j] = (a, b]$, then, after perhaps relabeling the index j, we must have $a = a_1 < b_1 = a_2 < b_2 = \ldots < b_n = b$, so $\sum_1^n [F(b_j) - F(a_j)] =$

h-intervals such that $\bigcup_1^n I_i = \bigcup_1^n J_j$, this reasoning shows that

$$\sum_i \mu_0(I_i) = \sum_{i,j} \mu_0(I_i \cap J_j) = \sum_j \mu_0(J_j).$$

Thus μ_0 is well defined, and it is finitely additive by construction.

It remains to show that if $\{I_j\}_1^\infty$ is a sequence of disjoint h-intervals with $\bigcup_1^\infty I_j \in \mathcal{A}$ then $\mu_0(\bigcup_1^\infty I_j) = \sum_1^\infty \mu_0(I_j)$. Since $\bigcup_1^\infty I_j$ is a finite union of h-intervals, the sequence $\{I_j\}_1^\infty$ can be partitioned into finitely many subsequences such that the union of the intervals in each subsequence is a single h-interval. By considering each subsequence separately and using the finite additivity of μ_0, we may assume that $\bigcup_1^\infty I_j$ is an h-interval $I = (a, b]$. In this case, we have

$$\mu_0(I) = \mu_0\Big(\bigcup_1^n I_j\Big) + \mu_0\Big(I \setminus \bigcup_1^n I_j\Big) \geq \mu_0\Big(\bigcup_1^n I_j\Big) = \sum_1^n \mu_0(I_j).$$

Letting $n \to \infty$, we obtain $\mu_0(I) \geq \sum_1^\infty \mu_0(I_j)$. To prove the reverse inequality, let us suppose first that a and b are finite, and let us fix $\epsilon > 0$. Since F is right continuous, there exists $\delta > 0$ such that $F(a+\delta) - F(a) < \epsilon$, and if $I_j = (a_j, b_j]$, for each j there exists $\delta_j > 0$ such that $F(b_j + \delta_j) - F(b_j) < \epsilon 2^{-j}$. The open intervals $(a_j,\, b_j + \delta_j)$ cover the compact set $[a+\delta,\, b]$, so there is a finite subcover. By discarding any $(a_j,\, b_j + \delta_j)$ that is contained in a larger one and relabeling the index j, we may assume that

- the intervals $(a_1,\, b_1 + \delta_1), \ldots, (a_N,\, b_N + \delta_N)$ cover $[a+\delta,\, b]$,
- $b_j + \delta_j \in (a_{j+1},\, b_{j+1} + \delta_{j+1})$ for $j = 1, \ldots, N-1$.

But then

$$\begin{aligned}
\mu_0(I) &< F(b) - F(a+\delta) + \epsilon \\
&\leq F(b_N + \delta_N) - F(a_1) + \epsilon \\
&= F(b_N + \delta_N) - F(a_N) + \sum_1^{N-1} [F(a_{j+1}) - F(a_j)] + \epsilon \\
&\leq F(b_N + \delta_N) - F(a_N) + \sum_1^{N-1} [F(b_j + \delta_j) - F(a_j)] + \epsilon \\
&< \sum_1^N [F(b_j) + \epsilon 2^{-j} - F(a_j)] + \epsilon \\
&< \sum_1^\infty \mu_0(I_j) + 2\epsilon.
\end{aligned}$$

Since ϵ is arbitrary, we are done when a and b are finite. If $a = -\infty$, for any $M < \infty$ the intervals $(a_j,\, b_j + \delta_j)$ cover $[-M, b]$, so the same reasoning gives $F(b) - F(-M) \leq \sum_1^\infty \mu_0(I_j) + 2\epsilon$, whereas if $b = \infty$, for any $M < \infty$ we likewise obtain $F(M) - F(a) \leq \sum_1^\infty \mu_0(I_j) + 2\epsilon$. The desired result then follows

1.16 Theorem. *If $F : \mathbb{R} \to \mathbb{R}$ is any increasing, right continuous function, there is a unique Borel measure μ_F on $\mathbb{R}$ such that $\mu_F((a,b]) = F(b) - F(a)$ for all a, b. If G is another such function, we have $\mu_F = \mu_G$ iff $F - G$ is constant. Conversely, if μ is a Borel measure on $\mathbb{R}$ that is finite on all bounded Borel sets and we define*

$$F(x) = \begin{cases} \mu((0,x]) & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -\mu((x,0]) & \text{if } x < 0, \end{cases}$$

then F is increasing and right continuous, and $\mu = \mu_F$.

Proof. Each F induces a premeasure on $\mathcal{A}$ by Proposition 1.15. It is clear that F and G induce the same premeasure iff $F - G$ is constant, and that these premeasures are σ-finite (since $\mathbb{R} = \bigcup_{-\infty}^\infty (j, j+1]$). The first two assertions therefore follow from Theorem 1.14. As for the last one, the monotonicity of μ implies the monotonicity of F, and the continuity of μ from above and below implies the right continuity of F for $x \geq 0$ and $x < 0$. It is evident that $\mu = \mu_F$ on $\mathcal{A}$, and hence $\mu = \mu_F$ on $\mathcal{B}_\mathbb{R}$ by the uniqueness in Theorem 1.14. ∎

Several remarks are in order. First, this theory could equally well be developed by using intervals of the form $[a, b)$ and left continuous functions F. Second, if μ is a finite Borel measure on $\mathbb{R}$, then $\mu = \mu_F$ where $F(x) = \mu((-\infty, x])$ is the cumulative distribution function of μ; this differs from the F specified in Theorem 1.16 by the constant $\mu((-\infty, 0])$. Third, the theory of §1.4 gives, for each increasing and right continuous F, not only the Borel measure μ_F but a complete measure $\overline{\mu}_F$ whose domain includes $\mathcal{B}_\mathbb{R}$. In fact, $\overline{\mu}_F$ is just the completion of μ_F (Exercise 22a or Theorem 1.19 below), and one can show that its domain is always strictly larger than $\mathcal{B}_\mathbb{R}$. We shall usually denote this complete measure also by μ_F; it is called the **Lebesgue-Stieltjes measure** associated to F.

Lebesgue-Stieltjes measures enjoy some useful regularity properties that we now investigate. In this discussion we fix a complete Lebesgue-Stieltjes measure μ on $\mathbb{R}$ associated to the increasing, right continuous function F, and we denote by $\mathcal{M}_\mu$ the domain of μ. Thus, for any $E \in \mathcal{M}_\mu$,

$$\begin{aligned}
\mu(E) &= \inf\Big\{\sum_1^\infty [F(b_j) - F(a_j)] : E \subset \bigcup_1^\infty (a_j, b_j]\Big\} \\
&= \inf\Big\{\sum_1^\infty \mu((a_j, b_j]) : E \subset \bigcup_1^\infty (a_j, b_j]\Big\}.
\end{aligned}$$

We first observe that in the second formula for $\mu(E)$ we can replace h-intervals by open h-intervals:

1.17 Lemma. *For any $E \in \mathcal{M}_\mu$,*

$$\mu(E) = \inf\Big\{\sum_1^\infty \mu((a_j, b_j)) : E \subset \bigcup_1^\infty (a_j, b_j)\Big\}.$$

Proof. Let us call the quantity on the right $\nu(E)$. Suppose $E \subset \bigcup_1^\infty (a_j, b_j)$. Each (a_j, b_j) is a countable disjoint union of h-intervals I_j^k ($k = 1, 2, \ldots$); specifically, $I_j^k = (c_j^k, c_j^{k+1}]$ where $\{c_j\}$ is any sequence such that $c_j^1 = a_j$ and c_j^k increases to b_j as $k \to \infty$. Thus $E \subset \bigcup_{j,k=1}^\infty I_j^k$, so

$$\sum_1^\infty \mu((a_j, b_j)) = \sum_{j,k=1}^\infty \mu(I_j^k) \geq \mu(E),$$

and hence $\nu(E) \geq \mu(E)$. On the other hand, given $\epsilon > 0$ there exists $\{(a_j, b_j]\}_1^\infty$ with $E \subset \bigcup_1^\infty (a_j, b_j]$ and $\sum_1^\infty \mu((a_j, b_j]) \leq \mu(E) + \epsilon$, and for each j there exists $\delta_j > 0$ such that $F(b_j + \delta_j) - F(b_j) < \epsilon 2^{-j}$. Then $E \subset \bigcup_1^\infty (a_j,\, b_j + \delta_j)$ and

$$\sum_1^\infty \mu((a_j,\, b_j + \delta_j)) \leq \sum_1^\infty \mu((a_j, b_j]) + \epsilon \leq \mu(E) + 2\epsilon,$$

so that $\nu(E) \leq \mu(E)$. ∎

1.18 Theorem. *If $E \in \mathcal{M}_\mu$, then*

$$\mu(E) = \inf\{\mu(U) : U \supset E \text{ and } U \text{ is open}\}$$
$$= \sup\{\mu(K) : K \subset E \text{ and } K \text{ is compact}\}.$$

Proof. By Lemma 1.17, for any $\epsilon > 0$ there exist intervals (a_j, b_j) such that $E \subset \bigcup_1^\infty (a_j, b_j)$ and $\sum_1^\infty \mu((a_j, b_j)) \leq \mu(E) + \epsilon$. If $U = \bigcup_1^\infty (a_j, b_j)$ then U is open, $U \supset E$, and $\mu(U) \leq \mu(E) + \epsilon$. On the other hand, $\mu(U) \geq \mu(E)$ whenever $U \supset E$, so the first equality is valid. For the second one, suppose first that E is bounded. If E is closed, then E is compact and the equality is obvious. Otherwise, given $\epsilon > 0$ we can choose an open $U \supset \overline{E} \setminus E$ such that $\mu(U) \leq \mu(\overline{E} \setminus E) + \epsilon$. Let $K = \overline{E} \setminus U$. Then K is compact, $K \subset E$, and

$$\mu(K) = \mu(E) - \mu(E \cap U) = \mu(E) - [\mu(U) - \mu(U \setminus E)]$$
$$\geq \mu(E) - \mu(U) + \mu(\overline{E} \setminus E) \geq \mu(E) - \epsilon.$$

If E is unbounded, let $E_j = E \cap (j, j+1]$. By the preceding argument, for any $\epsilon > 0$ there exist compact $K_j \subset E_j$ with $\mu(K_j) \geq \mu(E_j) - \epsilon 2^{-|j|}/3$. Let $H_n = \bigcup_{-n}^n K_j$. Then H_n is compact, $H_n \subset E$, and $\mu(H_n) \geq \mu(\bigcup_{-n}^n E_j) - \epsilon$. Since $\mu(E) = \lim_{n\to\infty} \mu(\bigcup_{-n}^n E_j)$, the result follows. ∎

1.19 Theorem. *If $E \subset \mathbb{R}$, the following are equivalent.*

a. $E \in \mathcal{M}_\mu$.

b. $E = V \setminus N_1$ *where V is a G_δ set and* $\mu(N_1) = 0$.

Proof. Obviously (b) and (c) each imply (a) since μ is complete on $\mathcal{M}_\mu$. Suppose $E \in \mathcal{M}_\mu$ and $\mu(E) < \infty$. By Theorem 1.18, for $j \in \mathbb{N}$ we can choose an open $U_j \supset E$ and a compact $K_j \subset E$ such that

$$\mu(U_j) - 2^{-j} \leq \mu(E) \leq \mu(K_j) + 2^{-j}.$$

Let $V = \bigcap_1^\infty U_j$ and $H = \bigcup_1^\infty K_j$. Then $H \subset E \subset V$ and $\mu(V) = \mu(H) = \mu(E) < \infty$, so $\mu(V \setminus E) = \mu(E \setminus H) = 0$. The result is thus proved when $\mu(E) < \infty$; the extension to the general case is left to the reader (Exercise 25). ∎

The significance of Theorem 1.19 is that all Borel sets (or, more generally, all sets in $\mathcal{M}_\mu$) are of a reasonably simple form modulo sets of measure zero. This contrasts markedly with the machinations necessary to construct the Borel sets from the open sets when null sets are not excepted; see Proposition 1.23 below. Another version of the idea that general measurable sets can be approximated by "simple" sets is contained in the following proposition, whose proof is left to the reader (Exercise 26):

1.20 Proposition. *If $E \in \mathcal{M}_\mu$ and $\mu(E) < \infty$, then for every $\epsilon > 0$ there is a set A that is a finite union of open intervals such that $\mu(E \triangle A) < \epsilon$.*

We now examine the most important measure on $\mathbb{R}$, namely, **Lebesgue measure**: This is the complete measure μ_F associated to the function $F(x) = x$, for which the measure of an interval is simply its length. We shall denote it by m. The domain of m is called the class of **Lebesgue measurable** sets, and we shall denote it by $\mathcal{L}$. We shall also refer to the restriction of m to $\mathcal{B}_\mathbb{R}$ as Lebesgue measure.

Among the most significant properties of Lebesgue measure are its invariance under translations and simple behavior under dilations. If $E \subset \mathbb{R}$ and $s, r \in \mathbb{R}$, we define

$$E + s = \{x + s : x \in E\}, \qquad rE = \{rx : x \in E\}.$$

1.21 Theorem. *If $E \in \mathcal{L}$, then $E + s \in \mathcal{L}$ and $rE \in \mathcal{L}$ for all $s, r \in \mathbb{R}$. Moreover, $m(E + s) = m(E)$ and $m(rE) = |r|m(E)$.*

Proof. Since the collection of open intervals is invariant under translations and dilations, the same is true of $\mathcal{B}_\mathbb{R}$. For $E \in \mathcal{B}_\mathbb{R}$, let $m_s(E) = m(E + s)$ and $m^r(E) = m(rE)$. Then m_s and m^r clearly agree with m and $|r|m$ on finite unions of intervals, hence on $\mathcal{B}_\mathbb{R}$ by Theorem 1.14. In particular, if $E \in \mathcal{B}_\mathbb{R}$ and $m(E) = 0$, then $m(E + s) = m(rE) = 0$, from which it follows that the class of sets of Lebesgue measure zero is preserved by translations and dilations. It follows that $\mathcal{L}$ (the members of which are a union of a Borel set and a Lebesgue null set) is preserved by translation and dilations and that $m(E + s) = m(E)$ and $m(rE) = |r|m(E)$ for all $E \in \mathcal{L}$. ∎

The relation between the measure-theoretic and topological properties of subsets of $\mathbb{R}$ is delicate and contains some surprises. Consider the following facts. Every

In particular, $m(\mathbb{Q}) = 0$. Let $\{r_j\}_1^\infty$ be an enumeration of the rational numbers in $[0,1]$, and given $\epsilon > 0$, let I_j be the open interval centered at r_j of length $\epsilon 2^{-j}$. Then the set $U = (0,1) \cap \bigcup_1^\infty I_j$ is open and dense in $[0,1]$, but $m(U) \le \sum_1^\infty \epsilon 2^{-j} = \epsilon$; its complement $K = [0,1] \setminus U$ is closed and nowhere dense, but $m(K) \ge 1 - \epsilon$. Thus a set that is open and dense, and hence topologically "large," can be measure-theoretically small, and a set that is nowhere dense, and hence topologically "small," can be measure-theoretically large. (A nonempty open set cannot have Lebesgue measure zero, however.)

The Lebesgue null sets include not only all countable sets but many sets having the cardinality of the continuum. We now present the standard example, the Cantor set, which is also of interest for other reasons.

Each $x \in [0,1]$ has a base-3 decimal expansion $x = \sum_1^\infty a_j 3^{-j}$ where $a_j = 0, 1,$ or 2. This expansion is unique unless x is of the form $p3^{-k}$ for some integers p, k, in which case x has two expansions: one with $a_j = 0$ for $j > k$ and one with $a_j = 2$ for $j > k$. Assuming p is not divisible by 3, one of these expansions will have $a_k = 1$ and the other will have $a_k = 0$ or 2. If we agree always to use the latter expansion, we see that

$$a_1 = 1 \text{ iff } \tfrac{1}{3} < x < \tfrac{2}{3},$$
$$a_1 \neq 1 \text{ and } a_2 = 1 \text{ iff } \tfrac{1}{9} < x < \tfrac{2}{9} \text{ or } \tfrac{7}{9} < x < \tfrac{8}{9},$$

and so forth. It will also be useful to observe that if $x = \sum a_j 3^{-j}$ and $y = \sum b_j 3^{-j}$, then $x < y$ iff there exists an n such that $a_n < b_n$ and $a_j = b_j$ for $j < n$.

The **Cantor set** C is the set of all $x \in [0,1]$ that have a base-3 expansion $x = \sum a_j 3^{-j}$ with $a_j \neq 1$ for all j. Thus C is obtained from [0,1] by removing the open middle third $(\frac{1}{3}, \frac{2}{3})$, then removing the open middle thirds $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$ of the two remaining intervals, and so forth. The basic properties of C are summarized as follows:

1.22 Proposition. *Let C be the Cantor set.*

a. C is compact, nowhere dense, and totally disconnected (i.e., the only connected subsets of C are single points). Moreover, C has no isolated points.

b. $m(C) = 0$.

c. $\mathrm{card}(C) = \mathfrak{c}$.

Proof. We leave the proof of (a) to the reader (Exercise 27). As for (b), C is obtained from $[0,1]$ by removing one interval of length $\frac{1}{3}$, two intervals of length $\frac{1}{9}$, and so forth. Thus

$$m(C) = 1 - \sum_0^\infty \frac{2^j}{3^{j+1}} = 1 - \frac{1}{3} \cdot \frac{1}{1 - (2/3)} = 0.$$

Lastly, suppose $x \in C$, so that $x = \sum_0^\infty a_j 3^{-j}$ where $a_j = 0$ or 2 for all j. Let $f(x) = \sum_1^\infty b_j 2^{-j}$ where $b_j = a_j/2$. The series defining $f(x)$ is the base-2 expansion of a number in [0,1], and any number in [0,1] can be obtained in this way.

Let us examine the map f in the preceding proof more closely. One readily sees that if $x, y \in C$ and $x < y$, then $f(x) < f(y)$ unless x and y are the two endpoints of one of the intervals removed from [0,1] to obtain C. In this case $f(x) = p2^{-k}$ for some integers p, k, and $f(x)$ and $f(y)$ are the two base-2 expansions of this number. We can therefore extend f to a map from [0,1] to itself by declaring it to be constant on each interval missing from C. This extended f is still increasing, and since its range is all of [0,1] it cannot have any jump discontinuities; hence it is continuous. f is called the **Cantor function** or **Cantor-Lebesgue function.**

The construction of the Cantor set by starting with [0,1] and successively removing open middle thirds of intervals has an obvious generalization. If I is a bounded interval and $\alpha \in (0,1)$, let us call the open interval with the same midpoint as I and length equal to α times the length of I the "open middle αth" of I. If $\{\alpha_j\}_1^\infty$ is any sequence of numbers in $(0,1)$, then, we can define a decreasing sequence $\{K_j\}$ of closed sets as follows: $K_0 = [0,1]$, and K_j is obtained by removing the open middle α_jth from each of the intervals that make up K_{j-1}. The resulting limiting set $K = \bigcap_1^\infty K_j$ is called a **generalized Cantor set.** Generalized Cantor sets all share with the ordinary Cantor set the properties (a) and (c) in Proposition 1.22. As for their Lebesgue measure, clearly $m(K_j) = (1 - \alpha_j)m(K_{j-1})$, so $m(K)$ is the infinite product $\prod_1^\infty (1-\alpha_j) = \lim_{n\to\infty} \prod_1^n (1-\alpha_j)$. If the α_j are all equal to a fixed $\alpha \in (0,1)$ (for example, $\alpha = \frac{1}{3}$ for the ordinary Cantor set), we have $m(K) = 0$. However, if $\alpha_j \to 0$ sufficiently rapidly as $j \to \infty$, $m(K)$ will be positive, and for any $\beta \in (0,1)$ one can choose α_j so that $m(K)$ will equal β; see Exercise 32. This gives another way of constructing nowhere dense sets of positive measure.

Not every Lebesgue measurable set is a Borel set. One can display examples of sets in $\mathcal{L} \setminus \mathcal{B}_{\mathbb{R}}$ by using the Cantor function; see Exercise 9 in Chapter 2. Alternatively, one can observe that since every subset of the Cantor set is Lebesgue measurable, we have $\mathrm{card}(\mathcal{L}) = \mathrm{card}(\mathcal{P}(\mathbb{R})) > \mathfrak{c}$, whereas $\mathrm{card}(\mathcal{B}_{\mathbb{R}}) = \mathfrak{c}$. The latter fact follows from Proposition 1.23 below.

Exercises

25. Complete the proof of Theorem 1.19.

26. Prove Proposition 1.20. (Use Theorem 1.18.)

27. Prove Proposition 1.22a. (Show that if $x, y \in C$ and $x < y$, there exists $z \notin C$ such that $x < z < y$.)

28. Let F be increasing and right continuous, and let μ_F be the associated measure. Then $\mu_F(\{a\}) = F(a) - F(a-)$, $\mu_F([a,b)) = F(b-) - F(a-)$, $\mu_F([a,b]) = F(b) - F(a-)$, and $\mu_F((a,b)) = F(b-) - F(a)$.

29. Let E be a Lebesgue measurable set.

a. If $E \subset N$ where N is the nonmeasurable set described in §1.1, then $m(E) = 0$.

b. If $m(E) > 0$, then E contains a nonmeasurable set. (It suffices to assume

30. If $E \in \mathcal{L}$ and $m(E) > 0$, for any $\alpha < 1$ there is an open interval I such that $m(E \cap I) > \alpha m(I)$.

31. If $E \in \mathcal{L}$ and $m(E) > 0$, the set $E - E = \{x - y : x, y \in E\}$ contains an interval centered at 0. (If I is as in Exercise 30 with $\alpha > \frac{3}{4}$, then $E - E$ contains $(-\frac{1}{2}m(I), \frac{1}{2}m(I))$.)

32. Suppose $\{\alpha_j\}_1^\infty \subset (0,1)$.

a. $\prod_1^\infty (1 - \alpha_j) > 0$ iff $\sum_1^\infty \alpha_j < \infty$. (Compare $\sum_1^\infty \log(1 - \alpha_j)$ to $\sum \alpha_j$.)

b. Given $\beta \in (0,1)$, exhibit a sequence $\{\alpha_j\}$ such that $\prod_1^\infty (1 - \alpha_j) = \beta$.

33. There exists a Borel set $A \subset [0,1]$ such that $0 < m(A \cap I) < m(I)$ for every subinterval I of $[0,1]$. (Hint: Every subinterval of $[0,1]$ contains Cantor-type sets of positive measure.)

1.6 NOTES AND REFERENCES

The history of measure theory is intimately connected with the history of integration theory, comments on which will be made in §2.7.

§1.1: The Banach-Tarski paradox appeared first in [11], but the following variant goes back to Hausdorff [68]:

> The unit sphere in $\mathbb{R}^3$, $\{x \in \mathbb{R}^3 : |x| = 1\}$, is the disjoint union of four sets $E_1, \ldots, E_4$ such that (a) E_1 is countable and (b) the sets E_2, E_3, E_4, and $E_3 \cup E_4$ are all images of each other under rotations.

An elementary exposition of the Banach-Tarski paradox and Hausdorff's result can be found in Stromberg [146].

§1.2: Our characterization of the σ-algebra $\mathcal{M}(\mathcal{E})$ generated by a family $\mathcal{E} \subset \mathcal{P}(X)$ is nonconstructive, and one might ask how to obtain $\mathcal{M}(\mathcal{E})$ explicitly from $\mathcal{E}$. The answer is rather complicated. One can begin as follows: Let $\mathcal{E}_1 = \mathcal{E} \cup \{E^c : E \in \mathcal{E}\}$, and for $j > 1$ define $\mathcal{E}_j$ to be the collection of all sets that are countable unions of sets in $\mathcal{E}_{j-1}$ or complements of such. Let $\mathcal{E}_\omega = \bigcup_1^\infty \mathcal{E}_j$: is $\mathcal{E}_\omega = \mathcal{M}(\mathcal{E})$? In general, no. $\mathcal{E}_\omega$ is closed under complements, but if $E_j \in \mathcal{E}_j \setminus \mathcal{E}_{j-1}$ for each j, there is no reason for $\bigcup_1^\infty E_j$ to be in $\mathcal{E}_\omega$. So one must start all over again. More precisely, one must define $\mathcal{E}_\alpha$ for every countable ordinal α by transfinite induction: If α has an immediate predecesor β, $\mathcal{E}_\alpha$ is the collection of sets that are countable unions of sets in $\mathcal{E}_\beta$ or complements of such; otherwise, $\mathcal{E}_\alpha = \bigcup_{\beta<\alpha} \mathcal{E}_\beta$. Then:

1.23 Proposition. $\mathcal{M}(\mathcal{E}) = \bigcup_{\alpha\in\Omega} \mathcal{E}_\alpha$, *where Ω is the set of countable ordinals.*

Proof. Transfinite induction shows that $\mathcal{E}_\alpha \subset \mathcal{M}(\mathcal{E})$ for all $\alpha \in \Omega$, and hence $\bigcup_{\alpha\in\Omega} \mathcal{E}_\alpha \subset \mathcal{M}(\mathcal{E})$. The reverse inclusion follows from the fact that any sequence in Ω has a supremum in Ω (Proposition 0.19): If $E_j \in \mathcal{E}_{\alpha_j}$, for $j \in \mathbb{N}$ and $\alpha = \sup\{\alpha_j\}$,

Combining this with Proposition 0.14, we see that if $\text{card}(\mathbb{N}) \leq \text{card}(\mathcal{E}) \leq \mathfrak{c}$, then $\text{card}(\mathcal{M}(\mathcal{E})) = \mathfrak{c}$. (Cf. Exercise 3.)

§1.3: Some authors prefer to take the domains of measures to be σ-rings rather than σ-algebras (see Exercise 1). The reason is that in dealing with "very large" spaces one can avoid certain pathologies by not attempting to measure "very large" sets. However, this point of view also has technical disadvantages, and it is no longer much in favor.

§1.4: Carathéodory's theorem appears in his treatise [22]. Theorem 1.14 has been attributed in the literature to Hahn, Carathéodory, and E. Hopf, but it is originally due to Fréchet [54]. The proof via Carathéodory's theorem was discovered independently by Hahn [60] and Kolmogorov [85].

See König [86] for a deeper study of the problem of constructing measures from more primitive data.

§1.5: Lebesgue originally defined the outer measure $m^*(E)$ of a set $E \subset \mathbb{R}$ in terms of countable coverings by intervals, as we have done. He then defined a bounded set E to be measurable if $m^*(E) + m^*((a,b) \setminus E) = b - a$, where (a,b) is an interval containing E, and an unbounded set to be measurable if its intersection with any bounded interval is measurable. Carathéodory's characterization of measurability, which is technically eaiser to work with, came later. For the equivalence of the two definitions, see Exercise 19.

One should convince oneself that the remarkably fussy proof of Proposition 1.15 is necessary by contemplating the complicated ways in which an h-interval can be decomposed into a disjoint union of h-subintervals. In any such decomposition the collection of right endpoints of the subintervals, when ordered from right to left, is a well ordered set, but it can be order isomorphic to any initial segment of the set of countable ordinals.

Lebesgue measure can be extended to a translation-invariant measure on σ-algebras that properly include $\mathcal{L}$; see Kakutani and Oxtoby [81]. Of course, such σ-algebras can never contain the nonmeasurable set discussed in §1. However, Lebesgue measure can be extended to a translation-invariant finitely additive measure on $\mathcal{P}(\mathbb{R})$, and its 2-dimensional analogue (see §2.6) can be extended to a finitely additive measure on $\mathcal{P}(\mathbb{R}^2)$ that is invariant under translations and rotations; see Banach [8]. The Banach-Tarski paradox prevents this result from being extended to higher dimensions.

In connection with the existence of nonmeasurable sets, Solovay [138] has proved a remarkable theorem which says in effect that it is impossible to prove the existence of Lebesgue nonmeasurable sets without using the axiom of choice. (The precise statement of the theorem involves some technical points of axiomatic set theory, which we shall not discuss here.) From the point of view of the working analyst, the effect of Solovay's theorem is to reaffirm the adequacy of the Lebesgue theory for all practical purposes.

See Rudin [124] for a terse solution of Exercise 33.